Energy landscape of various macromolecules revealed by transition network analysis

Jun Li[1], Pan Tan[1, 2], Runbang Li[3], Liang Hong[1, 2, *]

[1]*School of Physics and Astronomy, Shanghai Jiao Tong University, Shanghai 200240, China*

[2]*Institute of Natural Sciences, Shanghai Jiao Tong University, Shanghai 200240, China*

[3]*Zhiyuan College, Shanghai Jiao Tong University, Shanghai 200240, China*

ABSTRACT

The energy landscape of macromolecules holds responsible for their specific functions and underlying dynamics. To overcome the main drawbacks of reaction-coordinate-based methods, we visualize the underlying potential energy surface (PES) of various molecules from all-atom molecular dynamics (MD) simulations, represented by a complex network of conformations transitions between the corresponding metastable states. The topological and geometrical property of energy landscape for six molecules, e.g., globular protein (SHP2 and PGK), intrinsically disordered protein (IDP), double-stranded DNA (dsDNA), single-stranded DNA (ssDNA), and polymer (poly N-diethylacrylamide, PDEA), are investigated by explicit water MD simulations and transition network analysis. The complex network features characterized by the methods of box covering algorithm, loop length, and average path length, indicate that the dsDNA and a globular protein's energy landscape are high-dimensional and fractal. Whereas the non-compact molecules (ssDNA, IDP, polymer) are one-dimensional and straightforward. This difference between folded proteins, dsDNA, and coil-like molecules suggests the folded molecules have a "funnel-shaped" global energy minimum. In contrast, the disordered molecules have multiple local energy minima separated by small barriers and relatively flat.

Keywords: Energy landscape, Molecular dynamics simulations, Transition network analysis

INTRODUCTION

Macromolecules are commonly accepted vibrating under the control of the energy landscape held responsible for their complex dynamics (*1-7*). The approximately harmonic energy funnel with multiple local minima, and its overall complexity directly influences the rich dynamics and essential function (*8, 9*). However, the exact structure of the potential energy surface (PES) is still vague. Thus, visualization of transitions between local minima is a very engaging way of understanding the underlying features of energy basin. Among the motivations to study the energy surface details and its visualization is the extra help in understanding metastable states' role, a rare event, transition pathway, kinetic routes, and conformational changes associated with their function (*10-12*). The visualization of potential and free energy surfaces is not only crucial for revealing any dynamic or thermodynamic properties. Still, it can further help guide what those properties might be (*13-15*).

The simple reaction coordinates to depict the complicated energy landscape of a molecule is often not accurate. The network language provisions another way to understand it. For the visualization, one approach is "transition network" (TN), which encodes the system's transitions between a set of discrete states. The energy landscape can be visualized in terms of a complex network of transitions between the corresponding metastable conformational substates , which can help to investigate the dynamics of macromolecules, such as, peptides, proteins, RNA, and so on (*16-19*). Such a transition network, which overcomes the fundamental drawbacks of reaction-coordinate-based methods, can be directly constructed from all-atom molecular dynamics simulations.

In the transition network, a vertex corresponds to a discrete conformational state in a transition network, and an edge represents a transition between two states. The associated transition rate or probability may weight each edge. In graph-theoretical terms, the conformational states are represented by nodes or vertices, whereas the transitions correspond to the edges. Energy-landscape-based networks can now be visualized for macromolecules, such as proteins, proteins, nucleic acids, the polymer (*16-22*). This makes transition networks powerful tools for understanding large-scale conformational changes. TNs have found several applications in protein folding (*16, 17, 23-27*), enzyme catalysis (*28, 29*), and studies of electron spin resonance (*30*). Moreover, energy landscape based transition networks has been applied to small systems, such as atom clusters and glasses, e.g. (*31, 32*) or peptides (*24, 33, 34*).

Here, we characterize the energy landscape by transition network analysis from molecular dynamics simulation of six macromolecules, including globular protein (SHP2 and PGK), intrinsically disordered protein (IDP), double-stranded DNA (dsDNA), single-stranded DNA (ssDNA), and polymer (poly N-diethylacrylamide, PDEA). We found that non-shaped molecules (ssDNA, polymer, IDP) have falt energy basins with small barriers, while fractal, hierarchical, and complicated energy landscape contributes to globular protein and double-stranded DNA kinetics and thermodynamics.

METHODS

1. Simulation details and molecular system

1.1. SHP2 (SRC homology two domain-containing protein tyrosine phosphatase 2)

Here, the crystal structure of SHP2 with E76A mutation (Fig. 1A) can be found in the PDB data bank, 5XZR (*35*). The molecular mass of this multi-domain phosphatase is about 64.24kD with 519 amino acids. One SHP2 simulation was carried out using GROMACS (Version 2016.3) with the CHARMM27 force field for this enzyme (*36, 37*). The system was solvated in rectangular water boxes (edge lengths 8.5*9.5*10.5 nm^3) with periodic boundary conditions (PBC), leading to a total system size of about 83000 atoms with one SHP2 molecule, ~25000 water molecules. 77 Na^+, Cl^- ion were added to balance the charge. A simulation was produced using the TIP3P water model (*38*) in the NVT ensemble using a Nosé-Hoover thermostat to the reference temperatures 300K (*39*). The pressure coupling algorithm was performed using Parrinello-Rahman with a coupling time of τ =2ps (*40*). Van der Waals interaction was truncated at 1.2nm, with the LJ potential switched to zero gradually at 1.0nm. The short-range electrostatic interactions within the cut-off distance of r_c = 12 Å were treated (*41*) as Coulombic (*42, 43*). All bonds involving hydrogen atoms were constrained with the LINCS algorithm to allow a time step of 2 fs. The system was first energy minimized using steepest descent steps with a maximum force of 10.0 $kJ*mol^{-1}*nm^{-1}$ and a maximum of $5*10^4$ steps, then equilibrated in the NVT ensemble at T=300K for 1ns, and then in the NPT ensemble at p=1bar for 100ns. A one-microsecond long simulation was produced with a 2-fs time step, and the coordinates of the protein molecule are saved at every 1ps.

