JPCBFK

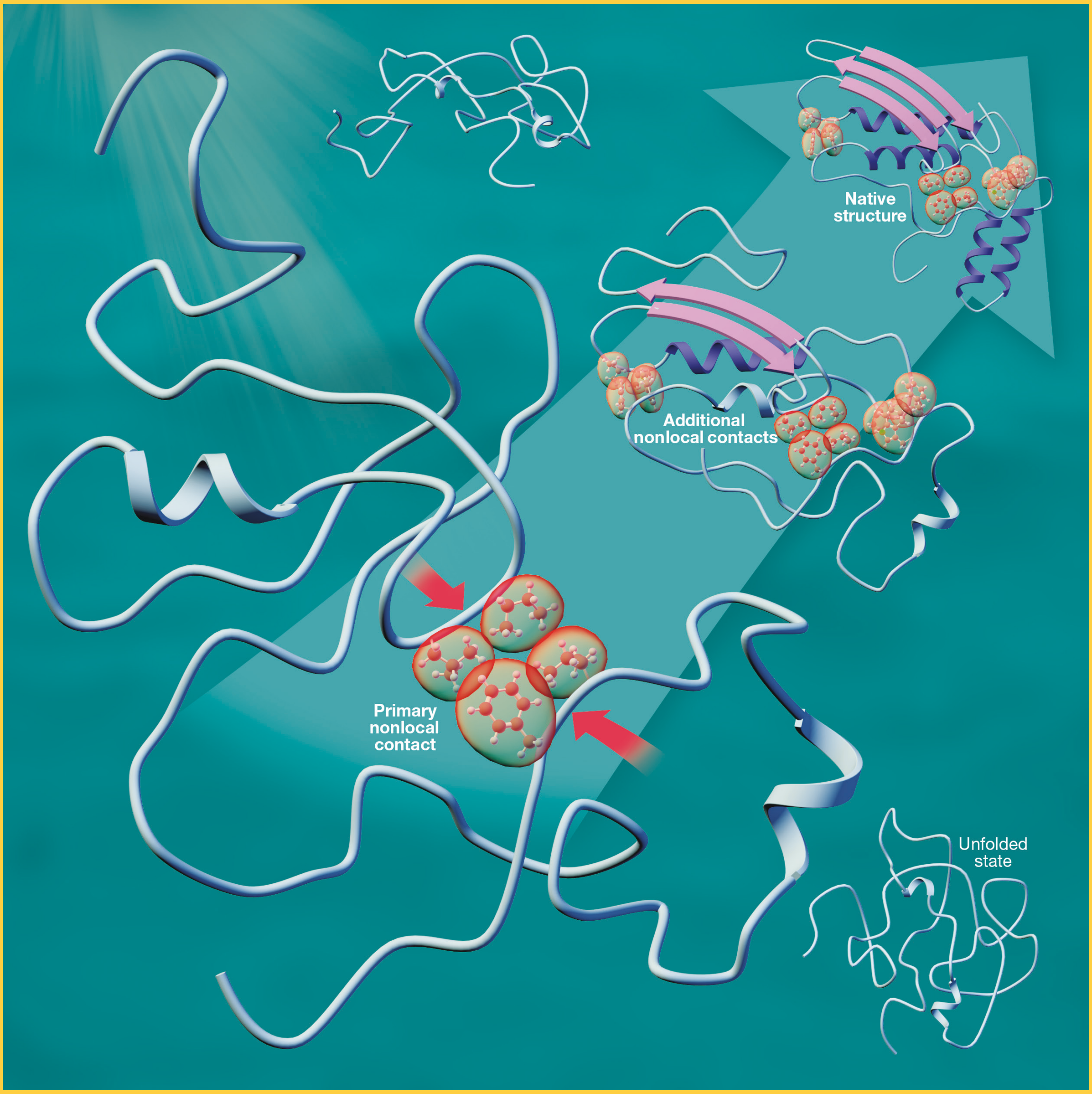
MAY 30, 2019

VOLUME 123

NUMBER 21

pubs.acs.org/JPCB

THE JOURNAL OF PHYSICAL CHEMISTRY B



Nature's Shortcut to Protein Folding Enabled by the Early Formation of Specific Nonlocal Contacts

BIOPHYSICS, BIOMATERIALS, LIQUIDS, SOFT MATTER

ACS Publications
Most Trusted. Most Cited. Most Read.

www.acs.org

THE JOURNAL OF PHYSICAL CHEMISTRY B

Cite This: J. Phys. Chem. B 2019, 123, 4463−4476

Feature Article

pubs.acs.org/JPCB

Nature’s Shortcut to Protein Folding

Fernando Bergasa-Caceres,[*,†] Elisha Haas,[‡,⊥] and Herschel A. Rabitz[§,#]

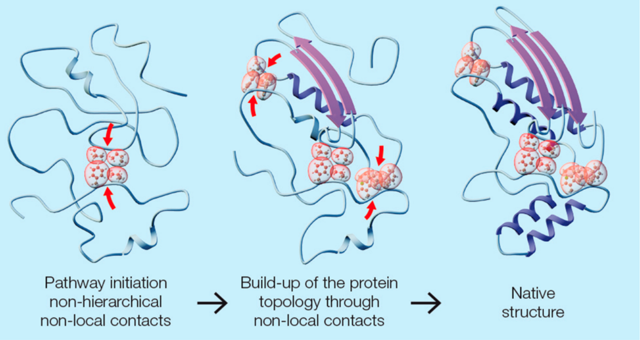
[†]Universidad Autonoma de Madrid, Cantoblanco 28049, Spain

[‡]The Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat Gan 52900, Israel

[§]Princeton University, Princeton, New Jersey 08544, United States

Supporting Information

ABSTRACT: This Feature Article presents a view of the protein folding transition based on the hypothesis that Nature has built features within the sequences that enable a Shortcut to efficient folding. Nature’s Shortcut is proposed to be the early establishment of a set of nonlocal weak contacts, constituting protein loops that significantly constrain regions of the collapsed disordered protein into a native-like low-resolution fluctuating topology of major sections of the backbone. Nature’s establishment of this scaffold of nonlocal contacts is claimed to bypass what would otherwise be a nearly hopeless unaided search for the final three-dimensional structure in proteins longer than ∼100 amino acids. To support this main contention of the Feature Article, the loop hypothesis (LH) description of early folding events is experimentally tested with time-resolved Förster resonance energy transfer techniques for adenylate kinase, and the data are shown to be consistent with theoretical predictions from the sequential collapse model (SCM). The experimentally based LH and the theoretically founded SCM are argued to provide a unified picture of the role of nonlocal contacts as constituting Nature’s Shortcut to protein folding. Importantly, the SCM is shown to reliably predict key nonlocal contacts utilizing only primary sequence information. This view on Nature’s Shortcut is open to the protein community for further detailed assessment, including its practical consequences, by suitable application of advanced experimental and computational techniques.



INTRODUCTION

A central goal of molecular biology has been to understand the principles that govern the protein folding transition, that is, the self-assembly process that leads from an unstructured protein chain to a functional protein.[1−3] Under physiological conditions, most globular proteins fold within seconds and even microseconds into a compact structure.[4] An ensemble of disordered protein molecules, which successfully undergoes a transition to ordered molecules, must, in principle, explore an astronomically large number of pathways, due to the many degrees of freedom of the backbone and side chains.[5] Yet, in the laboratory the folding transition is extremely fast and efficient for almost all globular proteins. This is the source of the common hypothesis that the mechanism of the folding transition of globular proteins is not a random search, but rather, a constrained random process that gradually reduces the conformational space available to the protein.[6−9] Experiments show that the folding time increases significantly with chain length,[10] as the number of possible configurations increases exponentially, but even so, the folding time still remains in the range of a few seconds in the laboratory,[4,11] which is an infinitesimal fraction of the time required to stochastically sample all possible chain configurations. Over the years, several proposals have been put forth to explain the broader features of the folding mechanism.[8,9,12−16]

The elementary steps of the folding transition, like any transition in biopolymers, are stochastic, driven by thermal fluctuations. The Anfinsen experiment led to the “thermodynamic hypothesis”, and the emergent original view of the mechanism of folding was a “bottom up” type of mechanism, where short segments could form native structures that eventually coalesce to form folding of the full-length chain. Later the Leventhal paradox[5] led to the notion that the mechanism of folding should include intermediate states, which form a sequential pathway.[7,17] Several versions of the dominant pathway model were proposed (e.g., framework model,[7,16,18] the diffusion collision,[19] and the nucleation-condensation models[20]), which differ in the details of the formation of the intermediate structures.

In contrast, the energy landscape theory (ELT)[21] (the “new view” 1995), which is based on considerations of statistical mechanics, assumes that proteins fold to their unique native state through multiple stochastic microscopic paths with a bias

Received: December 3, 2018
Revised: March 20, 2019
Published: March 22, 2019

DOI: 10.1021/acs.jpcb.8b11634

toward the native interactions. A funnel-shaped energy landscape represents these multiple pathways,[14,22] while it is assumed that no intermediate states exist and that none of the multiple pathways is dominant.[23] The ELT is supported by lattice and coarse-grained simulations[22,24,25] as well as by recent force-induced unfolding and refolding experiments.[26−28] A common feature for both types of models is the notion that the initial phase of the folding transition must include multiple microscopic paths due to the large number of conformations populated in the unfolded ensemble. However, the dominant pathway-type models assume that the paths toward the transition-state ensemble (TSE) converge to one major dominant pathway, while available alternative pathways are populated when the sequence or the folding conditions are perturbed.[26]

The specific pathway models do not assume a single obligatory order of subdomain transitions. Several alternative pathways are available, and if the dominant one is blocked, other pathways can become dominant, and folding can proceed.[4] The bulk of the folding research over the past 60 years focused at studies of the rate-limiting step of the transition, that is, the rate of formation of the transition-state ensemble. Moreover, it was shown that native-like conformers populate this ensemble. That observation calls for search for structural transitions that lead to the native-like TSE.

However, despite the collective prior work, the physical basis of the efficiency of protein folding has remained elusive, thereby forming a challenge to understand that deserves exploration. The need for detailed mapping of distributions of intramolecular distances in the transient ensembles of refolding proteins prior to the transition-state ensemble was recognized. This can be achieved by detailed Förster resonance energy transfer (FRET) experiments (either time-resolved (tr) at the ensemble level[29] or single molecule detection[14,30]) as well as by detailed simulations.

Advanced molecular dynamics simulations (MD) have attempted to mimic the natural stochastic search of the stable native structures by means of free energy calculations using increasingly refined force fields.[31,32] Ab initio native-like structures have been reliably predicted through MD for proteins of less than ∼100 residues, but predictions have not been satisfactory for longer proteins.[32−34]

The ab initio treatment of the protein folding problem, by any approach, is further complicated by the very nature of protein structure and dynamics: The global free energy difference between the folded and unfolded states is very small (often as low as ∼15−20 kT),[35,36] and the native structures are stabilized by the cooperative contribution of many very weak interactions that are dependent on the local physical properties of the sequence.[37−39] Often, the complete establishment of the native structure depends on the binding of specific ligands or prosthetic groups, and many proteins appear disordered in the absence of such specific interactions.[40] Likewise, early intermediate states along the folding process tend to be only marginally stable and context-dependent, and thus their characterization remains a significant experimental and simulation challenge.

However, it is also well-established that the sequence information that controls the folding transition is relatively robust to conservative substitutions,[41,42] which underlies the broad and successful employment of sequence homology to predict protein structures. The measured degree of mutational robustness[43,44] is an indication that a few basic common physical principles may form the foundation of the dynamics of the folding process. A protein can avoid searching irrelevant conformations and fold quickly by making a few critical folding steps early in the process, well before the rate-limiting folding step, that is, the formation of the TSE.

This Feature Article will draw out that Nature's Shortcut to folding appears to involve (a) specific very weak early dynamical events along the folding pathway that form nonlocal contacts (i.e., contacts established between amino acids more than ∼10−15 amino acids along the sequence), constraining the configurational search from the initial folding stages, (b) relative robustness, depending on some key dominant properties of the sequence and thermodynamic considerations, and (c) common principles across different classes of globular proteins. The resultant reduction in the search space enables fast and parallel folding of subdomain elements consistent with the experimentally observed time frames.

The stability of this early set of nonlocal contacts is likely very low, as the overall stability of the three-dimensional (3D) structure of globular proteins is low. But the stability of the early nonlocal contacts should be beyond the thermal noise, to have a significant impact on the folding mechanism.[45] Also, because of the weak nature of such nonlocal contacts, they likely form only transiently in subsets of the overall population of folding molecules. Thus, the experimental detection of such weak nonlocal contacts can be a challenge. However, even a small population of conformers with such specific contacts can prescribe the folding pathways.[45]

To support the view that Nature employs a Shortcut to folding, the results of two independent but related approaches will be presented. First, the experimental results obtained within the loop hypothesis (LH) context employing FRET techniques by Haas and collaborators (i.e., with an emphasis on recent results for the early kinetics of *Escherichia coli* adenylate kinase (AK) folding) will be shown to lend support to the main proposal of this Feature Article.[46,47] Then, the theoretical basis for the view of protein folding presented here will be seen to naturally emerge from the so-called sequential collapse model (SCM) of Bergasa-Caceres and Rabitz,[48] based on concepts first presented in Bergasa-Caceres' Ph.D. thesis.[49] A comparison also will be made between the SCM prediction and the LH experimental results for early nonlocal contact formation in adenylate kinase.

The resulting model of the folding process, either expressed as LH or via SCM, is then shown to be conceptually and experimentally consistent with additional protein folding experimental results and theoretical expectations. From an experimental Feature Article, it is well-established that (a) loops can form very fast in proteins;[50,51] (b) substantial weak interactions exist in highly disordered states of proteins;[52,53] and (c) nonlocal interactions in proteins are more conserved than short-range ones.[54] From a theoretical Feature Article, the view presented here shares, along with early pathway-based approaches, the general idea that native subdomain structural elements are folded sequentially along the pathway. Pathway sequentiality is common to models such as diffusion-collision,[8] framework,[7] foldon,[9] or nucleation-condensation.[13] Beyond such general considerations, the SCM aims to identify, from primary sequence information alone, specific nonlocal contacts that are important to initiate Nature's Shortcut to the folding of proteins. In this fashion, the SCM strives to be an effective, semiempirical predictive tool of specific folding transitions. Also, the SCM does not prescribe that folding initiation must always proceed from a unique initial event. Rather, the model is flexible enough to allow for several possible pathways to exist when allowed by the sequence. The prospect of having several possible pathways has also been argued to be generally consistent with the funnel view.[55]

Concerning the application of ab initio MD to the folding of long proteins (i.e., longer than ~100 residues), it is proposed that the understanding of the rules governing the establishment of the early scaffold of nonlocal contacts could be an important added component to facilitate such MD simulations, by significantly constraining of the initial state and, thus, opening the avenue toward successfully calculating protein tertiary structure and dynamics. In this regard, successful computational prediction of 3D structures for proteins longer than ~100 amino acids has recently been obtained, when empirical information (i.e., correlated mutations in an evolutionary series of specific proteins) was introduced as constraints to the initial unfolded state.[56,57] These results lend support to the contention that early topological constraints appear to be Nature's Shortcut for constraining the random search toward the native structure. We remark that Nature must be successfully and rapidly locating the early nonlocal contacts through its use of an intrinsic form of "MD". As computational MD has not been successful alone for such long proteins, even with sufficient computer time, some subtle issues with MD remain to be resolved to reconcile this apparent paradox, which remains as a challenge to understand protein folding. The need for the discovery of specific rules that will reduce the computational search in MD simulations was expressed recently within the MD community.[34,58]

Section 2 of the Feature Article will cover the experimental work that underpins the LH, thereby explaining how those results bolster the Shortcut hypothesis of this Feature Article. In Section 3, the SCM will be presented in summarized form, with a focus on early sequential nonlocal contact formation, considered to be the key to Nature's Shortcut, and a summary of contact predictions obtained to date is given in the supporting material. The SCM's predictions for adenylate kinase, employing just primary sequence information, also will be shown to match the experimental results presented in Section 2. Section 3 will briefly discuss the Nature's Shortcut view here within the current context of the protein folding field. Finally, Section 4 will present conclusions and directions for further research.

