Building blocks of protein structures: Physics meets biology

Tatjana Škrbić,[1,2] Amos Maritan,[3] Achille Giacometti,[2,4] George D. Rose,[5] and Jayanth R. Banavar[1,*]

[1]*Department of Physics and Institute for Fundamental Science, University of Oregon,*
Eugene, Oregon 97403, USA
[2]*Dipartimento di Scienze Molecolari e Nanosistemi, Università Ca' Foscari Venezia,*
Campus Scientifico, Edificio Alfa, via Torino 155, 30170 Venezia Mestre, Italy
[3]*Dipartimento di Fisica e Astronomia, Università di Padova and INFN, via Marzolo 8, 35131 Padova, Italy*
[4]*European Center for Living Technologies (ECLT), Ca' Bottacin, Dorsoduro 3911, Calle Crosera, 30123 Venezia, Italy*
[5]*T. C. Jenkins Department of Biophysics, Johns Hopkins University, 3400 North Charles Street,*
Baltimore, Maryland 21218-2683, USA

(Received 21 January 2021; accepted 22 March 2021; published 8 July 2021; corrected 16 September 2021)

The native state structures of globular proteins are stable and well packed indicating that self-interactions are favored over protein-solvent interactions under folding conditions. We use this as a guiding principle to derive the geometry of the building blocks of protein structures—α helices and strands assembled into β sheets—with no adjustable parameters, no amino acid sequence information, and no chemistry. There is an almost perfect fit between the dictates of mathematics and physics and the rules of quantum chemistry. Protein evolution is facilitated by sequence-independent platforms, which can elaborate sequence-dependent functional diversity. Our work highlights the vital role of discreteness in life and may have implications for the creation of artificial life and on the nature of life elsewhere in the cosmos.

DOI: 10.1103/PhysRevE.104.014402

Proteins [1–41], the molecular machines of life, are formidably complex [42]. They have myriad degrees of freedom, an astronomical number of possible sequences for even a moderate length chain, and are stabilized by thousands of interactions, both intramolecular and with solvent. Yet, many proteins (here, we do not consider disordered proteins or structural proteins) adopt their native conformation spontaneously under physiological conditions [5]. The native state structures of globular proteins are space filling and maximize self-interaction [6,7,9]. Here we use this as a constructive hypothesis to predict the building blocks of protein native state structures. The folded structures [21,27,33,36] are modular and built on scaffolds of α helices [2] and strands of β sheet [3], the only two conformers that can be extended indefinitely without steric interference while providing hydrogen-bonding partners for their own backbone polar groups [4,10,29]. Proteins are digital molecules: Nature's exclusion of α-β hybrid segments [28]—part α helix, part β strand—is built into proteins at the covalent level and restricts the topology of single domain proteins to a few thousand distinct folds at most [8,14,20,23].

Helices are ubiquitous in biomolecular structures. They are also found in everyday life, e.g., a garden hose (or a flexible tube) is often wound into a helix. Figure 1(a) is a sketch of a segment of a protein helix shown with a tube envelope. A uniform, flexible, self-avoiding solid tube, whose axis is a line, is a geometrical generalization of a sphere. A sphere is a region carving out space around a point, its center. Analogously, all points within the tube are at a distance from the tube axis smaller than or equal to the tube thickness, which is measured by the tube radius, Δ. A flexible tube is an extended object with uniaxial symmetry and is not plagued by symmetry conflicts, unlike the simple model of a chain of tethered spheres for which the uniaxial symmetry inherent to a chain clashes with the spherical symmetry of the constituent objects.

Here we model a protein as a discretized tube with a set of equally spaced points, analogous to the C_α atoms along the protein backbone, defining its axis. The coordinates of these points are described using two angles: θ and μ (see Fig. 2). The simplest repeating geometry of the axis of a tube of radius Δ is a helix of pitch P, wrapped around a straight cylinder of radius R, taken to be the helix radius. The helix is parametrized by a variable t and is defined by

$$\mathbf{r}(t) = [R\cos(t),\ R\sin(t),\ Pt/(2\pi)]. \tag{1}$$

As t advances by an integer multiple of 2π, the helix repeats periodically along the z axis, with an increment equal to the pitch. The helical tube geometry is characterized by three dimensionless quantities: Δ/R, $\eta = P/(2R\pi)$, and ε_0, the rotation angle between successive points along the axis. Our initial goal parallels the seminal work of Pauling *et al.* [2], who sought rotation angles that allowed for the optimal placement of hydrogen bonds in a helix. The crucial difference here is that we do not need to invoke quantum chemistry, covalent bonds, the planarity of peptide bonds, or hydrogen bonds.