1.2. PGK (phosphoglycerate kinase)

A typical cartoon structure of yeast enzyme phosphoglycerate kinase (PGK) presents in Fig. 1B, which contains N- and C-terminal domains (residues 1-185, residues 200-389) linked by a helix hinge (residues 186-199 & 390-415). This three-domain protein's molecular weight is about 45kD with 415 residues (PDB ID: 3PGK) (*44*). One 1μs MD simulation was performed for further examining the underlying energy landscape of a protein. The

simulation started from the initial crystal configuration, and the MD software GROMACS (Version 2016.3) was used (*45*), as it is the most used tool in the study of biomolecular dynamics. The force field used for protein was AMBER99SB-ILDN on a local computing cluster (*46*). PGK, which was placed at least 1.0 nm from the box edge, was simulated inside cubic cell center (7*7*7 nm^3) using periodic boundary conditions (PBC), leading to this total system size of about 101000 atoms, ~95000 TIP3P water molecules filled into the box (*38*). In the isothermal−isobaric (NPT) ensemble using a Nosé-Hoover thermostat algorithm(*39*) at 300K and the pressure to 1 bar using Parrinello-Rahman barostat(12). The non-bonded Van der Waals (VdW) interaction was truncated at 1.2nm, at which the VdW interactions reach zero with the LJ potential gradually at 1.0nm. The Coulomb cut-off distance of short-range electrostatic interactions was r_c = 12 Å. For a distance beyond 1.2nm, the Particle Mesh Ewald (PME) method (*42, 43*) was used. The VdW interactions were treated using a cut-off of 1 nm. The simulation coordinates of the system were recorded every 1ps with at least 100ps equilibration time.

1.3. SNAP-25A (Symbiosomal nerve-associated protein 25 isoform A)

Symbiosomal nerve-associated protein 25 isoform A (SNAP-25A), encoding by the SNAP-25A gene in humans, consists of N- and C-terminal α-helix connected by a small random coil linker (*47*). The molecular weight of this intrinsically disordered protein (IDP) SNAP-25A is ~23 kD with a primary sequence length of 204 residues. SNAP-25A, together with syntaxin and synaptobrevin, composes an exceptionally stable four-helix bundle, SNARE complex, an intracellular membrane fusion protein, by pulling the two membranes tightly together to exert the force required for fusion (*48, 49*). Therefore, the dynamics of SNAP-25A protein is considered to be tremendously crucial for intracellular trafficking and vesicle disassembling.

The three-dimensional structure of this disordered protein SNAP-25A (see Fig. 1C) was generated by homology modeling methods (*50*), based on the native amino acids sequence(*51*). This building structure shows 98.53% sequence identity with template protein (a single chain in the PDB ID 6MDM). We performed a single 1μs long trajectory for network analysis. All the simulations were carried by using the MD engine GROMACS (version 2016.3) (*45*) with the specific IDP force field (*52*) for the protein on a local computing cluster. The constant temperature was set to 300 K by using a modified Berendsen thermostat (*39*) and the pressure to 1 bar using Parrinello-Rahman barostat (*40*). The protein is centered in a triclinic water box (~ 6.5 × 6.5 × 14.2 nm^3) using the TIP3P explicit water model (*38*) and placed at least 1.0 nm from the box edge. The full system contains slightly over 58000 atoms and 14 sodium ions to neutralize the total charge. Electrostatic interactions were treated with a cut-off of 1 nm, beyond which the PME method was used (*42, 43*). The VdW interactions were treated using a cut-off of 1 nm. For all the simulations, system coordinates were written into the trajectory file every 1ps.

1.4. dsDNA (Double-stranded DNA)

A three-dimensional cartoon structure of double-stranded DNA shows in Fig. 1D, which is an 8 base-pairs B-form nucleic acid with the sequence of (G5TCCGCTG3-C3AGGGCGAC5), and the initial configuration was generated by x3dna module (http://web.x3dna.org). One 1μs MD simulation was carried out by the simulation engine GROMACS (Version 2016.3) (*45*) and the AMBER99SB-ILDN 96 force field for DNA (*46*). The study's simulation system was comprised of a small dsDNA and about 2500 TIP3P water molecules (*38*) together with the appropriate amount of sodium counterions to balance the negative phosphate charges, leading to this total system size of about 7800 atoms. The temperature was set to 300K using the Nosé-Hoover methods (*39*). The

pressure coupling was kept constant by using the Parrinello-Rahman algorithm (*40*) to the reference pressure one atmospheric (1 atm) in an aqueous solution. Van der Waals forces were evaluated using a 10-12 Å switching scheme (*42, 43*). Long-range electrostatic forces were computed using the particle-Mesh Ewald (PME) methods, 1.5-Å Fourier-space grid, and a 12-Å cut-off for the real-space Coulomb interaction (*42, 43*). The system coordinates were written into the trajectory file every 1 ps and integrated every 2 fs.

1.5. ssDNA (Single-stranded DNA poly dA_{40})

The single-stranded DNA (ssDNA) studied here is polyadenylic acid with 40 monomers (poly dA_{40}), and the cartoon conformation presents in Fig. 1E. The starting configuration for the MD study is generated by the x3dna module (http://web.x3dna.org). The GROMACS (version 2016.3) tools package (*45*) was used in conjunction with the AMBER99SB-ILDN 96 force field (*46*) to model ssDNA dynamics. Poly dA_{40} was solvated in a periodic box of dimensions 15.5*15.5*15.6 nm^3 at 300K temperature and 1 bar pressure in an aqueous solution. This contained approximately 168,000 TIP3P models (*38*) water molecules with the appropriate amount of Na+ counterions to neutralize the negative phosphate charges, and the total system size was about 504000 atoms. Periodic boundary conditions (PBC) were applied in all dimensions with long-range electrostatic interactions characterized by Particle Mesh Ewald (PME) method (*42, 43*). The pressure/temperature coupling was performed using the Parrinello-Rahman algorithm (*40*) and Nosé-Hoover methods (*39*), respectively. MD simulation was carried out using the leap-frog algorithm for integrating Newton's equation of motion for 100 ns at constant temperature (300 K) and pressure (1bar). Van der Waals interaction was truncated at 1.2nm, with the LJ potential switched to zero gradually at 1.0nm (*43*). The short-range electrostatic interactions within the cut-off distance of r_c = 12 Å were treated as Coulombic (*42, 43*). Following that, a 100ns long simulation was produced with a time step of 2 fs integration, and the coordinates of the system coordinates were saved at every 1ps.

