
Understanding Protein Dynamics with L_1 -Regularized Reversible Hidden Markov Models

Robert T. McGibbon

RMCGIBBO@STANFORD.EDU

Department of Chemistry, Stanford University, Stanford CA 94305, USA

Bharath Ramsundar

RBHARATH@STANFORD.EDU

Department of Computer Science, Stanford University, Stanford CA 94305, USA

Mohammad M. Sultan

MSULTAN@STANFORD.EDU

Department of Chemistry, Stanford University, Stanford CA 94305, USA

Gert Kiss

GKISS@STANFORD.EDU

Department of Chemistry, Stanford University, Stanford CA 94305, USA

Vijay S. Pande

PANDE@STANFORD.EDU

Department of Chemistry, Stanford University, Stanford CA 94305, USA

Abstract

We present a machine learning framework for modeling protein dynamics. Our approach uses L_1 -regularized, reversible hidden Markov models to understand large protein datasets generated via molecular dynamics simulations. Our model is motivated by three design principles: (1) the requirement of massive scalability; (2) the need to adhere to relevant physical law; and (3) the necessity of providing accessible interpretations, critical for both cellular biology and rational drug design. We present an EM algorithm for learning and introduce a model selection criteria based on the physical notion of convergence in relaxation timescales. We contrast our model with standard methods in biophysics and demonstrate improved robustness. We implement our algorithm on GPUs and apply the method to two large protein simulation datasets generated respectively on the NCSA Bluewaters supercomputer and the Folding@Home distributed computing network. Our analysis identifies the conformational dynamics of the ubiquitin protein critical to cellular signaling, and elucidates the stepwise activation mechanism of the c-Src kinase protein.

1. Introduction

Protein folding and conformational change are grand challenge problems, relevant to a multitude of human diseases, including Alzheimer's disease, Huntington's disease and cancer. These problems entail the characterization of the process and pathways by which proteins fold to their energetically optimal configuration and the dynamics between multiple long-lived, or "metastable," configurations on the potential energy surface. Proteins are biology's molecular machines; a solution to the folding and conformational change problem would deepen our understanding of the mechanism by which microscopic information in the genome is manifested in the macroscopic phenotype of organisms. Furthermore, an understanding of the structure and dynamics of proteins is increasingly important for the rational design of targeted drugs (Wong & McCammon, 2003).

Molecular dynamics (MD) simulations provide a computational microscope by which protein dynamics can be studied with atomic resolution (Dill et al., 1995). These simulations entail the forward integration of Newton's equations of motion on a classical potential energy surface. The potential energy functions in use, called forcefields, are semi-empirical approximations to the true quantum mechanical Born-Oppenheimer surface, designed to reproduce experimental observables (Beauchamp et al., 2012). For moderately sized proteins, this computation can involve the propagation of more than a million physical degrees of freedom. Furthermore, while folding events can take milliseconds

(10^{-3} s) or longer, the simulations must be integrated with femtosecond (10^{-15} s) timesteps, requiring the collection of datasets containing trillions of data points.

While the computational burden of performing MD simulations has been a central challenge in the field, significant progress has been achieved recently with the development of three independent technologies: ANTON, a special-purpose supercomputer using a custom ASIC to accelerate MD (Shaw, 2007); Folding@Home, a distributed computing network harnessing the desktop computers of more than 240,000 volunteers (Shirts & Pande, 2000); and Google Exacycle, an initiative utilizing the spare cycles on Google’s production infrastructure for science (Kohlhoff et al., 2014).

The analysis of these massive simulation datasets now represents a major difficulty: how do we turn data into knowledge (Lane et al., 2013)? In contrast to some other machine learning problems, the central goal here is not merely prediction. Instead, we view analysis – often in the form of probabilistic models generated from MD datasets – as a tool for generating scientific insight about protein dynamics.

Useful probabilistic models must embody the appropriate physics. The guiding physical paradigm by which chemical dynamics are understood is one of *states* and *rates*. States correspond to metastable regions in the configuration space of the protein and can often be visualized as wells on the potential energy surface. Fluctuations within each metastable state are rapid; the dominant, long time-scale dynamics can be understood as a jump process moving with various rates between the states. This paradigm motivates probabilistic models based on a discrete-state Markov chain. *A priori*, the location of the metastable states are unknown. As a result, each metastable state should correspond to a latent variable in the model. Hidden Markov models (HMMs) thus provide the natural framework.

Classical mechanics at thermal equilibrium satisfy a symmetry with respect to time: a microscopic process and its time-reversed version obey the same laws of motion. The stochastic analogue of this property is reversibility (also called detailed balance): the equilibrium flux between any two states X and Y is equal in both directions. Probabilistic models which fail to capture this essential property will assign positive probability to systems that violate the second law of thermodynamics (Prinz et al., 2011). Hence, we enforce detailed balance in our HMMs.

In addition to the constraints motivated by adherence to physical laws, suitable probabilistic models should, in broad strokes, incorporate knowledge from prior experimental and theoretical studies of proteins. Numerous studies indicate that only a subset of the degrees of freedom

are essential for describing the protein’s dominant long time-scale dynamics (see Cho et al. (2006) and references therein). Furthermore, substantial prior work indicates that protein folding occurs via a sequence of localized shifts (Maity et al., 2005). Together, these pieces of evidence motivate the imposition of L_1 -fusion regularization (Tibshirani et al., 2005). The L_1 term penalizes deviations amongst states along uninformative degrees of freedom, thereby suppressing their effect on the model. Furthermore, the pairwise structure of the fusion penalty minimizes the number of transitions which involve global changes: many pairs of states will only differ along a reduced subset of the dimensions.

