JCTC Journal of Chemical Theory and Computation
Article
Cite This: J. Chem. Theory Comput. 2018, 14, 453−460
pubs.acs.org/JCTC

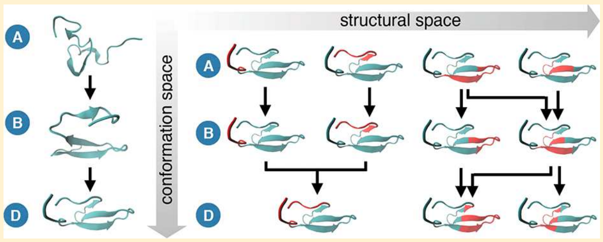
A Data-Driven Perspective on the Hierarchical Assembly of Molecular Structures

Lorenzo Boninsegna,[†] Ralf Banisch,[‡] and Cecilia Clementi*,[†,‡]

[†]Department of Chemistry, and Center for Theoretical Biological Physics, Rice University, 6100 Main Street, Houston, Texas 77005, United States

[‡]Department of Mathematics and Computer Science, Freie Universität Berlin, Arnimallee 6, 14195 Berlin, Germany

Supporting Information



ABSTRACT: Macromolecular systems are composed of a very large number of atomic degrees of freedom. There is strong evidence suggesting that structural changes occurring in large biomolecular systems at long time scale dynamics may be captured by models coarser than atomistic, although a suitable or optimal coarse-graining is a priori unknown. Here we propose a systematic approach to learning a coarse representation of a macromolecule from microscopic simulation data. In particular, the definition of effective coarse variables is achieved by partitioning the degrees of freedom both in the structural (physical) space and in the conformational space. The identification of groups of microscopic particles forming dynamical coherent states in different metastable states leads to a multiscale description of the system, in space and time. The application of this approach to the folding dynamics of two proteins provides a revised view of the classical idea of prestructured regions (foldons) that combine during a protein-folding process and suggests a hierarchical characterization of the assembly process of folded structures.

INTRODUCTION

The last several years have seen an immense increase in high-throughput and high-performance techniques to simulate molecular systems at a microscopic level,[1−3] which has, in turn, stimulated the surge of powerful data analysis techniques to extract essential features, collective variables, or representative states from simulations[4−6] in order to reconcile them with experimental data. These new techniques have provided tremendous help in advancing our understanding of macromolecular processes (e.g., see refs 7 and 8). However, even if the simulation of considerably sized molecular systems over milliseconds is now feasible, the same approach is not possible for large macromolecular complexes, thus leaving a gap when attempting to scale to cellular signaling.

Empirical and theoretical results indicate that for most macromolecular processes only a limited fraction of the phase space is relevant, and most of the dynamics is not "interesting", as it consists of fast and local fluctuations around long-lived (metastable) conformational states.[9−12] In contrast, the rarely observed transitions between metastable states are crucial, as they govern the switching between biologically relevant functions. Recent work[13,14] has shown that, for medium/long time scales (generally on the range $\gtrsim$10 ns), a very small number of parameters are enough to describe the coarse dynamics of a large macromolecular system. These results suggest that it may be possible to reproduce the thermodynamics and long time scale kinetics of a macromolecular system by means of reduced models, using a significantly smaller number of degrees of freedom. Simulation of coarse models may have a significantly reduced computational burden and allow the study of larger systems on longer time scales.[15−17] Moreover, by filtering out nonessential details, coarse models allow a more direct identification of the essential physical ingredients needed to reproduce a macromolecular process, a key step toward the formulation of the "rules" regulating the behavior of biomolecular processes at different scales.

However, there is still, as yet, no general solution on how to select an optimal set of effective degrees of freedom to reproduce the long time scale dynamics of a given system. The choice of the coarse coordinates is usually made by replacing a group of atoms by one effective particle, usually based on physical and chemical intuition. Because of the local geometric regularity of a protein backbone or a DNA structure, popular models reduce the complexity of a macromolecule to a few interaction sites per residue or nucleotide, e.g., the C_α and C_β atoms for a protein.[18−22]

Complementary mathematical work has suggested how to define optimal collective variables of macromolecular systems, that is, descriptors that can identify interesting collective phenomena over long time scales, and separate macroscopically different structures or aggregation states.[5,6] In a sense, the description of the macromolecular dynamics in terms of very few descriptors can be considered as a form of extreme coarse-graining, where the thermodynamics and kinetics of the slowest processes are well described by a few variables. However, such a

Received: September 22, 2017
Published: December 5, 2017

mathematically rigorous definition does not readily provide a physically meaningful coarse-grained representation of a macromolecular system in terms of clusters of atoms or collections of internal coordinates (angles, distances). In practice, one would like to use a coarse representation of the system that is both physically interpretable and satisfies some criterion of optimality in reproducing the microscopic dynamics.

We propose here to use mathematical ideas similar to what has been developed to identify slow variables[6] but extend the analysis to include the chemical structure of the molecule. The definition of collective variables associated with the slow processes allows one to simplify the complex geometry of the high dimensional configurational space in terms of a few metastable regions. In the present work, the geometry we want to represent in minimal form is not only the high dimensional configurational space but includes the molecular structure itself: we combine the partitioning of the overall configurational space visited by the molecule with the partitioning of the three-dimensional structure of the molecule into groups of atoms that display dynamical coherence.

The minimal representation of the system dynamics as a network of few metastable states, their relative population, and the transition rates between them is usually referred to as a Markov state model (MSM).[4,11,23−25] For each metastable state of a MSM, we seek an optimal coarse-graining of the molecular structure in terms of groups of atoms that can be considered effective dynamical "blocks". The resulting groups of atoms preserve some structural integrity during the system dynamics and at the same time capture the slow processes of the system. We name this approach as the *structure and state space decomposition* (S^{3}D). We show that S^{3}D provides different coarse-grained representations of a molecular structure in different metastable regions of the system: the optimal resolution to describe the system dynamics changes as a macromolecule visits different metastable states.