2. THE ROLE OF NONLOCAL CONTACTS FORMED AT THE INITIATION OF THE FOLDING TRANSITION OF GLOBULAR PROTEINS STUDIED BY TIME-RESOLVED FRET-BASED METHODS

Experiments aimed to test theories of the mechanism of protein folding should ideally yield a series of "snapshots" of ensembles of folding protein molecules, at least with microsecond resolution, to resolve the order of formation of specific features (e.g., local or nonlocal contacts, subdomain transitions, etc.) in the background of multiple nonspecific interactions during the initial and late phases of the folding transition. The observed relative order of appearance of local and nonlocal contacts and formation of subdomain folding transitions prior to the cooperative TSE would provide a direct test of the hypothesis presented in this Feature Article.

Such an ideal kinetic experiment could be combined with site-directed mutagenesis, to search for the key sequence sites and their residues that stabilize and lock structural elements along the gradually narrowing pathway options toward reaching the native fully folded state. A series of experiments could enable identification of the basic principles that define the nature of the constraints on the stochastic search through formation of specific nonlocal contacts and formation of subdomain partially folded elements such as 3D structure loops and foldons.[2] According to this view, the key sequence information can be viewed as prescribing the formation of native nonlocal contacts that shorten the time needed for formation of the TSE, thereby corresponding to Nature's Shortcut to folding whose elementary steps still remains subject to stochastic fluctuations.

The challenges inherent in any attempt to implement such a combined mutagenesis and kinetic approach are formidable. At a minimum they include (a) the characterization of transient very short-lived ensembles in terms of distributions of intramolecular distances at microsecond time resolution under conditions where these ensembles include a wide range of fluctuating conformations; (b) the detection of small subpopulations of conformers where specific (likely native) contacts are formed within the broader ensemble of overall disordered molecules; (c) the obligation to work with complete proteins, as any globular protein is a highly cooperative system, with context-dependent interactions.[39,59] Reductionist approaches studying the folding of isolated subdomain elements are likely to be only partially informative, or possibly even misleading.

Experimental methodologies based on FRET[60] and, in particular, on time-resolved FRET (trFRET),[61] are adequate to achieve the goal of characterizing the transient ensembles of unfolded, collapsed, and partially folded globular protein molecules during the full span of the folding transition.[29,62] These methodologies rely on determining the distribution of intramolecular distances between specifically labeled donor–acceptor probes through trFRET measurements. This approach enables the monitoring of minute changes of the average spatial distance between specific selected subdomain chain elements (e.g., the loops' contact segments, or secondary structure elements, within the entire molecule). Such high spatial resolution, combined with very fast data collection, enables, in principle, the direct observation of the sequential establishment of distance-constraining contacts that guide the folding pathways.[47] Also, subpopulations of folded and disordered chain segments, and the time course of their evolution along the folding pathway, can be resolved.[63] Thus, trFRET experiments, in the ensemble as well as operating in the single molecule mode,[64] can yield meaningful information describing selected specific subdomain conformational transitions and the order of their occurrence along the folding pathways.[62,65−67]

In practice, however, trFRET kinetics experiments are still far from reaching the ideal comprehensive picture described above. The main issues associated with the current trFRET and smFRET methodologies are (a) the experiments are often limited to a few pairs of suitable sites where probes can be inserted, and (b) the probes inserted in the protein sequence can contribute non-native interactions and bias the results. Thus, the unambiguous application of trFRET techniques to the study of fast protein folding remains a challenge, and considerable effort must be put into making sure that the results are both comprehensive and meaningful for the particular protein.

In the "double-kinetics" method employed in the experiments discussed here, fast initiation of the folding transition is achieved by rapid dilution of a denaturant utilizing a mixing device (e.g., stopped flow or continuous flow).[68−72] Two time regimes are involved in these experiments: (1) the duration of the folding transition, from the ensemble of fully disordered molecules to the ensemble of native structures, referred to as the *chemical time regime* (t_c) from microseconds to seconds, and (2) the *spectroscopic time regime* (t_s), (nanoseconds) defined as the rate of fluorescence decay of the probes. The product of a double-kinetics experiment is a series of snapshots of the distributions of distances between pairs of sites along the protein backbone selected for labeling. A set of such experiments can enable the determination of the rates of specific nonlocal contact formation events at the initiation of the folding transition.

2.1. Probing Nonlocal Contact Formation in the Initial Compact Ensemble of Disordered Adenylate Kinase Molecules.

On the basis of existing evidence, it is generally accepted that the starting point of the folding pathway is a very short-lived transient ensemble of disordered, apparently nonspecifically compact molecules under folding conditions.[73−75] This ensemble is expected to be different from the collapsed ensembles studied under equilibrium at low denaturant conditions,[76] because such ensembles can contain a substantial number of specific interactions.[77] This Feature Article proposes that Nature's Shortcut to protein folding is initiated by the formation of specific nonlocal contacts in at least a fractional population of the initial disordered protein ensemble.

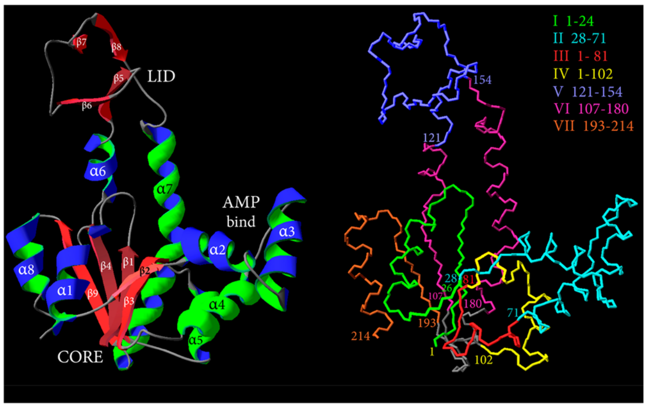
The existence of specific nonlocal contact formation events at the initiation of the folding transition has been investigated within the LH context, employing *E. coli* adenylate kinase (AK) as a model system. AK is a 214 residue bacterial protein that catalyzes the transfer of a phosphoryl group between adenosine triphosphate (ATP) and adenosine monophosphate (AMP).[78] The 3D structure of AK includes three domains, the CORE domain (residues 1−29, 60−121, and 160−214), the LID domain (residues 122−151), and the AMP binding domain (residues 30−59). The native structure of the AK molecule includes seven long loops defined by nonlocal contacts, eight helices, and nine strands.

Folding of Adenylate Kinase. Potential nonlocal contacts (i.e., loop nodes) and both ends of several helices and strands were double labeled (FRET pairs), one pair in a particular mutant, which was studied through double kinetics experiments as described above.[29] The long loops and the secondary structural elements of the AK molecule are presented (color coded) in Figure 1. The amino acid sequence of the AK molecule is also presented in Figure 1. The highlighted nonpolar residues are expected to possibly contribute interactions in the loop nodes, based on the 3D structure.

In a solution of 2 M guanidinium chloride (GndHcl) at pH 7 an ensemble of AK molecules is fully disordered. Ultrafast initiation of refolding of such an ensemble is made possible by fast dilution of the denaturant and formation of an ensemble of still disordered AK molecules under folding conditions.[29] The ensemble refolds to a native structure at a rate of 0.7 s^{-1}.

2.2. Microsecond Kinetics of Long Loop Closure Events in Adenylate Kinase.

Application of the trFRET-based double-kinetics experiment yielded results for the rate of closure of five of the loops evident in the folded structure. Loops IV and V were closed only at the scale of formation of the TSE (i.e., at a time corresponding to the global cooperative transition).[47] However, the three N terminal loops (I−III, Figure 1) were closed within the initial ∼200 μs of the refolding transition.[29,79]



MRIILLGAPG^{10} AGKGTQAQFI$^{\mathbf{20}}$ MEKYGIPQIS^{30} TGDMLRAAVK^{40}SGSELGKQAK^{50}DIMDAGKLVT^{60} DELVIAAAKE^{70} RIAQEDSRNG^{80} FLLDGFPRTI$^{\mathbf{90}}$ PQADAMKEAG^{100}INVDYVLEFD^{110}VPCELIVDRI$^{\mathbf{120}}$ VGRRVHAPSG^{130}RVYHVKFNPP^{140}KVEGKDDVTG^{150} EELTTRKDDQ^{160}EETVRKRLVE^{170}YHQMTAPLIG^{180} YYSKEAEAGN^{190}TKYAKVDGTK^{200}PVAEVRADLE^{210} KILC

Figure 1. Loop structures and secondary structure elements in the *E. coli* AK molecule in its native folded state. (left) The five-strand β structure forming the main body of the CORE domain and the nine helices. (right) The seven loops defined in the backbone. Loops I (green) and II (light blue) are included in loop III (red) and further extended in loop IV (yellow). Loop V (purple) is extended to form loop VI (light purple). (bottom) The amino acid sequence of the AK molecule highlighting the clusters of hydrophobic residues that can form loop nodes (PDB ID code: 4AKE).

The N terminal loop (loop I Figure 1) is closed within less than 60 μs from the initiation of refolding.[29] Loop II closes at a similar rate but shows an additional small reduction of the mean end-to-end distance in a second phase within a time of ∼200 μs (Figure 2). Loop III is also closed in the microsecond time regime, and final tight binding of the loop's ends is done in the slower folding transition. Loop IV, which includes the three fast-closing N terminal loops, was partially closed within the first 5 ms of the transition followed by slow full compaction to native-like end to end distance.

Notably, while the TSE is reached only ∼1.5 s after initiation of folding, specific loop closure events occur in a time regime that is 5 orders of magnitude faster. We conjecture that the closure of the selected loops is likely to have a major role in ensuring that the overall folding mechanism is fast and efficient.

The combination of site specifically labeled protein samples enabled the detection of the fast closure of selected long loops This study shows the very fast closure of long loops in the ensemble of nonspecifically collapsed disordered protein molecules and the fine-tuning of sequential order of formation of a few very early specific pretransition state nonlocal contacts.

2.3. Mutational Analysis of Potential Loop Nodes.

The information that directs the early loop closure in the collapsed ensemble is expected to be contained in the

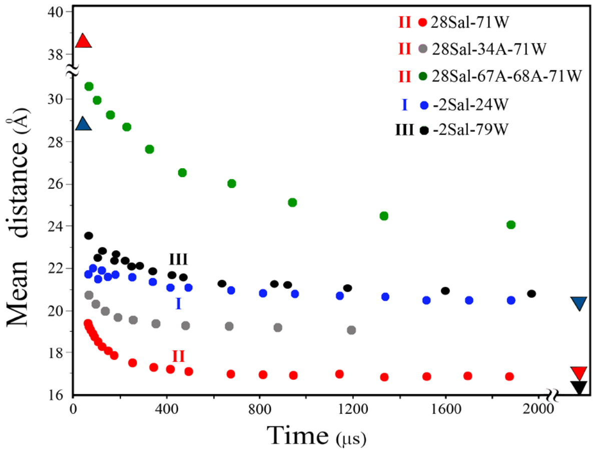


Figure 2. Kinetics of loop closure studied by trFRET-detected double-kinetics experiments in the microsecond time regime. Each color-coded dot represents the mean of the distribution of the distance between a Trp residue and a Cys residue, which is modified by labeling with 4-acetamidosalicylic acid located near the loop nodes. The triangles to the left represent the mean of the corresponding distributions under denaturing conditions (2 M GndHcl, pH 7.0), and the triangles at the right-hand side are the corresponding values under equilibrium. Color codes, Loop I (blue), loop II (red); loop II with a perturbation mutation (M34A, which shows no effect upon the rate of loop closure) (gray), Loop II with two replacements (Ala) at both residues 67 and 68, which are thought to be part of the loop node (green). Loop III (black).

sequences of few (∼3−5) residue segments on each loop's end. This hypothesis can be tested by site-directed mutagenesis of the relevant segments. An interesting example of a significant mutagenesis test is the AK N-terminal cluster of nonpolar residues (i.e., 1 and 3−6), which form several nonlocal interactions in the native state including closure of loops I and III. This test will be shown in Section 3 to be an essential element of the SCM analysis of the AK folding pathway. We conclude that this short segment seems to be critical for the formation of the native structure of AK. Several attempts to produce foldable AK mutants failed when residues R2, I3, or I4 were replaced by either Cys or Trp or Ala residues.[80] This outcome is probably due to lack of closure of loop I as well as loop III affecting both the folding mechanism and the native state stability.

A second example is perturbation of the sequence assumed to be the loop II node. Mutation L35A results in a negligible expression yield of this AK mutant, which is an indication of a perturbed folding mechanism and loss of most of the incompletely folded molecules by digestion. Mutation of residues 34, 67, or 68 to Ala (one at a time) had a minor effect on the fast closure of loop II and the stability of the folded mutants. However, simultaneous mutation of both residues 67 and 68 resulted in a mutant that showed slow closure of loop II but still folding to the native structure. Thus, residues 35, 67, and 68 seem to contribute to the specific nonlocal interaction that enables fast closure of loop II in the otherwise disordered molecule. Perhaps the side chain of LEU35 is inserted as a wedge between residues 67 and 68.

2.4. Additional Studies of AK Folding. Following the first 1996 trFRET experiments AK became a popular model for folding and dynamic studies. Several groups studied the conformational dynamics of native AK.[81−86] Haran and co-workers studied AK folding by smFRET methods.[30] These studies were done at equilibrium under partial unfolding conditions. Six metastable states were observed by means of hidden Markov model analysis of detected single-molecule trajectories.[87] Further analysis[87,88] showed that the multiple folding states (at low ∼0.5 M urea) converged to one sequential dominant mechanism at higher concentrations (∼1 M urea), which get closer to the cooperative transition. Destabilizing mutations at the interface between two domains perturbed the folding pathways of wt-AK and populated alternative pathways.[89] Mantulin and co-workers[90] studied the kinetics of folding of AK by monitoring tryptophan emission and found sequential folding phases in agreement with our time scales.

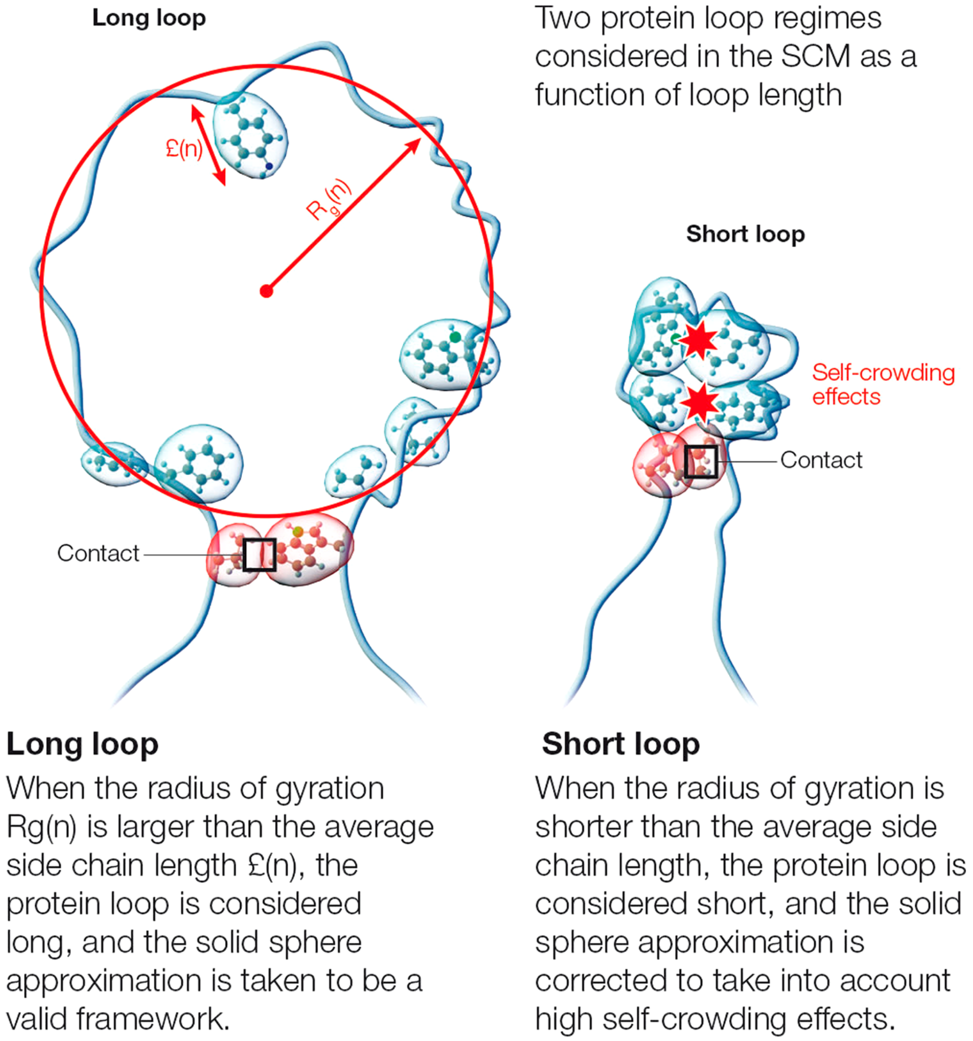
Several coarse-grained simulations of AK folding pathway were reported. Nussinov and co-workers[91] identified the N terminal loop (loop I above) as a building block critical for the folding of AK. Gosavi[92] used coarse-grained simulation of the folding pathway of AK and concluded that his simulation closely reproduced the sequential folding steps detected by our first set of double-kinetics experiments. Coarse-grained simulation by Li et al.[93] reproduced the folding rate detected by the double-kinetics experiments and detected successive transitions among five substates in the CORE domain.