The main results of this paper are the formulation of the L_1 -regularized reversible HMM and the introduction of a simple and scalable learning algorithm to fit the model. We contrast our approach against standard frameworks for the analysis of MD data and demonstrate improved robustness and physical interpretability.

This paper is organized as follows. Section 2 describes prior work. Section 3 introduces the model and associated learning algorithm. Section 4 applies the model to three systems: a toy double well potential; ubiquitin, a human signaling protein; and c-Src kinase, a critical regulatory protein involved in cancer genesis. Section 5 provides discussion and indicates future directions.

2. Prior Work

Earlier studies have applied machine learning techniques to investigate protein structure prediction – the problem of discovering a protein’s energetically optimal configuration – using CRFs, belief propagation, deep learning, and other general ML methods (Sontag et al., 2012; Di Lena et al., 2012; Chu et al., 2006; Baldi & Pollastri, 2003). But proteins are fundamentally dynamic systems, and none of these approaches offer insight into kinetics; rather, they are concerned with extracting static information about protein structure.

The dominant computational tool for studying protein dynamics is MD. Traditional analyses of MD datasets are primarily visual and non-quantitative. Standard approaches include watching movies of a protein’s structural dynamics along simulation trajectories, and inspecting the time evolution of a small number of pre-specified degrees of freedom (Humphrey et al., 1996; Karplus & Kuriyan, 2005). While these methods have been successfully applied to smaller proteins, they struggle to characterize the dynamics of the large and complex biomolecules critical to biological function. Quantitative methods like PCA can elucidate important (high variance) degrees of freedom, but fail to capture the rich temporal structure in MD datasets.

Markov state models (MSMs) are a simple class of probabilistic models, recently introduced to capture the temporal dynamics of the folding process. In an MSM, protein dynamics are modeled by the evolution of a Markov chain on a discrete state space. The finite set of states is generated by clustering the set of configurations in the MD trajectories (Beauchamp et al., 2011). MSMs can be viewed as fully observable HMMs. More recently, HMMs with multinomial emission distributions have been employed on this discrete state space (Noé et al., 2013).

Although MSMs have had a number of notable successes (Voelz et al., 2012; Sadiq et al., 2012), they are brittle and complex. Traditional MSMs lack complete data likelihood functions, and learning cannot be easily characterized by a single optimization problem. For these reasons, MSM learning requires significant manual tuning. For example, because clustering is purely a preprocessing step, the likelihood function contains no guidance on the choice of the metastable states. Moreover, the lack of uncertainty in the observation model necessitates the introduction of a very large number of states, typically more than ten thousand, in order to cover the protein’s phase space at sufficient resolution. This abundance of states is statistically inefficient, as millions of pairwise transition parameters must be estimated in typically-sized models, and renders interpretation of learned MSMs challenging.

3. Fusion L_1 -Regularized Reversible HMM

We introduce the L_1 -regularized reversible HMM with Gaussian emissions, a generative probabilistic model over multivariate discrete-time continuous-space time series. As discussed in Section 1, we integrate necessary physical constraints on top of the core hidden Markov model (Rabiner, 1989).

Let $\{Y_t\}$ be the observed time series in \mathbb{R}^D of length T (i.e., the input simulation data), and let $\{X_t\}$ be the corresponding latent time series in $\{1, \dots, K\}$, where K is a hyperparameter indicating the number of hidden states in the model. Each hidden variable x_t corresponds to a metastable state of the physical system. The emission distribution given $X_t = k$ is a multivariate normal distribution parameterized by mean $\mu_k \in \mathbb{R}^D$ and diagonal covariance matrix $\text{Diag}(\sigma_k^2) \in \mathbb{R}^{D \times D}$ (where $\sigma_k^2 \in \mathbb{R}^D$ is the vector of diagonal covariance elements). We use the notation $(\mu_k)_j$ to indicate the j th element of the vector μ_k .

Controlling the means $\{\mu_k\}$ is critical for achieving physically interpretable models. As discussed in Section 1, we wish to minimize the differences between μ_k and $\mu_{k'}$ to the extent possible. Consequently, we place a fusion L_1 penalty (Tibshirani et al., 2005) on our log likelihood func-

tion, which adds the following pairwise cost:

$$\lambda \sum_{k, k'} \sum_j \tau_{k, k'}^{(j)} |(\mu_k)_j - (\mu_{k'})_j|.$$

Here, λ governs the overall strength of the penalty, while the adaptive fusion weights, $\{\tau_{k, k'}^{(j)}\}$, control the contribution from each pair of states (Guo et al., 2010). During learning, the adaptive fusion weights are computed as

$$\tau_{k, k'}^{(j)} = |(\tilde{\mu}_k)_j - (\tilde{\mu}_{k'})_j|^{-1},$$

where the $\{\tilde{\mu}_k\}$ are the learned metastable state means in the absence of the penalty. The intuition motivating the adaptive strength of the penalty is that if degree of freedom j is informative for separating states k and k' , the corresponding fusion penalty should be applied lightly.

