nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-13307-x OPEN

Learning dynamical information from static protein and sequencing data

Philip Pearce[1], Francis G. Woodhouse[2], Aden Forrow[1,2], Ashley Kelly[3], Halim Kusumaatmaja[3*] & Jörn Dunkel[1*]

Many complex processes, from protein folding to neuronal network dynamics, can be described as stochastic exploration of a high-dimensional energy landscape. Although efficient algorithms for cluster detection in high-dimensional spaces have been developed over the last two decades, considerably less is known about the reliable inference of state transition dynamics in such settings. Here we introduce a flexible and robust numerical framework to infer Markovian transition networks directly from time-independent data sampled from stationary equilibrium distributions. We demonstrate the practical potential of the inference scheme by reconstructing the network dynamics for several protein-folding transitions, gene-regulatory network motifs, and HIV evolution pathways. The predicted network topologies and relative transition time scales agree well with direct estimates from time-dependent molecular dynamics data, stochastic simulations, and phylogenetic trees, respectively. Owing to its generic structure, the framework introduced here will be applicable to high-throughput RNA and protein-sequencing datasets, and future cryo-electron microscopy (cryo-EM) data.

[1] Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139-4307, USA. [2] Mathematical Institute, University of Oxford, Andrew Wiles Building, Woodstock Road, Oxford OX2 6GG, UK. [3] Department of Physics, Durham University, South Road, Durham DH1 3LE, UK. *email: halim.kusumaatmaja@durham.ac.uk; dunkel@mit.edu

Energy landscapes encapsulate the effective dynamics of a wide variety of physical, biological, and chemical systems[1,2]. Well-known examples include a myriad of biophysical processes[3–7], multi-phase systems[2], thermally activated hopping in optical traps[8,9], chemical reactions[1,10], brain neuronal expression[11], cellular development[12–16], and social networks[17]. Energetic concepts have also been connected to machine learning[18] and to viral fitness landscapes, where pathways with the lowest energy barriers may explain typical mutational evolutionary trajectories of viruses between fitness peaks[19,20]. Recent advances in experimental techniques including cryo-electron microscopy (cryo-EM)[3,21,22] and single-cell RNA-sequencing[23], as well as new online social interaction datasets[24], are producing an unprecedented wealth of high-dimensional instantaneous snapshots of biophysical and social systems. Although much progress has been made in dimensionality reduction[25–27] and the reconstruction of effective energy landscapes in these settings[3,13,16,17,28], the problem of inferring dynamical information such as protein-folding or mutation pathways and rates from instantaneous ensemble data remains a major challenge.

To address this practically important question, we introduce here an integrated computational framework for identifying metastable states on reconstructed high-dimensional energy landscapes and for predicting the relative mean first passage times (MFPTs) between those states, without requiring explicitly time-dependent data. Our inference scheme employs an analytic representation of the data based on a Gaussian mixture model (GMM)[29] to enable efficient identification of minimum-energy transition pathways[30–32]. We show how the estimation of transition networks can be optimized by reducing the dimension of a high-dimensional landscape while preserving its topology. Our algorithm utilizes experimentally validated analytical results[8,9] for transition rates[1,33–35]. Thus, it is applicable whenever the time evolution of the underlying system can be approximated by a Fokker–Planck-type Markovian dynamics, as is the case for a wide range of physical, chemical, and biological processes[1,34].

Specifically, we illustrate the practical potential by inferring protein-folding transitions, state-switching in gene-regulatory networks, and HIV evolution pathways. Current standard methods for coarse-graining the conformational dynamics of biophysical structures[36,37] typically estimate Markovian transition rates from time-dependent trajectory data in large-scale molecular dynamics (MD) simulations[38,39]. By contrast, we show here that protein-folding pathways and rates can be recovered without explicit knowledge of the time-dependent trajectories, provided the system is sufficiently ergodic and equilibrium distributions are sampled accurately. Furthermore, we show that the dynamics of state-switching or phenotype-switching in gene-regulatory networks[40] can be inferred directly from static snapshots of protein abundances in regimes where deterministic modeling only captures a single steady state[41,42]. The agreement of our inferred results with two separate sets of time-dependent measurements suggests that the inference of complex transition networks via reconstructed energy landscapes can provide a viable and often more efficient alternative to traditional time-series estimates, particularly as new experimental techniques will offer unprecedented access to high-dimensional ensemble data.

Results

Minimum-energy-path network reconstruction. The equilibrium distribution $p(\mathbf{x})$ of a particle diffusing over a potential energy landscape $E(\mathbf{x})$ is the Boltzmann distribution $p(\mathbf{x}) = \exp[-E(\mathbf{x})/k_\mathrm{B}T]/Z$, where k_B is the Boltzmann constant, T is the temperature, and Z is a normalization constant. Given the probability density function (PDF) $p(\mathbf{x})$, the effective energy can be inferred from

$$E(\mathbf{x}) = -k_\mathrm{B}T\ln[p(\mathbf{x})/p_\mathrm{max}], \qquad (1)$$

where p_max is the maximum value of the PDF, included to fix the minimum energy at zero. Our goal is to estimate the MFPTs between minima on the landscape using only sampled data. We divide this task into three steps, as illustrated in Fig. 1 for test data (Supplementary Methods). In the first step, we approximate the empirical PDF by using the expectation maximization algorithm to fit a GMM in a space of sufficiently large dimension d (Methods and Fig. 1a). Mixtures with a bounded number of components can be recovered in time polynomial in both the dimension d and the required accuracy[43]. The resulting GMM yields an analytical expression for $E(\mathbf{x})$ via Eq. (1).