1.6. PDEA, poly N-diethylacrylamide

The polymer shown in Fig. 1F is poly N-diethylacrylamide (PDEA), with molecular weight ~12000 g/mol (96 monomers). The build mono polymer tool of the Material Studio Packages was used to produce initial configurations of PDEA. We performed a 100ns all-atom molecular dynamics simulation on the local cluster with GROMACS 2016 package (*45*) and was carried out using the OPLS-AA force field (*53*) for PDEA. A single chain of atactic PDEA was placed at the center of cubic boxes (15.2*15.2*15.2 nm3) with periodic boundary conditions (PBC). The total system size is about 470000 atoms comprised of a PDEA molecule and 467900 TIP4PEW water molecules (*54*). The system was set to keep constant pressure (1 bar) and 293K reference temperature using the Parrinello-Rahman borastat (*40*) (τ= 2.0 ps) and V-rescale thermostat method (*39*) (τ= 0.1 ps), respectively. The LINCS algorithm (*41*) was used to constrain H-bonds. Both coulomb radius and van der Waals cut-off were set as 1.5 nm, and the long-range electrostatic forces were controlled using the particle mesh Ewald (PME) algorithm (*42, 43*). The steepest descent method was used to minimize the system, and then equilibrated 100ps in the canonical ensemble (NVT) and following isothermal-isobaric ensemble (NPT) before the 100ns simulation production. An integration time step of 2 fs was used, and snapshots were saved every 1ps.

2. Building of the conformational transition cluster network (CCTN)

To grouping the different molecular conformations in the configuration space into clusters of data points representing local minima, a common way uses geometric clustering methods depending on structure similarity (*55, 56*), which are fast and easy to use. These clusters are associated with different conformational states sampled from MD simulation (*18*). Here, based on the algorithm reported before (*56*), we construct the conformational cluster transition network (CCTN) of different macromolecules. Briefly, we first count the number of neighbors below the parameter cut-off threshold, and take a structure with the largest number of neighbors with its all neighbors as a cluster and then eliminate it from the pool of snapshots. Repeat for remaining structures in the pool (*56*). The molecule configurations sampled during a simulation trajectory were sorted into clusters based on the structural similarity parameter: root mean square deviation (RMSD) of the heavy atoms. Using a 1~4.0 Å cut-off for the cluster methods leads to 100~300 RMSD-based states. The network analysis was carried out by sampling a sufficiently large set of snapshots (100,000) from the MD trajectory, equidistant in time. These snapshots were grouped into clusters for further analysis.

After the clustering procedure, the snapshots of the molecule are discretized into n conformational substrates (typically a few thousand), and then the associated $n \times n$ transition count matrix $C(\tau)$ is computed, where each element $C_{ij}(\tau)$ represents the probability observed counts that state i changes to state j within a sampling time interval τ. Ideally, this matrix should be symmetric for the equilibrium system to satisfy ergodicity, i.e., $C_{ij}(\tau) = C_{ji}(\tau)$. However, the transition count matrix is only nearly symmetric (*16, 27, 57*) due to the finite observation (simulations length). We can further compute the so-called transition matrix (TM) based on the count matrix via:

$$T_{ij}(\tau) = C_{ij}(\tau) / \sum_j C_{ij}(\tau) \quad (1)$$

The elements T_{ij} represents the probability that transitions occur from in state i at time t to state j at a time $t+\tau$. It's clear that TM is row normalized. The diagonal elements T_{ii} provide the population density of the associated state at the observation time window, and the overall probability of jumping out of the current state is $1-T_{ii}$. As an illustration, we produced the transition matrix as a complex network (*16*), where the verities correspond to substates, and the edges represent transitions between these conformational states. We used Matlab for data analysis, python module graph-tool (http://graph-tool.skewed.de) for visualization of the network.

Networks similar to CCTN have been used to describe the structural dynamics of complex systems, *e.g.*, refs. (*16, 18*). Similar conformations are located close to each other on the energy landscape, and therefore a structural cluster represents a metastable local well on the landscape, which is visualized by the vertices (nodes) of the network. Then we can depict a coarse-grained version of the energy landscape onto a complex network (or a directed graph). This transition network can be used to process molecule structural change, which envisions the evolution of a system controlling by the transitions between a set of discrete conformational states (*16*). In this mapped conformational cluster transition complex network (CCTN) (*18*), a vertex (nodes) corresponds to one cluster of a set of similar structures, and an edge corresponds to a transition probability between two states (nodes). The vertices and edges are weighted by the associated occurrence rate or probability.

3. Characterizing the features of energy landscape represented by CCTN

A box covering method, namely compact-box-burning (cbb) algorithm (*58, 59*), was performed to examine the fractal scaling of the conformational cluster transition network. For a given CCTN, which has a fixed number of nodes, the length of each edge connecting two neighbor nodes is assumed to be 1.

Step (1), a node is randomly chosen from the CCTN to be the center of a box, and all the nearby nodes, whose shortest distance from the center node is smaller than the length of the box, l_b (an integer), are counted to belong to this box and then get removed from the CCTN. Here, the shortest distance between two nodes is the smallest number of edges that connect them.

Step (2), repeat the same process in Step (1) for the remaining nodes in the CCTN. Repeat the above process until no more nodes are left. Then the number of boxes used to cover the CCTN is noted as N_b, which will strongly depend on the box's size, l_b.

If the geometry of the CCTN network is intrinsically fractal, the number of boxes, N_b, required to cover it should scale with the box size, l_b, as:

$$\langle M \rangle \sim \frac{N_b(l_b)}{N_v} \sim l_b^{-d_f} \tag{2}$$

N_v is the total number of nodes in the network, and the power-law exponent d_f represents the fractal dimension.

The average path length (APL) is the mean shortest path between two nodes in a network. The mean path length is the shortest path length average, averaged over all pairs of nodes. For an undirected graph of N nodes, the mean path length is calculated by using the following formula:

$$L = \frac{1}{N(N-1)} \sum_{i=j} d_{ij} \tag{3}$$

d_{ij}, the length of the shortest path between nodes i and j. Here $d(v_i, v_j)$ represents the length of the shortest path that exists between two vertices. So, we take the sum of all shortest paths between all vertices and divide the number of all possible paths. The average path length is one of the three most robust measures of network topology measures, along with its clustering coefficient and degree distribution. In a network, the mean path length is the average of the shortest path length, averaged over all pairs of nodes. We can calculate the average path length of a graph by using the following formula: Here, d_{ij}, represents the length of the shortest path exists between nodes i and j. where the sum is over all pairs of distinct nodes, N is the total number of vertices. It is a measure of the efficiency of information or mass transport on a network.