We begin with a brief account of earlier work on maximizing the self-interaction of a *continuum* tube [43–48] by

[*] banavar@uoregon.edu

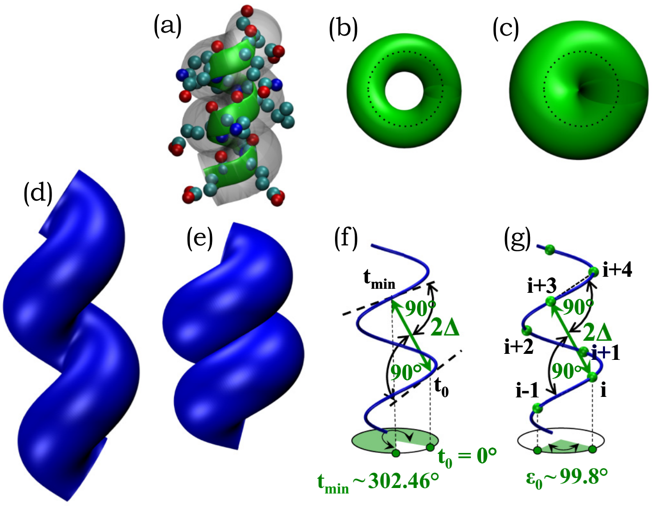

FIG. 1. Optimal geometry of space-filling helix. (a) A segment of ten residues of a helix from phage T4 lysozyme protein 1L56 (residues 61–70). The ribbon represents the helical trace formed by the C_α atoms, the spheres denote the heavy backbone and side chain atoms in the helix, and the transparent tube is a guide to the eye. (b) and (c) Top views of two continuum helices, both with a helix pitch P to helix radius R ratio $\eta = (P/2\pi R) \sim 0.4$ and a local radius of curvature of the helix, $R_{\rm local} = R(1+\eta^2) \sim 1.16R$. The tube radii Δ in the two cases are different: $\Delta/R_{\rm local} = 1/2$ and 1, respectively. (b) When Δ is less than $R_{\rm local}$, there is empty space in the interior. When Δ is bigger than $R_{\rm local}$, the turn is too tight leading to a kink, as is sometimes observed in a garden hose (not shown). (c) The sweet spot occurs when $\Delta = R_{\rm local}$, leading to maximization of the *local* self-interaction. (d) and (e) Side views of two helices with η values of 0.8 and ~0.4, respectively. In both cases, Δ has been chosen to be the local radius of curvature of the latter helix ~$1.16R$. (d) When η is larger than ~ 0.4, there is empty space between successive turns and the *nonlocal* self-interaction is not maximized. In the other limit of small η (not shown), successive turns of the tube overlap and this is forbidden sterically. (e) A Goldilocks situation here is when η is tuned just right to ~ 0.4 yielding $(\Delta/R) \sim 1.16$ for a continuum space-filling helix maximizing both local and nonlocal self-interaction. The top and side views of the optimal continuum helix are shown in (c) and (e) respectively. Panels (f) and (g) show how these results can be captured analytically (see text) for a continuum and a discrete tube, respectively.

winding the tube as tightly as possible, subject to the excluded volume constraint that the tube cannot penetrate itself. Such space filling ensures the expulsion of the solvent (water) from the core of the folded protein and is driven by hydrophobicity. We ensure local space filling of the helix by equating the tube radius to the local radius of curvature [Fig. 1(c)], which, in turn, is equal to $R(1+\eta^2)$ [47], yielding

$$\Delta = R(1+\eta^2). \tag{2}$$

The successive turns of a space-filling helix need to be parallel and alongside each other [Fig. 1(e)]. The square of the distance between a reference point in the continuum helix (denoted by $t_0 = 0°$) and an arbitrary point t is given by