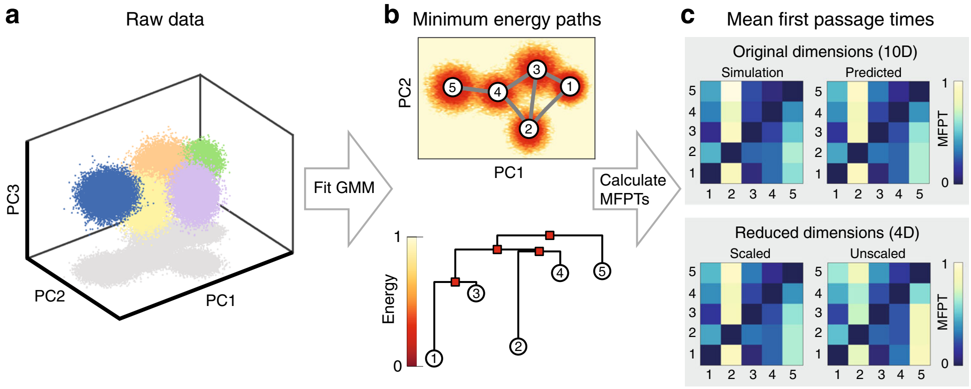

Fig. 1 Inference scheme for estimating transition networks and mean first passage times (MFPTs). We apply the protocol to test data generated from a Gaussian Mixture Model (GMM; Supplementary Methods). **a** Inputs are the instantaneously measured data, sampled here from a ten-dimensional GMM with five Gaussians, plotted in the first three principal components (PCs); colors denote the Gaussian that a point was sampled from. **b** Top: a GMM is fit to the samples to construct the empirical probability distribution, which is then converted to the energy landscape using Eq. (1). Background color indicates the projection of the empirical energy landscape onto the first two PCs. Minimum-energy paths (MEPs, gray lines) between minima 1–5 on the landscape are calculated using the NEB algorithm (Supplementary Methods). Bottom: disconnectivity graph illustrating minima on the energy landscape (circles) and saddle points between them (squares). **c** A Markov state model (MSM) is constructed with transition rates given by Eq. (2) and solved to predict the MFPTs between discrete states (top right; Methods). MFPTs predicted by the MSM agree with direct estimates from Brownian dynamics simulations in the inferred energy landscape (top left; Supplementary Methods). MFPTs calculated in a reduced four-dimensional space using the scaling given in Eq. (3) recover the MFPTs accurately (bottom left). Without the appropriate scaling, the predicted MFPTs are inaccurate (bottom right).

In the second step, the inferred energy landscape $E(\mathbf{x})$ is reduced to a minimum-energy-path (MEP) network whose nodes (states) are the minima of $E(\mathbf{x})$ (Fig. 1b top). Each edge represents an MEP that connects two adjacent minima and passes through an intermediate saddle point (Fig. 1b). The MEPs are found using the nudged elastic band (NEB) algorithm[30,31], which discretizes paths with a series of bead-spring segments (Supplementary Methods).

Markov state model. Given the MEP network, the final step is to infer the rates for transitioning from a minimum α to an adjacent minimum β. Assuming overdamped Brownian dynamics, the directed transition $\alpha \to \beta$ can be characterized by the generalized Kramers transition rate[1]

$$k_{\alpha\beta} = \frac{\omega_\mathrm{b}}{2\pi\gamma}\frac{\prod_i \omega_i^\alpha}{\prod_i' \omega_i^S}\exp(-E_\mathrm{b}/k_\mathrm{B}T), \tag{2}$$

where γ is the effective friction, E_b is the energy difference between the saddle point S on the MEP (over the energy barrier) and the minimum α, ω_i^α are the stable angular frequencies at the minimum α, and ω_i^S and ω_b are the stable and unstable angular frequencies at the saddle, respectively. Equation (2) assumes isotropic friction but can be generalized to a tensorial form[1] if anisotropies are relevant. In most practical applications, the error from assuming γ to be isotropic is likely negligible compared with other experimental noise sources. In principle, Eq. (2) can be refined further by including quartic (or higher) corrections to the prefactor ω_b/γ to account for details of the saddle shape[1]. Such corrections can be significant for GMMs (Supplementary Methods).

Each edge $(\alpha\beta)$ has two weights, $k_{\alpha\beta}$ and $k_{\beta\alpha}$, assigned to it. The rate matrix $(k_{\alpha\beta})$ completely specifies the Markov state model (MSM) on the network. Solving the MSM yields the matrix of pairwise MFPTs between states (Fig. 1c and Methods). In a simple two-state system, the MFPTs are determined up to a time scale by detailed balance, but for three or more states the influence of landscape topography and the associated state network topology (Methods) can lead to interesting hierarchical ordering of passage times. Identifying these hierarchies and ways to manipulate them is the key to controlling protein-folding or viral evolution pathways.

Topology-preserving dimensionality reduction. To ensure that the inference protocol can be efficiently applied to larger systems with a high-dimensional energy landscape, we derive a general method for reducing the dimension D of an energy landscape while preserving its topology. A PDF with C well-separated Gaussians in D dimensions can be projected onto the $d = C - 1$ dimensional hyperplane spanning the Gaussian means using principal component analysis (PCA); projecting onto a hyperplane of dimension $d - 1$ risks losing information about the relative positions of the Gaussian means and, in general, does not allow a correct recovery of the MFPTs (Supplementary Methods). In practice, it suffices to choose C to be larger than the number of energy minima if their number is not known in advance.

To preserve the topology under such a transformation—which is essential for the correct preservation of energy barriers and MEPs in the reduced-dimensional space—one needs to rescale GMM components in the low-dimensional space depending on the covariances of the Gaussians in the $D - d$ neglected dimensions (Fig. 1c). Explicitly, one finds that within the subspace spanned by the retained principal components (Supplementary Methods)

$$p(\mathbf{x}_D) = \sum_{i=1}^{C}\phi_i\, p_i^d(\mathbf{x}_d)\,\frac{\sqrt{\det\left(2\pi\mathbf{U}_d^T\boldsymbol{\Sigma}_i\mathbf{U}_d\right)}}{\sqrt{\det(2\pi\boldsymbol{\Sigma}_i)}} \tag{3}$$

as long as p satisfies certain minimally restrictive conditions (Supplementary Methods). Here, $\mathbf{U}_d$ denotes the first $d = C - 1$ columns of the matrix of sorted eigenvectors $\mathbf{U}$ of the covariance matrix of the Gaussian means, and ϕ_i, p_i^d and $\boldsymbol{\Sigma}_i$ are the mixing components, reduced-dimensional PDF, and the covariance matrix of each individual Gaussian in the mixture, respectively (Supplementary Methods). Neglecting the determinant scale factors in Eq. (3), as is often done when GMM models are fitted to PCA-projected data, leads to inaccurate MFPT estimates (Fig. 1c, bottom). It is noteworthy that Eq. (3) does not represent inversion of the transformation performed on the data by PCA, unless all D dimensions are retained; if some dimensions are neglected, Eq. (3) represents a rescaling of the marginal distribution in the retained dimensions to reconstruct the PDF in the original dimension. In other words, the transition rates are best recovered from the conditional—not marginal—distributions, which are given by Eq. (3) up to a constant factor that does not affect energy differences.

Dimensionality reduction can substantially improve the efficiency of the NEB algorithm step as follows: when the MEPs in the reduced d-dimensional space have been computed, the identified minima and saddles can be transformed back into the original data dimension D to calculate the Hessian matrices at these points, allowing Kramers' rates to be calculated as usual (Fig. 1c and Supplementary Methods). Alternatively, in specific situations where the MEPs lie outside the hyperplane spanning the means (Supplementary Methods), the MEP in the reduced d-dimensional space can be transformed back to the D-dimensional space and can be used as an initial condition in that space, significantly reducing computational cost. These results present a step towards a general protocol for identifying reaction coordinates or collective variables for projection of a high-dimensional landscape onto a reduced space, while quantitatively preserving the topology of the landscape.