RESULTS

The macromolecules are necessary for life that are commonly built by the polymerization of smaller molecules (monomers), typically composed of hundreds of atoms or more. For example, proteins, nucleic acids, and polymers, and each is a necessary component of life and produces a wide range of functions. Here, we study five major classes of macromolecules to characterize the underlying energy landscape, i.e., globular proteins, intrinsically disordered proteins (IDP), double-stranded DNA (dsDNA), single-stranded DNA (ssDNA), and classical polymer. Globular proteins, which are abundant proteins in nature, are polypeptide chains spontaneously folded sphere-like, which are functional significant (*35, 44*). Here, we selected two representative globular proteins (PGK, SHP2) to study. SHP2 (Fig. 1A), a non-receptor enzyme, plays a vital role in the progression of human diseases, particularly cancer (*35*). Phosphoglycerate kinase (PGK) is a 415-residue metabolic enzyme that

produces ATP and comprises two roughly equally sized subunits connected by a flexible hinge (*44*). However, many proteins lack a primary low-energy funnel and instead sample a broad and distinct ensemble of conformations. Such intrinsically disordered proteins (IDPs) lack a well-folded structure. Nonetheless, IDPs are now considered to play pivotal roles in cell signaling pathways and regulatory networks. Here, the studied IDP is Symbiosomal nerve-associated protein 25 isoform A (SNAP-25A) (*47*). Another kind of biopolymers is DNA, or deoxyribonucleic acid, which is the genetic material in humans and almost all living things. The double-stranded DNA (dsDNA) molecule wind around one another forms a shape known as a helix. Here, we carried out MD simulations on an 8-base pairs small dsDNA with the sequential sequence G5TCCGCTG3-C3AGGGCGAC5. Single-stranded DNA (ssDNA, Fig. 1E), as a linear chain of nucleotides with a thin diameter and high flexibility and can be stretched to a greater length at high force because it no longer forms a helix. Here, we produced all-atom MD simulations on an ssDNA with 40 Adenine. Moreover, all-atom MD simulations of a classical polymer, poly N-diethylacrylamide (PDEA), in water solutions were also performed on a local computing cluster.

Within the molecule's free energy landscape, there are many small local minima separated by different heights of energy barriers. Thus, the dynamics of macromolecules can be imagined as a fictitious particle diffusing on a rugged energy surface undergoing many walls of various depths (*18, 60*). Herein, to understand the subtle difference and capture the essential features of the underlying landscape from the MD simulation trajectories of various macromolecules introduced above, we used a geometrical approach (*56*): we sort the conformational snapshots into different clusters based on root mean square deviation (RMSD) of the heavy atom by a given cut-off, typically ranging from 1.0-4.0 Å. The molecule's snapshots sampled during a trajectory are grouped into different states when the RMSD value of two configurations is larger than the given threshold. Then, a coarse-grained version of the energy landscape can be projected (see Methods), which mapped a single MD trajectory onto a complex network (or a directed graph) depending on the transitions between different metastable conformational states. Similar conformations are located close to each other on the energy surface. Therefore, a structural cluster represents a local minimum on the landscape, which are the nodes (vertices) of the network. When the molecule transits between two states during the trajectory, a directed edge is developed between the two nodes, thus the formed network we referred to as the *conformational cluster transition network* (CCTN) (*18*).

Similar analysis has been reported previously to study complex dynamics, *e.g.,* Ref. (*16, 18, 20*). A similar network analysis of CCTN has been studied to examine protein structural dynamics in Ref. (*18*). This transition network can be used to process molecule conformational change, which visualizes the evolution of a system controlling by the transitions between a set of discrete conformational states (*16, 18*). In this mapped conformational cluster transition complex network (CCTN) (*18*), a vertex (nodes) corresponds to one cluster of a set of similar structures, and an edge corresponds to a transition between two states. The vertices and edges are weighted by the associated occurrence rate or probability.

Examples of CCTN sampled from all-atom MD trajectories of six macromolecules are presented in Fig. 2A-F. Using the method developed in Ref. (*56*), we define states from 10000 structural snapshots for those six molecules. As presented in Fig. 2 of these six complex networks, the network of dsDNA (green) and globular protein (blue, SHP2; red, PGK) are highly inhomogeneous and tends to form densely connected hubs around a few nodes with the highest ranks, while outside the hub, the connectivity is rather sparse and there has a relatively large distance between the hubs. The densely connected hubs are a community of complex networks and represent larger metastable states of the energy surface. Moreover, such networks indicate high energy barriers between

two metastable states, suggesting it has a low probability of escaping current energy minima (*20*). The connectivity of the CCTN resembles the neighborhoods of the conformational cluster in the transition network. Insights into the free energy landscape's hierarchical geometry exist when networks for the macromolecules with the tertiary structure are connected in a network hierarchy. Such a network yields a complex picture of the molecule's internal dynamics. While the visualized energy landscape of non-compact molecules in Fig. 2 (gray, ssDNA; purple, polymer PDEA) is string-like and having the favor of going back to the previous state (node) to form 2 steps loops. Indicating the molecule's energy surface is flat and can quickly transfer to new conformational states because of the low energy walls.

The energy landscape is the key to the molecule atomic configuration. Similar conformations are located close to each other on the energy surface. Thus, the grouped cluster (see Methods) represents a local energy minimum. Then the networks are produced by connecting the cluster using the transitions (representing by edges) between these minima, *i.e.,* the CCTN. Thus, characterizing the geometrical and topological features of the energy landscape is equivalent to examining the properties of the CCTN properties. Particularly, we are most appealing to the fractal geometry of the network. Therefore, a version box covering algorithm (*59, 61*), the so-called compact-box-burning (cbb) algorithm, which detailed protocol of algorithm is presented in Methods, was applied to the CCTN to obtain the fractal geometry of this complex network.