The reversible time evolution of the model is parameterized by an irreducible, aperiodic, row-normalized K by K stochastic matrix \mathbf{T} , which satisfies detailed balance. Mathematically, the detailed balance constraint is

$$\forall k, k', \pi_k \mathbf{T}_{k, k'} = \pi_{k'} \mathbf{T}_{k', k},$$

where row vector π is the stationary distribution of \mathbf{T} . The stationary distribution π also parameterizes the initial distribution over the metastable states. By the Perron–Frobenius theorem, π is the dominant left eigenvector of \mathbf{T} with eigenvalue 1 and is not an independent parameter in this model.

The initial distributions and evolution of $\{X_t, Y_t\}$ satisfy the following equations:

$$\begin{aligned} X_0 &\sim \sum_{k=1}^K \pi_k \delta_k, \\ X_{t+1} &\sim \sum_{k=1}^K \mathbf{T}_{X_t, k} \delta_k, \\ Y_t &\sim \mathcal{N}(\mu_{X_t}, \sigma_{X_t}^2). \end{aligned}$$

The complete data likelihood $\{x_t, y_t\}$ is

$$\begin{aligned} \mathcal{L}(\{x_t\}, \{y_t\} | \mathbf{T}, \mu, \sigma) = \\ \pi_{x_0} \prod_{t=1}^{T-1} \mathbf{T}_{x_{t-1}, x_t} \prod_{t=0}^{T-1} \mathcal{N}(y_t; \mu_{x_t}, \sigma_{x_t}^2). \end{aligned}$$

The hyperparameter Δ controls the discretization interval at which a protein’s coordinates are sampled to obtain $\{y_t\}$. In the absence of downsampling by Δ , subsequent samples y_t, y_{t+1} would be highly correlated. On the other hand, subsequent samples from an HMM are conditionally independent given the hidden state. Choice of Δ large enough recovers this conditional independence (*vide infra*).

3.1. Learning

The model is fit using expectation-maximization. The E-step is standard, while the M-step requires modification to enforce the detailed balance constraint on \mathbf{T} and the adaptive fusion penalty on the $\{\mu_k\}$.

3.1.1. E-STEP

Inference is identical to that for the standard HMM, using the forward-backward algorithm (Rabiner, 1989) to compute the following quantities:

$$\begin{aligned}\gamma_i(t) &= \mathbb{P}(X_t = i | \{y_t\}), \\ \xi_{ij}(t) &= \mathbb{P}(X_t = i, X_{t+1} = j | \{y_t\}).\end{aligned}$$

3.1.2. M-STEP

Both the penalty on $\{\mu_k\}$ and the reversibility constraint affect only the M-step. The M-step update to the means in the t -th iteration of EM consists of maximizing the penalized log-likelihood function

$$\begin{aligned}\mu_k^{(t+1)} &= \underset{\mu_k}{\operatorname{argmin}} \sum_i \sum_k \gamma_k(i) \frac{(x_i - \mu_k)^2}{2(\sigma_k^2)^{(t)}} \\ &\quad + \lambda \sum_{k,k'} \sum_j \tau_{k,k'}^{(j)} |\mu_{k,j} - \mu_{k',j}|.\end{aligned}$$

The $\{\mu_k\}$ update is a quadratic program, which can be solved by a variety of methods. We compute $\{\mu_k\}$ by iterated ridge regression. Following Guo et al. (2010) and Fan & Li (2001), we use the local quadratic approximation

$$\begin{aligned}\left| \mu_{k,j}^{(t,s+1)} - \mu_{k',j}^{(t,s+1)} \right| &\approx \\ &\frac{\left(\mu_{k,j}^{(t,s+1)} - \mu_{k',j}^{(t,s+1)} \right)^2}{2 \left| \mu_{k,j}^{(t,s)} - \mu_{k',j}^{(t,s)} \right|} + \frac{1}{2} \left| \mu_{k,j}^{(t,s)} - \mu_{k',j}^{(t,s)} \right|.\end{aligned}$$

where s is the iteration index for this procedure within the t -th M-step. This approximation is based on the identity

$$|x - y| = \frac{(x - y)^2}{2|x - y|} + \frac{1}{2}|x - y|.$$

Under the approximation, we obtain a generalized ridge regression problem which can be solved in closed form during each iteration s . Note that this approximation is only valid when $|\mu_{k,j}^{(t,s)} - \mu_{k',j}^{(t,s)}| > 0$. For numerical stability, we threshold $|\mu_{k,j}^{(t,s)} - \mu_{k',j}^{(t,s)}|$ to zero at values less than 10^{-10} .

The variance update is standard:

$$\sigma_k^2 = \frac{\sum_t \gamma_k(t) (y_t - \mu_k)^T (y_t - \mu_k)}{\sum_t \gamma_k(t)}.$$

The transition matrix update is

$$\mathbf{T} = \arg \max_{\mathbf{T}} \sum_{ij} \log(\mathbf{T}_{ij}) \sum_t \xi_{ij}(t).$$

Because the Gaussian emission distributions have infinite support, \mathbf{T} is irreducible and aperiodic by construction. However, we must explicitly constrain \mathbf{T} to satisfy detailed balance.

Lemma 1. \mathbf{T} satisfies detailed balance if and only if $\mathbf{T}_{ij} = \frac{\mathbf{W}_{ij}}{\sum_k \mathbf{W}_{ik}}$, where $\mathbf{W} = \mathbf{W}^T$.