$$d^2 = R^2[2(1-\cos t) + \eta^2 t^2]. \tag{3}$$

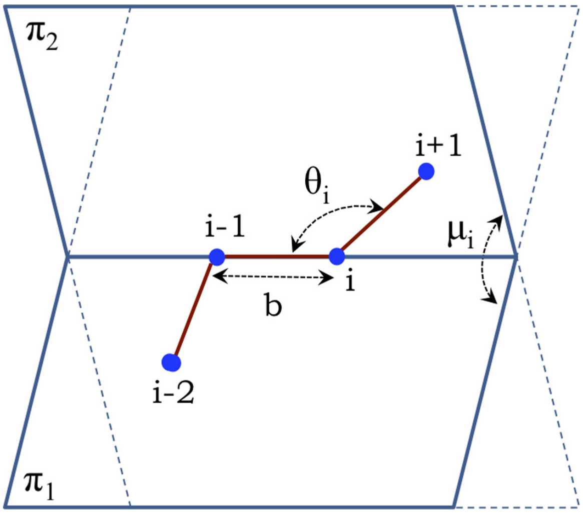

FIG. 2. Coordinate system at a discrete location i along the tube axis. The bond length b, assumed here to be a constant, is the distance between successive points. The angle θ_i is the angle subtended at i by points $(i-1)$ and $(i+1)$ along the tube axis. θ_i was taken to be the magnitude of the angle between neighboring bonds $(i-1, i)$ and $(i, i+1)$ along the chain. μ_i is the dihedral angle between the planes π_1 and π_2 formed by $[(i-2), (i-1), i]$ and $[(i-1), i, (i+1)]$, respectively, or equivalently the angle between the binormals in a Frenet reference frame at points $(i-1)$ and i. Knowledge of the coordinates of the previous three points $(i-2, i-1, i)$ and the variables (θ_i, μ_i) are sufficient to uniquely specify the coordinates of the point $(i+1)$.

We determine the parameter value $t_{\rm min}$ for which d^2 is a minimum and set this minimum distance equal to the square of the tube diameter, $4\Delta^2$, thereby ensuring nonlocal space filling [Fig. 1(f)]. The minimization condition is

$$\sin t_{\rm min} + \eta^2 t_{\rm min} = 0, \tag{4}$$

and the distance constraint is

$$4\Delta^2 = R^2\big[2(1-\cos t_{\rm min}) + \eta^2 t_{\rm min}^2\big]. \tag{5}$$

We solve Eqs. (2), (4), and (5) simultaneously to obtain the unique geometry of the continuum space-filling helix [Figs. 1(c), 1(e), and 1(f)]: $\eta \sim 0.4$, $\Delta/R \sim 1.16$, and $t_{\rm min} \sim 302°$.

The idealized continuum tube does not take into account discreteness, a common ingredient to all matter, which is crucial at small length scales. A unique benefit of discreteness is the emergence of a second building block (besides the space-filling helix): a two-dimensional strand with a zigzag tube axis [Fig. 3(a)], the rotation angle ε_0 of 180°, and $\mu = 180°$. The existence of two building blocks is *required* for the rich diversity of topologically distinct folds, necessary for the versatile functioning of the molecular machines. A helix is defined by a repeat of (θ, μ) values and a planar strand by a repeat of $\mu = 180°$. For repeat μ values close to 180°, one obtains a twisted planar strand, a geometrical feature often observed in protein structures.