Protein folding. To illustrate the vast practical potential of the above scheme, we demonstrate the successful recovery of several protein-folding pathways, using data from previous large-scale MD simulations[38]. The protein trajectories, consisting of the time-dependent coordinates of the alpha carbon backbone, were pre-processed, subsampled by a factor of 5, treated as a set of static equilibrium measurements, and reduced in dimension before fitting a GMM (Methods). As is typical for high-dimensional parameter estimation with few structural assumptions, the fitting error due to a finite sample size n in d dimensions scales approximately as $\sqrt{d/n}$ (Supplementary Methods); see refs. [44–46] for advanced techniques tackling sample-size limitations. Here, $d < 10$ so the sample size $n \sim 10^5$ suffices for effective recovery; indeed, our results were found to be robust for trajectories further subsampled by up to a factor of 25, leaving around 500 samples per Gaussian (Supplementary Fig. 3).

For each of the four analyzed proteins Villin, BBA, NTL9, and WW, the reconstructed energy landscapes reveal multiple states including a clear global minimum corresponding to the folded state (Fig. 2a, b). To estimate MFPTs, we determined the effective friction γ in Eq. (2) for each protein from the condition that the line of best fit through the predicted vs. measured MFPTs has unit gradient. Although not usually known, γ could in principle be calculated by incorporating time-dependent information from MD simulations or experimental data. Our MFPT predictions

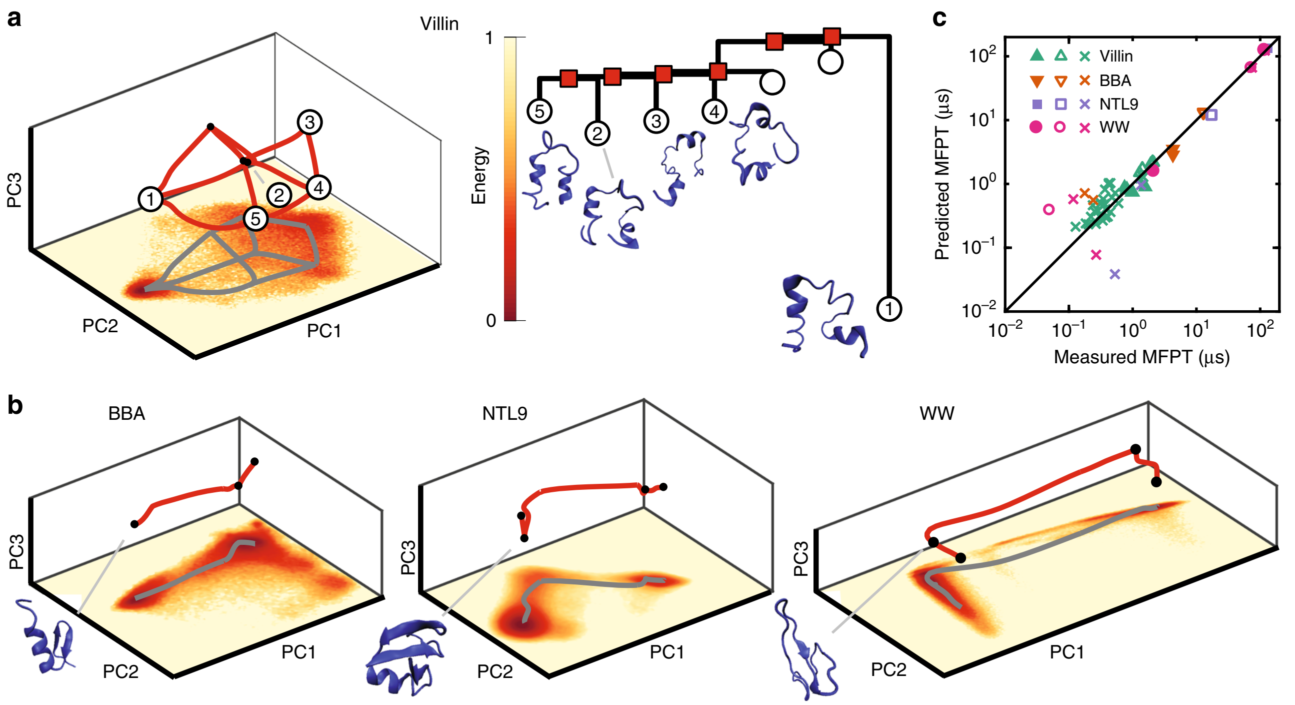

Fig. 2 Reconstructed MEP networks for protein-folding transitions. We compare predicted MFPTs with direct estimates from molecular dynamics (MD) simulations (Supplementary Methods). **a** Left: low-energy states and transition network in the first three principal components (PCs) for Villin including predicted transition paths between states (red lines); bottom coloring shows two-dimensional projection of the empirical energy landscape onto the first two PCs. Right: associated disconnectivity graph and illustrations of the five lowest energy states, with state 1 corresponding to the folded state. **b** Low-energy states, transition paths, and empirical energy landscape for BBA, WW, and NTL9 proteins, and sketches of their folded states. **c** Predicted MFPTs agree well with estimates from MD simulations when energy minima are well separated and become less accurate for fast transitions with small MFPTs. Filled shapes correspond to transitions ending in the unfolded state and unfilled shapes correspond to transitions ending in the folded state (Supplementary Methods), for Villin, BBA, NTL9, and WW. Crosses correspond to transitions between intermediate states.

agree well with direct estimates (Supplementary Methods) from the time-dependent MD trajectories (Fig. 2c). Detailed analysis confirms that the MFPT estimates are robust under variations of the number of Gaussians used in the mixture (Supplementary Fig. 1). Also, the estimated MEPs are in good agreement with the typical transition paths observed in the MD trajectories (Supplementary Fig. 2).