Our trFRET double-kinetics experiments enable us to get closer to the ideal folding experiment portrayed (Section 2) above. We obtain transient distributions of intramolecular distances between multiple pairs of sites specifically labeled by site-directed labeling. Unlike the new and popular smFRET experiments (so far done mainly under equilibrium conditions), trFRET experiments produce strong enough flux of photons that enable microsecond double-kinetics experiments. Our experiments employ probes of small size and natural amino acyl FRET pairs, which were inserted at multiple pairs of sites in the AK molecule with minimal structural perturbation. This enables us to probe the folding of each subdomain element and obtain a global time-dependent map of the multiple transitions from multiple different labeled mutants. These probes also allow selection of small Förster critical distances in the range from 10 to 40 Å, which enable fine resolution of distances and detection of close contacts.

We achieve highly detailed time-dependent mapping of the evolution of the ensemble of refolding AK molecules with microsecond time resolution and subdomain spatial resolution. We detect the transition of each selected (by labeling) subdomain element and test specific hypotheses related to the stepwise self-assembly of the AK native structure. We are testing the role of selected short clusters of residues that contribute to each step in the global transition. We focus on the pretransition state steps by the ultrafast mixing by the microfluidic device starting from the ensemble of fully disordered molecules. We already monitored the kinetics of closure of six long loops and several secondary structure elements in the AK molecule. This work is in progress, and we are on our way to obtaining a full time-resolved map of the subdomain transitions in the AK molecule. We tested the loop hypothesis in the case of AK, and we tested some of the theoretical predictions obtained by the application of the SCM to the AK case.

2.5. Do the Nonlocal Contacts Depend on Secondary Structure? An interesting question is whether the initial nonlocal interaction defining a protein loop, which form helices or β strands in the native structure, depends on prior or simultaneous establishment of the respective nodes' segments'

DOI: 10.1021/acs.jpcb.8b11634



Two protein loop regimes considered in the SCM as a function of loop length

Long loop

When the radius of gyration Rg(n) is larger than the average side chain length £(n), the protein loop is considered long, and the solid sphere approximation is taken to be a valid framework.

Short loop

When the radius of gyration is shorter than the average side chain length, the protein loop is considered short, and the solid sphere approximation is corrected to take into account high self-crowding effects.

Figure 3. A schematic of the two protein loop regimes considered in the SCM as a function of loop length n. (left) When the radius of gyration $R_g(n)$ is larger than the average side-chain length $£(n)$, the protein loop is considered long, and the solid sphere approximation is taken to be a valid framework. (right) When the radius of gyration $R_g(n)$ is shorter than the average side-chain length $£(n)$, the protein loop is considered short, and the solid sphere approximation is corrected to take into account high self-crowding effects.

secondary structures. This issue remains an open question.[80] Fersht[94] concludes that, only when the propensity of establishing stable secondary structures is high, do they form first, followed by their assembly. But, for the majority of proteins, the unstable secondary structures are also stabilized by nonlocal interactions. In the case of AK, the kinetics of local folding of five secondary structure elements in the CORE domain (β_1, β_3, β_9, α_2, and α_7) were studied by the double-kinetics experiment. The short helix, α_2, at the N terminal end of loop II was folded in the microseconds time regime, while strands β_1 and β_3 (i.e., part of the nodes of loops I and III) were disordered at the time of the loop closure transition.[29,95] Strand β_4, which forms the C terminal side of the node of loop IV (with strand β_1 on the N terminal side) is also disordered at the time of the loop IV initial closure.[29] Thus, the closure of loops I, III, and IV seems to be independent of the secondary structures of their nodes. Similarly, in the folding of the B domain of protein A two early loops are closed between two helix-forming segments, but in both cases nonlocal contacts were formed between the folded helix and unfolded helix-forming segment.[67] Thus, our experiments suggest that, at least in some cases, loops are closed independently of the formation of local secondary structure elements, particularly in the case of β strands.

3. NONLOCAL INITIATION OF THE FOLDING PATHWAY: THE SEQUENTIAL COLLAPSE MODEL (SCM)

3.1. Early Nonlocal Contacts Constrain the Random Search toward the Native Structure. The core physical elements of the hypothesis presented in this Feature Article have been extensively studied theoretically within the SCM, independent of the experimental evidence gathered within the LH context and presented in Section 2. This section will give an overall summary of the SCM with further details found in the references. The SCM was developed starting ~25 years ago to investigate protein folding pathways, with a focus on intermediate states whose main element consisted of early nonlocal contact formation leading to protein loops. The present form of the SCM relies on a combination of coarse-grained physics and empirical results to predict the location of the earliest possible nonlocal contacts along the folding pathway of proteins longer than ~100 amino acids, from primary sequence information alone.[48,96−99] It has also been successfully applied to the study of (a) the kinetics of two-state transitions,[100,101] (b) the effects of macromolecular crowding on protein folding thermokinetics,[102−104] and (c) nonlocal contact formation in α-synuclein.[105] Here we will just focus on the earliest nonlocal contact formation events in protein folding, and it is bearing on the subsequent folding pathways.

In the SCM, the entropic cost of forming protein contacts is studied at low resolution, assuming that, to a reasonable approximation, the degrees of freedom of the unfolded protein can be separated into the contributions of (a) the protein backbone and (b) the chemical properties of the side chains of the amino acids involved in the contact and the intervening loop. This assumption has recent support from some MD simulations across the protein database.[106] Thus, the entropic cost ΔS_{loop} of forming a protein loop of n amino acids can be written to a first approximation as

$$\Delta S_{\text{loop}} \approx \Delta S_{\text{contact}} + \Delta S_{\text{loop,backbone}} + \Delta S_{\text{loop,sidechains}} \tag{1}$$

The term $\Delta S_{\text{contact}}$ represents the entropic cost of constraining the side chains of the amino acids forming the contact, and it opposes contact formation when $\Delta S_{\text{contact}} > 0$. A precise sequence-specific determination of the value $\Delta S_{\text{contact}}$ for a given contact would require advanced MD techniques; however, its determination is not critical if very weak early contacts are consistently detected in experiments (see section 2). For a coarse-grained model, it is reasonable to assume that the value of $\Delta S_{\text{contact}}$ can be taken as roughly constant for any contact defined by relatively similar amino acid types, based on the observation that protein folding is usually resilient to conservative amino acid substitution.[107] The term $\Delta S_{\text{loop,backbone}}$ represents the entropic cost of constraining the backbone into a loop conformation. In the SCM it is taken to have the classical Jacobson-Stockmeyer (JS) form, which yields a result for the entropic cost of forming loops in polymers with excluded volume, when the monomer volume is taken in Flory style as solid balls,[108] such that

$$\Delta S_{\text{loop,backbone}} \approx -3/2\ln(n) + C \tag{2}$$

where n is the loop length, and C is a constant. Kinetic loop closure experiments both with short peptides[109] and in protein loops[110] have provided significant support to the proposition that eq 1 represents a good approximation to the entropic cost of forming loops in globular proteins as a function of loop size. Because the JS entropic cost of forming loops grows monotonically with loop size, these observations were taken generally to imply that there was a strong preference for short-range contacts (i.e., short loops) at all stages of the folding process. Finally, the term $\Delta S_{\text{loop,side chains}}$ represents the entropic cost of constraining the side chains of the amino acids (i.e., included in a protein loop) into a much smaller volume than in the unfolded chain, even if those side chains are not involved in the contact itself. The side-chain term is then expected to satisfy $\Delta S_{\text{loop,side chains}} < 0$ and oppose formation of the contact. Also, because the side chains contain many more degrees of freedom than the backbone, within the SCM it is assumed that $|\Delta S_{\text{loop,side chains}}| > |\Delta S_{\text{loop,backbone}}|$, for loops small enough to induce a significant increase of steric hindrance between side chains not close along the sequence upon loop formation.[48]

Author: In the SCM, the side-chain entropic term $\Delta S_{\text{loop,side chains}}$ is assumed to drive the formation of nonlocal contacts in the earliest folding stages based on a coarse-grained analysis of the excluded volume effects of forming protein loops: Two types of nonlocal contacts are distinguished as a function of the resulting loop length: (a) short-range nonlocal contacts are those for which the Gaussian radius of gyration of the resulting loop is smaller than the average side-chain length $£(n)$, such that $R_g(n) < £(n)$; for these short loops the JS Flory-like solid sphere approximation is inadequate, as loop formation will induce increased interpenetration in the absence of significant constraints on the available configurations; (b) long-range nonlocal contacts are those for which the Gaussian radius of gyration of the resulting loop is larger than side-chain length $R_g(n) \gtrsim £(n)$. For such long loops, the amino acid side chains do not need to interpenetrate more than in the unfolded chain, and a Flory-like approach is likely to be satisfactory. Both loop formation regimes are represented in Figure 3. It is then reasonable to expect that the transition from short to long loops takes place when $R_{\text{g,loop}}(n) \approx £(n)$. For a typical proteins sequence $R_{\text{g,loop}}(n) \approx £(n)$ for $n \approx 65$ amino acids, and this loop length is called the optimal length, n_{op}. However, the exact value of $£(n)$ will be sequence-dependent. Moreover, a fully detailed model should go beyond consideration of the coarse-grained $£(n)$ measure of side-chain size and calculate the optimal loop length as a function of the detailed structure of the side chains in the loop.

The long-loop regime is physically equivalent to the original JS picture, and the entropic cost of forming protein loops is well-represented by eq 1 in the solid ball limit. In the short-loop regime, however, the internal degrees of freedom of the side chains cannot be neglected, and the entropic cost of forming short loops must be higher than would arise from treating the amino acids as solid spheres. The difference in the physical regimes for short and long loops in the SCM is depicted in Figure 3. Moreover, because most of the degrees of freedom are in the side chains, we expect the contribution of the side chains to the overall entropic cost to be dominant with respect to that of the backbone. Thus, in the SCM, it is expected that, for a short loop, the entropic cost of loop formation may be approximately expressed as

$$\Delta S_{\text{loop}} \approx \Delta S_{\text{contact}} - 3/2\ln(n) + \Delta S_{\text{loop,sidechains}}(n, £) + C \tag{3}$$

with $\Delta S_{\text{loop,side chains}} \ll 0$, and opposing folding. When $R_g(n) \gtrsim £(n)$, we have $\Delta S_{\text{loop,side chains}} \approx 0$, and the fully JS regime is recovered. The side-chain crowding term $\Delta S_{\text{loop,side chains}}$ will appear as a thermodynamic correction to the JS results for shorter loops. Experimental kinetic observations show that the behavior of the entropic loss due to short-loop formation is roughly JS-like.[109,110] We do not expect a direct correlation between these experiments and the SCM explanation for nonlocal early contact formation, as the term $\Delta S_{\text{loop,side chains}}(n, £)$ appears in the model as a thermodynamic consequence of stable contact formation. Given the complexity of considering the entropy of a full protein for long time scales, it is only reasonable to expect that unveiling the general properties of $\Delta S_{\text{loop,side chains}}(n, £)$ and its precise sequence dependence will require a fuller consideration of side-chain geometry and dynamics.

Thus, in the SCM, Nature's Shortcut to the folding of relatively long proteins (i.e., $\gtrsim$100 amino acids), is likely to involve early nonlocal contacts, with initial preference in most proteins for contacts separated by $n \gtrsim 65$ amino acids, called primary contacts in the model, with the values of n modulated by the sequence through the $£(n)$-dependent term. The

DOI: 10.1021/acs.jpcb.8b11634

entropy of loop closure, and the attractive interactions that stabilize the nonlocal contact, determine the location of the primary contact. The initial formation of the nonlocal contacts in the SCM nucleates a specific cooperative molecular collapse (i.e., a TSE) dominated by short-range nonlocal loops,[100] similar to the general nucleation-condensation mechanism.[20] Because primary contacts are established between segments located at distances comparable to the overall protein length, only a few early primary contacts can be established, and most of the native nonlocal contacts will form at $n < n_{\text{op}}$, usually with n closer to the minimum loop size that allows for contact formation, which in the model is taken to be $n_{\min} \approx 10-15$ amino acids, likely depending as well on the excluded volume-driven persistence length of the unfolded chain in the region.[48,100] The specifically collapsed state including the set of nonlocal interactions that broadly define the native topology is then expected to guide a constrained search of the final native structure through a series of slower processes involving the full exclusion of water from the protein core and the establishment of the detailed interactions characteristic of native structures. Thus, the SCM provides the theoretical justification for the view presented here and contains the necessary ingredients to develop from the primary sequence the successive establishment of the early loops. The SCM is in the same spirit as that of theoretical efforts to develop simple models able to describe the intermediates along the folding pathway in the context of the funnel view of the folding process.[111,112] The SCM, however, also shares much of the "classical" view of the folding process, in which the protein descends toward the free energy minimum through a few highly deterministic intermediate steps. Both views have been shown to be compatible, in principle.[24] Also, the possibility that nonlocal contacts might play a role in folding nucleation is not exclusive of the SCM. Earlier models such as Diffusion-Collision[8] naturally include the possibility that diffusing pieces of the protein chain, separated along the sequence, might collide and nucleate folding. The novelty in the SCM is the provision of additional physics through a fuller consideration of the entropic consequences of forming loops, that rest on utilizing primary sequence information to unambiguously predict the location of the nonlocal contacts that nucleate the native folding pathway.

After a nonlocal contact is formed, thus initiating Nature's Shortcut, in the SCM the folding process proceeds through the two-state topological (i.e., establishment of a set of specific loops) transition of the region included in the intervening loop.

3.3. Two-State Molecular Collapse Transitions Are Kinetically Determined by Loop Dynamics.

In the SCM, formation of the overall topology of the folded subregion (i.e., occurring naturally through random search) occurs prior to the full apparent two-state transition. Then, the rate of the two-state transition for any protein region is governed in the SCM by a kinetic rate equation[100,101]

$$\ln \kappa_{\text{f}} \approx \ln g - \text{RCO}\Delta G_{\text{conf}} \tag{4}$$

where κ_{f} is the rate of two-state collapse, g is the characteristic diffusional frequency, RCO is the relative contact order (i.e., the average loop size of the final topology), and ΔG_{conf} is the entropic free energy cost of folding, which can be simply written in Boltzmann-Gibbs form as $\Delta G_{\text{conf}} = -kTL\ \ln(f_{\text{f}}/f_0)$, where L is the length in terms of the number of amino acids of the collapsing region, and f_{f} and f_0 are the average number of configurations accessible to the amino acids in the folded and the unfolded state, respectively. Equation 1 reproduces the experimentally observed Plaxco-Baker dependence of two-state rates with contact order,[35] when we employ the experimental value $g \approx 1 \times 10^7\ \text{s}^{-1}$, which is close to the observed folding speed limit.[113] From the SCM and LH Feature Articles, in accordance with the view presented here, the native topology of proteins larger than ~100 amino acids is defined by an earlier configurational search in the unfolded state for the correct topology as defined by loops determined from excluded volume effects and attractive hydrophobic interactions, nucleated by one or a few nonlocal contacts. There is theoretical evidence showing that the two-state folding rates of globular proteins can be obtained, if the assumption is made that the protein folds from a topologically native-like unfolded, yet compact, state,[114−117] consistent with the main contention of the Feature Article.