We illustrate S^{3}D by applying it to molecular simulation trajectories of two different protein systems over *millisecond* time scales (generated on the Anton supercomputer[2,26]). The results illustrate the mechanism of transitions between protein models at different resolutions associated with different macroscopic states of the system. The formation and disassembly of different groups of atoms into *coherent domains* in different metastable states present a multiscale characterization of the system dynamics, both in conformation space and chemical structure.

Interestingly, the results provide a quantification and new interpretation (in terms of coherent dynamical structures) of the idea of modular units in proteins (foldons),[27] and their hierarchical formation and assembly.[28] We show that, while groups of atoms forming constitutive blocks can be clearly detected in the dynamics, these groups are highly heterogeneous in size and composition, and not always associated with the formation of secondary structure. The comparison of the S^{3}D results with what is obtained by using the thermodynamic definition of foldons, and the consequences for coarse-graining methods are discussed.

■ RESULTS AND DISCUSSION

Dynamical Coherent Groups of Atoms in a Macromolecule. We turn to recent results in dynamical system theory to define a general and robust approach to identify a minimal number of dynamically coherent domains in proteins or protein complexes and obtain a faithful description of macromolecular conformational rearrangements over long time scales in a reductionist fashion.

The notion of *coherence* has attracted considerable attention in the mathematics community in the recent past.[31,32] As a typical example, consider a number of particles released in close proximity to each other. The goal is to identify the particles that will remain mutually close to each other for some time. For instance, when caught in the same current stream, drifters released in the ocean will move together as a group, even if the current carries them around the globe. These groups of particles form *coherent sets*.[32] Here we demonstrate that this idea can also be used in the context of a macromolecular system, where the role of the ocean drifters is played by the individual atoms, and we want to determine the groups of atoms that move coherently. In order to preserve the long time scale processes of a macromolecular system, we combine the partitioning of the coherent groups of atoms in a molecular structure with the partitioning of the conformation space (by means of a MSM analysis[4]).

The diffusion map approach[33] has been shown to provide a clear geometric interpretation[14] in the dimensionality reduction of high-dimensional systems and to obtain low-dimensional energy landscapes of macromolecules.[12,29,34−37] In order to solve the problem of finding coherent sets of atoms from molecular dynamics trajectories, we employ a version of the diffusion map extended to the time dimension ("time-averaged diffusion map") that has been recently proposed in the context of dynamical systems.[31] In a nutshell, the diffusion map approach applied to the atomic positions in a single molecular configuration constructs a Markov probability matrix based on which pairs of atoms are close to each other. The dominant eigenfunctions of this Markov matrix then capture the geometry of the given molecular configuration. The *time-averaged diffusion map* considers an ensemble of configurations generated by molecular dynamics and constructs a Markov probability matrix based on which pairs of atoms stay close to each other on average over all configurations in the ensemble. The ensemble we consider here consists of all configurations that are in a certain partition of the conformational space (metastable state), as identified by a MSM analysis. The configurations could be sampled either from one long or multiple short trajectories. The eigenfunctions of the time-averaged diffusion map Markov matrix are dynamical coordinates, and geometric clustering in their space returns groups of atoms that are mutually close over the whole time range considered and are, therefore, coherent. Details on the time-averaged diffusion map method and its implementation are provided in the Supporting Information.

We apply this approach to the subset of molecular dynamics trajectories within each metastable state, as found by a Markov model analysis, resulting in the definition of a strategy for the S^{3}D of the macromolecular dynamics.

The identification of coherent structural "domains" in different regions of the conformational landscape of a macromolecule allows us to identify the minimal structural "units" that remain coherent in every region of the landscape and to illustrate the assembly or disassembly of these units to form the different functional states. In the following, we describe the results from the application of S^{3}D to obtain a minimal representation (both in configuration and structure space) of the folding dynamics of two different proteins for which long equilibrium all-atom trajectories are available.[2,26]