Gene-regulatory networks. Next, we demonstrate the ability of our protocol to infer state-switching pathways in multistable gene-regulatory networks. Using a Gillespie stochastic simulation algorithm (SSA; Methods), we simulated three repressilator-type gene-regulatory network motifs[47] with self-activation. Gene network motifs with features such as these have been studied extensively in recent years, owing to their ability to exhibit precise oscillations[48] and to their possible importance in the determination of multiple cell fates[49] in the appropriate parameter regimes, although the role of noise in such networks is not well understood. In our simulated gene networks, each gene encodes a protein that activates the expression of its associated gene and represses another, with $D=2$, 3, and 4 dimensions at low molecule numbers (Fig. 3a and Supplementary Methods). In each case, parameters were chosen to preclude oscillatory dynamics (Fig. 3a). The energy landscapes reconstructed from the simulation datasets in protein molecule-number space (with time-dependence removed) revealed multiple metastable states for each network (Fig. 3b and Supplementary Fig. 5). Broadly, we found each state to correspond to a mixture of low and high abundances for each separate protein, with the two most common states in $D=4$ dimensions consisting of two abundant and two depleted proteins (Fig. 3b). In agreement with previous studies[41,42], the identified metastable states were not recovered from deterministic simulations of the governing ordinary differential equations (Supplementary Methods), but could only be identified directly from the stochastic data (Fig. 3a, b). We determined the effective friction γ in Eq. (2) for each D as in the protein example. The predicted MFPTs and MEPs between each metastable state were found to be accurate in comparison with time-dependent measurements (Fig. 3c and Supplementary Fig. 5b) and were robust to measurement noise typically encountered in single-cell sequencing (Supplementary Fig. 6). Our framework also correctly predicted MFPTs for a 5D asymmetric gene network (Supplementary Fig. 7). These results demonstrate the utility of our protocol for gene-regulatory network datasets and, more generally, energy landscapes in discrete spaces.

Viral evolution. As a final proof-of-concept application, we demonstrate that our inference scheme recovers the expected evolution pathways between HIV sequences as well as the key features of a distance-based phylogenetic tree (Fig. 4). To this end, we reconstructed an effective energy landscape from publicly available HIV sequences sampled longitudinally at several points in time from multiple patients[50], assuming that the frequency of an observed genotype is proportional to its probability of fixation and that the high-dimensional discrete sequence space can be projected onto a continuous reduced-dimensional phenotype space (Fig. 4a and Supplementary Methods). First, a Gaussian was fit to each patient and then combined in a GMM with equal weights, to avoid bias in the fitness landscape towards sequences

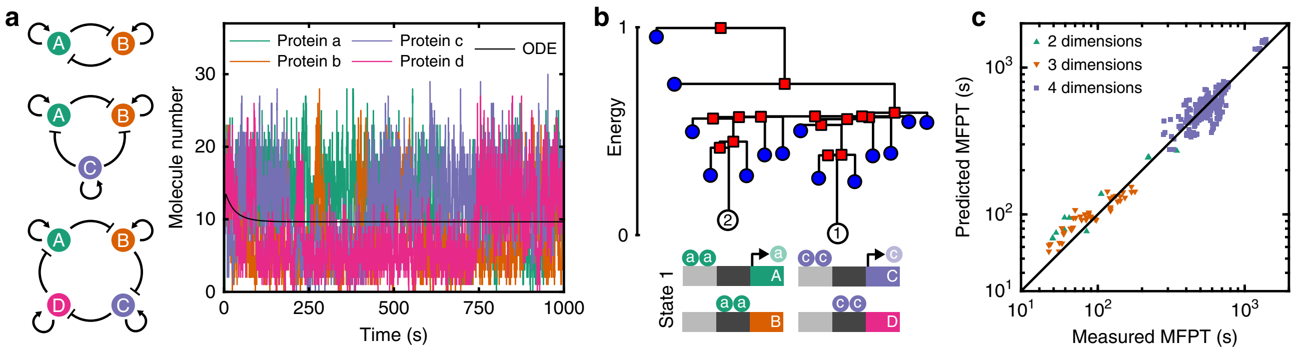

Fig. 3 Reconstructed MEP networks for multistable gene-regulatory networks. We compare predicted MFPTs with direct estimates from stochastic simulations (Methods). **a** We performed simulations of three repressilator-type gene-regulatory network motifs with self-activation (left), consisting of two (top), three (middle), and four (bottom) genes, denoted A, B, C, and D. In the stochastic simulations (right), the numbers of each protein fluctuate between metastable states, but deterministic simulations of the system of ordinary differential equations (ODEs) at low molecule numbers are not able to identify the states, instead converging to a single steady state where all protein numbers are equal. **b** Disconnectivity graph and illustration of the identified lowest energy state for the four-dimensional system. In the lowest energy state (State 1), larger numbers of proteins a and c are present, leading to the activation of their associated genes and repressing the expression of proteins b and d. The next lowest energy state (State 2) is the converse scenario, with large numbers of proteins b and d present. **c** Predicted MFPTs agree well with those calculated from the time-dependent stochastic simulations for all three of the network motifs.

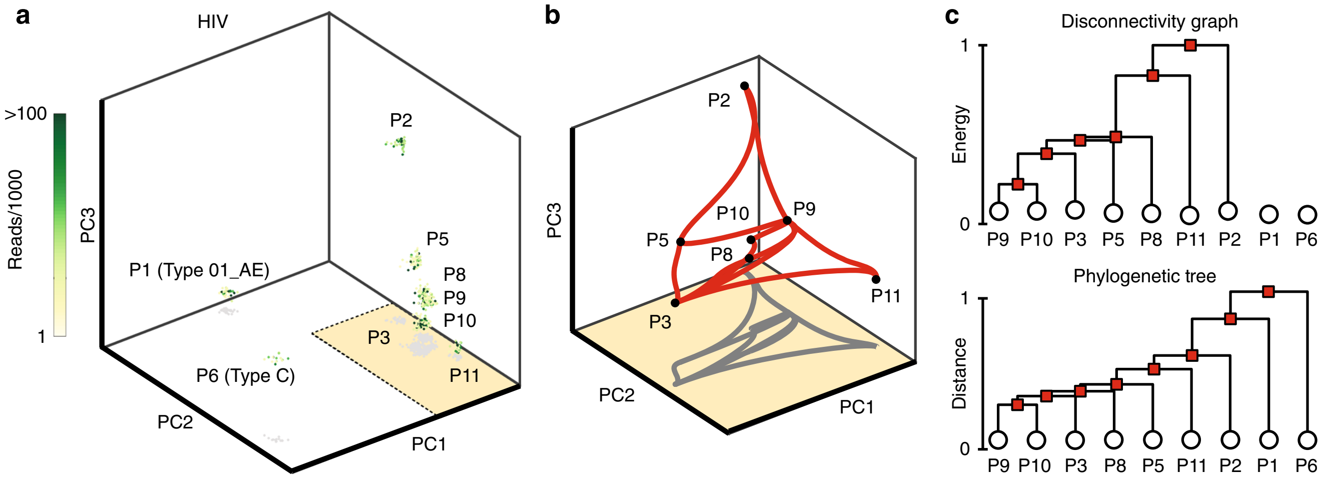

Fig. 4 MEPs on viral fitness landscapes reconstructed from publicly available HIV sequencing data[50]. **a** Longitudinal samples of the HIV virus are binarized after multiple sequence alignment (Supplementary Methods) and plotted in the first three PCs. Samples of the same HIV subtype are closer in PC-space. Patient labels correspond to those used in ref. [50]. **b** MEPs between minima corresponding to patients infected with Type B HIV, plotted in the first three PCs. Paths between minima indicate likely evolutionary pathways. Minima corresponding to patients with Type 01_AE and Type C HIV were unconnected to the other minima. **c** Disconnectivity graph for connected minima, where vertical evolution frequency is assumed to be proportional to the normalized energy barriers (top). The disconnectivity graph reproduces the majority of the structure of a distance-based phylogenetic tree (bottom), where the lengths of vertical lines are proportional to the Jukes-Cantor sequence distance (scaled to $[0, 1]$).