Proof. If \mathbf{T} satisfies detailed balance, then let $\mathbf{W}_{ij} = \pi_i \mathbf{T}_{ij} = \pi_j \mathbf{T}_{ji} = \mathbf{W}_{ji}$. Then note

$$\frac{\mathbf{W}_{ij}}{\sum_k \mathbf{W}_{ik}} = \frac{\pi_i \mathbf{T}_{ij}}{\sum_k \pi_i \mathbf{T}_{ik}} = \frac{\mathbf{T}_{ij}}{\sum_k \mathbf{T}_{ik}} = \mathbf{T}_{ij}$$

To prove the converse, assume $\mathbf{T}_{ij} = \frac{\mathbf{W}_{ij}}{\sum_k \mathbf{W}_{ik}}$, with $\mathbf{W} = \mathbf{W}^T$. Let $\pi_i = \sum_k \mathbf{W}_{ik}$. Then $\pi_i \mathbf{T}_{ij} = \mathbf{W}_{ij} = \mathbf{W}_{ji} = \pi_j \mathbf{T}_{ji}$. \square

Substituting the results of Lemma 1, we rewrite the transition matrix update as

$$\mathbf{W} = \arg \max_{\mathbf{W}} \left(\left[\sum_{ij} \log(\mathbf{W}_{ij}) - \log \pi_i \right] \sum_t \xi_{ij}(t) \right).$$

We compute the derivative of the inner term with respect to $\log \mathbf{W}_{ij}$ and optimize with L-BFGS (Nocedal & Wright, 2006).

3.2. Model Selection

There are two free model parameters: K and Δ . The number of metastable states, K , is expected to be small – at most a few dozen. To choose K , we can use the AIC or BIC selection criteria, or alternatively enumerate a few small values.

The choice of Δ is more difficult than the choice of K , as changing the discretization interval alters the support of the likelihood function. Recall that choosing Δ too small results in subsequent samples y_t, y_{t+1} becoming highly correlated, while the model satisfies the conditional independence assumption $Y_t \perp\!\!\!\perp Y_{t+1} | X_t$. Moreover, small Δ increases data-storage requirements, while Δ too large will needlessly discard data. Thus a balance between these two conflicting directives is necessary.

We use the physical criterion of convergence in the relaxation timescales to evaluate when Δ is large enough. The propagation of the dynamics from an initial distribution over the hidden states, X_t , can be described by

$$P(y_{t+n} | x_t) = \sum_{k=0}^{K-1} \mathcal{N}(y_{t+n}; \mu_k, \sigma_k^2) (x_t^T \mathbf{T}^n)_k.$$

Diagonalize \mathbf{T} in terms of its left eigenvectors ϕ_i , right eigenvectors ψ_i , and eigenvalues λ_i (such a diagonalization is always possible since \mathbf{T} is a stochastic matrix).

$$P(y_{t+n} | x_t) = \sum_{k=1}^K \left[\mathcal{N}(y_{t+n}; \mu_k, \sigma_k^2) \left(\sum_{i=1}^K \lambda_i^n \langle x_t, \psi_i \rangle \phi_i \right) \right]_k$$

Since π is the stationary left eigenvector of \mathbf{T} , and the remaining eigenvalues lie in the interval $-1 < \lambda_i < 1$, the collective dynamics can be interpreted as a sum of exponential relaxation processes.

$$P(y_{t+n} | x_t) = \sum_{k=1}^K \left[f_k(y_{t+n}) \left(\pi + \sum_{i=2}^K e^{-n/\tau_i} \langle x_t, \psi_i \rangle \phi_i \right) \right]_k$$

In the equation, we define $f_k(y) = \mathcal{N}(y; \mu_k, \sigma_k^2)$. Each eigenvector of \mathbf{T} (except the first) describes a dynamical mode with characteristic relaxation timescale

$$\tau_i = -\frac{1}{\ln \lambda_i}.$$

The longest timescales, τ_i , are of central interest from a molecular modeling perspective because they describe dynamical modes visible in time-resolved protein experiments (Zhuang et al., 2011) and are robust against perturbations (Weber & Pande, 2012). We choose Δ large enough to converge the τ_i : for adequately large Δ , we expect $\tau_i(\Delta)$ to asymptotically converge to the true relaxation timescale τ_i^* . For simple systems, we may evaluate τ_i^* explicitly, while for larger systems, we choose Δ large enough so that $\tau_i(\Delta)$ no longer changes with further increase in the discretization interval.

3.3. Implementation

We implement learning for both multithreaded CPU and NVIDIA GPU platforms. In the CPU implementation, we parallelize across trajectories during the E-step using OpenMP. The largest portion of the run time is spent in LOG-SUM-EXP operations, which we manually vectorize with SSE2 intrinsics for SIMD architectures. Parallelism on the GPU is more fine grained. The E-step populates two $T \times K \times K$ arrays with forward and backwards sweeps respectively. To fully utilize the GPU's massive parallelism, each trajectory has a team of threads which cooperate on updating the $K \times K$ matrix at each time step. Specialized CUDA kernels were written for $K = 4, 8, 16$ and 32 along with a generic kernel for $K > 32$.

Even in log space, for long trajectories, the forward-backward algorithm can suffer from an accumulation of floating point errors which lead to catastrophic cancellation during the computation of $\gamma_i(t)$. This risk requires that the forward-backward matrices be accumulated in double precision, whereas the rest of the calculation is safe in single precision.