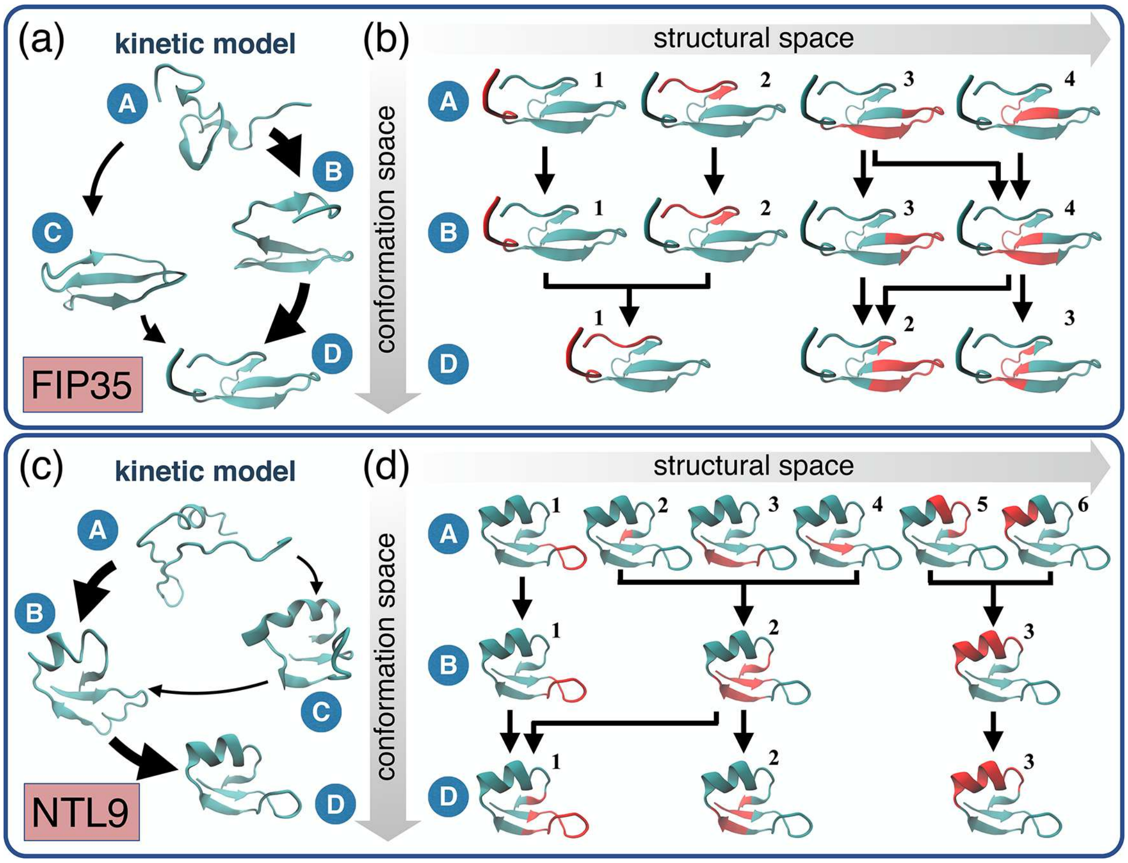


Figure 1. Kinetic network and coherent set analysis results obtained by S^3D for proteins FIP35 (a, b) and NTL9 (c, d). (a and c) Schematic representation of the kinetic models; transition fluxes between states are indicated with arrows of widths proportional to the intensity of transition probabilities. FIP35 (a) folds through the sequence $A \rightarrow B \rightarrow D$ or $A \rightarrow C \rightarrow D$, where state A is the unfolded state, D is the folded state, and B and D are two different intermediate states. Protein NTL9 (c) folds along the sequence $A \rightarrow B \rightarrow D$ or $A \rightarrow C \rightarrow B \rightarrow D$. State A is the unfolded state, D is the folded state, B is an on-pathway intermediate, and C is a misfolded intermediate (not kinetically connected to the folded state). For both proteins, the first pathway ($A \rightarrow B \rightarrow D$) dominates the folding dynamics, as its transition flux is about 3 orders of magnitude larger than that of the other pathway. These kinetics models are consistent with previous studies.[29,30] (b and d) Results from the time-averaged diffusion map analysis along the most likely folding pathways. The different structural coherent domains labeled with numbers and highlighted in red on the folded structure in each metastable state are the coherent sets identified by S^3D as different metastable states. Black arrows indicate the changes in the structural clusters as a molecule transitions between the different metastable states. Results for the less likely folding pathways are reported in the Supporting Information.

Details on the implementation of the Markov model, the time-averaged diffusion map, the spectral clustering, the choice of parameters, and the validation are provided in the Supporting Information.

Coarse-Grained Kinetic Models of Proteins. We use S^3D to discover the hierarchical dynamical processes in the folding of two proteins, for which long equilibrium all-atom trajectories are available: FIP35 WW-domain[2] and NTL9.[26]

The resulting coarse-grained kinetic model is illustrated in Figure 1, both for FIP35 (a,b) and NTL9 (c,d). Figure 1a,c shows that, for both proteins, the folding dynamics mainly proceeds through two possible pathways. Figure 1b,d illustrates the decomposition of the structures in each state into coherent domains (red) and how these domains change along the most likely folding pathway (see the Supporting Information for the less probably pathways). The arrows connecting coherent domains in different metastable states indicate if such domains preserve their identity, split into separate ones, or merge together as the proteins proceed along the primary folding pathways. In the following, we indicate as Xi the ith cluster in metastable state $X \in \{A, B, C, D\}$, with $i \in \{1, ..., n_X\}$, where n_X is the total number of coherent domains for state X.

The dynamics of the coherent domains in FIP35 (Figure 1b) shows that the N and C terminal tails of the proteins move independently in the unfolded (clusters $A1$ and $A2$) and intermediate state ($B1$ and $B2$) and merge into a single coherent domain ($D1$) in the folded state. In the FIP35 protein core, two separate clusters are detected in the unfolded state ($A3$ and $A4$), which undergo a "domain exchange" process during folding: a piece of $A3$ corresponding to a hairpin segment detaches and is absorbed by $A4$ to generate coherent domains $B3$ and $B4$ in the intermediate state, which, in turn, split unevenly into domains $D2$ and $D3$ in the folded state. In essence, the major folding pathway of FIP35 consists of the splitting and merging of clusters $A3$ and $A4$ and the assembly of clusters $A1$ and $A2$. A similar scenario occurs in the major folding pathway of the NTL9 protein. The transition from the unfolded state to the intermediate corresponds to the assembly of different coherent clusters into larger coherent domains: $A2$ + $A3$ + $A4$ merge into $B2$, and $A5$ + $A6$ into $B3$, while $A1$ remains an independent domain. The transition from the intermediate to the folded state involves the splitting and merging of coherent domains $B1$ + $B2$ into $C1$ + $C2$, while the previously assembled α-helical domain $B3$ is maintained.