infecting any specific patient (Supplementary Methods). Thereafter, we applied our inference protocol to reconstruct the effective energy landscape, transition network (Fig. 4b), and disconnectivity graph (Fig. 4c), where each state is associated to a separate patient. As expected, states corresponding to patients infected with different HIV subtypes are not connected by MEPs (Fig. 4a, b). The disconnectivity graph reproduces the key features of a coarse-grained patient-level representation of the phylogenetic tree (Fig. 4c). Using our inference scheme, vertical evolution in the tree can be tracked along the MEPs in a reduced-dimensional sequence space (Fig. 4b). The energy barriers, represented by the lengths of the vertical lines in the disconnectivity graph (Fig. 4c), provide an estimate for the relative likelihood of evolution to fixation via point mutations between fitness peaks (energy minima). If mutation rates are known, the MEPs can also be used to estimate the time for evolution to fixation from one fitness peak to another[51].

Discussion

Finding the appropriate number of collective macro-variables to describe an energy landscape is a generic problem relevant to many fields. For example, although some proteins can be described through effective one-dimensional reaction coordinates[5,7,52,53], the accurate description of their diffusive dynamics over the full microscopic energy landscape requires many degrees of freedom[54,55]. Whenever dynamics are inherently high-dimensional, topology-preserving dimensionality reduction can enable a much faster search of the energy landscape for minima and MEPs. In practice, data dimension is often reduced with PCA or similar methods before constructing an energy landscape[55–62]. The extent to which commonly used dimensionality reduction techniques alter MEP network topology or quantitatively preserve energy barriers is not well understood. Equation (3) suggests that reducing dimensions using PCA should not introduce significant errors if the variance of the landscape around each state (energy minimum) in the neglected dimensions is similar. For instance, we found that the protein-folding data could be reduced to five dimensions while maintaining accuracy (Supplementary Fig. 1), although additional higher energy states may become evident in higher dimensions. As an alternative to using Eq. (2) in the last stage of our approach, a method such as maximum caliber[63–65], which does not take the derivatives of landscape topology into account, could be supplied with the sizes of the energy barriers and used to infer MFPTs. However, we found that owing to the dependence of the MFPTs on the prefactors in Eq. (2) for different transitions, this technique could not recover all transition rates accurately for either proteins or gene-regulatory networks (Supplementary Fig. 4). Overall, our theoretical results demonstrate the benefits of combining an analytical PDF with a linear dimensionality reduction technique so that the neglected dimensions can be accounted for explicitly.

Rapidly advancing imaging techniques, such as cryo-EM, will allow many snapshots of biophysical structures to be taken at the atomic level in the near future[3,21,22,28,66,67]. A biologically and biophysically important task will be to infer dynamical information from such instantaneous static ensemble measurements. The protein-folding example in Fig. 2 suggests that the framework introduced here can help overcome this major challenge; in principle, the framework requires only the pairwise distances between recognizable features of the protein as input (here we used the carbon alpha coordinates). Another promising area of future application is the analysis of single-cell RNA-sequencing data quantifying the expression within individual cells[23]. Related to this application, Fig. 3 demonstrates that our protocol recovers state-switching pathways in multistable gene-regulatory networks, which are thought to underlie cell-fate decisions. These results are most relevant in low-molecule-number regimes, in which noise is known to be an important factor[68]. In relevant recent work, an effective non-parametric energy landscape of single-cell expression snapshots was inferred using the Laplacian of a k-nearest neighbor graph on the data, allowing lineage information to be derived via a Markov chain[15]. The GMM-based framework here provides a complementary parametric approach for reconstructing faithful low-dimensional transition state dynamics from such high-dimensional data.

Furthermore, the proof-of-concept results in Fig. 4 suggest that our inference scheme for Markovian network dynamics can be useful for studying viral and bacterial evolution, which are often modeled as movements through a series of DNA or protein sequences[69]. The fitness landscape of an organism in sequence space is analogous to the negative of an effective energy landscape. The process of fixation by a succession of mutants in a population, whereby each mutant replaces the previous lineage as the population's most recent common ancestor, has been modeled as a Markov process[70]. Successive sweeps to fixation have been observed in long-term evolution experiments, promising groundbreaking data for future analysis as whole-genome sequencing technologies improve[71].

The inference protocol opens the possibility to analyze previously intractable multi-phase systems: many high-dimensional physical, chemical, and other stochastic processes can be described by a Fokker–Planck dynamics[1], with phase equilibria corresponding to maxima of the stationary distribution. By taking near-simultaneous measurements of many subsystems within a large multistable Fokker–Planck system, the above scheme allows the inference of coexisting equilibria and transition rates between them. Other possible applications may include neuronal expression[11] and social networks[17,24], which have been described in terms of effective energy landscapes.

Although we focused here on normal white-noise diffusive behavior, as is typical of protein-folding dynamics, the above ideas can in principle be generalized to other classes of stochastic exploration processes. Such extensions will require replacing Eq. (2) through suitable generalized rate formulas, as have been derived for correlated noise[1]. Conversely, the present framework provides a means to test for diffusive dynamics: if the MFPTs of an observed system differ markedly from those inferred by the above protocol, then either important degrees of freedom have not been measured, the system is out of equilibrium on measurement time scales, or the system does not have Brownian transition statistics, necessitating further careful investigation of its time dependence.

By construction, the above framework is applicable to systems whose steady-state dynamics is approximately Markovian and can be described by a Fokker–Planck-type dynamics. This broad class includes thermal equilibrium systems as well as non-equilibrium systems that can be approximated by effective equilibrium theories[72,73]. However, such approximations can become inaccurate if probabilistic non-equilibrium fluxes dominate the system dynamics[74]. For example, reconstructing dynamical gene-expression information from static snapshots is sometimes possible in the presence of oscillatory dynamics caused by processes such as the cell cycle, but can fail for gene networks with large oscillations that are not orthogonal to the processes of interest[15]. Adapting the above protocol to reconstruct the dynamics in the latter case, and of far-from-equilibrium systems in general, will require incorporating more sophisticated theories that include time-resolved information[75–78] and improved expressions for non-equilibrium transition rates[79], and account for probabilistic fluxes[80].