It is now easy to show that formation of primary contacts greatly accelerates folding in the SCM. We assume that collapse is preceded by formation of a nonlocal contact that constrains the region consisting of n amino acids in a protein loop. The volume $V_{\text{loop}}(n)$ of a protein loop of length $\sim n$ is in the Gaussian approximation $V_{\text{loop}}(n) \approx (1/2)^{3/2} V_{\text{open}}(n)$, where $V_{\text{open}}(n)$ is the volume of an equivalent unfolded region. This large reduction in the available freedom due to formation of a nonlocal contact can then be introduced in eq 1, assuming that the conformational freedom of the fully unfolded chain undergoes an equivalent reduction upon formation of a nonlocal contact that constrains the region n in a loop, that is, $f_0(\text{loop}) \approx 0.354 f_0(\text{open})$, and the conformational cost of folding then becomes $\Delta G_{\text{conf}}\ (n, \text{loop}) \approx \Delta G_{\text{conf}}\ (n, \text{open}) - 1.04n$. For a protein of more than ~100 amino acids, within the Gaussian approximation, the effect of establishing an early nonlocal contact between residues 65 amino acids apart along the sequence is a reduction of the entropic barrier to collapse of ~68 kT. The effect on the rate is equally dramatic, $\kappa_{\text{f}}(\text{loop}) \approx \exp(-\text{RCO } 68)\kappa_{\text{f}}(\text{open})$, which for typical values of RCO from 0.1 to 0.2 gives $\kappa_{\text{f}}(\text{loop}) \approx 9 \times 10^{-3} - 8 \times 10^{-6}\ \kappa_{\text{f}}(\text{open})$. Thus, nucleation by primary contacts is an effective mechanism to greatly facilitate the stochastic search for the native topology, thereby supporting the SCM's contention that primary contact formation is the key to Nature's Shortcut to the 3D structure. This calculation is a rough estimate, as the Gaussian approximation does not take into account nonspecific hydrophobic effects, which might lead to a higher degree of initial compaction, and chain self-avoidance that would counterbalance such an effect to some extent.

The large reduction in conformational entropy, attainable through the formation of just a few early nonlocal contacts, lends support to the existence of a minimal entropy variational principle implied by the SCM. According to this "least entropy pathway" principle,[49] the protein folding pathway is driven by minimization of the entropic cost of folding initiated by a heavily constrained initial scaffold of nonlocal contacts that determines the optimal set of loops driving the protein at the fastest possible rate toward the 3D structure. Unveiling the details of the SCM entropic variational principle governing folding remains an open challenge.

3.4. Practical Testing of the SCM Predictions from Primary Sequences, and Particular Results for Adenylate Kinase.

Controlled experimental tests of the above predictions can include (a) persistent structural elements in largely unfolded states of globular proteins and (b) kinetic experimental evidence on the earliest folding stages. Evidence of substantial native-like topological elements in heavily denatured states (up to 8 M urea) of a truncated form of staphyloccocal nuclease was discovered by Ackerman and Shortle.[118] Several observations of nonlocal contacts in denatured states of globular proteins were reported,[52,119−121] including nonlocal interactions in lysozyme,[52] which were disrupted by mutation of Trp62 to gly. Regarding kinetic experimental evidence, the SCM results for the earliest folding events have been compared mostly with proton exchange data and shown to be consistent for the proteins studied.[48,96−98,101] However, such comparisons might not be conclusive, as the experimental time scales involved (~ms) are long enough that substantial structure may already have formed, and purely local interpretations of the observed interactions are possible.[122−127]

The SCM has been applied to a substantial set of proteins to reveal the earliest nonlocal contacts, employing only primary sequence information, and the results have been shown to be consistent with experimental evidence.[48,96−98,101] In particular, initial nonlocal contacts, consistent with the structural and kinetic data available, were predicted for cythochrome *c*, apomyoglobin, barnase, and ribonuclease A.[48] Additionally, the SCM prediction for the earliest events along the folding pathways of two structurally and kinetically related proteins, namely, lysozyme and α-lactalbumin, were compared,[96] particularly also showing that the model predicts that major rearrangements of the protein core are entropically unfavorable and that the formation of more than one early nonlocal contact likely leads to the formation of autonomous folding domains.[96] Interestingly, independent experiments on protein SH2 have shown that the entropic cost of forming 60−80 loops is less than theoretically expected from classical polymer theory suggesting that loops of this length play a significant role in domain formation.[128] The model was then shown to be able to reproduce the observed large differences in the observed proton exchange folding pathways of two structurally very similar proteins, apomyoglobin and apoleghemoglobin.[97] The issue of pathway degeneracy emerging from alternative initial contacts was studied for β-lactoglobulin[101] and staphylococcal nuclease.[98] As described in Section 3.2, the model was employed to derive an equation describing the kinetics of molecular two-state collapse, both for proteins that had undergone an initial collapse nucleated by a nonlocal contact[100] and for small proteins with apparent two-state kinetics.[101] The resulting equations were shown to be consistent with the observed Plaxco-Baker relationship between a protein's topology and folding rates, thus underscoring the relevance of considering the dynamics of protein loops for a description of overall folding kinetics. The existing SCM predictions for primary contacts for 10 proteins are presented in Tables S1 and S2 in the Supporting Information, together with their representation on the accompanying the 3D structure figures. The excellent agreement between the SCM predictions and the location on the 3D structure of the contacts, in all cases, lends support to the hypothesis that the primary contact is the key to Nature's Shortcut.

Considering the growing importance of research into pathogenic proteins, the SCM was applied to calculate the entropic barrier to folding of the murine prion protein *m*PrP(121−231). It was shown that the model predicts that there is a direct correlation between average protein flexibility and folding rate.[129] This work was done while developing a model for the effects of macromolecular crowding on protein stability and folding kinetics, and the results were applied to study the effects of macromolecular crowding on the stability and kinetics of *m*PrP(121−231).[103,104,130] Also, the model was recently applied to predict the location of the nonlocal contacts in the intrinsically disordered α-synuclein.[72]

We applied the SCM to the prediction of the best (i.e., the most thermodynamically favorable) initial nonlocal contacts in AK and compared the results to the trFRET studies of the Haas lab (see Section 2). The best primary contact for AK is predicted to form between segments I3-G7, centered at L5, and N79-L83, centered at D81. The second best possible contact is established between segments Y105−F109 centered at L107 and L178-Y182, centered at G180.

Only the best possible contact was so far probed in the trFRET experiment (see Section 2), corresponding closely to loop III in the 3D structure. It appears to form within the experimental time of ~50 μs, in agreement with the SCM prediction. The contact also includes most of the mutation-ultrasensitive segment M1-L6 that seems to be critical for the folding of AK, an observation consistent with the SCM proposal that formation of the best primary contact (I3-G7, N79-L83) constitutes an essential nucleation point for the early folding of the N-terminal domain of AK. Whether the best primary contact (I3-G7, N79-L83) also constitutes a key nucleation event for the formation of the shorter loops that also appear folded in the trFRET experiment could only be determined by higher time resolution or mutagenesis. Interestingly, the region around residue 66 is by far the most hydrophobic within the best predicted primary loop. This observation suggests that it might play a key role in nucleating the specific set of subsequent contacts within the primary loop, in accordance with the experimental results presented in section 2.

Because the SCM postulates that the earliest nonlocal contacts are weak, it is not expected that these initial contacts translate unequivocally into native nonlocal contacts. It is however reasonable to expect that they are spatially near in the 3D structures. The segments defining the predicted best contact are shown on the 3D structure in Figure 4.

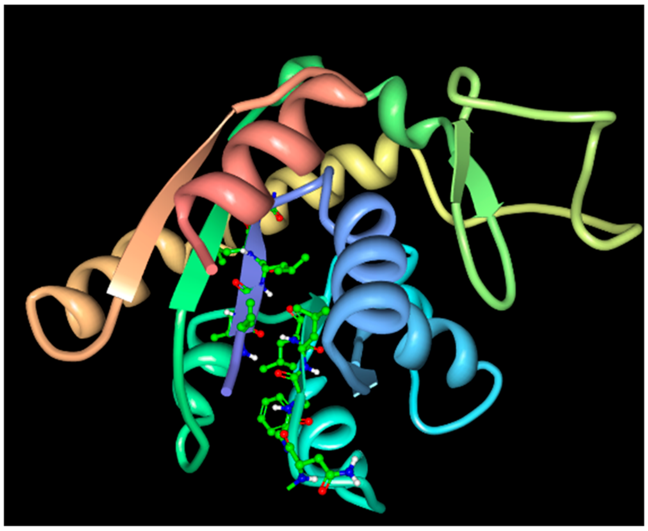


Figure 4. Best predicted SCM primary contact for *E. coli* AK (I3-G7, N79-L83), displayed on the ribbon 3D structure. The side chains of the amino acids involved in the contact are displayed in atomic detail, the N-terminus of the protein is represented in blue, while the C-terminus is represented in pink.

4. CONCLUSION

In this Feature Article, we suggest that the ability to locate pairs of clusters of residues that can interact and form nonlocal contacts early along the folding pathway constitute Nature's Shortcut to the efficient folding of globular proteins. Also, we argued that identification of the early nonlocal contacts by SCM from the primary sequence, and their utilization in MD calculations, can bolster the capability to ab initio predict the 3D structures of long proteins. We believe that further development of the SCM with more detailed calculations, and the study of early loop closure in many proteins with spectroscopic methods (e.g., trFRET), can fully assess the Feature Article claimed to be Nature's Shortcut to folding. The final goal of these efforts would be the development of a complete understanding of a common set of rules determining the establishment of the early set of nonlocal contacts and their influence on the later detailed folding events.

This Feature Article on protein folding presented here opens a new set of questions that will likely call for the application of advanced experimental and theoretical techniques for their elucidation. On the experimental side, if the set of early nonlocal contacts is established on the microsecond time scale, its dynamics can only be probed fully by refining existing ultrafast techniques of protein dynamics characterization, such as with trFRET, and possibly developing new spectroscopic tools specifically tailored to the task. From a theoretical point of view, a more detailed version of the SCM that takes into fuller account the microscopic details of the chain constituents at each stage is needed to completely elucidate the early dynamics of nonlocal loops. Moreover, theoretical developments within the SCM could also provide additional insight into determining the set of loops that constrain the early folding stages, employing just primary sequence information, thereby possibly guiding future experiments. As remarked earlier, an important goal is an essentially ab initio model based on an entropy optimization principle.

We propose that a thorough program, both experimental and theoretical, based on the concepts presented in this Feature Article, will lead to an elucidation of Nature's Shortcut to the fast and efficient folding of long proteins. Further efforts are needed to (a) reveal and characterize early nonlocal contacts to assess their decisive role for the initial folding nucleation, as well as (b) discover the rules by which the scaffold of early nonlocal contacts translates into the proteins' complete 3D structure. We suggest that the notions in this Feature Article and these further advances could provide a basis to better understand and predict the process of protein folding.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jpcb.8b11634.

> The results obtained in previous work for the primary contacts of 10 proteins within the SCM are summarized in condensed form, including graphical representations of the predicted contacts on the 3D structures (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: Bergasa@princeton.edu.

ORCID

Herschel A. Rabitz: 0000-0002-4433-6142

Notes

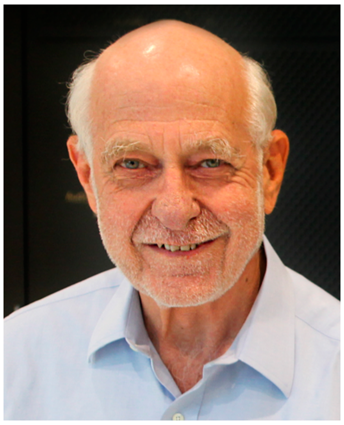
The authors declare no competing financial interest.

⊥E-mail: Elisha.Haas@gmail.com. (E.H.)
#E-mail: HRabitz@Princeton.edu. (H.A.R.)

Biographies

Fernando Bergasa-Caceres obtained a B.S. in Biochemistry and Molecular Biology at the Universidad Autonoma de Madrid (UAM) in 1990. While at the UAM he did research in scanning probe microscopy of proteins under the supervision of Prof. Juan J. Saenz. After spending 1991 at Tufts University studying protein folding with Prof. David L. Weaver, he joined Prof. Herschel A. Rabitz's lab at Princeton in 1992, receiving his Ph.D. in Chemistry from Princeton University in 1996. In collaboration with Herschel A. Rabitz, Fernando Bergasa-Caceres has continued to publish in the field of protein folding. He serves on the Alumni Advisory Board and the Advisory Board at the Instituto de Fisica de la Materia Condensada, both at the UAM. Fernando Bergasa-Caceres is currently the Chairman and CEO of Redexis in Spain.

Elisha Haas received his Ph.D. in Physical Chemistry from the Feinberg Graduate School of the Weizmann Institute in 1976. In 1979, Professor Haas joined the faculty of the Bar-Ilan University Department of Life Sciences. From 1992 he is the chairman of Bar-Ilan University's Biophysics program, and since 2015 he is the president of the Israel Biophysical Society. Professor Haas's research interests lie at the interface of physics, chemistry, and biology with principal focus on the problem of protein folding and dynamics, biochemical kinetics, and fluorescence spectroscopy.



Herschel A. Rabitz received his Ph.D. in Chemical Physics from Harvard University in 1970. In 1971, Professor Rabitz joined the faculty of the Princeton University Department of Chemistry, and from 1993 to 1996, he was chair of the department. He is also an affiliated member of Princeton University's Program in Applied and Computational Mathematics. Professor Rabitz's research interests lie at the interface of chemistry, physics, and engineering, with principal areas of focus including molecular dynamics, biophysical chemistry, chemical kinetics, optical interactions with matter, and molecular scale systems analysis.

■ ACKNOWLEDGMENTS

This study was supported by grants from the Israel Science Foundation (ISF1464/10 and the I-CORE 1902/12) and NSF IDBR1353942 (to O.B.). The collaborative work of E.H. and O.B. was enhanced by a Research Coordination Network grant, MCB1051344, from the National Science Foundation. H.R. acknowledges support from the NSF (Grant No. CHE-1763198). We thank D. Amir, T. Orevi, O. Bilsel, and G. Rahamim who performed the experimental work described in Section 2.

■ REFERENCES

(1) Anfinsen, C. B.; Haber, E.; Sela, M.; White, F. H., Jr. The Kinetics of Formation of Native Ribonuclease During Oxidation of the Reduced Polypeptide Chain. *Proc. Natl. Acad. Sci. U. S. A.* **1961**, *47*, 1309−1314.

(2) Englander, S. W.; Mayne, L. The Case for Defined Protein Folding Pathways. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 8253−8258.

(3) Chung, H. S.; Eaton, W. A. Protein Folding Transition Path Times from Single Molecule Fret. *Curr. Opin. Struct. Biol.* **2018**, *48*, 30−39.

(4) Gelman, H.; Gruebele, M. Fast Protein Folding Kinetics. *Q. Rev. Biophys.* **2014**, *47*, 95−142.

(5) Levinthal, C. Are There Pathways of Protein Folding. *J. Chim. Phys. Phys.-Chim. Biol.* **1968**, *65*, 44−45.

(6) Anfinsen, C. B.; Scheraga, H. A. Experimental and Theoretical Aspects of Protein Folding. *Adv. Protein Chem.* **1975**, *29*, 205−300.

(7) Kim, P. S.; Baldwin, R. L. Specific Intermediates in the Folding Reactions of Small Proteins and the Mechanism of Protein Folding. *Annu. Rev. Biochem.* **1982**, *51*, 459−489.

(8) Karplus, M.; Weaver, D. L. Diffusion−Collision Model for Protein Folding. *Biopolymers* **1979**, *18*, 1421−1437.

(9) Englander, S. W.; Mayne, L. The Nature of Protein Folding Pathways. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 15873−15880.

(10) Naganathan, A. N.; Munoz, V. Scaling of Folding Times with Protein Size. *J. Am. Chem. Soc.* **2005**, *127*, 480−481.

(11) Eaton, W. A.; Munoz, V.; Hagen, S. J.; Jas, G. S.; Lapidus, L. J.; Henry, E. R.; Hofrichter, J. Fast Kinetics and Mechanisms in Protein Folding. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 327−359.

(12) Baldwin, R. L. The Search for Folding Intermediates and the Mechanism of Protein Folding. *Annu. Rev. Biophys.* **2008**, *37*, 1−21.

(13) Fersht, A. R. Nucleation Mechanisms in Protein Folding. *Curr. Opin. Struct. Biol.* **1997**, *7*, 3−9.