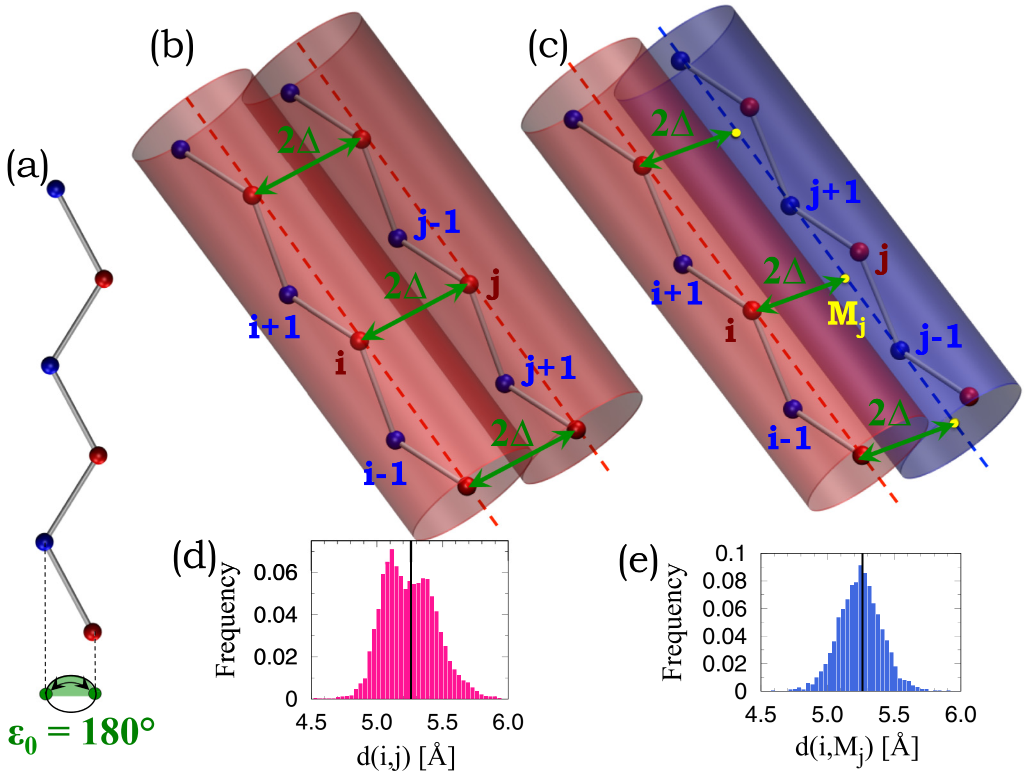

FIG. 3. Optimal packing of strands. (a) A single two-dimensional zigzag strand (with a rotation angle of 180°) lying in the plane of the paper. This planarity can only occur for a discrete tube and is forbidden for a tube in the continuum. Alternate points along a strand are colored red and blue (right and left points). There are two equivalent choices for a straight tube axis, one lying along the line of blue (left) points (blue or the left axis) or the line of red points (red or the right axis). Two distinct space-filling arrangements for strand packing are shown corresponding to (b) red axis–red axis (right axis–right axis) tubes [or equivalently blue axis–blue axis (left axis–left axis) tubes (not shown)] and (c) red axis–blue axis (right axis–left axis) (not shown). The two cases correspond to antiparallel and parallel β sheets with distinct distance constraints. The yellow point M_j lies midway between the blue (left) points $j-1$ and $j+1$. The maximization of self-interaction dictates that the distances (i, j) in (b) and (i, M_j) in (c) ought to be $2\Delta \sim 5.26$ Å to ensure space filling. (d) and (e) show the histograms of the distances (i, j) and (i, M_j) in the interior of antiparallel and parallel β sheets in protein structures. The black vertical lines show the theoretical prediction of $2\Delta \sim 5.26$ Å. The mean values of both histograms are the same as the theoretical prediction (see Table I).

Figure 1(g) shows the space-filling discrete helix with $\eta \sim 0.4$ and $\Delta/R \sim 1.16$, the geometrical characteristics of the continuum space-filling helix. The discretization requires the specification of the rotation angle ε_0 between successive points that retains the space-filling conditions for the discrete case. This choice of ε_0 is made (in direct analogy with the continuum case) by requiring that the distance between points i (analogous to $t_0 = 0°$) and $i+m$ with integer m (analogous to t_{min}) is equal to the tube diameter and the angles $(i-1,\ i,\ i+m)$ and $(i,\ i+m,\ i+m+1)$ are both equal to 90° (analogous to the minimization condition). The smallest value of m for which these conditions are satisfied is $m = 3$ and $\varepsilon_0 \sim 99.8°$ (the ratio of the distance to the tube diameter is found to be 1.00… and both the angles are 90.0…° for this value of ε_0). Upon defining the length scale to match the mean C_α-C_α distance along the protein backbone of 3.81 Å, the tube radius is found to be $\Delta \sim 2.63$ Å. Using these basic results, one may derive many attributes of the space-filling discrete helix, which are in excellent accord with the α helix building block of protein structures (see Figs. 4 and 5 and Table I).

A space-filling helix maximizes self-interaction through local interactions, whereas the nonlocal interactions of strands assembled into sheets lead to space filling. We build on the insights gained from the helix analysis to make predictions of the geometrical arrangements for strand pairing [Figs. 3(b) and 3(c)]. First, the strands need to be in phase with each other mimicking the behavior of adjoining turns in the continuum helix, placed parallel to and alongside each other. Second, there are two distinct ways [Figs. 3(b) and 3(c)] of accomplishing space filling of assembled strands corresponding to antiparallel and parallel β sheet hydrogen-bonding patterns, first predicted by Pauling and Corey [3] based on hydrogen bonding. The space-filling packing requires that the distances (i, j) in Fig. 3(b) (antiparallel arrangement) and (i, M_j) in Fig. 3(c) (parallel arrangement), which are measures of the closest approach of two parallel tube segments, both ought to be $2\Delta \sim 5.26$ Å [see Figs. 3(d) and 3(e) and Table I).