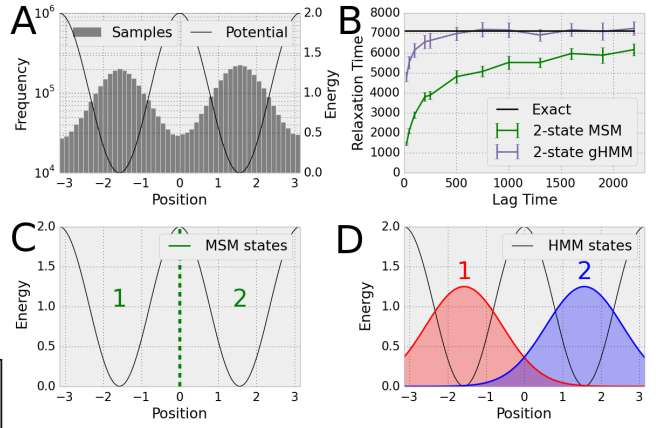


Figure 1. Simulations of Brownian dynamics on a double well potential (A) illustrate the advantages of the HMM over the MSM. When the dynamics are discretized at a time interval of > 500 steps, the 2-state HMM, unlike the 2-state MSM achieves a quantitatively accurate prediction of the first relaxation timescale (B). The MSM (C) features hard cutoffs between the states whereas the HMM (D) each have infinite support.

The speedup using our GPU implementation is $15\times$ compared to our optimized CPU implementation and $75\times$ with respect to a standard numpy implementation using $K = 16$ states on a NVIDIA GTX TITAN GPU / Intel Core i7 4 core Sandy Bridge CPU platform. Further scaling of the implementation could be achieved by splitting the computation over multiple GPUs with MPI.

4. Experiments

4.1. Double Well Potential

We first consider a one-dimensional diffusion process y_t governed by Brownian dynamics. The process is described by the stochastic differential equation

$$\frac{dy_t}{dt} = -\nabla V(y_t) + \sqrt{2D}R(t)$$

where V is the reduced potential energy, D is the diffusion constant, and $R(t)$ is a zero-mean delta-correlated stationary Gaussian process. For simplicity, we set $D = 1$ and consider the double well potential

$$V(y) = 1 + \cos(2y)$$

with reflecting boundary conditions at $y = -\pi$ and $y = \pi$. Using the Euler-Maruyama method and a time step of $\Delta t = 10^{-3}$, we produced ten simulation trajectories of length 5×10^5 steps each. The histogrammed trajectories are shown in Fig. 1(A). The exact value of the first relaxation timescale was computed by a finite element discretization of the corresponding Fokker-Planck equation (Higham, 2001).

We applied both a two-state MSM and two-state HMM, with fusion L_1 regularization parameter $\lambda = 0$, to the simulation trajectories. The MSM states were fixed, with a dividing surface at $y = 0$, as shown in Fig. 1(C). The HMM states were learned, as shown in Fig. 1(D). Both the MSM and the HMM display some sensitivity with respect to the discretization interval, with more accurate predictions of the relaxation timescale at longer lag times.

The two-state MSM is unable to accurately learn the longest timescale, τ_1 , even with large lag times, while the two-state HMM succeeds in identifying τ_1 with $\Delta \geq 500$ Fig. 1(B).

4.2. Ubiquitin

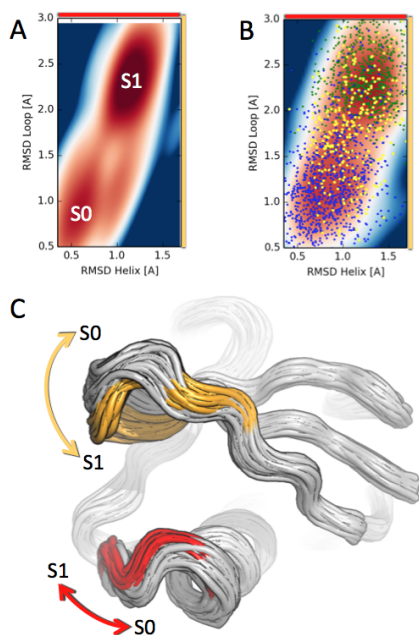


Figure 2. Dynamics of Ub. (A) The HMM identifies the two metastable states of Ub, varying primarily in the loop and helix regions (axes in yellow and red respectively). (B) The MSM fails to cleanly separate the two underlying physical states. Three post-processed macrostates from the MSM are shown (in blue, green, and yellow). (C) A structural rendering of the conformational states of the Ub system. S0, shown in grey, binds to the UCH family of proteins, and S1 (with characteristic structural differences to S0 in red and yellow) binds to the USP family.

Ubiquitin (Ub) is a regulatory hub protein at the intersection of many signaling pathways in the human body (Hershko & Ciechanover, 1998). Among its many tasks are the regulation of inflammation, repair of DNA, and the breakdown and recycling of waste proteins. Ubiquitin interacts with close to 5000 human signaling proteins. Understanding the link between structure and function in ubiquitin would elucidate the underlying framework of the human

signaling network.

We obtained a dataset of MD simulations of human Ub consisting of 3.5 million data points. The protein, shown in Fig. 2, is composed of 75 amino acids. The simulations were performed on the NCSA Blue Waters supercomputer. The resulting structures were featurized by extracting the distance from each amino acid’s central carbon atom to its position in the simulations’ starting configurations. HMMs were constructed with 2 to 6 states. We chose Δ by monitoring the convergence of the relaxation timescales as discussed in Sec. 3.2, and set the L_1 fusion penalty heuristically to a default value of $\lambda = 0.01$. In agreement with existing biophysical data (Zhang et al., 2012), the HMMs correctly determined that Ub was best modeled with 2 states (Fig. 2A). For ease of representation, the learned HMM is shown projected onto two critical degrees of freedom (discussed below).