(14) Eaton, W. A.; Wolynes, P. G. Theory, Simulations, and Experiments Show That Proteins Fold by Multiple Pathways. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, E9759−E9760.

(15) Dill, K. A.; Ozkan, S. B.; Shell, M. S.; Weikl, T. R. The Protein Folding Problem. *Annu. Rev. Biophys.* **2008**, *37*, 289−316.

(16) Baldwin, R. L.; Rose, G. D. Is Protein Folding Hierarchic? Ii. Folding Intermediates and Transition States. *Trends Biochem. Sci.* **1999**, *24*, 77−83.

(17) Baldwin, R. L. Clash between Energy Landscape Theory and Foldon-Dependent Protein Folding. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 8442−8443.

(18) Kim, P. S.; Baldwin, R. L. Intermediates in the Folding Reactions of Small Proteins. *Annu. Rev. Biochem.* **1990**, *59*, 631−660.

(19) Karplus, M.; Weaver, D. L. Protein-Folding Dynamics. *Nature* **1976**, *260*, 404−406.

(20) Daggett, V.; Fersht, A. R. Is There a Unifying Mechanism for Protein Folding? *Trends Biochem. Sci.* **2003**, *28*, 18−25.

(21) Ferreiro, D. U.; Komives, E. A.; Wolynes, P. G. Frustration, Function and Folding. *Curr. Opin. Struct. Biol.* **2018**, *48*, 68−73.

(22) Wolynes, P. G.; Eaton, W. A.; Fersht, A. R. Chemical Physics of Protein Folding. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 17770−17771.

(23) Wolynes, P. G.; Onuchic, J. N.; Thirumalai, D. Navigating the Folding Routes. *Science* **1995**, *267*, 1619−1620.

(24) Lazaridis, T.; Karplus, M. "New View" of Protein Folding Reconciled with the Old through Multiple Unfolding Simulations. *Science* **1997**, *278*, 1928−1931.

(25) Weinkam, P.; Zong, C.; Wolynes, P. G. A Funneled Energy Landscape for Cytochrome C Directly Predicts the Sequential Folding Route Inferred from Hydrogen Exchange Experiments. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 12401−12406.

(26) Guinn, E. J.; Jagannathan, B.; Marqusee, S. Single-Molecule Chemo-Mechanical Unfolding Reveals Multiple Transition State Barriers in a Small Single-Domain Protein. *Nat. Commun.* **2015**, *6*, 6861.

(27) Zoldak, G.; Rief, M. Force as a Single Molecule Probe of Multidimensional Protein Energy Landscapes. *Curr. Opin. Struct. Biol.* **2013**, *23*, 48−57.

(28) Fernandez, J. M.; Li, H. Force-Clamp Spectroscopy Monitors the Folding Trajectory of a Single Protein. *Science* **2004**, *303*, 1674−1678.

(29) Orevi, T.; Rahamim, G.; Amir, D.; Kathuria, S.; Bilsel, O.; Matthews, C. R.; Haas, E. Sequential Closure of Loop Structures Forms the Folding Nucleus During the Refolding Transition of the Escherichia Coli Adenylate Kinase Molecule. *Biochemistry* **2016**, *55*, 79−91.

(30) Rhoades, E.; Cohen, M.; Gussakovsky, E.; Schuler, B.; Haran, G. Single Molecule Protein Folding. *Biophys. J.* **2004**, *86*, 616a−616a.

(31) Levitt, M.; Warshel, A. Computer Simulation of Protein Folding. *Nature* **1975**, *253*, 694−698.

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

(33) Piana, S.; Klepeis, J. L.; Shaw, D. E. Assessing the Accuracy of Physical Models Used in Protein-Folding Simulations: Quantitative Evidence from Long Molecular Dynamics Simulations. *Curr. Opin. Struct. Biol.* **2014**, *24*, 98−105.

(34) Perez, A.; Morrone, J. A.; Simmerling, C.; Dill, K. A. Advances in Free-Energy-Based Simulations of Protein Folding and Ligand Binding. *Curr. Opin. Struct. Biol.* **2016**, *36*, 25−31.

(35) Baker, D. A Surprising Simplicity to Protein Folding. *Nature* **2000**, *405*, 39−42.

DOI: 10.1021/acs.jpcb.8b11634

(36) Privalov, P. L. Stability of Proteins: Small Globular Proteins. *Adv. Protein Chem.* **1979**, *33*, 167−241.

(37) Halgren, T. A.; Damm, W. Polarizable Force Fields. *Curr. Opin. Struct. Biol.* **2001**, *11*, 236−242.

(38) Warshel, A.; Kato, M.; Pisliakov, A. V. Polarizable Force Fields: History, Test Cases, and Prospects. *J. Chem. Theory Comput.* **2007**, *3*, 2034−2045.

(39) Jacob, M. H.; Amir, D.; Ratner, V.; Gussakowsky, E.; Haas, E. Predicting Reactivities of Protein Surface Cysteines as Part of a Strategy for Selective Multiple Labeling. *Biochemistry* **2005**, *44*, 13664−13672.

(40) Berezovsky, I. N. The Diversity of Physical Forces and Mechanisms in Intermolecular Interactions. *Phys. Biol.* **2011**, *8*, 035002.

(41) Azia, A.; Uversky, V. N.; Horovitz, A.; Unger, R. The Effects of Mutations on Protein Function: A Comparative Study of Three Databases of Mutations in Humans. *Isr. J. Chem.* **2013**, *53*, 217−226.

(42) Hecht, M. H.; Nelson, H. C.; Sauer, R. T. Mutations in Lambda Repressor's Amino-Terminal Domain: Implications for Protein Stability and DNA Binding. *Proc. Natl. Acad. Sci. U. S. A.* **1983**, *80*, 2676−2680.

(43) Baase, W. A.; Eriksson, A. E.; Zhang, X. J.; Heinz, D. W.; Sauer, U.; Blaber, M.; Baldwin, E. P.; Wozniak, J. A.; Matthews, B. W. Dissection of Protein Structure and Folding by Directed Mutagenesis. *Faraday Discuss.* **1992**, *93*, 173−181.

(44) Sauer, R. T. Mutagenic Dissection of the Sequence Determinants of Protein Folding, Recognition, and Machine Function. *Protein Sci.* **2013**, *22*, 1675−1687.

(45) Zwanzig, R.; Szabo, A.; Bagchi, B. Levinthal's Paradox. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 20−22.

(46) Ittah, V.; Haas, E. Nonlocal Interactions Stabilize Long Range Loops in the Initial Folding Intermediates of Reduced Bovine Pancreatic Trypsin Inhibitor. *Biochemistry* **1995**, *34*, 4493−4506.

(47) Orevi, T.; Rahamim, G.; Hazan, G.; Amir, D.; Haas, E. The Loop Hypothesis: Contribution of Early Formed Specific Non-Local Interactions to the Determination of Protein Folding Pathways. *Biophys. Rev.* **2013**, *5*, 85−96.

(48) Bergasa-Caceres, F.; Ronneberg, T. A.; Rabitz, H. Sequential Collapse Model for Protein Folding Pathways. *J. Phys. Chem. B* **1999**, *103*, 9749−9758.

(49) Bergasa-Caceres, F. *Least Entropy Pathway for Protein Folding Pathways*. Ph.D. Thesis, Princeton University,1999.

(50) Lapidus, L. J.; Eaton, W. A.; Hofrichter, J. Dynamics of Intramolecular Contact Formation in Polypeptides: Distance Dependence of Quenching Rates in a Room-Temperature Glass. *Phys. Rev. Lett.* **2001**, *87*, 258101.

(51) Fierz, B.; Satzger, H.; Root, C.; Gilch, P.; Zinth, W.; Kiefhaber, T. Loop Formation in Unfolded Polypeptide Chains on the Picoseconds to Microseconds Time Scale. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 2163−2168.

(52) Klein-Seetharaman, J.; Oikawa, M.; Grimshaw, S. B.; Wirmer, J.; Duchardt, E.; Ueda, T.; Imoto, T.; Smith, L. J.; Dobson, C. M.; Schwalbe, H. Long-Range Interactions within a Nonnative Protein. *Science* **2002**, *295*, 1719−1722.

(53) Meng, W.; Lyle, N.; Luan, B.; Raleigh, D. P.; Pappu, R. V. Experiments and Simulations Show How Long-Range Contacts Can Form in Expanded Unfolded Proteins with Negligible Secondary Structure. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 2123−2128.

(54) Noivirt-Brik, O.; Hazan, G.; Unger, R.; Ofran, Y. Non Local Residue-Residue Contacts in Proteins Are More Conserved Than Local Ones. *Bioinformatics* **2013**, *29*, 331−337.

(55) 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.

(56) Ovchinnikov, S.; Park, H.; Varghese, N.; Huang, P. S.; Pavlopoulos, G. A.; Kim, D. E.; Kamisetty, H.; Kyrpides, N. C.; Baker, D. Protein Structure Determination Using Metagenome Sequence Data. *Science* **2017**, *355*, 294−298.

(57) Marks, D. S.; Hopf, T. A.; Sander, C. Protein Structure Prediction from Sequence Variation. *Nat. Biotechnol.* **2012**, *30*, 1072−1080.

(58) Raval, A.; Piana, S.; Eastwood, M. P.; Shaw, D. E. Assessment of the Utility of Contact-Based Restraints in Accelerating the Prediction of Protein Structure Using Molecular Dynamics Simulations. *Protein Sci.* **2016**, *25*, 19−29.

(59) Warshel, A. Multiscale Modeling of Biological Functions: From Enzymes to Molecular Machines (Nobel Lecture). *Angew. Chem., Int. Ed.* **2014**, *53*, 10020−10031.

(60) Förster, T. Zwischenmolekulare Energiewanderung Und Fluoreszenz. *Ann. Phys.* **1948**, *437*, 55−75.

(61) Haas, E. The Study of Protein Folding and Dynamics by Determination of Intramolecular Distance Distributions and Their Fluctuations Using Ensemble and Single-Molecule Fret Measurements. *ChemPhysChem* **2005**, *6*, 858−870.

(62) Sinha, K. K.; Udgaonkar, J. B. Dissecting the Non-Specific and Specific Components of the Initial Folding Reaction of Barstar by Multi-Site Fret Measurements. *J. Mol. Biol.* **2007**, *370*, 385−405.

(63) Rahamim, G.; Chemerovski-Glikman, M.; Rahimipour, S.; Amir, D.; Haas, E. Resolution of Two Sub-Populations of Conformers and Their Individual Dynamics by Time Resolved Ensemble Level Fret Measurements. *PLoS One* **2015**, *10*, No. e0143732.

(64) Rhoades, E.; Gussakovsky, E.; Haran, G. Watching Proteins Fold One Molecule at a Time. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 3197−3202.

(65) Hillger, F.; Nettels, D.; Dorsch, S.; Schuler, B. Detection and Analysis of Protein Aggregation with Confocal Single Molecule Fluorescence Spectroscopy. *J. Fluoresc.* **2007**, *17*, 759−765.

(66) Welker, E.; Maki, K.; Shastry, M. C.; Juminaga, D.; Bhat, R.; Scheraga, H. A.; Roder, H. Ultrarapid Mixing Experiments Shed New Light on the Characteristics of the Initial Conformational Ensemble During the Folding of Ribonuclease A. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 17681−17686.

(67) Rahamim, G.; Amir, D.; Haas, E. Simultaneous Determination of Two Subdomain Folding Rates Using the "Transfer-Quench" Method. *Biophys. J.* **2017**, *112*, 1786−1796.

(68) Teilum, K.; Maki, K.; Kragelund, B. B.; Poulsen, F. M.; Roder, H. Early Kinetic Intermediate in the Folding of Acyl-Coa Binding Protein Detected by Fluorescence Labeling and Ultrarapid Mixing. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 9807−9812.

(69) Ben Ishay, E.; Hazan, G.; Rahamim, G.; Amir, D.; Haas, E. An Instrument for Fast Acquisition of Fluorescence Decay Curves at Picosecond Resolution Designed for "Double Kinetics" Experiments: Application to Fret Study of Protein Folding. *Rev. Sci. Instrum.* **2012**, *83*, 084301.

(70) Kathuria, S. V.; Chan, A.; Graceffa, R.; Paul Nobrega, R.; Robert Matthews, C.; Irving, T. C.; Perot, B.; Bilsel, O. Advances in Turbulent Mixing Techniques to Study Microsecond Protein Folding Reactions. *Biopolymers* **2013**, *99*, 888−896.

(71) Gambin, Y.; VanDelinder, V.; Ferreon, A. C.; Lemke, E. A.; Groisman, A.; Deniz, A. A. Visualizing a One-Way Protein Encounter Complex by Ultrafast Single-Molecule Mixing. *Nat. Methods* **2011**, *8*, 239−241.

(72) Dingfelder, F.; Wunderlich, B.; Benke, S.; Zosel, F.; Zijlstra, N.; Nettels, D.; Schuler, B. Rapid Microfluidic Double-Jump Mixing Device for Single-Molecule Spectroscopy. *J. Am. Chem. Soc.* **2017**, *139*, 6062−6065.

(73) Raleigh, D. P.; Plaxco, K. W. The Protein Folding Transition State: What Are Phi-Values Really Telling Us? *Protein Pept. Lett.* **2005**, *12*, 117−122.

(74) Riback, J. A.; Bowman, M. A.; Zmyslowski, A. M.; Knoverek, C. R.; Jumper, J. M.; Hinshaw, J. R.; Kaye, E. B.; Freed, K. F.; Clark, P. L.; Sosnick, T. R. Innovative Scattering Analysis Shows That Hydrophobic Disordered Proteins Are Expanded in Water. *Science* **2017**, *358*, 238−241.

(75) Kathuria, S. V.; Kayatekin, C.; Barrea, R.; Kondrashkina, E.; Graceffa, R.; Guo, L.; Nobrega, R. P.; Chakravarthy, S.; Matthews, C. R.; Irving, T. C.; Bilsel, O. Microsecond Barrier-Limited Chain

Collapse Observed by Time-Resolved Fret and Saxs. *J. Mol. Biol.* **2014**, *426*, 1980−1994.

(76) Ziv, G.; Thirumalai, D.; Haran, G. Collapse Transition in Proteins. *Phys. Chem. Chem. Phys.* **2009**, *11*, 83−93.

(77) Fazelinia, H.; Xu, M.; Cheng, H.; Roder, H. Ultrafast Hydrogen Exchange Reveals Specific Structural Events During the Initial Stages of Folding of Cytochrome C. *J. Am. Chem. Soc.* **2014**, *136*, 733−740.

(78) Muller, C. W.; Schulz, G. E. Structure of the Complex of Adenylate Kinase from Escherichia Coli with the Inhibitor P1,P5-Di(Adenosine-5′-)Pentaphosphate. *J. Mol. Biol.* **1988**, *202*, 909−912.

(79) Orevi, T.; Ben Ishay, E.; Gershanov, S. L.; Dalak, M. B.; Amir, D.; Haas, E. Fast Closure of N-Terminal Long Loops but Slow Formation of Beta Strands Precedes the Folding Transition State of Escherichia Coli Adenylate Kinase. *Biochemistry* **2014**, *53*, 3169−3178.

(80) Orevi, T.; Rahamim, G.; Shemesh, S.; Ben Ishay, E.; Amir, D.; Haas, E. Fast Closure of Long Loops at the Initiation of the Folding Transition of Globular Proteins Studied by Time Resolved Fret-Based Methods. *Bio-Algorithms and Med-Systems* **2014**, *10*, 169−193.

(81) Henzler-Wildman, K. A.; Lei, M.; Thai, V.; Kerns, S. J.; Karplus, M.; Kern, D. A Hierarchy of Timescales in Protein Dynamics Is Linked to Enzyme Catalysis. *Nature* **2007**, *450*, 913−916.

(82) Olsson, U.; Wolf-Watz, M. Overlap between Folding and Functional Energy Landscapes for Adenylate Kinase Conformational Change. *Nat. Commun.* **2010**, *1*, 1−8.

(83) Whitford, P. C.; Miyashita, O.; Levy, Y.; Onuchic, J. N. Conformational Transitions of Adenylate Kinase: Switching by Cracking. *J. Mol. Biol.* **2007**, *366*, 1661−1671.

(84) Whitford, P. C.; Gosavi, S.; Onuchic, J. N. Conformational Transitions in Adenylate Kinase. Allosteric Communication Reduces Misligation. *J. Biol. Chem.* **2008**, *283*, 2042−2048.