It is important to note that, for both helices and sheets, the side chains do not clash sterically unlike in a well-packed compact arrangement of parallel strands in a hexagonal array. In addition to helices and strands, chain turns are needed to interconnect these building blocks. In proteins, the most abundant turns are β turns, tight, four-residue segments that approximately reverse the overall chain direction [13]. β turns

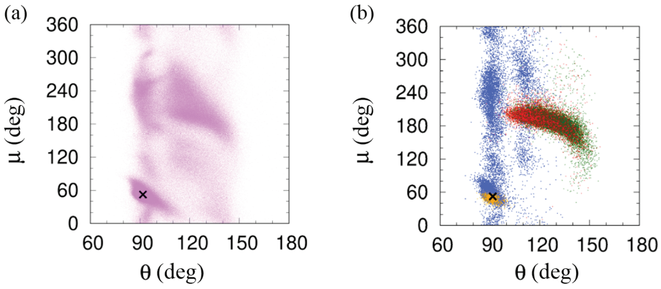

FIG. 4. Two views of the local structure representation of proteins. (a) (θ, μ) plot of the PDB data set (see Table S1 of the Supplemental Material) comprising 4416 proteins and 972 519 residues. Here, the local conformations of residues are shown in the (θ, μ) plane. For strands, a μ value that deviates from ∼180° is the signature of a twisted strand, which is still locally planar. The plot shows chiral symmetry breaking; i.e., the points are not symmetrically placed around $\mu = 180°$. Our simplified analysis does not attempt to account for this. (b) (θ, μ) coordinates of random samples of 12 000 points each from the interior of α helices (orange); antiparallel (green) and parallel (red) β sheets; and β turns [the two interior sites of ($i, i+3$) hydrogen-bonded residues with no helical residues] (blue). The tight turns have θ values similar to those of helices. Unlike for helices and turns, the θ values of strands are not constrained. The black X in both panels shows our prediction of the geometry of the space-filling helix.

are tightly wound like an α helix, and therefore are predicted to have similar θ angles as in the α helix (Fig. 4).

Figure 4(b) shows the (θ, μ) coordinates for four classes of residues: those that participate in α helices, parallel β sheets, antiparallel β sheets, and β turns. The black X marks the coordinates of the predicted space-filling helix. Unsurprisingly, α helix μ values $(49.7 \pm 3.9)°$ are a bit lower than the theoretical prediction of 52.4° because the distance between a hydrogen-bonded donor and acceptor (N-H$\cdots$O=C) can be less than their summed van der Waals radii. As predicted, the tight turns predominantly have a θ value close to that of the α helix. The β strands are twisted with a μ angle around 180° and have a spread of θ angles.

The accord between our prediction and structural data from the protein data bank underscores the consilience [49] between mathematics and physics on one hand and quantum chemistry on the other and shows how self-interaction is maximized through a space-filling arrangement of individual helices and sheets (Fig. 6). The large but finite number of protein native state folds [8,14,20,23] sculpted by geometry and symmetry [25,26] is reminiscent of the restriction of the number of space groups of Bravais lattices of three-dimensional crystals to exactly 230 due to periodicity and space-filling requirements [50].

Our theory shows convincingly that the structure space and sequence space of proteins are separable, yielding sequence-independent forms [22] that are Platonic and immutable, and not subject to Darwinian evolution. Sequences can then populate these forms resulting in the evolution of the functional diversity of life. The evolution [41,51,52] of biological macromolecules can be framed as a random walk in an inordinately vast sequence space, with selection guided by fitness. Our formalism imposes an important constraint on protein evolution. A consequence is that the repertoire of possible folds is generated from presculpted α helices and β strands, and, of necessity, accessible folds are mix and match constructs of these fundamental forms. This diversity of structural scaffolds provides a platform for elaborating functional diversity.