Supposing the network total "mass" M is proportional to the number of nodes N_v, if the geometry of the network is intrinsically fractal, the least number of boxes (N_b) required to fully cover the network should scales with the length of the box (l_b) as a power-law relation $M \sim \frac{N_b(l_b)}{N_v} \sim l_b^{-d_f}$. N_v is the total number of nodes for normalization. The exponent d_f representing the fractal dimension of the network is found to be about 2~2.7 for dsDNA and globular protein (Fig. 3A), while number boxes (N_b) show liner relation with box length (l_b) for non-shaped molecules (ssDNA, polymer PDEA and IDP), that indicate the CCTN of non-compact molecules are one-diemnsional rather than fractal. Hence, the underlying energy landscape is a self-similar fractal with a dimension 2~2.7 for well-folded macromolecules (proteins, dsDNA). On the contrary, the energy landscape of unfolded macromolecules is relatively flat with the one-dimensional geometry of CCTN. A similar analysis was performed on protein PGK, where a fractal dimension of 2.4 was reported (*18*). The fractal energy landscape could be a general feature for proteins. Still, the exact value of the fractal dimension might depend on the system and the reaction coordinates used to construct the network.

Next, we examined the more properties of the graph. We are especially interested in understanding whether the network will also show the self-transition loop. Thus, we count the number of loops (N_l) with the increasing lengths of a loop (assuming the physical length of each edge is the same). Due to each transition network have a different total number of nodes, N_v (number of vertices) was used to normalize the number of loops (N_l). The results of all macromolecules studied here are presented in Fig. 3B, which show proteins and dsDNA exist longer transition loops than string-like molecules (ssDNA, IDP, and PDEA). Besides, even for the same length of the loop for those two class molecules, the well-folded macromolecules have a relatively higher probability of forming. Those behaviors indicate metastable states of packaging molecules are hard to escape current local wall, easy to develop more and longer loop than polymer, resulting from high energy barriers. By contrast, non-packaging molecules can easily overcome energy barriers, leading to change to arbitrary conformation.

The last analysis of the transition network is the average path length (APL), which detailed definition is introduced in Methods. Briefly, APL average all pairs short path length between any two nodes. In Fig. 3B, polymer molecules (IDP, ssDNA, polymer PDEA) show a more considerable APL value than compact molecules (protein and dsDNA), and the APL error bar corresponding to variance suggests large fluctuation existing in the polymer transition network. The variance is calculated from three different RMSD cut-off complex networks. Those results indicate the protein and dsDNA transition network are easy to form densely connected hubs, furtherly to appear community in a graph. In energy-landscaped language, the polymer will sample wider regions of phase space than constrained molecules. The motions of protein and dsDNA are constrained to native states, which is functionally essential.

Knowledge of the transition rates between conformational states or energy landscape representing by the network in various macromolecules is computationally captured here. The observed difference above between well-folded molecules (globular proteins, dsDNA) and non-shaped molecules (IDP, ssDNA, polymer) can be interpreted based on the topological and geometrical features of their potential energy surfaces (Fig. 4). Shaped molecules have a funnel-like global energy minimum, where the well-folded structure has the lowest energy on the landscape bottom (Fig. 4A). Conformational change of those molecules appears to involve transitions to overcome higher energy barriers. Non-shaped molecules have extensive local minima separated by low barriers. Escaping from the local energy minima one to another is quick and easy, leading to extensively distinct conformation, which has small energy barriers with the approximately same height between them (Fig. 4B).

CONCLUSION AND DISCUSSION

All-atom molecular dynamics simulations of six macromolecules (folded proteins SHP2 and PGK, IDP, dsDNA, ssDNA, polymer) were carried out to understand the features of their undergoing energy landscape. The structural similarity parameter, the RMSD approach used here, allows the sorting of all simulation snapshots into several conformational clusters, represented the metastable states of macromolecules. Then, we performed the network-based model to obtain the properties of the free energy surfaces by visualizing them onto a graph depending on the structural clusters, which were sampled during the MD trajectory and the transition between them. One of the main findings in the present paper is that the resulting networks are fractal self-similar for packaging molecules (globular proteins and dsDNA) in terms of geometrical organization. Simultaneously, polymer-like macromolecules (IDP, ssDNA, PDEA) are randomly danced and show one-dimensional dynamic behavior on the free energy surfaces. By counting the loop number and computing the average path length, those results further approve such differences in the underlying free energy landscape between those two class molecules. This difference between folded proteins, dsDNA, and purely polymeric molecules (IDP, ssDNA, and polymer PDEA) suggests the folded molecules have a "funnel-shaped" global energy minimum, appear to involve transitions with high energy barriers. In contrast, the disordered molecules have multiple local energy minima separated by small barriers and relatively flat. We contend that disordered molecules (absence of secondary conformational content and tertiary structure) contain wide arrays of conformers, which can distribute broad, populate on flat energy surfaces, within which barriers to conformational transition are extremely low.

ACKNOWLEDGMENTS

The authors acknowledge NSFC grants 11974239, 31630002, the Innovation Program of Shanghai Municipal Education Commission, and Shanghai Jiao Tong university Multidisciplinary research fund of medicine and engineering YG 2016QN13. JCS acknowledges funding from project ERKP300 funded by the Office of Biological & Environmental Research in the U.S. Department of Energy (DOE) Office of Science (BER). The authors acknowledge the Center for High-Performance Computing at Shanghai Jiao Tong University for computing resources and the student innovation center at Shanghai Jiao Tong University. AG gratefully acknowledges the financial support from the German Research Foundation (DFG) through the Emmy Noether Program GO 2762/1-1.

FIGURES

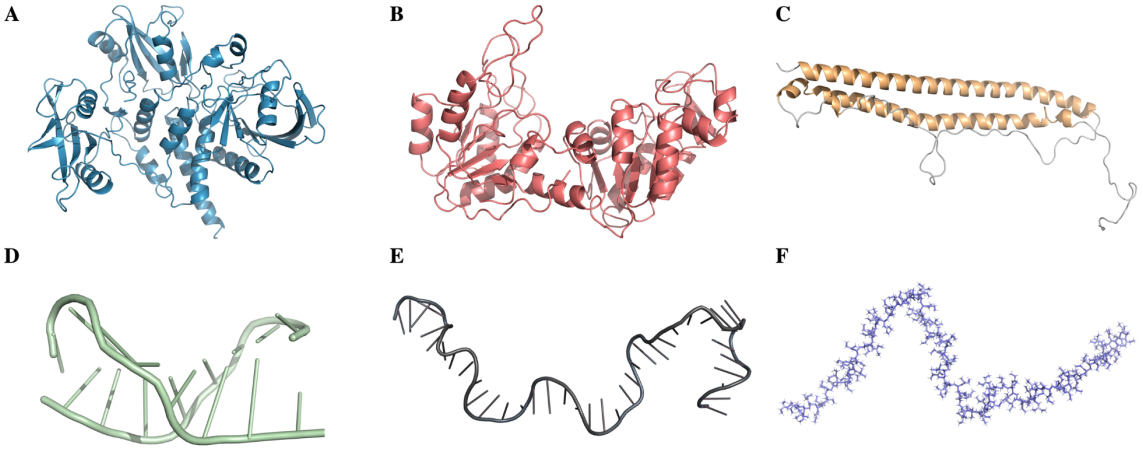

Figure 1 | The cartoon structure of different macromolecules in the network analysis.