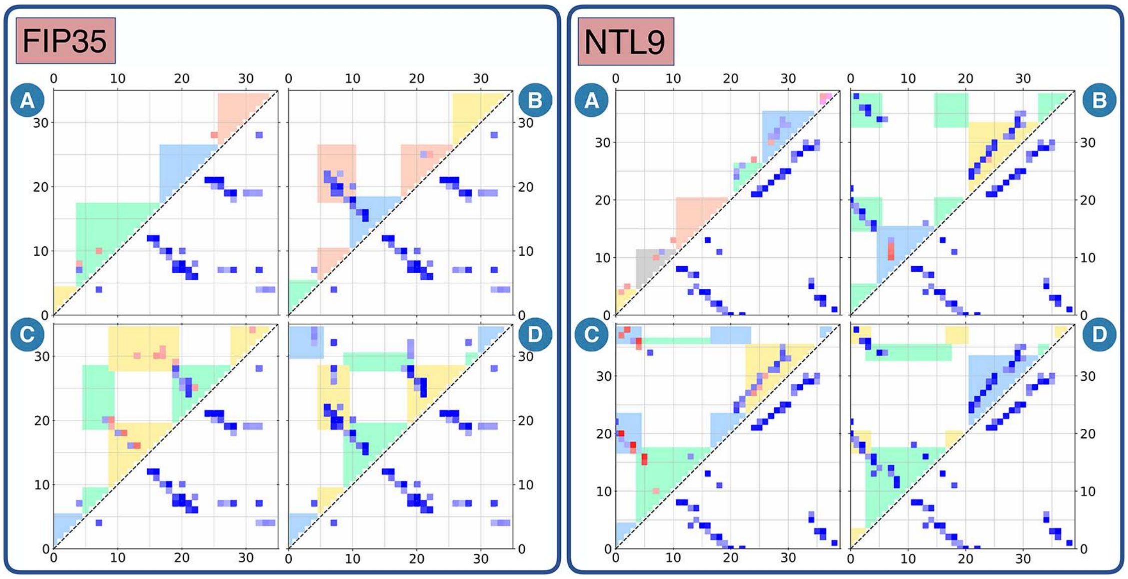


Figure 2. Contact probability versus coherent domains in each metastable state of protein FIP35 (left) and NTL9 (right). Labels (*A*, *B*, *C*, *D*) identify the states as in Figure 1. Areas shaded in different colors indicate different coherent domains in a metastable state. The color correspondence between different panels is arbitrary: colors are used to distinguish the different coherent domains in each metastable state independently. Native and non-native contacts are shown in blue and red, respectively. The color intensity indicates the value of the contact probability, from white to dark blue or red. In each panel, the lower triangular matrix always shows the contact map of the folded state of the corresponding protein, as a comparison. A residue−residue contact is considered formed if the shortest distance among all pairs of heavy atoms is shorter than a threshold value of 0.35 nm. The contact probability is estimated as the frequency of the contact in a given metastable state. For clarity, only contacts with probability higher than 0.3 are shown. The contact maps of the different metastable states are consistent with what found in previous studies, both for FIP35[29] and NTL9.[38]

Not surprisingly, for both proteins the unfolded state is decomposed into a larger number of coherent domains than the intermediate or folded states. Coherent domains in the unfolded states always correspond to connected regions along the protein sequence, indicating that although the proteins appear partitioned into stretches of sequence that move coherently, the different stretches move independently from each other, as expected in the absence of collective motions. On the other hand, coherent domains in the intermediate and native states comprise disconnected regions (i.e., groups of atoms far apart along the sequence are assigned to the same coherent domain), signaling the presence of a tertiary structure and long-range order.

Figure 2 confirms that these coherent domains capture most of the tertiary structure of a protein in a given metastable state. The probability of contact formation in states *A*, *B*, *C*, and *D* is shown for proteins FIP35 (left) and NTL9 (right). In each metastable state, the colored areas on the upper contact matrix indicate all the contacts that can be formed by pairs of residues inside a given coherent domain, and different colors correspond to different structural domains. In each metastable state, all contacts with non-negligible probability of formation essentially fall inside a colored region of the contact space in Figure 2. That is, each domain captures a set of contacts occurring in the metastable state (either native or non-native), and almost no contact is formed between atoms belonging to separate coherent domains.

These results offer a complementary view of what is presented in Figures 1 and S6 (Supporting Information). In both proteins, metastable state *C* presents a larger number of non-native contacts than the other states. The comparison between the contact maps of state *C* and the folded state *D* shows that the shift in the coherent domain boundaries from *C* to *D* corresponds to a rearrangement from out of register to native structural packing.

Although the coherent domains capture most of the secondary and tertiary structure formed in the different metastable states, the structural content is not distributed equally and the domain partitioning cannot be easily inferred from the contact maps alone: some coherent domains contain only a marginal number of contacts, whereas others capture a massive number of contacts. This result suggests that the coherent clusters are amenable to a hierarchical interpretation, very similar to the interpretation of the diffusion coordinates in a standard diffusion map.[12,14] In the latter, diffusion coordinates of higher order resolve dynamical details at a finer time resolution, and clustering in this coordinate space may return states with different levels of metastability. Similarly, in the present context, dynamical coordinates of higher order encode higher "coherence resolution", and the structural coherent domains obtained by clustering in this space are expected to a have different level of coherence.

Minimal Assembly Units as Dynamic Building Blocks. In order to investigate the variations in the coherent domain decomposition in the different metastable states and how the different domains split and merge in the transition between metastable states, we define the *minimal assembly units*, $\{\mathcal{U}_i\}$ as the smallest set of complete and disjoint structural units that can be composed to form any coherent domain in any state: $\mathcal{U}_i \cap \mathcal{U}_j = \varnothing$ and $\cup_i \mathcal{U}_i = \mathcal{P}$, where $\mathcal{P}$ represents the whole protein, and every coherent domain $Xi = \cup_{j \in J} U_j$ for some index set J.