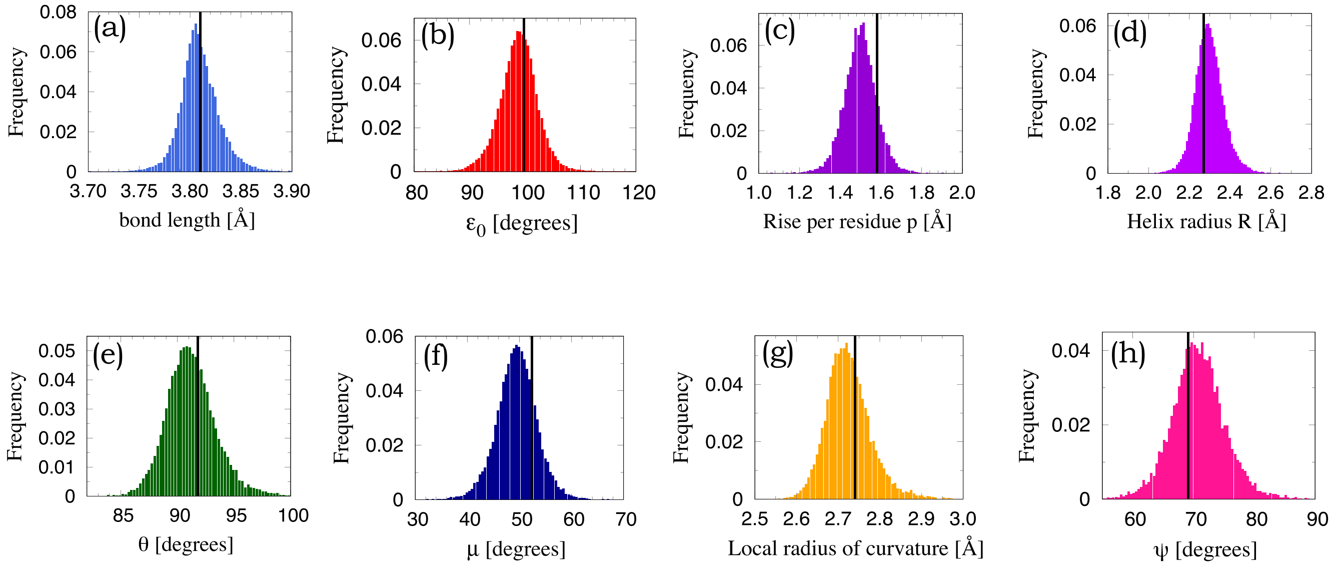

FIG. 5. Distribution of α helix characteristics. (a) Distribution of the experimentally determined bond lengths (consecutive C_α-C_α distances). The bond length in the theory was chosen to be the mean bond length of 3.81 Å and sets the characteristic length scale. The other panels show the distributions of (b) the rotation angle, (c) the rise per residue, (d) the helix radius, (e) θ, (f) μ, (g) the local radius of curvature, and (h) $\psi = 180° - \angle[\pi(i-1, i, i+3), \pi(i-1, i+2, i+3)]$, where the angle $\angle[\pi(i-1, i, i+3), \pi(i-1, i+2, i+3)]$ is the dihedral angle between the planes defined by the points $(i-1, i, i+3)$ and $(i-1, i+2, i+3)$ in Fig. 1(g). The triangles formed by the two triplets ought to be congruent but they are not coplanar. The black line in each of the panels (except the first) shows the zero parameter theoretical prediction. Overall, there is excellent accord between theory and observations from protein structures.

TABLE I. Quantitative comparison between theory and data from the Protein Data Bank (PDB). We choose the bond length to match the experimentally determined mean distance between successive C_α atoms of 3.81(±0.02) Å. The chain is defined by discrete points denoted by 1, 2, 3,... i...; $d(i,j)$ is the distance between the points i and j. $\psi = 180^\circ - \angle[\pi(i-1,i,i+3),\pi(i-1,i+2,i+3)]$, where the angle $\angle[\pi(i-1,i,i+3),\pi(i-1,i+2,i+3)]$ is the dihedral angle between the planes defined by the points $(i-1,i,i+3)$ and $(i-1,i+2,i+3)$ in Fig. 1(g). M_j is defined to be the geometrical center of the points $j-1$ and $j+1$. The agreement between theory and data is striking considering that the theory is parameter free.