For comparison, we generated MSM models with 500 microstates (Fig. 2B) and projected upon the same critical degrees of freedom. We used a standard kinetic lumping post-processing step to identify 3 macrostates (shown in green, blue, and yellow respectively); the lumping algorithm collapsed when asked to identify 2 macrostates (Bowman, 2012). Contrast the simple, clean output of the 2 state HMM in Fig. 2(A) with the standard MSM of Fig. 2(B). Note how significant post-processing and manual tuning would be required to piece together the true two-state structural dynamics of Ub from the MSM output.

We display a structural rendering of the Ub system in Fig. 2(C). The imposed L_1 penalty of the HMM suppresses differences among the uninformative degrees of freedom depicted in grey. The remaining portions of the protein (shown in color) reveal the two critical axes of motion of the Ub system: the hinge dynamics of the loop region displayed in yellow and a kink in the lower helix shown in red. We use these axes in the simplified representations shown in Figs. 2(A,B).

The states S0 and S1 identified by the HMM have direct biological interpretations. Comparison to earlier experimental work reveals that configuration S0 binds to the UCH family of proteins, while configuration S1 binds to the USP family instead (Komander et al., 2009). The families play differing roles in the vital task of regenerating active Ub for the cell-signaling cycle.

Together, MD and the HMM analysis provide atomic insight into the effect of protein structure on ubiquitin’s role in the signaling network. Our analysis approach may have significant value for protein biology and for the further study of cellular signaling networks. Although experimental studies of protein signaling provide the gold standard for hard data, they struggle to provide structural explanations

— knowing why a certain protein is more suited for certain signaling functions is challenging at best. In contrast, the MD/HMM approach can provide a direct link between structure and function and give a causal basis for observed protein activity.

4.3. c-Src Tyrosine Kinase

Protein kinases are a family of enzymes that are critical for regulating cellular growth whose aberrant activation can lead to uncontrolled cellular proliferation. Because of their central role in cell proliferation, kinases are a critical target for anti-cancer therapeutics. The c-Src tyrosine kinase is a prominent member of this family that has been implicated in numerous human malignancies (Aleshin & Finn, 2010).

Due to the protein’s size and complexity, performing MD simulations of the c-Src kinase is a formidable task. The protein, shown in Fig. 3A, consists of 262 amino acids; when surrounding water molecules – necessary for accurate simulation – are taken into account, the system has over 40,000 atoms. Furthermore, transition between the active and inactive states takes hundred of microseconds. Adequate sampling of these processes therefore requires hundreds of billions of MD integrator steps. Simulations of the c-Src kinase were performed on the Folding@Home distributed computing network, collecting a dataset of 4.7 million configurations from 550 μs of sampling, for a total of 108 GB of data (Shukla et al., 2014).

In order to understand the molecular activation mechanism of the c-Src kinase, we analyzed this dataset using the L_1 regularized reversible HMM. We featurized the configurations by extracting the distance from each amino acid’s central carbon atom to its position in an experimentally determined inactive configuration. We built HMMs with 2 to 6 states, and singled out the 3 state model for achieving a balance of complexity and interpretability. As with Ub, we chose Δ by monitoring the convergence of the relaxation timescales, same default L_1 fusion penalty of $\lambda = 0.01$.

The L_1 -regularized reversible HMM elucidates the c-Src kinase activation pathway, revealing a stepwise mechanism of the dynamics. A projection of the learned HMM states onto two key degrees of freedom is shown in Fig. 3B. Fig. 3C shows a structural representation of the means of the three states, highlighting a sequential activation mechanism. The transformation from the inactive to the intermediate state occurs first by the unfolding of the A-loop (the subsection of the protein highlighted in red). Activation is completed by the inward rotation of the C-helix (highlighted in orange) and rupture of a critical side chain interaction between two amino acids on the C-helix and the A-loop respectively.

Although the protein structure is complex, the activation

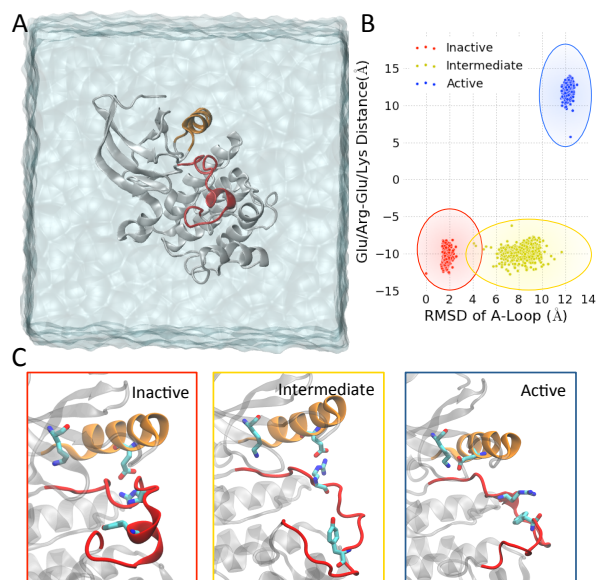


Figure 3. Activation of the c-Src Kinase. (A) Structure of the protein system. (B) The 3 state HMM, projected onto two degrees of freedom representing the positions of the A-loop (shown in red) and C-helix (shown in orange) respectively. (C) Structural renderings of the means of the hidden states showing atomistic details of the activation pathway.

process takes place only in a small portion of the overall protein; the random fluctuations of the remaining degrees of freedom are largely uncoupled from the activation process. As with Ub, the L_1 penalty suppresses the signal from unimportant degrees of freedom shown in grey. In contrast to the simplicity of HMM approach, a recent MSM analysis of this dataset found similar results, but required 2,000 microstates and significant post-processing of the models to generate physical insight into the activation mechanism (Shukla et al., 2014).