A. E76A mutation crystal structure of protein tyrosine phosphatase (PTP) SHP2 (blue) contains two Src homology-2 domains and a PTP domain, PDB code: 5XZR.

B. The yeast enzyme phosphoglycerate kinase (PGK) has two fist-like and equally weighted domains (N-terminal domain; C-terminal domain) and a hinge (red).

C. Cartoon diagram of intrinsically disordered protein (orange), Symbiosomal nerve-associated protein 25 isoform A (SNAP-25A).

D. The structure of an 8 base-pairs double-stranded DNA (dsDNA) from the simulation snapshots (green).

E. The coordinate of 40-mer single-stranded DNA (ssDNA) poly dA_{40} is used for all-atom simulation (gray).

F. A single chain of atactic polymer poly N-diethylacrylamide (PDEA) with 96 monomers (purple) in our all-atom MD study.

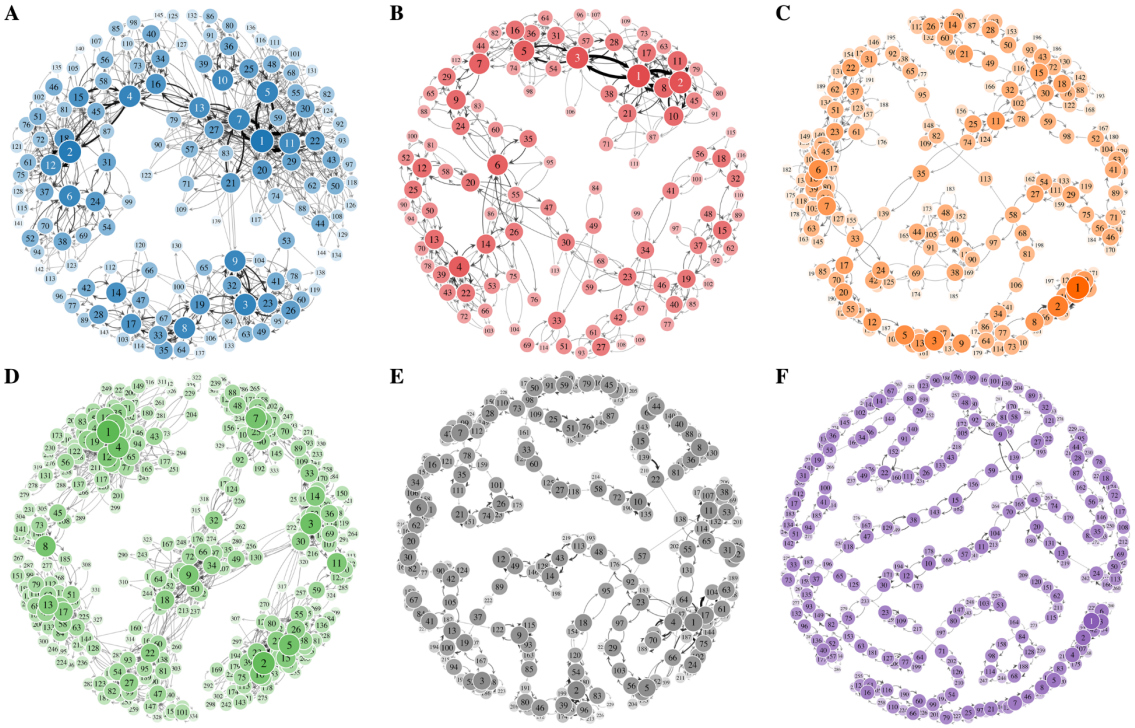

Figure 2 | Energy landscape of various macromolecules represented by the conformational cluster transition network (CCTN). A Graphical illustration of SHP2 conformational clusters transition network (CCTN) from one all-atom MD simulation trajectory (blue) at 300K has 145 nodes and 852 edges. Each vertex represents conformational clusters. The diameter and color shades of circles indicate its occurrence rate, calculated by counting the total number of snapshots belonging to the cluster. The vertices mark with an integer in terms of the rank of occurrence probability. The directed edges denote a transition between two conformational states and are weighted by the associated transition probability. The visualized networks representing energy landscape were produced using the Python module graph-tool (http://graph-tool.skewed.de/). **B**, a network is representing conformational transition derived from protein PGK simulation (red), containing 116 vertices and 461 edges. The corresponding globular protein structure is presented in Fig. 1B. **C**, Conformational transition network from intrinsically disordered protein (SNAP-25A) computational simulation (orange), form 199 nodes, and 626 transitions are representing by edges. **D**. A high-resolution version of double-stranded DNA (dsDNA) network illustrations show community and obvious metastable state, containing 334 vertices and 1, 968 edges. **E**. Network representation of conformational transitions in single-stranded DNA (ssDNA) simulation (gray), this string-liked network contains 229 nodes and 586 transitions/changes. **F**, Polymer PDEA all-atom simulation trajectory visualizes to complex network, with 290 vertices and 686 edges.

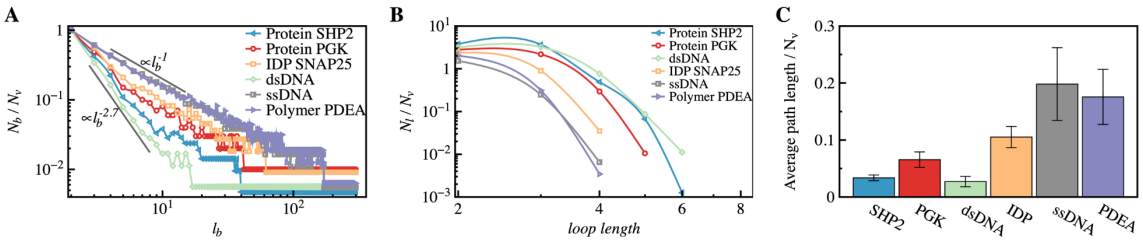

Figure 3 | Characterize the properties of transition networks.