Continuum tube diameter from theory $2\Delta = 5.26\ldots$ Å		
Quantity	Theory	PDB data
HELIX		
Rotation angle ε_0 (deg)	99.8	99.1 ± 3.4
Number of residues per turn	3.61	3.63 ± 0.13
Helix radius R (Å)	2.27	2.30 ± 0.07
Rise per residue p (Å)	1.58	1.51 ± 0.08
Helix pitch P (Å)	5.69	5.47 ± 0.49
Pitch to radius ratio $\eta = P/(2R\pi)$	0.400	0.377 ± 0.046
ψ (deg)	69.1	71.0 ± 4.4
Local radius of curvature (Å)	2.74	2.73 ± 0.05
θ (deg)	91.8	91.3 ± 2.2
μ (deg)	52.4	49.7 ± 3.9
SHEET		
Type I β-sheet: Parallel		
θ (deg)	Flexible	121 ± 10
μ (deg)	~180	191 ± 17
$d(i,M_j)$ (Å)	$2\Delta = 5.26$	5.26 ± 0.16
Type II β-sheet: Antiparallel		
θ (deg)	Flexible	127 ± 10
μ (deg)	~180	186 ± 20
$d(i,j)$ (Å)	$2\Delta = 5.26$	5.26 ± 0.20

In a seminal work, Anfinsen [5] demonstrated that proteins fold rapidly and reproducibly into their native state structures. This naturally led to the text book wisdom [36] that *the amino acid sequence of a protein determines its three-dimensional structure*, leading to much effort in finding the energy minimum of a many-body complex system of a protein in its solvent with a huge number of degrees of freedom and with myriad interactions. Subsequent work by Matthews [16] and others showed that protein structure is nevertheless *very tolerant of amino acid replacement.*

Our results here conclusively demonstrate a simple two-step process for understanding proteins. First, a menu of putative native state structures is created without regard to amino acid sequence and chemistry. In the second step, a given protein selects its native state from this menu. Thus the horrendous problem of working out the native state structure of a given protein from knowledge of its sequence by finding, from scratch, the conformation, which minimizes the net energy of myriad imperfectly known microscopic interactions, is replaced by the much simpler task of finding the best fit of the sequence to one among the library of geometrically sculpted folds determined in a sequence-independent and chemistry-independent manner. This best-fit process, also exploited in the threading algorithm [15], is where the role of the amino acid sequence becomes paramount. Indeed, in an influential series of papers [12,17–19], it has been highlighted that the amino acid side chains must be able to fit into the native state fold with minimal frustration, thereby creating a landscape akin to a folding funnel.

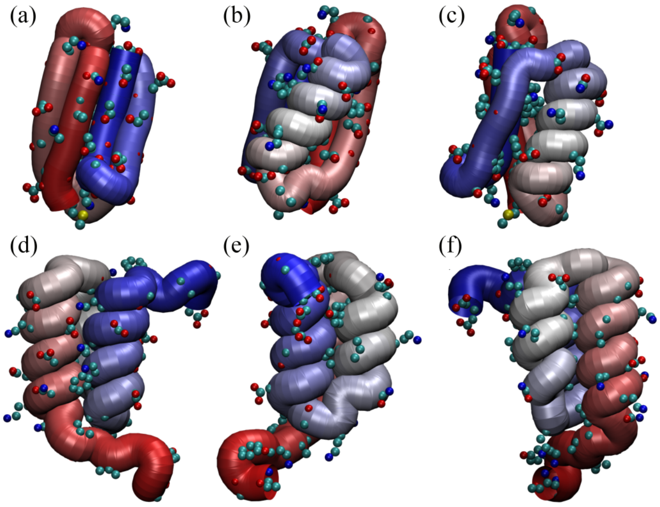

FIG. 6. Consilience between mathematics and biochemistry. The figure shows three views each of two short proteins. (a)–(c) is the 56-residue long protein 3GB1 comprising four strands assembled into sheets along with a single helix. (d)–(f) is a protein of the same length, 2KDL, comprised of a three-helix bundle. Each panel shows a uniform tube, with the theoretically predicted radius of 2.6 Å, whose axis passes through the C_α atoms. The sole exception is the β sheet for which hydrogen bonding was identified using DSSP [11]), where every other C_α atom is considered [as explained in Figs. 3(b) and 3(c)]. The tube color varies continuously from red to blue (via gray) as its axis moves from the N terminal to the C terminal (in the black and white image, the tube is darkest at the two ends). The heavy atoms of the side chains sticking outside the tube are shown. The maximization of the self-interaction through space-filling is evident.