As these units never split into subcomponents across all the metastable states, they can be considered the elementary building blocks of the protein, which can assemble in different ways to form different structural ensembles in different regions of the energy landscape. A synoptic representation of the

DOI: 10.1021/acs.jctc.7b00990

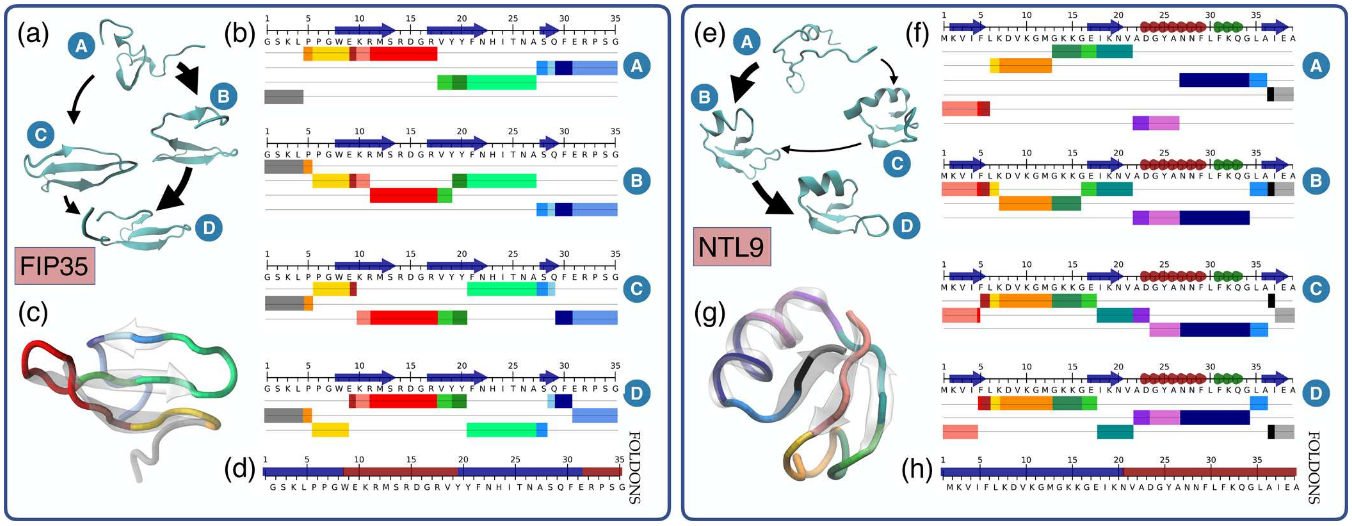

(a)
A
B
C
D
FIP35
(b)
G S K L P P G W E K R M S R D G R V Y Y F N H I T N A S Q F E R P S G
(c)
(d)
FOLDONS
(e)
(f)
M K V I F L K D V K G M G K K G E I K N V A D G Y A N N F L F K Q G L A I E A
NTL9
(g)
(h)

Figure 3. Minimal assembly units for FIP35 (a−c) and NTL9 (e−g) proteins. The same kinetic network models as in Figure 1 are reported in parts a and e, as a reference. (b and f) The minimal assembly units are illustrated for each protein; different colors are used to distinguish between the different units inside a given state, and the same color is used to identify the same unit across different states. In each metastable state (*A*, *B*, *C*, *D*, as in Figure 1), the units are assembled in the corresponding coherent domains along the protein primary sequences. Thin black horizontal lines indicate the different domains in each metastable state, along which the units are placed in a notes-on-the-staff fashion. The locations of the native secondary structure on the protein are also reported on the sketch of the protein sequence on top of the coherent domain representation for every state. An α- or 3_{10}-helix is marked by dark red or green chained dots and a strand by a dark blue arrow. The secondary structure assignment has been obtained by parts c and g: The same color palettes as in parts b and f are used to highlight the minimal units on the protein native structures. As a comparison, the results from a foldon analysis on the two proteins are reported in parts d and h. Four foldons are identified in protein FIP35 (d) and two in NTL9 (h). The regions corresponding to the different foldons are marked by alternating dark blue and red colors.

coherent structural domains in terms of their minimal assembly units is given for both proteins and all states in Figure 3. The decomposition of each state (*A*, *B*, *C*, and *D*) into its coherent domains and the composition of each domain in terms of minimal assembly units is illustrated by the schematics for FIP35 (Figure 3b) and for NTL9 (Figure 3f). Coherent domains are shown linearly along the protein sequence, with minimal units indicated as differently colored stripes, along with the protein sequence and location of the secondary structure. The positions of the minimal units on the native structure are shown in Figure 3c (FIP35) and Figure 3g (NTL9) [with the same colors as in Figure 3b (FIP35) and Figure 3f (NTL9)].

Overall, protein FIP35 consists of 13 minimal units and NTL9 of 14. It is clear from Figure 3 that the units have very nonuniform sizes. In protein NTL9, one unit comprises a single peptide plane (between Ile4 and Phe5), and several units in both proteins are defined by a single residue (Pro6, Arg11, Tyr19, Tyr20, Ser28, Phe30 in FIP35 and Leu6, Ile27 in NTL9) or even a single side chain (Lys10, Gln29 in FIP35 and Phe5, Ile27 in NTL9). On the other hand, in both proteins there are units encompassing multiple consecutive residues (up to seven). The detailed atomic composition of each unit is provided in the Supporting Information.

Interestingly, the larger units do not necessarily correspond to the location of the secondary structure in the proteins. While the structural coherent domains in the native state include the whole helices and most of the strands, assembly units smaller than the full secondary structure elements need to be considered in order to accommodate the coherent domains formed in different metastable states. For instance, the largest domain in the native structure of FIP35 (domain *D*2 in Figure 1b and the corresponding area shaded in green in state *D* in Figure 2) encompasses the bulk of the protein β-sheet. However, this coherent domain is composed of seven different minimal units that can assemble in different ways to form different coherent domains in other metastable states. A similar picture appears for NTL9, for which a large coherent domain in the native state (domain *D*3 in Figure 1b, corresponding to the blue area in state *D* in Figure 2) contains both the α-helix and the consecutive 3_{10}-helix, but it decomposes into three different minimal units that combine to form other coherent domains in other metastable states.