To conclude, the conformational dynamics of biophysical structures such as viruses and proteins, and the state-switching dynamics of noisy gene-regulatory networks, are characterized by their metastable states and associated transition networks, and can often be captured through Markovian models. Current experimental techniques, such as cryo-EM or RNA-sequencing, provide limited dynamical information. In these cases, transition networks must be inferred from static snapshots. Here we have introduced and tested a numerical framework for inferring Markovian state transition networks via reconstructed energy landscapes from high-dimensional static data. The successful application to protein-folding, gene-regulatory network, and viral evolution pathways illustrates that high-dimensional energy landscapes can be reduced in dimension without losing relevant topological information. In general, the inference scheme presented here is applicable whenever the dynamics of a high-dimensional physical, biological, or social system can be approximated by diffusion in an effective energy landscape.

Methods

Population landscapes. A GMM was used to represent the PDF, or population landscape, of samples. The PDF at position $\mathbf{x}$ of a GMM with C mixture components in d dimensions is

$$p(\mathbf{x}) = \sum_{i=1}^{C} \phi_i p_i(\mathbf{x})$$
$$p_i(\mathbf{x}) = \frac{\exp\left(-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_i\right)^T \boldsymbol{\Sigma}_i^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_i\right)\right)}{\sqrt{\det(2\pi\boldsymbol{\Sigma}_i)}},$$

where ϕ_i are the weights of each component, $\boldsymbol{\mu}_i$ are the means, and $\boldsymbol{\Sigma}_i$ are the covariance matrices. More details on GMMs and how they were fit to data are given in the Supplementary Methods.

Mean first passage times. We form a discrete-state continuous-time Markov chain on states given by the minima of the energy landscape. For a pair of states α and β directly connected by a minimum-energy pathway via a saddle, we approximate the transition rate $\alpha \to \beta$ by the Kramers rate $k_{\alpha\beta}$ in Eq. (2), whereas if α and β are not directly connected we set $k_{\alpha\beta} = 0$. Given these rates, the Markov chain has generator matrix $M_{\alpha\beta}$ where $M_{\alpha\beta} = k_{\alpha\beta}$ for $\alpha \neq \beta$ and $M_{\alpha\alpha} = -\sum_{\beta:\beta\neq\alpha} k_{\alpha\beta}$. Then the matrix $\tau_{\alpha\beta}$ of MFPTs (hitting times) for transitions $\alpha \to \beta$ satisfies

$$\sum_{\gamma} M_{\alpha\gamma}\tau_{\gamma\beta} = -1 \ \text{ for } \ \alpha \neq \beta, \quad \tau_{\alpha\alpha} = 0.$$

Protein data pre-processing. Protein-folding trajectories were obtained from all-atom MD simulations performed by D.E. Shaw Research[38]. Data were subsampled by a factor of 5 to reduce the size. For some proteins, residues at the flexible tails of proteins were removed from the dataset to reduce noise. Pairwise distances between carbon alpha atoms on the protein backbone were taken, with a cutoff of 6–8 Å, depending on the size of the protein; any distance above the threshold was taken to be equal to the threshold. This vector of pairwise distances was used as input to PCA, to reduce dimension. The first five principle components of the protein data were found to be sufficient for inference of energy landscapes and transition networks (Supplementary Fig. 1).

Gene-regulatory network simulations. Gene-regulatory network motifs were simulated using a Gillespie SSA in the SimBiology toolbox in Matlab. A full list of the reactions simulated for each motif, as well as the values of the parameters used, is given in the Supplementary Methods and in the simulation code, which is available from Github (see Code availability).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Two publicly available datasets were used in this study. Protein-folding trajectories[38] are available from D.E. Shaw Research (https://www.deshawresearch.com/). HIV sequences[50] are available from https://hiv.biozentrum.unibas.ch/. Gene-regulatory network simulation data are available upon request, or can be generated by running the simulation code available from Github (see Code availability).

Code availability

The source code used in this study to learn a dynamical transition network and mean first passage times from a Gaussian mixture model is publicly available from Github (https://github.com/philip-pearce/learning-dynamical). Also included are all data processing codes required to convert the raw data used in this study into the appropriate format and the simulation code for the gene-regulatory network simulations.

Received: 5 September 2018; Accepted: 24 October 2019;
Published online: 26 November 2019