The identification of the intermediate state along the activation pathway has substantial implications in the field of rational drug design. Chemotherapy drugs often have harmful side effects because they target portions of proteins that are common across entire families, interfering with both the uncontrolled behavior of tumor proteins as well as the critical cellular function of healthy proteins. Intermediate states, such as the one identified by the HMM, are more likely to be unique to each kinase protein; future therapeutics that target these intermediate states could have significantly fewer deleterious side effects (Fang et al., 2013).

5. Discussion and Conclusion

Currently, MSMs are a dominant framework for analyzing protein dynamics datasets. We propose replacing this

methodology with L_1 -regularized reversible HMMs. We show that HMMs have significant advantages over MSMs: whereas the MSM state decomposition is a preprocessing procedure without guidance from a complete-data likelihood function, the HMM couples the identification of metastable states with the estimation of transition probabilities. As such, accurate models require fewer states, aiding interpretability from a physical perspective.

The switch is not without tradeoffs. MSMs are backed by a significant body of theoretical work: the MSM is a direct discretization of an integral operator which formally controls the long timescale dynamics known as the transfer operator. This connection enables the quantification of approximation error in the MSM framework (Prinz et al., 2011). No such theoretical guarantees yet exist for the L_1 -regularized reversible HMM because the evolution of Y_t is no longer unconditionally Markovian. However, because the HMM can be viewed as a generalized hidden MSM, there is reason to believe that analogues of MSM theoretical guarantees extend to the HMM framework.

While the L_1 -regularized reversible hidden Markov model represents an improvement over previous methods for analyzing MD datasets, future work will likely confront a number of remaining challenges. For example, the current model does not learn the featurization and treats Δ as a hyperparameter. Bringing these two aspects of the model into the optimization framework would reduce the required amount of manual tuning. Adapting techniques from Bayesian nonparametrics, unsupervised feature learning and linear dynamical systems may facilitate the achievement of these goals.

Our results show that structured statistical analysis of massive protein datasets is now possible. We reduce complex dynamical systems with thousands of physical degrees of freedom to simple statistical models characterized by a small number of metastable states and transition rates. The HMM framework is a tool for turning raw molecular dynamics data into scientific knowledge about protein structure, dynamics and function. Our experiments on the ubiquitin and c-Src kinase proteins extract insight that may further the state of the art in cellular biology and rational drug design.

Acknowledgments

We thank Diwakar Shukla for kindly providing the c-Src kinase MD trajectories. B.R. was supported by the Fannie and John Hertz Foundation. G.K and V.S.P acknowledge support from the Simbios NIH Center for Biomedical Computation (NIH U54 Roadmap GM072970). V.S.P. acknowledges NIH R01-GM62868 and NSF MCB-0954714.

References

- Aleshin, A. and Finn, R. S. SRC: a century of science brought to the clinic. *Neoplasia*, 12(8):599–607, 2010.
- Baldi, Pierre and Pollastri, Gianluca. The principled design of large-scale recursive neural network architectures—DAG-RNNs and the protein structure prediction problem. *J. Mach. Learn. Res.*, 4:575–602, 2003.
- Beauchamp, Kyle A., Bowman, Gregory R., Lane, Thomas J., Maibaum, Lutz, Haque, Imran S., and Pande, Vijay S. MSMBuilder2: Modeling conformational dynamics on the picosecond to millisecond scale. *J. Chem. Theory Comput.*, 7(10):3412–3419, 2011.
- Beauchamp, Kyle A., Lin, Yu-Shan, Das, Rhiju, and Pande, Vijay S. Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements. *J. Chem. Theory Comput.*, 8(4):1409–1414, 2012.
- Bowman, Gregory R. Improved coarse-graining of markov state models via explicit consideration of statistical uncertainty. *J. Chem. Phys.*, 137(13):134111, 2012.
- Cho, Samuel S., Levy, Yaakov, and Wolynes, Peter G. P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes. *Proc. Natl. Acad. Sci. U.S.A.*, 103(3):586–591, 2006.
- Chu, Wei, Ghahramani, Zoubin, Podtelezhnikov, Alexei, and Wild, David L. Bayesian segmental models with multiple sequence alignment profiles for protein secondary structure and contact map prediction. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 3(2):98–113, 2006.
- Di Lena, Pietro, Baldi, Pierre, and Nagata, Ken. Deep spatio-temporal architectures and learning for protein structure prediction. In *Adv. Neural Inf. Process. Syst.* 25, NIPS ’12, pp. 521–529, 2012.
- Dill, Ken A, Bromberg, Sarina, Yue, Kaizhi, Chan, Hue Sun, Ftebig, Klaus M, Yee, David P, and Thomas, Paul D. Principles of protein folding – a perspective from simple exact models. *Protein Sci.*, 4(4):561–602, 1995.
- Fan, Jianqing and Li, Runze. Variable selection via non-concave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.*, 96(456):1348–1360, 2001.
- Fang, Zhizhou, Grtter, Christian, and Rauh, Daniel. Strategies for the selective regulation of kinases with allosteric modulators: Exploiting exclusive structural features. *ACS Chem. Biol.*, 8(1):58–70, 2013.
- Guo, Jian, Levina, Elizaveta, Michailidis, George, and Zhu, Ji. Pairwise variable selection for high-dimensional model-based clustering. *Biometrics*, 66(3):793–804, 2010.