The existence of “building blocks” in proteins has been previously proposed multiple times and explored with different approaches (e.g., see refs 39−42). Previous studies, however, have mostly focused on energetic or structural factors. Most existing methods for decomposing proteins into domains are based on the analysis of fluctuations of interatomic distances as indicators of rigidity of different parts of the macromolecule. Coordinate fluctuations are usually computed over short trajectories or sets of homologous structures, or by normal mode analysis. Previous approaches have not included temporal information or the fact that multiple metastable states can be visited during the protein dynamics. It is well-known that structural similarity does not necessarily correlate with kinetic similarity;[29,43,44] therefore, domains based solely on structural considerations may not capture important dynamical information. Additionally, two-point correlations have been shown to encode partial correlation content only.[45]

In contrast, S^3D provides a coherence-based domain decomposition by combining the notions of structural and kinetic similarity. Although S^3D shares some resemblance with parts of previously proposed methods, there are significant differences in the formulation and implementation. In particular, the use of the diffusion map construction[33] to build the similarity matrix allows us to capture nonlinear effects and is supported by mathematical theory,[31,32] establishing a rigorous link with the concept of coherence in dynamical

DOI: 10.1021/acs.jctc.7b00990

systems and providing an optimality criterion that is missing from other more heuristic formulations. Additionally, the combination of structural and space state analysis of S^3D allows a state-dependent definition of coherent domains, each of them adequately describing subsets of the configuration space. This is very close in spirit to the idea of "ultra-coarse-graining",[46] which argues that state-dependent coarse-graining variables can provide a more suitable protein model and may improve the performance and the interpretation of the results. Indeed, the final output of S^3D is a kinetic network where each node represents a metastable state-specific protein decomposition into coherent domains, which is amenable to a dynamic interpretation: the conformational transition from one state to the other is here translated in a stepwise assembly or disassembly of the minimal units.

Foldons Revisited. The existence of kinetically competent, quasi-independently folding units of a protein was first suggested on the base of geometrical considerations,[47] tested by means of an energy landscape analysis,[27] and appears in agreement with experimental results on some proteins.[48,49] The idea of a hierarchical dynamics was also implemented as a practical method ("zip and assembly") to speed up protein-folding simulation.[28]

So called "foldons" have been identified on a number of proteins by considering different protein segments, $\{j\}$ (of length N_j), and estimating the ratio, $\theta_j = \frac{\Delta E_j}{\delta E_j \sqrt{N_j}}$, between the energy stability gap, ΔE_j, of segment j in the folded configuration (with respect to misfolded alternatives) and its energy variance in the misfolded state(s), δE_j. The energy landscape theory of protein folding[50] relates this quantity θ_j to the ratio between the folding temperature and the glass transition temperature, therefore providing a measure of the relative foldability of a protein segment. By invoking the minimal frustration principle,[50] it was proposed[27] that contiguous protein regions that maximize θ_j could be considered fundamental units of protein folding.

We apply this idea to the two proteins considered here (see the Supporting Information for details). The results are reported in Figure 3d,h and show that foldons include the secondary structure in both proteins and provide a coarser structural decomposition than S^3D. Interestingly, foldons identified by a purely *thermodynamic* criterion correlate with the largest assembly units identified by S^3D on the basis of *dynamic* considerations. This result is not entirely surprising in the light of the energy landscape theory,[50] yet it provides an independent validation for S^3D, which is a data-driven approach. Additionally, it is worth noting that the foldons are identified here by using an all-atom force field to evaluate the energy of the protein fragments in the different states, while the approach was proposed and previously used only with coarse-grained energy models that may significantly smooth out the energetic roughness of a protein-folding landscape.

The minimal units shown in Figure 3b,f complement and revisit the foldons idea, by providing a finer discretization of the protein structure and adding a dynamical interpretation to the foldability criterion.

■ CONCLUSIONS

We propose S^3D, a rigorous approach to identify dynamically coherent structural domains in macromolecules, i.e., groups of molecular components which move collectively and maintain their identity during the system dynamics. By partitioning both in structural and state space, S^3D extends to the physical space the idea of data-driven coarse-graining that has been proposed and used in the macromolecular conformation space for the definition of optimal reaction coordinates. In principle, S^3D can also be tuned to operate on spaces different from that of the atomic Cartesian coordinates used here, e.g., the space of contacts, or other physical observables, which could provide alternative points of view of the dynamics.

The S^3D analysis of the folding mechanism of two proteins, FIP35 and NTL9, show that although different coherent domains are formed in different metastable regions along the folding process, minimal assembly units can be identified. These structural units can be considered building blocks of the macromolecular dynamics, as all the relevant regions of the molecule state space are formed by their assembly and disassembly. As such, S^3D provides a truly multiscale characterization of the system dynamics both in conformational space and physical space. Interestingly, this rigorous analysis also revisits the classic idea of foldons, as "maximally foldable" segments that assemble during a protein-folding process.[27]

While we have illustrated S^3D with a protein-folding application, it can be used in general to learn the minimal dynamical units in large conformational changes in macromolecules or in the assembly of supermolecular complexes. As such, it can offer a link between different resolutions, for a systematic upscaling of biophysical models.

The picture emerging from the S^3D analysis suggests that a global coarse representation of a macromolecule may be inadequate and that different minimalist models should be considered as different metastable states are visited, as has been advocated in ref 46. A logical consequence is that the minimal assembly units identified here could provide natural candidates for a state-dependent coarse-graining approach: the collections of atoms that preserve their geometric integrity as a function of time across all the relevant states visited by the molecule could be considered as effective "beads" in a coarse-grained model, as pictorially presented in Figure 4. Such a data-driven coarse-graining could overcome limitations that affect most structural coarse-grained models, where the choice of the collective degrees of freedom is mostly guided by intuition.

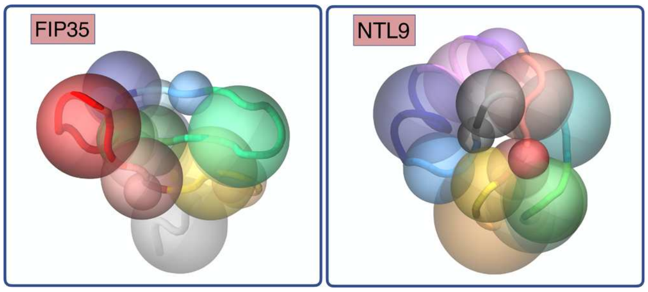


Figure 4. Beads of different colors identify the coarse-grained units for FIP35 (left) and NTL9 (right). The units are defined as the minimal set of structural components that can be composed to form all the different coherent domains in all metastable states of a protein. The backbone of the folded structure is shown in the background as a reference.