A fractal dimension of network estimated by box covering methods. The number of the boxes (N_b) required to cover the network normalized by the number of the vertex (N_v) is plotted against the box's length, l_b. Power-law fit suggests the fractal scaling in dsDNA and globular proteins energy-landscaped based network, while one-dimensional geometry in ssDNA, polymer, and IDP, which are string-like and atactic macromolecules.

B, the normalized loop number (N_l) versus the loop length, IDP, ssDNA, and polymer contains less and shorter loop than well-folded biomolecules (dsDNA, protein SHP2, and PGK).

C, the average path length (APL), shortest path length averaged over all pairs of nodes, of different macromolecules. The atactic molecules show larger values (orange, IDP; gray, ssDNA; purple, polymer PDEA).

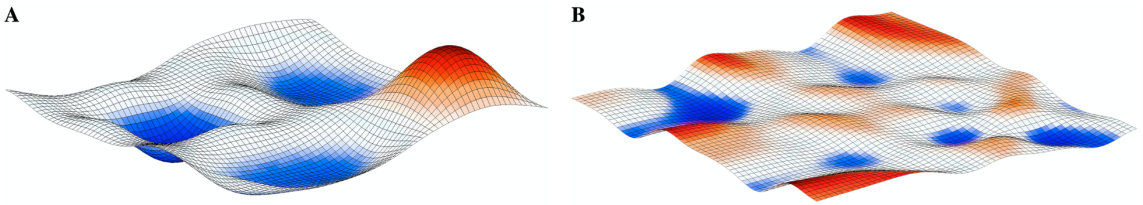

Figure 4 | Schematic diagram of two energy landscapes.

A, Energy landscape of globular proteins or double-stranded DNA, which are separated by high energy barriers, and consist of relatively deep local energy well, and it's hard to escape current metastable states, but still have some transition path to jump out. The gradient change from red to blue, corresponding to high to low energy.

B, an example of free energy surface for intrinsically disordered protein, single stranded DNA and polymer PDEA, which is flat with wide arrays.

REFERENCES

1. H. Frauenfelder, S. G. Sligar, P. G. Wolynes, The energy landscapes and motions of proteins. *Science* **254**, 1598-1603 (1991).
2. J. Sabelko, J. Ervin, M. Gruebele, Observation of strange kinetics in protein folding. *Proceedings of the National Academy of Sciences* **96**, 6031-6036 (1999).
3. J. Brujić, K. A. Walther, J. M. Fernandez, Single-molecule force spectroscopy reveals signatures of glassy dynamics in the energy landscape of ubiquitin. *Nature Physics* **2**, 282 (2006).
4. P. Senet, G. G. Maisuradze, C. Foulie, P. Delarue, H. A. Scheraga, How main-chains of proteins explore the free-energy landscape in native states. *Proceedings of the National Academy of Sciences*, pnas. 0810679105 (2008).
5. K. Henzler-Wildman, D. Kern, Dynamic personalities of proteins. *Nature* **450**, 964 (2007).
6. W. Min, G. Luo, B. J. Cherayil, S. Kou, X. S. Xie, Observation of a power-law memory kernel for fluctuations within a single protein molecule. *Physical review letters* **94**, 198302 (2005).
7. X. Yu, D. M. Leitner, Anomalous diffusion of vibrational energy in proteins. *The Journal of chemical physics* **119**, 12673-12679 (2003).
8. H. Yang *et al.*, Protein conformational dynamics probed by single-molecule electron transfer. *Science* **302**, 262-266 (2003).
9. S. Kou, X. S. Xie, Generalized Langevin equation with fractional Gaussian noise: subdiffusion within a single protein molecule. *Physical review letters* **93**, 180603 (2004).
10. Y. Shan, A. Arkhipov, E. T. Kim, A. C. Pan, D. E. Shaw, Transitions to catalytically inactive conformations in EGFR kinase. *Proceedings of the National Academy of Sciences* **110**, 7270-7275 (2013).
11. C. M. Dobson, Protein folding and misfolding. *Nature* **426**, 884-890 (2003).
12. A. S. Reddy *et al.*, Stable and metastable states of human amylin in solution. *Biophysical journal* **99**, 2208-2216 (2010).
13. D. J. Wales, Energy landscapes: some new horizons. *Current opinion in structural biology* **20**, 3-10 (2010).
14. D. J. Wales, Decoding the energy landscape: extracting structure, dynamics and thermodynamics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **370**, 2877-2899 (2012).
15. D. J. Wales, T. V. Bogdan. (ACS Publications, 2006).
16. F. Noé, I. Horenko, C. Schütte, J. C. Smith, Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. *The Journal of chemical physics* **126**, 04B617 (2007).
17. F. Noé, D. Krachtus, J. C. Smith, S. Fischer, Transition networks for the comprehensive characterization of complex conformational change in proteins. *Journal of chemical theory and computation* **2**, 840-857 (2006).
18. X. Hu *et al.*, The dynamics of single protein molecules is non-equilibrium and self-similar over thirteen decades in time. *Nature Physics* **12**, 171 (2016).
19. G. M. Giambaşu, T.-S. Lee, W. G. Scott, D. M. York, Mapping L1 ligase ribozyme conformational switch. *Journal of molecular biology* **423**, 106-122 (2012).
20. F. Noé, S. Fischer, Transition networks for modeling the kinetics of conformational change in macromolecules. *Current opinion in structural biology* **18**, 154-162 (2008).
21. K. Klenin, B. Strodel, D. J. Wales, W. Wenzel, Modelling proteins: Conformational sampling and reconstruction of folding kinetics. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **1814**, 977-1000 (2011).
22. D. Nerukh, C. H. Jensen, R. C. Glen, Identifying and correcting non-Markov states in peptide conformational dynamics. *The Journal of chemical physics* **132**, 084104 (2010).
23. S. Muff, A. Caflisch, Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a β-sheet miniprotein. *Proteins: Structure, Function, and Bioinformatics* **70**, 1185-1195 (2008).
24. S. V. Krivov, M. Karplus, Hidden complexity of free energy surfaces for peptide (protein) folding. *Proceedings of the National Academy of Sciences* **101**, 14766-14770 (2004).
25. N. Singhal, C. D. Snow, V. S. Pande, Using path sampling to build better Markovian state models: predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *The Journal of chemical physics* **121**, 415-425 (2004).
26. J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, W. C. Swope, Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *The Journal of chemical physics* **126**, 04B616 (2007).
27. W. C. Swope, J. W. Pitera, F. Suits, Describing protein folding kinetics by molecular dynamics simulations. 1. Theory. *The Journal of Physical Chemistry B* **108**, 6571-6581 (2004).