References

1. Hänggi, P., Talkner, P. & Borkovec, M. Reaction-rate theory: fifty years after Kramers. *Rev. Mod. Phys.* **62**, 251–341 (1990).
2. Yukalov, V. Phase transitions and heterophase fluctuations. *Phys. Rep.* **208**, 395–489 (1991).
3. Dashti, A. et al. Trajectories of the ribosome as a Brownian nanomachine. *Proc. Natl Acad. Sci. USA* **111**, 17492–17497 (2014).
4. Chung, H. S., Piana-Agostinetti, S., Shaw, D. E. & Eaton, W. A. Structural origin of slow diffusion in protein folding. *Science* **349**, 1504–1510 (2015).
5. Neupane, K., Manuel, A. P. & Woodside, M. T. Protein folding trajectories can be described quantitatively by one-dimensional diffusion over measured energy landscapes. *Nat. Phys.* **12**, 700–703 (2016).
6. Hosseinizadeh, A. et al. Conformational landscape of a virus by single-particle X-ray scattering. *Nat. Methods* **14**, 877–881 (2017).
7. Best, R. B. & Hummer, G. Diffusive model of protein folding dynamics with Kramers turnover in rate. *Phys. Rev. Lett.* **96**, 228104 (2006).
8. McCann, L. I., Dykman, M. & Golding, B. Thermally activated transitions in a bistable three-dimensional optical trap. *Nature* **402**, 785–787 (1999).
9. Rondin, L. et al. Direct measurement of Kramers turnover with a levitated nanoparticle. *Nat. Nanotechnol.* **12**, 1130–1133 (2017).
10. García-Müller, P. L., Borondo, F., Hernandez, R. & Benito, R. M. Solvent-induced acceleration of the rate of activation of a molecular reaction. *Phys. Rev. Lett.* **101**, 178302 (2008).
11. Ezaki, T., Watanabe, T., Ohzeki, M. & Masuda, N. Energy landscape analysis of neuroimaging data. *Philos. Trans. R. Soc. A* **375**, 20160287 (2016).
12. Corson, F. & Siggia, E. D. Geometry, epistasis, and developmental patterning. *Proc. Natl Acad. Sci. USA* **109**, 5568–5575 (2012).
13. Lang, A. H., Li, H., Collins, J. J. & Mehta, P. Epigenetic landscapes explain partially reprogrammed cells and identify key reprogramming genes. *PLoS Comput. Biol.* **10**, e1003734 (2014).
14. Pusuluri, S. T., Lang, A. H., Mehta, P. & Castillo, H. E. Cellular reprogramming dynamics follow a simple 1D reaction coordinate. *Phys. Biol.* **15**, 016001 (2017).
15. Weinreb, C., Wolock, S., Tusi, B. K., Socolovsky, M. & Klein, A. M. Fundamental limits on dynamic inference from single-cell snapshots. *Proc. Natl Acad. Sci. USA* **115**, E2467–E2476 (2018).
16. Jin, S., MacLean, A. L., Peng, T. & Nie, Q. scEpath: energy landscape-based inference of transition probabilities and cellular trajectories from single-cell transcriptomic data. *Bioinformatics* **34**, 2077–2086 (2018).
17. Facchetti, G., Iacono, G. & Altafini, C. Exploring the low-energy landscape of large-scale signed social networks. *Phys. Rev. E* **86**, 036116 (2012).
18. Ballard, A. J. et al. Energy landscapes for machine learning. *Phys. Chem. Chem. Phys.* **19**, 12585–12603 (2017).
19. Ferguson, A. L. et al. Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity* **38**, 606–617 (2013).
20. Ebeling, W. & Feistel, R. Studies on Manfred Eigen's model for the self-organization of information processing. *Eur. Biophys. J.* **47**, 395–401 (2018).
21. Fischer, N., Konevega, A. L., Wintermeyer, W., Rodnina, M. V. & Stark, H. Ribosome dynamics and tRNA movement by time-resolved electron cryomicroscopy. *Nature* **466**, 329–333 (2010).
22. Bai, X. C. et al. An atomic structure of human γ-secretase. *Nature* **525**, 212–217 (2015).
23. Shalek, A. K. et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).
24. Kunegis, J., Lommatzsch, A. & Bauckhage, C. In *Proc. 18th International World Wide Web Conference (WWW'09)* 741–750 (Madrid, 2009).
25. Chiavazzo, E. et al. Intrinsic map dynamics exploration for uncharted effective free-energy landscapes. *Proc. Natl Acad. Sci. USA* **114**, E5494–E5503 (2017).
26. Wasserman, L. Topological data analysis. *Annu. Rev. Stat. Appl.* **5**, 501–532 (2018).
27. Mattingly, H. H., Transtrum, M. K., Abbott, M. C. & Machta, B. B. Maximizing the information learned from finite data selects a simple model. *Proc. Natl Acad. Sci. USA* **115**, 1760–1765 (2018).
28. Frank, J. & Ourmazd, A. Continuous changes in structure mapped by manifold embedding of single-particle data in cryo-EM. *Methods* **100**, 61–67 (2016).
29. Westerlund, A. M., Harpole, T. J., Blau, C. & Delemotte, L. Inference of Calmodulin's Ca^{2+}-dependent free energy landscapes via Gaussian mixture model validation. *J. Chem. Theory Comput.* **14**, 63–71 (2018).
30. Jónsson, H., Mills, G. and Jacobsen, K. W. in *Classical and Quantum Dynamics in Condensed Phase Simulations* 385–404 (World Scientific, 1998).
31. Trygubenko, S. A. & Wales, D. J. A doubly nudged elastic band method for finding transition states. *J. Chem. Phys.* **120**, 2082–2094 (2004).
32. Kusumaatmaja, H. Surveying the free energy landscapes of continuum models: application to soft matter systems. *J. Chem. Phys.* **142**, 124112 (2015).
33. Kramers, H. A. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica* **7**, 284–304 (1940).
34. Malakhov, A. N. & Pankratov, A. L. Evolution times of probability distributions and averages - exact solutions of the Kramers' problem. *Adv. Chem. Phys.* **121**, 357–438 (2002).
35. Dunkel, J., Ebeling, W., Schimansky-Geier, L. & Hänggi, P. Kramers problem in evolutionary strategies. *Phys. Rev. E* **67**, 061118 (2003).
36. Chodera, J. D. & Noé, F. Markov state models of biomolecular conformational dynamics. *Curr. Opin. Struct. Biol.* **25**, 135–144 (2014).
37. Mardt, A., Pasquali, L., Wu, H. & Noé, F. VAMPnets for deep learning of molecular kinetics. *Nat. Commun.* **9**, 5 (2018).
38. Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science* **334**, 517–520 (2011).
39. Sborgi, L. et al. Interaction networks in protein folding via atomic-resolution experiments and long-time-scale molecular dynamics simulations. *J. Am. Chem. Soc.* **137**, 6506–6516 (2015).
40. Thomas, P., Popović, N. & Grima, R. Phenotypic switching in gene regulatory networks. *Proc. Natl Acad. Sci. USA* **111**, 6994–6999 (2014).
41. Schultz, D., Walczak, A. M., Onuchic, J. N. & Wolynes, P. G. Extinction and resurrection in gene networks. *Proc. Natl Acad. Sci. USA* **105**, 19165–19170 (2008).
42. Chu, B. K., Margaret, J. T., Sato, R. R. & Read, E. L. Markov State Models of gene regulatory networks. *BMC Syst. Biol.* **11**, 14 (2017).
43. Kalai, A. T., Moitra, A. & Valiant, G. Disentangling Gaussians. *Commun. ACM* **55**, 113–120 (2012).
44. Bühlmann, P., Kalisch, M. & Meier, L. High-dimensional statistics with a view toward applications in biology. *Annu. Rev. Stat. Appl.* **1**, 255–278 (2014).
45. Lee, A. A., Brenner, M. P. & Colwell, L. J. Optimal design of experiments by combining coarse and fine measurements. *Phys. Rev. Lett.* **119**, 208101 (2017).