At the level presented here, the results are system-dependent. The coarse-grained representation shown in Figure 4 is not immediately transferable to different molecules. Additional investigation of common features across a broad range of systems is needed to draw general conclusions and to understand if and what parts of the minimal assembly units could be transferable.

■ MATERIALS AND METHODS

Identification of Metastable States. The conformational space explored by molecular dynamics trajectories was reduced to a set of metastable states using a MSM approach.[4] All calculations were performed by using implementations available in the pyEMMA software package (pyemma.org).[51] Details on the implementation, the choice of the model parameters, and validation of the models are provided in the Supporting Information.

Clustering into Coherent Groups of Atoms. In each of the metastable states, coherent structural domains were identified by performing agglomerative clustering on the first several dominant eigenvectors. Validation of the clustering was performed by using both the silhouette scoring and the distribution of distance thresholds returned by the agglomerative clustering, and it provided a very clear and robust criterion for the selection of the number of clusters and the cluster assignment (see Figure S5, Supporting Information). Details on the clustering implementation and validation are provided in the Supporting Information.

Codes. A set of python codes for running the S^{3}D analysis are freely available for download on GitHub (https://github.com/ClementiGroup/S3D).

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.7b00990.

> Details about the model parameters, implementation, and validations (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: cecilia@rice.edu.

ORCID

Cecilia Clementi: 0000-0001-9221-2358

Funding

This work was supported by the National Science Foundation (CHE-1265929, CHE-1738990, and PHY-1427654), the Welch Foundation (C-1570), and the Deutsche Forschungsgemeinschaft through CRC1114 (Project B03). Simulations have been performed on the computer clusters of the Center for Research Computing at Rice University, supported in part by the Big-Data Private-Cloud Research Cyberinfrastructure MRI-award (NSF grant CNS-1338099), and on the clusters of the Department of Mathematics and Computer Science at Freie Universität, Berlin.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We are grateful to D. E. Shaw research for sharing the protein-folding simulation data. We are indebted to Peter Koltai (FU Berlin) for inspiring discussions and to Frank Noé (FU Berlin) for his insightful suggestions and his assistance with the MSM and HMM analysis.

■ REFERENCES

(1) Shirts, M.; Pande, V. S. Screen Savers of the World Unite! *Science* **2000**, *290*, 1903−1904.

(2) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R.; Eastwood, M.; Bank, J.; Jumper, J.; Salmon, J.; Shan, Y.; Wriggers, W. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **2010**, *330*, 341−346.

(3) Doerr, S.; Harvey, M. J.; Noé, F.; De Fabritiis, G. HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. *J. Chem. Theory Comput.* **2016**, *12*, 1845−1852.

(4) Bowman, G. R.; Pande, V. S.; Noé, F., Eds. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*; Advances in Experimental Medicine and Biology; Springer Heidelberg, 2014; Vol. 797.

(5) 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.* **2013**, *64*, 295−316.

(6) Noé, F.; Clementi, C. Collective variables for the study of long-time kinetics from molecular trajectories: Theory and methods. *Curr. Opin. Struct. Biol.* **2017**, *43*, 141−147.

(7) Plattner, N.; Doerr, S.; De Fabritiis, G.; Noé, F. Complete protein−protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nat. Chem.* **2017**, *9*, 1005.

(8) Shukla, D.; Meng, Y.; Roux, B.; Pande, V. S. Activation pathway of Src kinase reveals intermediate states as targets for drug design. *Nat. Commun.* **2014**, *5*, 3397.

(9) 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. U. S. A.* **2006**, *103*, 9885−9890.

(10) Kevrekidis, I. G.; Samaey, G. Equation-free multiscale computation: Algorithms and applications. *Annu. Rev. Phys. Chem.* **2009**, *60*, 321−344.

(11) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. Markov models of molecular kinetics: generation and validation. *J. Chem. Phys.* **2011**, *134*, 174105.

(12) Rohrdanz, M. A.; Zheng, W.; Maggioni, M.; Clementi, C. Determination of reaction coordinates via locally scaled diffusion map. *J. Chem. Phys.* **2011**, *134*, 124116.

(13) Schütte, C.; Fischer, A.; Huisinga, W.; Deuflhard, P. A Direct Approach to Conformational Dynamics Based on Hybrid Monte Carlo. *J. Comput. Phys.* **1999**, *151*, 146−168.

(14) Nadler, B.; Lafon, S.; Coifman, R. R.; Kevrekidis, I. G. *Adv. Neural Inf. Process. Syst.* **2005**, 955−962.

(15) Clementi, C. Coarse-grained models of protein folding: toy models or predictive tools? *Curr. Opin. Struct. Biol.* **2008**, *18*, 10−15.

(16) Saunders, M. G.; Voth, G. A. Coarse-graining methods for computational biology. *Annu. Rev. Biophys.* **2013**, *42*, 73−93.

(17) Noid, W. Perspective: Coarse-grained models for biomolecular systems. *J. Chem. Phys.* **2013**, *139*, 090901.

(18) Clementi, C.; Nymeyer, H.; Onuchic, J. N. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? Investigation for small globular proteins. *J. Mol. Biol.* **2000**, *298*, 937−953.

(19) Davtyan, A.; Schafer, N. P.; Zheng, W.; Clementi, C.; Wolynes, P. G.; Papoian, G. A. AWSEM-MD: Protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *J. Phys. Chem. B* **2012**, *116*, 8494−8503.