- Hershko, Avram and Ciechanover, Aaron. The ubiquitin system. *Annu. Rev. Biochem.*, 67(1):425–479, 1998.
- Higham, Desmond J. An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Rev.*, 43(3):525–546, 2001.
- Humphrey, William, Dalke, Andrew, and Schulten, Klaus. VMD: Visual molecular dynamics. *J. Mol. Graphics*, 14(1):33 – 38, 1996.
- Karplus, M. and Kuriyan, J. Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. U.S.A.*, 102(19): 6679–6685, 2005.
- Kohlhoff, Kai J, Shukla, Diwakar, Lawrenz, Morgan, Bowman, Gregory R, Konerding, David E, Belov, Dan, Altman, Russ B, and Pande, Vijay S. Cloud-based simulations on Google Exacycle reveal ligand modulation of gpcr activation pathways. *Nat. Chem.*, 6(1):15–21, 2014.
- Komander, David, Clague, Michael J, and Urbé, Sylvie. Breaking the chains: structure and function of the deubiquitinases. *Nat. Rev. Mol. Cell Biol.*, 10(8):550–563, 2009.
- Lane, Thomas J, Shukla, Diwakar, Beauchamp, Kyle A, and Pande, Vijay S. To milliseconds and beyond: challenges in the simulation of protein folding. *Curr. Opin. Struct. Biol.*, 23(1):58 – 65, 2013.
- Maity, Haripada, Maity, Mita, Krishna, Mallela M. G., Mayne, Leland, and Englander, S. Walter. Protein folding: The stepwise assembly of foldon units. *Proc. Natl. Acad. Sci. U.S.A.*, 102(13):4741–4746, 2005.
- Nocedal, J. and Wright, S. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2nd edition, 2006.
- Noé, Frank, Wu, Hao, Prinz, Jan-Hendrik, and Plattner, Nuria. Projected and hidden markov models for calculating kinetics and metastable states of complex molecules. *J. Chem. Phys.*, 139(18):184114, 2013.
- Prinz, Jan-Hendrik, Wu, Hao, Sarich, Marco, Keller, Bettina, Senne, Martin, Held, Martin, Chodera, John D., Schütte, Christof, and Noé, Frank. Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.*, 134(17):174105, 2011.
- Rabiner, Lawrence R. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, 1989.
- Sadiq, S. Kashif, Noé, Frank, and De Fabritiis, Gianni. Kinetic characterization of the critical step in HIV-1 protease maturation. *Proc. Natl. Acad. Sci. U.S.A.*, 109(50): 20449–20454, 2012.
- Shaw, David E. et al. Anton, a special-purpose machine for molecular dynamics simulation. In *ACM Comp. Ar. 34, ISCA '07*, pp. 1–12, 2007.
- Shirts, Michael and Pande, Vijay S. Screen savers of the world unite! *Science*, 290(5498):1903–1904, 2000.
- Shukla, Diwakar, Meng, Yilin, Roux, Benoît, and Pande, Vijay S. Activation pathway of Src kinase reveals intermediate states as novel targets for drug design. *Nat. Commun.*, 5, 2014.
- Sontag, David, Meltzer, Talya, Globerson, Amir, Jaakkola, Tommi S, and Weiss, Yair. Tightening LP relaxations for MAP using message passing. *arXiv preprint arXiv:1206.3288*, 2012.
- Tibshirani, Robert, Saunders, Michael, Rosset, Saharon, Zhu, Ji, and Knight, Keith. Sparsity and smoothness via the fused lasso. *J. R. Statistic. Soc. B*, 67(1):91–108, 2005.
- Voelz, Vincent A., Jäger, Marcus, Yao, Shuhuai, Chen, Yujie, Zhu, Li, Waldauer, Steven A., Bowman, Gregory R., Friedrichs, Mark, Bakajin, Olgica, Lapidus, Lisa J., Weiss, Shimon, and Pande, Vijay S. Slow unfolded-state structuring in Acyl-CoA binding protein folding revealed by simulation and experiment. *J. Am. Chem. Soc.*, 134(30):12565–12577, 2012.
- Weber, Jeffrey K and Pande, Vijay S. Protein folding is mechanistically robust. *Biophys. J.*, 102(4):859–867, 2012.
- Wong, Chung F. and McCammon, J. Andrew. Protein flexibility and computer-aided drug design. *Annu. Rev. Pharmacol. Toxicol.*, 43(1):31–45, 2003.
- Zhang, Yingnan, Zhou, Lijuan, Rouge, Lionel, Phillips, Aaron H, Lam, Cynthia, Liu, Peter, Sandoval, Wendy, Helgason, Elizabeth, Murray, Jeremy M, Wertz, Ingrid E, et al. Conformational stabilization of ubiquitin yields potent and selective inhibitors of USP7. *Nat. Chem. Biol.*, 9(1):51–58, 2012.
- Zhuang, Wei, Cui, Raymond Z., Silva, Daniel-Adriano, and Huang, Xuhui. Simulating the T-jump-triggered unfolding dynamics of trpzip2 peptide and its time-resolved IR and two-dimensional IR signals using the markov state model approach. *J. Phys. Chem. B*, 115(18):5415–5424, 2011.