46. Bolhuis, P. G. & Csányi, G. Nested transition path sampling. *Phys. Rev. Lett.* **120**, 250601 (2018).
47. Müller, S. et al. A generalized model of the repressilator. *J. Math. Biol.* **53**, 905–937 (2006).
48. Potvin-Trottier, L., Lord, N. D., Vinnicombe, G. & Paulsson, J. Synchronous long-term oscillations in a synthetic gene circuit. *Nature* **538**, 514–517 (2016).
49. Ferrell, J. E. Jr Bistability, bifurcations, and Waddington's epigenetic landscape. *Curr. Biol.* **22**, R458–R466 (2012).
50. Zanini, F. et al. Population genomics of intrapatient HIV-1 evolution. *eLife* **4**, e11282 (2015).
51. Gokhale, C. S., Iwasa, Y., Nowak, M. A. & Traulsen, A. The pace of evolution across fitness valleys. *J. Theor. Biol.* **259**, 613–620 (2009).
52. Socci, N. D., Onuchic, J. N. & Wolynes, P. G. Diffusive dynamics of the reaction coordinate for protein folding funnels. *J. Chem. Phys.* **104**, 5860–5868 (1996).
53. Zheng, W. & Best, R. B. Reduction of all-atom protein folding dynamics to one-dimensional diffusion. *J. Phys. Chem. B* **119**, 15247–15255 (2015).
54. Ceriotti, M., Tribello, G. A. & Parrinello, M. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc. Natl Acad. Sci. USA* **108**, 13023–13028 (2011).
55. Ferguson, A. L., Panagiotopoulos, A. Z., Kevrekidis, I. G. & Debenedetti, P. G. Nonlinear dimensionality reduction in molecular simulation: the diffusion map approach. *Chem. Phys. Lett.* **509**, 1–11 (2011).
56. Das, P., Moll, M., Stamati, H., Kavraki, L. E. & Clementi, C. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proc. Natl Acad. Sci. USA* **103**, 9885–9890 (2006).
57. Hegger, R., Altis, A., Nguyen, P. H. & Stock, G. How complex is the dynamics of peptide folding? *Phys. Rev. Lett.* **98**, 028102 (2007).
58. Zhuravlev, P. I., Materese, C. K., Papoian, G. A. & Carolina, N. Deconstructing the native state: Energy landscapes, function, and dynamics of globular proteins. *J. Phys. Chem. B* **113**, 8800–8812 (2009).
59. Rohrdanz, M. A., Zheng, W. & Clementi, C. Discovering mountain passes via torchlight: methods for the definition of reaction coordinates and pathways in complex macromolecular reactions. *Annu. Rev. Phys. Chem.* **64**, 295–316 (2013).
60. Krivov, S. V. On reaction coordinate optimality. *J. Chem. Theory Comput.* **9**, 135–146 (2013).
61. Best, R. B., Hummer, G. & Eaton, W. A. Native contacts determine protein folding mechanisms in atomistic simulations. *Proc. Natl Acad. Sci. USA* **110**, 17874–17879 (2013).
62. Ernst, M., Sittel, F. & Stock, G. Contact- and distance-based principal component analysis of protein dynamics. *J. Chem. Phys.* **143**, 244114 (2016).
63. Pressé, S., Ghosh, K., Lee, J. & Dill, K. A. Principles of maximum entropy and maximum caliber in statistical physics. *Rev. Mod. Phys.* **85**, 1115 (2013).
64. Dixit, P. D., Jain, A., Stock, G. & Dill, K. A. Inferring transition rates of networks from populations in continuous-time Markov processes. *J. Chem. Theory Comput.* **11**, 5464–5472 (2015).
65. Dixit, P. D. et al. Perspective: Maximum caliber is a general variational principle for dynamical systems. *J. Chem. Phys.* **148**, 010901 (2018).
66. Behrmann, E. et al. Structural snapshots of actively translating human ribosomes. *Cell* **161**, 845–857 (2015).
67. Fernandez-Leiro, R. & Scheres, S. H. Unravelling biological macromolecules with cryo-electron microscopy. *Nature* **537**, 339–346 (2016).
68. Bar-Even, A. et al. Noise in protein expression scales with natural protein abundance. *Nat. Genet.* **38**, 636–643 (2006).
69. Orr, H. A. Fitness and its role in evolutionary genetics. *Nat. Rev. Genet.* **10**, 531–539 (2009).
70. Sella, G. & Hirsh, A. E. The application of statistical physics to evolutionary biology. *Proc. Natl Acad. Sci. USA* **102**, 9541–9546 (2005).
71. Barrick, J. E. & Lenski, R. E. Genome dynamics during experimental evolution. *Nat. Rev. Genet.* **14**, 827–839 (2013).
72. Fodor, É. et al. How far from equilibrium is active matter? *Phys. Rev. Lett.* **117**, 038103 (2016).
73. Nardini, C. et al. Entropy production in field theories without time-reversal symmetry: quantifying the non-equilibrium character of active matter. *Phys. Rev. X* **7**, 021007 (2017).
74. Li, J., Horowitz, J. M., Gingrich, T. R. & Fakhri, N. Quantifying dissipation using fluctuating currents. *Nat. Commun.* **10**, 1666 (2019).
75. Gammaitoni, L., Hänggi, P., Jung, P. & Marchesoni, F. Stochastic resonance. *Rev. Mod. Phys.* **70**, 223 (1998).
76. Qian, H. Phosphorylation energy hypothesis: open chemical systems and their biological functions. *Annu. Rev. Phys. Chem.* **58**, 113–142 (2007).
77. Li, C. & Wang, J. Landscape and flux reveal a new global view and physical quantification of mammalian cell cycle. *Proc. Natl Acad. Sci. USA* **111**, 14130–14135 (2014).
78. Feng, H., Zhang, K. & Wang, J. Non-equilibrium transition state rate theory. *Chem. Sci.* **5**, 3761–3769 (2014).
79. Scacchi, A., Brader, J. M. & Sharma, A. Escape rate of transiently active brownian particle in one dimension. *Phys. Rev. E* **100**, 012601 (2019).
80. Li, C. & Wang, J. Quantifying cell fate decisions for differentiation and reprogramming of a human stem cell network: landscape and biological paths. *PLoS Comput. Biol.* **98**, e1003165 (2013).

Acknowledgements

We thank D.E. Shaw Research for protein-folding trajectories and Stefano Piana-Agostinetti of D.E. Shaw Research for helpful discussions. This work was supported by the Royal Society International Exchanges award IE160909 (H.K. and J.D.), a summer research studentship from the Biophysical Sciences Institute of Durham University (A.K.), and a Complex Systems Scholar Award from the James S. McDonnell Foundation (J.D.).

Author contributions

P.P., F.G.W., A.F., H.K. and J.D. developed the theory. P.P., F.G.W., A.K. and H.K. wrote the code for learning a transition network. P.P. wrote the simulation codes and applied the code for learning a transition network to datasets. P.P., F.G.W., A.F., H.K. and J.D. wrote the paper. H.K. and J.D. conceived the study and provided supervision.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-019-13307-x.

Correspondence and requests for materials should be addressed to H.K. or J.D.

Peer review information *Nature Communications* thanks Robert Best, Chunhe Li and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019