(20) Matysiak, S.; Clementi, C. Minimalist protein model as a diagnostic tool for misfolding and aggregation. *J. Mol. Biol.* **2006**, *363*, 297−308.

(21) Voth, G. A. In *Coarse-Graining of Condensed Phase and Biomolecular Systems*; CRC Press: Boca Raton, FL, 2008; Chapter 1, DOI: 10.1201/9781420059564.ch1.

(22) Potoyan, D. A.; Savelyev, A.; Papoian, G. A. Recent successes in coarse-grained modeling of DNA. *Wiley Interdiscip Rev. Comput. Mol. Sci.* **2013**, *3*, 69−83.

(23) Buchete, N.-V.; Hummer, G. Coarse master equations for peptide folding dynamics. *J. Phys. Chem. B* **2008**, *112*, 6057−6069.

(24) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Chem. Phys.* **2007**, *126*, 155101.

(25) Noé, F.; Wu, H.; Prinz, J. H.; Plattner, N. Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules. *J. Chem. Phys.* **2013**, *139*, 184114.

(26) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How Fast-Folding Proteins Fold. *Science* **2011**, *334*, 517−520.

(27) Panchenko, A. R.; Luthey-Schulten, Z.; Wolynes, P. G. Foldons, protein structural modules, and exons. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93*, 2008−2013.

(28) Ozkan, S. B.; Wu, G. A.; Chodera, J. D.; Dill, K. A. Protein folding by zipping and assembly. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 11987−11992.

(29) Boninsegna, L.; Gobbo, G.; Noé, F.; Clementi, C. Investigating Molecular Kinetics by Variationally Optimized Diffusion Maps. *J. Chem. Theory Comput.* **2015**, *11*, 5947−5960.

(30) Schwantes, C. R.; Pande, V. S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J. Chem. Theory Comput.* **2013**, *9*, 2000−2009.

(31) Banisch, R.; Koltai, P. Understanding the geometry of transport: Diffusion maps for Lagrangian trajectory data unravel coherent sets. *Chaos* **2017**, *27*, 035804.

(32) Froyland, G.; Lloyd, S.; Santitissadeekorn, N. Coherent sets for nonautonomous dynamical systems. *Phys. D* **2010**, *239*, 1527−1541.

(33) Coifman, R. R.; Lafon, S. Diffusion maps. *Appl. Comput. Harm. Anal.* **2006**, *21*, 5−30.

(34) Zheng, W.; Qi, B.; Rohrdanz, M. A.; Caflisch, A.; Dinner, A. R.; Clementi, C. Delineation of folding pathways of a β-sheet miniprotein. *J. Phys. Chem. B* **2011**, *115*, 13065−13074.

(35) Zheng, W.; Rohrdanz, M. A.; Maggioni, M.; Clementi, C. Polymer reversal rate calculated via locally scaled diffusion map. *J. Chem. Phys.* **2011**, *134*, 144109.

(36) Nedialkova, L. V.; Amat, M. A.; Kevrekidis, I. G.; Hummer, G. Diffusion maps, clustering and fuzzy Markov modeling in peptide folding transitions. *J. Chem. Phys.* **2014**, *141*, 114102.

(37) 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.* **2011**, *509*, 1−11.

(38) Beauchamp, K. a.; McGibbon, R.; Lin, Y.-S.; Pande, V. S. Simple few-state models reveal hidden complexity in protein folding. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 17807−17813.

(39) Hinsen, K. Analysis of domain motions by approximate normal mode calculations. *Proteins: Struct., Funct., Genet.* **1998**, *33*, 417−429.

(40) Ponzoni, L.; Polles, G.; Carnevale, V.; Micheletti, C. SPECTRUS: A Dimensionality Reduction Approach for Identifying Dynamical Domains in Protein Complexes from Limited Structural Datasets. *Structure* **2015**, *23*, 1516−1525.

(41) Yesylevskyy, S. O.; Kharkyanen, V. N.; Demchenko, A. P. Dynamic protein domains: identification, interdependence, and stability. *Biophys. J.* **2006**, *91*, 670−685.

(42) Sinitskiy, A. V.; Saunders, M. G.; Voth, G. A. Optimal number of coarse-grained sites in different components of large biomolecular complexes. *J. Phys. Chem. B* **2012**, *116*, 8363−8374.

(43) Noé, F.; Clementi, C. Kinetic distance and kinetic maps from molecular dynamics simulation. *J. Chem. Theory Comput.* **2015**, *11*, 5002−5011.

(44) Noé, F.; Banisch, R.; Clementi, C. Commute maps: separating slowly-mixing molecular configurations for kinetic modeling. *J. Chem. Theory Comput.* **2016**, *12*, 5620−5630.

(45) Lange, O. F.; Grubmüller, H. Generalized correlation for biomolecular dynamics. *Proteins: Struct., Funct., Genet.* **2006**, *62*, 1053−1061.

(46) Dama, J. F.; Sinitskiy, A. V.; McCullagh, M.; Weare, J.; Roux, B.; Dinner, A. R.; Voth, G. A. The theory of ultra-coarse-graining. 1. General principles. *J. Chem. Theory Comput.* **2013**, *9*, 2466−2480.

(47) Gō, M. Correlation of DNA exonic regions with protein structural units in haemoglobin. *Nature* **1981**, *291*, 90−92.

(48) Lindberg, M. O.; Oliveberg, M. Malleability of protein folding pathways: a simple reason for complex behaviour. *Curr. Opin. Struct. Biol.* **2007**, *17*, 21−29.

(49) Maity, H.; Maity, M.; Krishna, M. M.; Mayne, L.; Englander, S. W. Protein folding: the stepwise assembly of foldon units. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 4741−4746.

(50) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins: Struct., Funct., Genet.* **1995**, *21*, 167−195.

(51) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **2015**, *11*, 5525−5542.

DOI: 10.1021/acs.jctc.7b00990