(85) Aviram, H. Y.; Pirchi, M.; Mazal, H.; Barak, Y.; Riven, I.; Haran, G. Direct Observation of Ultrafast Large-Scale Dynamics of an Enzyme under Turnover Conditions. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, 3243−3248.

(86) Arora, K.; Brooks, C. L., 3rd. Large-Scale Allosteric Conformational Transitions of Adenylate Kinase Appear to Involve a Population-Shift Mechanism. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 18496−18501.

(87) Pirchi, M.; Ziv, G.; Riven, I.; Cohen, S. S.; Zohar, N.; Barak, Y.; Haran, G. Single-Molecule Fluorescence Spectroscopy Maps the Folding Landscape of a Large Protein. *Nat. Commun.* **2011**, *2*, 493.

(88) Taylor, J. N.; Pirchi, M.; Haran, G.; Komatsuzaki, T. Deciphering Hierarchical Features in the Energy Landscape of Adenylate Kinase Folding/Unfolding. *J. Chem. Phys.* **2018**, *148*, 123325.

(89) Kantaev, R.; Riven, I.; Goldenzweig, A.; Barak, Y.; Dym, O.; Peleg, Y.; Albeck, S.; Fleishman, S. J.; Haran, G. Manipulating the Folding Landscape of a Multidomain Protein. *J. Phys. Chem. B* **2018**, *122*, 11030−11038.

(90) Ruan, Q.; Ruan, K.; Balny, C.; Glaser, M.; Mantulin, W. W. Protein Folding Pathways of Adenylate Kinase from E. Coli: Hydrostatic Pressure and Stopped-Flow Studies. *Biochemistry* **2001**, *40*, 14706−14714.

(91) Kumar, S.; Sham, Y. Y.; Tsai, C. J.; Nussinov, R. Protein Folding and Function: The N-Terminal Fragment in Adenylate Kinase. *Biophys. J.* **2001**, *80*, 2439−2454.

(92) Giri Rao, V. V.; Gosavi, S. In the Multi-Domain Protein Adenylate Kinase, Domain Insertion Facilitates Cooperative Folding While Accommodating Function at Domain Interfaces. *PLoS Comput. Biol.* **2014**, *10*, No. e1003938.

(93) Li, W.; Terakawa, T.; Wang, W.; Takada, S. Energy Landscape and Multiroute Folding of Topologically Complex Proteins Adenylate Kinase and 2ouf-Knot. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 17789−17794.

(94) Daggett, V.; Fersht, A. The Present View of the Mechanism of Protein Folding. *Nat. Rev. Mol. Cell Biol.* **2003**, *4*, 497−502.

(95) Ratner, V.; Kahana, E.; Haas, E. The Natively Helical Chain Segment 169−188 of Escherichia Coli Adenylate Kinase Is Formed in the Latest Phase of the Refolding Transition. *J. Mol. Biol.* **2002**, *320*, 1135−1145.

(96) Bergasa-Caceres, F.; Rabitz, H. Role of Topology in the Cooperative Collapse of the Protein Core in the Sequential Collapse Model. Folding Pathway of R-Lactalbumin and Hen Lysoz. *J. Phys. Chem. B* **2001**, *105*, 2874−2880.

(97) Bergasa-Caceres, F.; Rabitz, H. Differences between the Sequential Collapse Folding Pathways of Apoleghemoglobin and Apomyoglobin. *J. Phys. Chem. B* **2002**, *106*, 4818−4822.

(98) Bergasa-Caceres, F.; Rabitz, H. Sequential Collapse Folding Pathway of Staphylococcal Nuclease: Entropic Activation Barriers to Hydrophobic Collapse of the Protein Core. *J. Phys. Chem. B* **2004**, *108*, 8023−8030.

(99) Bergasa-Caceres, F.; Rabitz, H. A. Sequential Collapse Folding Pathway of B-Lactoglobulin: Parallel Pathways and Non-Native Secondary Structure. *J. Phys. Chem. B* **2003**, *107*, 3606−3612.

(100) Bergasa-Caceres, F.; Rabitz, H. Relating Contact Order to the Rate of Cooperative Collapse in the Sequential Collapse Model for Protein Folding Pathways. *Chem. Phys. Lett.* **2003**, *376*, 612−617.

(101) Bergasa-Caceres, F.; Rabitz, H. Two-State Folding Kinetics of Small Proteins in the Sequential Collapse Model: Dependence of the Folding Rate on Contact Order and Temperature. *J. Phys. Chem. B* **2003**, *107*, 12874−12877.

(102) Bergasa-Caceres, F.; Rabitz, H. A. A Simple Quantitative Model of Macromolecular Crowding Effects on Protein Folding: Application to the Murine Prion Protein (121−231). *Chem. Phys. Lett.* **2013**, *574*, 112−115.

(103) Bergasa-Caceres, F.; Rabitz, H. Flexibility Damps Macromolecular Crowding Effects on Protein Folding Dynamics: Application to the Murine Prion Protein (121−231). *Chem. Phys. Lett.* **2014**, *591*, 207−211.

(104) Bergasa-Caceres, F.; Rabitz, H. A. Macromolecular Crowding Facilitates the Conformational Transition of Molten Globule States of the Prion Protein. *J. Phys. Chem. B* **2016**, *120*, 11093−11701.

(105) Bergasa-Caceres, F.; Rabitz, H. Predicting the Location of the Non-Local Contacts in Alpha-Synuclein Biochimica Et Biophysica Acta Proteins and Proteomics. *Biochim. Biophys. Acta, Proteins Proteomics* **2018**, *1866*, 1201−1208.

(106) Towse, C. L.; Akke, M.; Daggett, V. The Dynameomics Entropy Dictionary: A Large-Scale Assessment of Conformational Entropy across Protein Fold Space. *J. Phys. Chem. B* **2017**, *121*, 3933−3945.

(107) Riddle, D. S.; Santiago, J. V.; Bray-Hall, S. T.; Doshi, N.; Grantcharova, V. P.; Yi, Q.; Baker, D. Functional Rapidly Folding Proteins from Simplified Amino Acid Sequences. *Nat. Struct. Biol.* **1997**, *4*, 805−809.

(108) Jacobson, H.; Stockmayer, W. H. Intramolecular Reactions in Polycondensations. I. The Theory of Linear Systems. *J. Chem. Phys.* **1950**, *18*, 1600−1607.

(109) Lapidus, L. J.; Eaton, W. A.; Hofrichter, J. Measuring the Rate of Intramolecular Contact Formation in Polypeptides. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 7220−7225.

(110) Grupi, A.; Haas, E. Segmental Conformational Disorder and Dynamics in the Intrinsically Disordered Protein Alpha-Synuclein and Its Chain Length Dependence. *J. Mol. Biol.* **2011**, *405*, 1267−1283.

(111) Karanicolas, J.; Brooks, C. L., 3rd. Improved Go-Like Models Demonstrate the Robustness of Protein Folding Mechanisms Towards Non-Native Interactions. *J. Mol. Biol.* **2003**, *334*, 309−325.

(112) Taketomi, H.; Ueda, Y.; Go, N. Studies on Protein Folding, Unfolding and Fluctuations by Computer Simulation. I. The Effect of Specific Amino Acid Sequence Represented by Specific Inter-Unit Interactions. *Int. J. Pept. Protein Res.* **1975**, *7*, 445−459.

(113) Yang, W. Y.; Gruebele, M. Folding at the Speed Limit. *Nature* **2003**, *423*, 193−197.

(114) Sheinerman, F. B.; Brooks, C. L., 3rd. Molecular Picture of Folding of a Small Alpha/Beta Protein. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 1562−1567.

(115) Socci, N. D.; Onuchic, J. N.; Wolynes, P. G. Protein Folding Mechanisms and the Multidimensional Folding Funnel. *Proteins: Struct., Funct., Genet.* **1998**, *32*, 136−158.
(116) Debe, D. A.; Goddard, W. A., 3rd. First Principles Prediction of Protein Folding Rates. *J. Mol. Biol.* **1999**, *294*, 619−625.
(117) Makarov, D. E.; Plaxco, K. W. The Topomer Search Model: A Simple, Quantitative Theory of Two-State Protein Folding Kinetics. *Protein Sci.* **2003**, *12*, 17−26.
(118) Ackerman, M. S.; Shortle, D. Robustness of the Long-Range Structure in Denatured Staphylococcal Nuclease to Changes in Amino Acid Sequence. *Biochemistry* **2002**, *41*, 13791−13797.
(119) Evans, P. A.; Topping, K. D.; Woolfson, D. N.; Dobson, C. M. Hydrophobic Clustering in Nonnative States of a Protein: Interpretation of Chemical Shifts in Nmr Spectra of Denatured States of Lysozyme. *Proteins: Struct., Funct., Genet.* **1991**, *9*, 248−266.
(120) Neri, D.; Billeter, M.; Wider, G.; Wuthrich, K. Nmr Determination of Residual Structure in a Urea-Denatured Protein, the 434-Repressor. *Science* **1992**, *257*, 1559−1563.
(121) Guzman-Casado, M.; Parody-Morreale, A.; Robic, S.; Marqusee, S.; Sanchez-Ruiz, J. M. Energetic Evidence for Formation of a Ph-Dependent Hydrophobic Cluster in the Denatured State of Thermus Thermophilus Ribonuclease H. *J. Mol. Biol.* **2003**, *329*, 731−743.
(122) Jennings, P. A.; Wright, P. E. Formation of a Molten Globule Intermediate Early in the Kinetic Folding Pathway of Apomyoglobin. *Science* **1993**, *262*, 892−896.
(123) Jacobs, M. D.; Fox, R. O. Staphylococcal Nuclease Folding Intermediate Characterized by Hydrogen Exchange and Nmr Spectroscopy. *Proc. Natl. Acad. Sci. U. S. A.* **1994**, *91*, 449−453.
(124) Radford, S. E.; Dobson, C. M.; Evans, P. A. The Folding of Hen Lysozyme Involves Partially Structured Intermediates and Multiple Pathways. *Nature* **1992**, *358*, 302−307.
(125) Nishimura, C.; Prytulla, S.; Dyson, H. J.; Wright, P. E. Conservation of Folding Pathways in Evolutionarily Distant Globin Sequences. *Nat. Struct. Biol.* **2000**, *7*, 679−686.
(126) Loh, S. N.; Kay, M. S.; Baldwin, R. L. Structure and Stability of a Second Molten Globule Intermediate in the Apomyoglobin Folding Pathway. *Proc. Natl. Acad. Sci. U. S. A.* **1995**, *92*, 5446−5450.
(127) Roder, H.; Elove, G. A.; Englander, S. W. Structural Characterization of Folding Intermediates in Cytochrome C by H-Exchange Labelling and Proton Nmr. *Nature* **1988**, *335*, 700−704.
(128) Scalley-Kim, M.; Minard, P.; Baker, D. Low Free Energy Cost of Very Long Loop Insertions in Proteins. *Protein Sci.* **2003**, *12*, 197−206.
(129) Bergasa-Caceres, F.; Rabitz, H. A. Low Entropic Barrier to the Hydrophobic Collapse of the Prion Protein: Effects of Intermediate States and Conformational Flexibility. *J. Phys. Chem. A* **2010**, *114*, 6978−6982.
(130) Bergasa-Caceres, F.; Rabitz, H. A Simple Quantitative Model of Macromolecular Crowding Effects on Protein Folding: Application to the Murine Prion Protein(121−231). *Chem. Phys. Lett.* **2013**, *574*, 112−115.

■ NOTE ADDED AFTER ASAP PUBLICATION

This paper published ASAP on April 3, 2019 with errors in the caption of Figure 2. The corrected version reposted to the Web on April 4, 2019.

Nature's Shortcut to Protein Folding

Fernando Bergasa-Caceres*[1,4]

Elisha Haas[2,5]

Herschel A. Rabitz[3,6]

1. Universidad Autonoma de Madrid, Cantoblanco 28049, Spain
2. The Goodman Faculty of life sciences, Bar-Ilan University, Ramat Gan 52900, Israel
3. Princeton University, NJ 08544, USA
4. Bergasa@Princeton.edu
5. Elisha.Haas@biu.ac.il
6. HRabitz@Princeton.edu

*Corresponding author

SCM existing predictions for primary contacts

In this supplementary section, the results obtained in previous work for the primary contacts of ten proteins within the SCM (references 48 and 96-99 in the main article) are summarized in condensed form. The best and next to best predicted contacts are listed in tables S1 and S2, and the best predicted contacts are represented on the crystal structures in figures S1 to S9 (the best predicted contact for adenylate kinase was presented in the main body of the article). The figures have been elaborated employing Protein Workshop[2].

1. Calculational method to identify initial non-local contacts from primary sequence information

The methodology employed in all cases to predict the location of the primary contacts of the 10 proteins was explained in full in ref. 48 of the main text. The methodology relies only on the primary sequence properties to determine the location of the primary contact. Because the amino acid side chains are significantly larger than the typical peptide bond length, it is expected that early contacts, nucleated by a loop defined by any two amino acids, will immediately involve segments including several amino acids. Thus, the typical early contact segment size was taken to be ~ 5 amino acids. Since the hydrophobic stabilization energy of the contact is determined by the hydrophobicity of the segments involved, hydrophobicity values h_k were obtained from the Fauchere-Pliska scale[1] and were assigned to each residue. The N- and C-terminal residues carry a charge and their hydrophobicity should be much less than that assigned by the scale to amino acids within the chain. Thus, a value of zero was assigned to the hydrophobicity of the end residues. Then the hydrophobicity h_k of each residue was added over a segment contact window of five amino acids centered at residue *i*, resulting in a segment hydrophobicity $h_{i,5}$ (a value of ~ 0.5 is equivalent to a change in energy of *k*T (1)). In order to determine the highest propensity contact (i.e., the primary contact), the $h_{i,5}$ value of a segment centered at residue *i* was added to the $h_{j,5}$ value of a segment centered at residue *j*, located 65-85 amino acids apart along the sequence, to give a contact propensity P_{ij} ~ ($h_{i,5}$ + $h_{j,5}$), a difference in propensity of ~ 0.45 reflect a difference in energy of ~*k*T.

2. Results

Table S1: Best predicted primary contact.

Protein	PDB structure	Primary contact	Propensity
Cythochrome *c*[1]	1HRC	9-13 on 94-98	10.1
Myoglobin[1]	1MBN	28-32 on 111-115	12.6
Ribonuclease A	1KF5	43-47 on 116-120	9
Barnase[2]	1BNR	13-17 on 93-97	10.7
α-Lactalbumin	1A4V	27-31 on 101-105	12.7
Hen lysozyme	1DPX	28-32 on 107-111	11.4
Leghemoglobin[1,3]	1LH1	43-47 on 109-113	11.4
β-Lactoglobulin	1BEB	19-23 on 103-107	13.1
Staphyloccocal nuclease	1STN	34-38 on 111-115	10.6
Adenylate kinase	4AKE	3-7 on 79-83	10.6

Table S2. Second best predicted primary contact.

PDB structure	Second best contact	Propensity
1HRC	9-13 on 81-85	9.3
1MNB	7-11 on 72-76	12.2
1KF5	26-30 on 106-110	8.7
1BNR	3-7 on 88-92	9.0
1A4V	51-55 on 116-120	9.7
1DPX	54-58 on 120-124	10.4
1LH1	65-69 on 136-140	11.2
1BEB	29-33 on 103-107	12.4
1STN	11-15 on 89-93	10.5
4AKE	105-109 on 178-182	10.5

1. For additional visual clarity the protein has been represented without the heme group present in the crystal structure

2. The result for barnase is different than that presented in ref.48, which was a printing error (a duplication of the result for myoglobin)

3. The experiments to which the SCM predictions were compared in ref.97 were carried out on apoleghemoglobin. The heme group in leghemoglobin is "wedged" between the two segments defining the best primary contact (see figure S7).