28. S. Yang, B. Roux, Src kinase conformational activation: thermodynamics, pathways, and mechanisms. *PLoS Comput Biol* **4**, e1000047 (2008).
29. S. Yang, N. K. Banavali, B. Roux, Mapping the conformational transition in Src activation by cumulating the information from multiple molecular dynamics trajectories. *Proceedings of the National Academy of Sciences* **106**, 3776-3781 (2009).
30. D. Sezer, J. H. Freed, B. Roux, Using Markov models to simulate electron spin resonance spectra from molecular dynamics trajectories. *The Journal of Physical Chemistry B* **112**, 11014-11027 (2008).
31. D. J. Wales, Structure, dynamics, and thermodynamics of clusters: tales from topographic potential surfaces. *Science* **271**, 925-929 (1996).
32. F. Calvo, T. V. Bogdan, V. K. de Souza, D. J. Wales, Equilibrium density of states and thermodynamic properties of a model glass former. *The Journal of chemical physics* **127**, 044508 (2007).
33. O. M. Becker, M. Karplus, The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *The Journal of chemical physics* **106**, 1495-1517 (1997).
34. Y. Levy, O. M. Becker, Effect of conformational constraints on the topography of complex potential energy surfaces. *Physical review letters* **81**, 1126 (1998).
35. J. Xie *et al.*, Allosteric inhibitors of SHP2 with therapeutic potential for cancer treatment. *Journal of medicinal chemistry* **60**, 10205-10219 (2017).
36. A. D. MacKerell Jr *et al.*, All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The journal of physical chemistry B* **102**, 3586-3616 (1998).
37. A. D. Mackerell Jr, M. Feig, C. L. Brooks III, Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *Journal of computational chemistry* **25**, 1400-1415 (2004).
38. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics* **79**, 926-935 (1983).
39. H. J. Berendsen, J. v. Postma, W. F. van Gunsteren, A. DiNola, J. Haak, Molecular dynamics with coupling to an external bath. *The Journal of chemical physics* **81**, 3684-3690 (1984).
40. M. Parrinello, A. Rahman, Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics* **52**, 7182-7190 (1981).
41. B. Hess, P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *Journal of Chemical Theory and Computation* **4**, 116-122 (2008).
42. T. Darden, D. York, L. Pedersen, Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems. *The Journal of chemical physics* **98**, 10089-10092 (1993).
43. U. Essmann *et al.*, A smooth particle mesh Ewald method. *The Journal of chemical physics* **103**, 8577-8593 (1995).
44. H. Watson *et al.*, Sequence and structure of yeast phosphoglycerate kinase. *The EMBO journal* **1**, 1635-1640 (1982).
45. M. Abraham, D. van der Spoel, E. Lindahl, B. Hess, GROMACS User Manual, version 2016.3; GROMACS Development Team, 2017. *Google Scholar There is no corresponding record for this reference*.
46. K. Lindorff-Larsen *et al.*, Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function, and Bioinformatics* **78**, 1950-1958 (2010).
47. D. Fasshauer, R. B. Sutton, A. T. Brunger, R. Jahn, Conserved structural features of the synaptic fusion complex: SNARE proteins reclassified as Q-and R-SNAREs. *Proceedings of the national academy of sciences* **95**, 15781-15786 (1998).
48. P. Washbourne *et al.*, Genetic ablation of the t-SNARE SNAP-25 distinguishes mechanisms of neuroexocytosis. *Nature neuroscience* **5**, 19 (2002).
49. K. Weninger, M. E. Bowen, U. B. Choi, S. Chu, A. T. Brunger, Accessory proteins stabilize the acceptor complex for synaptobrevin, the 1: 1 syntaxin/SNAP-25 complex. *Structure* **16**, 308-320 (2008).
50. P. Benkert, M. Biasini, T. Schwede, Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* **27**, 343-350 (2010).
51. D. Fasshauer, W. K. Eliason, A. T. Brünger, R. Jahn, Identification of a minimal core of the synaptic SNARE complex sufficient for reversible assembly and disassembly. *Biochemistry* **37**, 10354-10362 (1998).
52. D. Song, R. Luo, H.-F. Chen, The IDP-specific force field ff14IDPSFF improves the conformer sampling of intrinsically disordered proteins. *Journal of chemical information and modeling* **57**, 1166-1178 (2017).
53. G. A. Kaminski, R. A. Friesner, J. Tirado-Rives, W. L. Jorgensen, Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *The Journal of Physical Chemistry B* **105**, 6474-6487 (2001).

54. H. W. Horn *et al.*, Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *The Journal of chemical physics* **120**, 9665-9678 (2004).
55. J. A. Hartigan, M. A. Wong, Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)* **28**, 100-108 (1979).
56. X. Daura *et al.*, Peptide folding: when simulation meets experiment. *Angewandte Chemie International Edition* **38**, 236-240 (1999).
57. N. Singhal, V. S. Pande, Error analysis and efficient sampling in Markovian state models for molecular dynamics. *The Journal of chemical physics* **123**, 204909 (2005).
58. C. Song, L. K. Gallos, S. Havlin, H. A. Makse, How to calculate the fractal dimension of a complex network: the box covering algorithm. *Journal of Statistical Mechanics: Theory and Experiment* **2007**, P03006 (2007).
59. L. K. Gallos, C. Song, S. Havlin, H. A. Makse, Scaling theory of transport in complex biological networks. *Proceedings of the National Academy of Sciences* **104**, 7746-7751 (2007).
60. H. Frauenfelder, D. T. Leeson, The energy landscape in non-biological and biological molecules. *Nature Structural Molecular Biology* **5**, 757 (1998).
61. C. Song, S. Havlin, H. A. Makse, Self-similarity of complex networks. *Nature* **433**, 392 (2005).