References

1. Fauchere, J. L.; Pliska, V. Hydrophobic parameters II of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides *Eur. J. Med. Chem.* 1983, *18*, 369-75

2. Moreland, J.L.; Granada, A.; Buzko, O. V.; Zhang, Q.; P.E. Bourne, P. E. The molecular biology toolkit (MBT): A modular platform for developing molecular visualization applications *BMC Bioinformatics*, 2005, 6, 21

Figures

Figure S1: Primary contact for cythochrome *c*

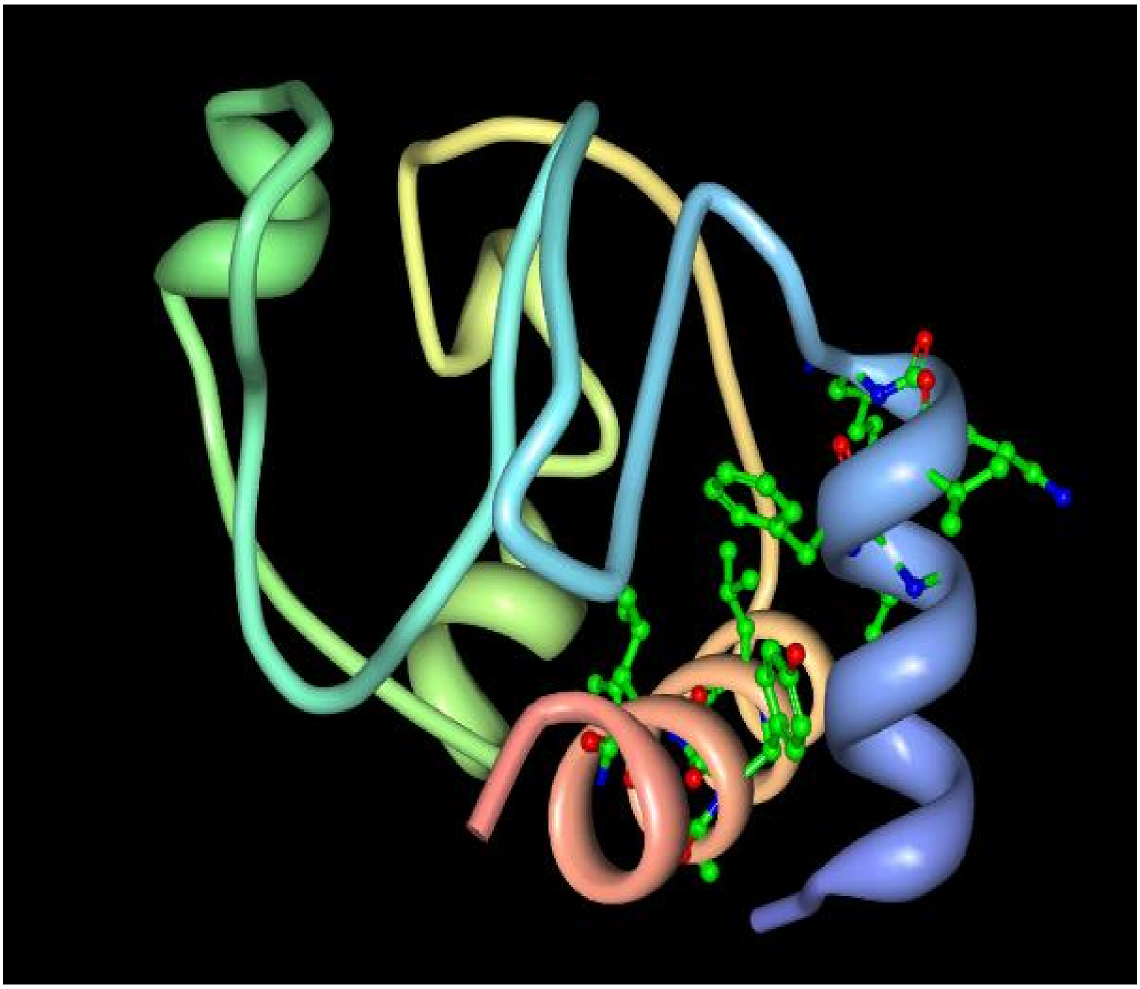


Figure S2: Primary contact for myoglobin

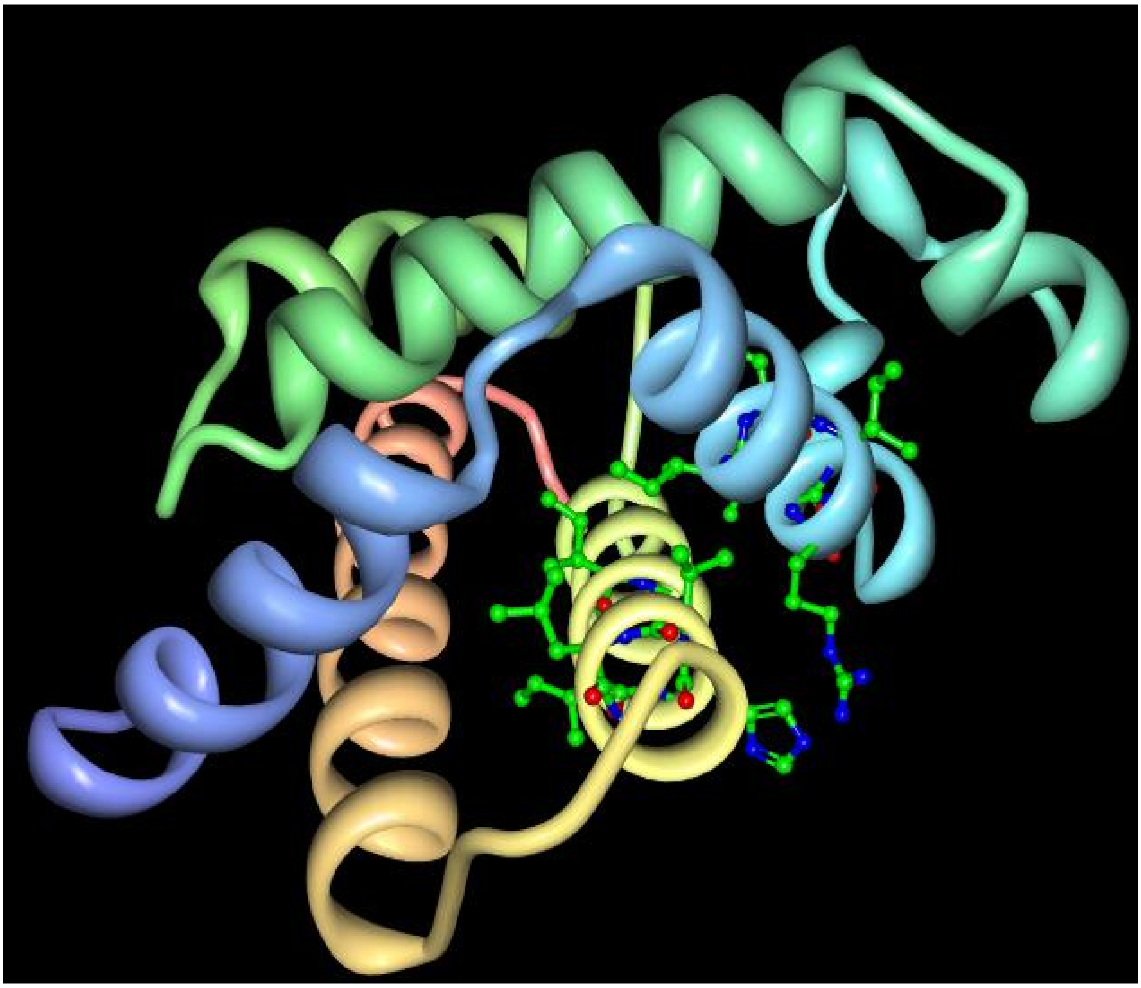


Figure S3: Primary contact for ribonuclease A

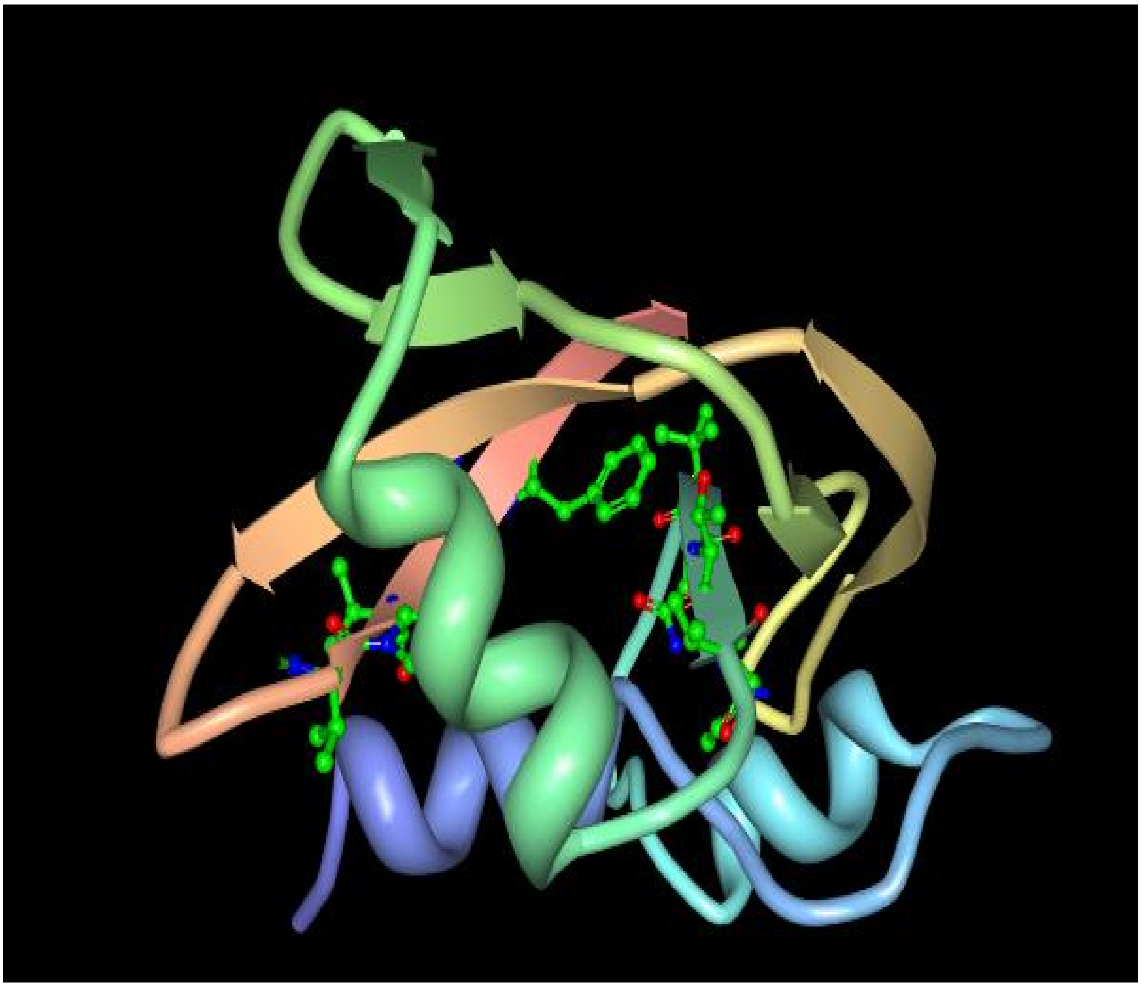


Figure S4: Primary contact for barnase

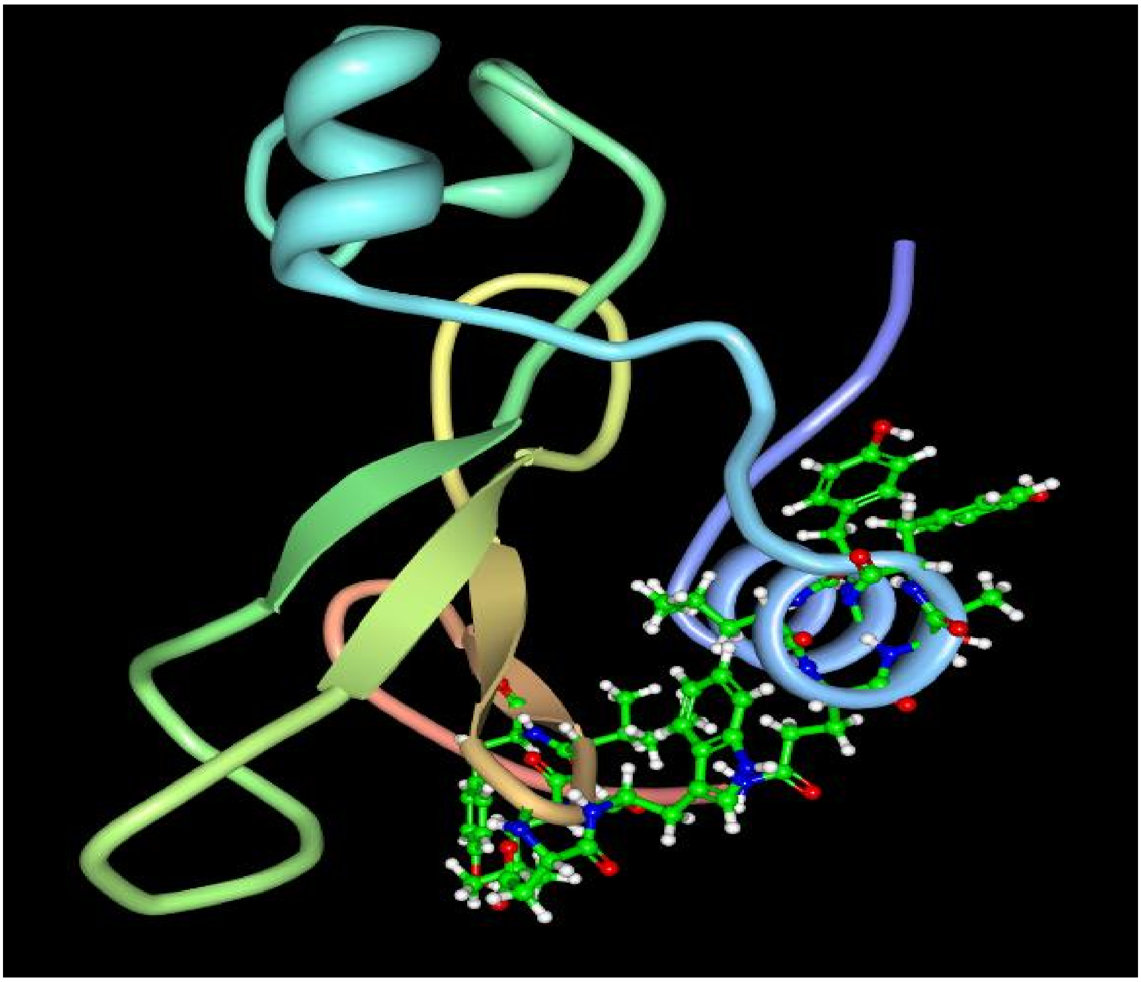


Figure S5. Primary contact for α-lactalbumin

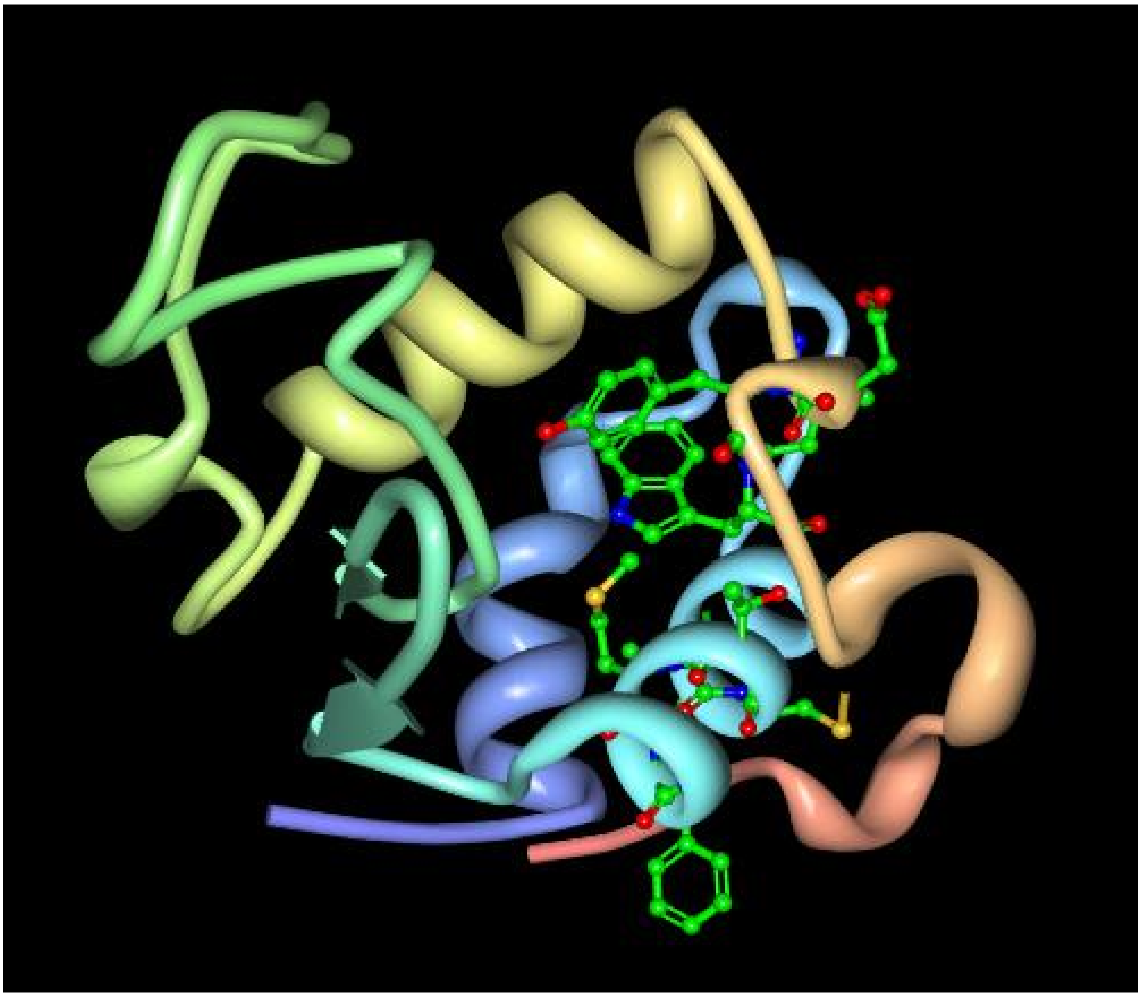


Figure S6. Primary contact for hen lysozyme

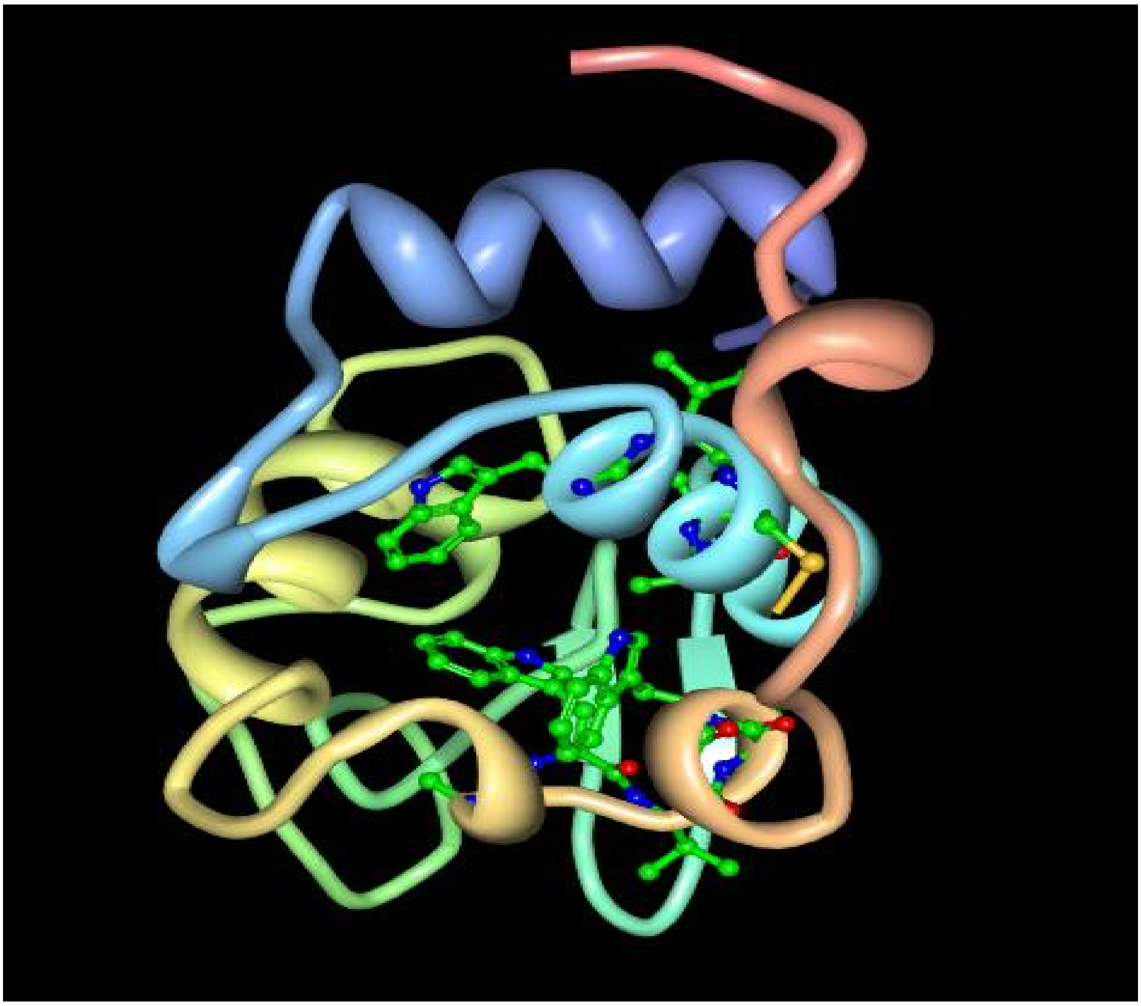
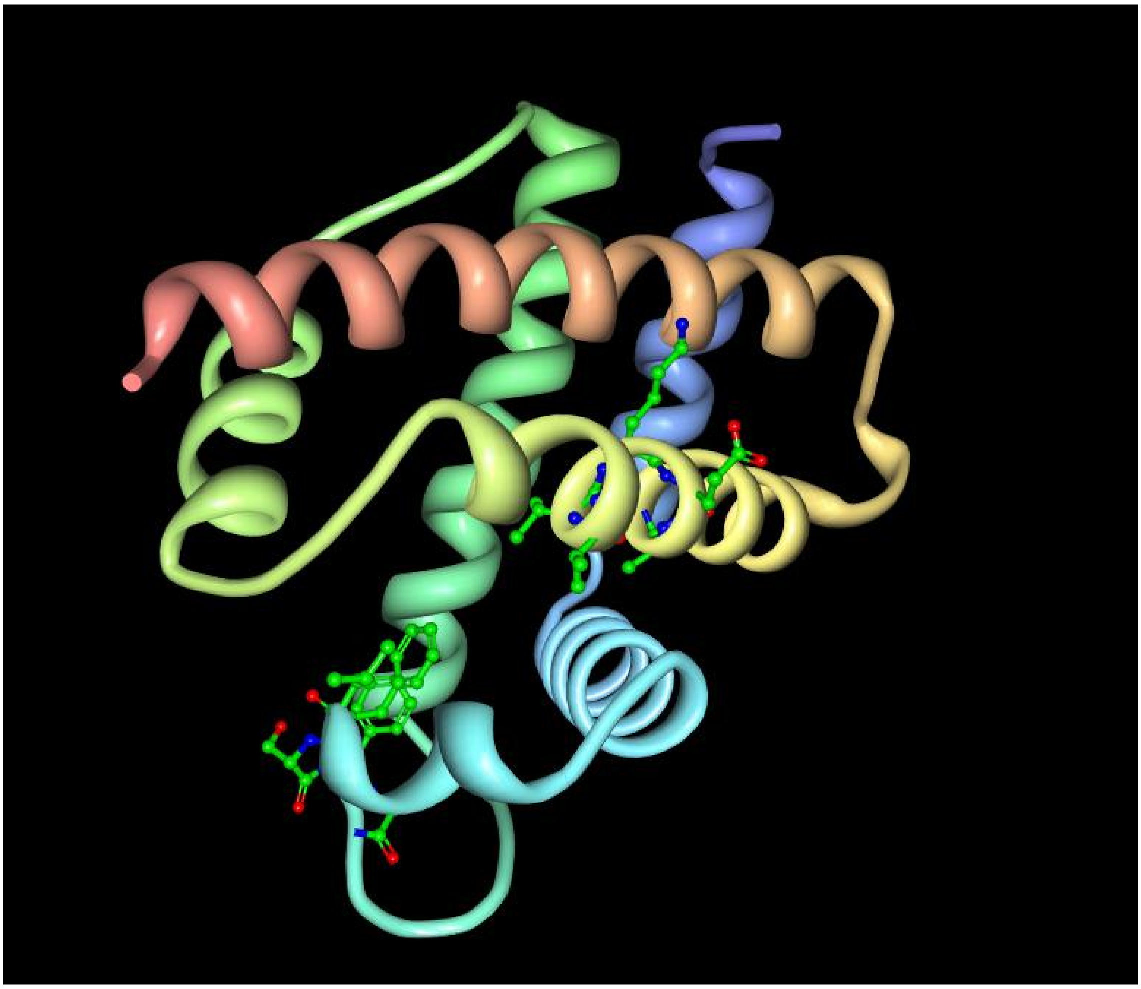


Figure S7. Primary contact for leghemoglobin: (a) not showing the heme group; (b) showing the heme group

(a)



(b)

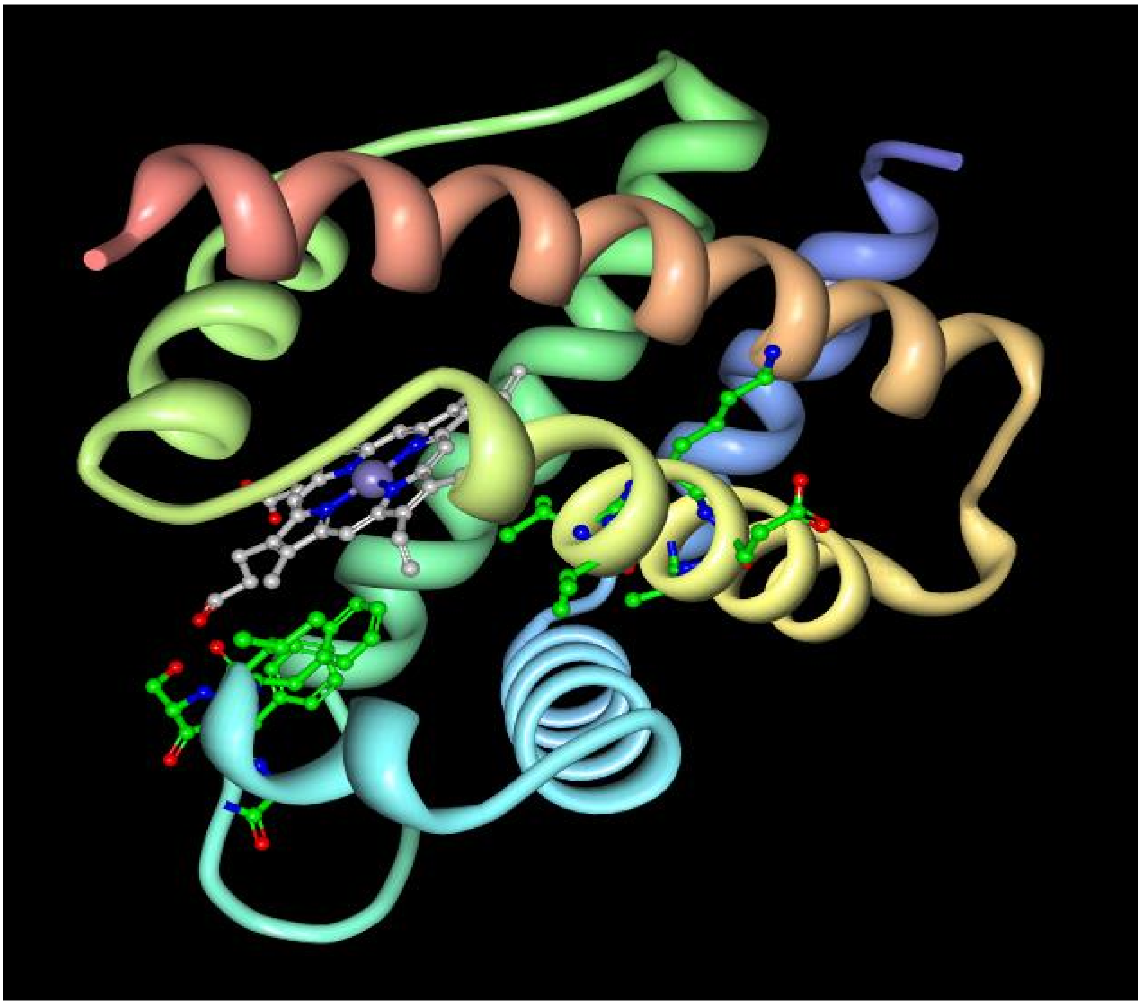


Figure S8. Primary contact for β-Lactoglobulin

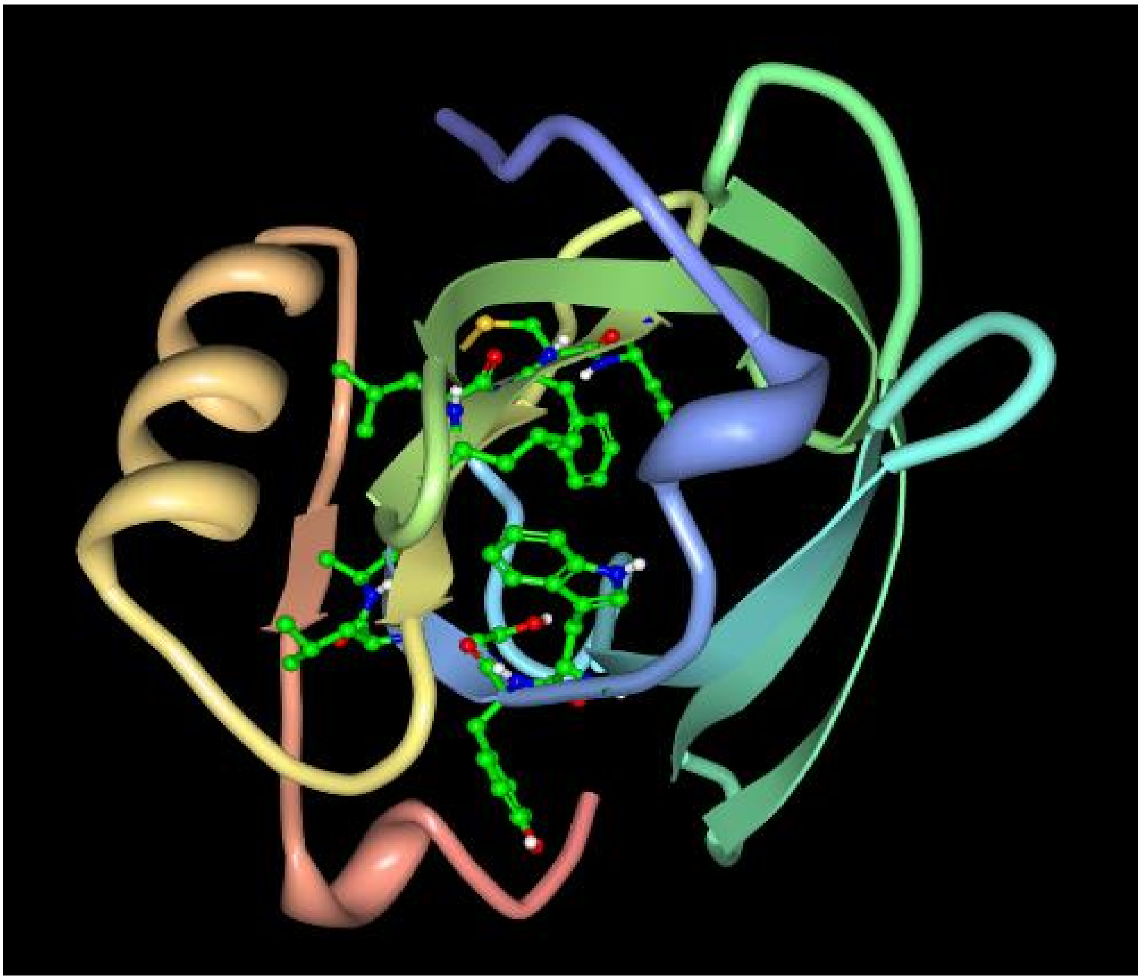


Figure S9. Primary contact for staphyloccocal nuclease