Some 80 years ago, Bernal [1] wrote “Any effective picture of protein structure must provide at the same time for the common character of all proteins as exemplified by their many chemical and physical similarities, and for the highly specific nature of each protein type. It is reasonable to believe, though impossible to prove, that the first of these depends on some common arrangement of the amino acids”. Indeed, our work here shows that the common character of all proteins originates from an appropriate tubelike geometrical description of just the backbone C_α atoms, which are common to all proteins, and results in the library of native state folds sculpted by geometry and symmetry, without a need for sequence specificity or chemistry. The highly specific nature of each protein type then arises from its distinctive amino acid side chains and their fit to one of the folds from the library. For a protein, the folded structure is central to its functionality. The situation is loosely analogous to a restaurant in which the chef (geometry and symmetry) creates a menu of items (the library of putative native state folds) that customers (protein sequences) can order from (fold into). The chef does not cater to the individual tastes of the customers. Rather, all patrons of the restaurant are satisfied picking an item from the menu. As in proteins, the total number of patrons can vastly exceed the number of menu items. If, in fact, the menu of protein structures itself evolved, then one would be confronted by an almost impossible situation for evolution and natural selection in which a protein and its interacting partners would have to coevolve their structures synergistically in order to maintain function. This situation is deftly avoided by the geometrically determined native state folds providing a fixed backdrop for evolution to shape protein sequences and functionalities.

Richard Feynman, in a lecture entitled "There's plenty of room in the bottom: An invitation to enter a new field of physics" at the annual American Physical Society Meeting at Caltech on December 29, 1959, suggested that tiny, nanoscale machines could be constructed by manipulating individual atoms. Proteins are precisely such machines [21,27,33,36]. Indeed, proteins as well as macroscopic machines establish a stable framework that can accommodate moving parts, which perform a function. Proteins are nature's implementation of the abstract forms presented here, a diversity of stable forms deduced entirely from mathematical considerations. These predictions—independent of any chemistry—have implications for life elsewhere in our cosmos [53] suggesting that there is no absolute need for carbon chemistry for life to exist. We look forward to other implementations in the lab, raising the prospect of powerful interacting machines, potentially leading to artificial life [54].

In summary, underlying life's evolving complexity [42] is a sequence-independent energy landscape with thousands of stable minima—a landscape formed from nature's scaffold building blocks, a protein grammar. In both natural and artificial languages, a grammar is a finite set of rules that can generate a large number of syntactically correct sentences or strings. The discretized tube model establishes an immutable grammar of life and "from so simple a beginning, endless"—protein sequences and functionalities—"most beautiful and most wonderful have been, and are being, evolved" [55].

PDB analysis. We have carried out a quantitative comparison between our predictions and protein structure. To develop a working set for comparison, Richardson's Top 8000 set of high-resolution, quality-filtered protein chains (resolution <2 Å, 70% PDB homology level) (see the web site [56]) was further filtered to exclude all structures with missing backbone atoms, yielding a working set of 4416 structures (listed in Table S1 of the Supplemental Material [57]). The working set was cross-checked against 478 proteins having a more stringent homology cutoff of 20%, taken from the Pisces database [24]; 205 entries are in common to both sets. Almost all bond lengths ($C_{\alpha(i)} - C_{\alpha(i+1)}$ distance) (∼99.7%) in the working set are clustered around 3.81 Å, as expected for a *trans* peptide. Those remaining have shorter bonds, ∼2.95 Å, predominantly from *cis* residues. For purposes of comparison, a fixed bond length of 3.81 Å is used. Hydrogen bonds were identified using DSSP [11]. Hydrogen-bonded conformers extracted from the working set include 3595 helices, 8473 antiparallel pairs, 4639 parallel pairs, and 58 820 turns. Helices were identified as 12-residue segments with intrahelical hydrogen bonds ($\mathrm{N_i}$-$\mathrm{H}\cdots\ \mathrm{O}_{i-4}$ and $\mathrm{O_i} \cdots$H-N_{i+4}) at each residue. Antiparallel strand pairs were identified by three interpair hydrogen bonds at (i, j), $(i+2,\ j-2)$, and $(i-2,\ j+2)$; $i \in$ strand 1, $j \in$ strand 2. To avoid possible end effects, only (i, j) residue pairs were used. Parallel strand pairs were identified by four interpair hydrogen bonds between $(i, j-1)$, $(i, j+1)$, $(i+2,\ j+1)$, and $(i-2,\ j-1)$; $i \in$ strand 1, $j \in$ strand 2. Again, only the i-th residue was retained. Double counting was assiduously avoided. β turns were identified by hydrogen bonds between $(i, i+3)$ with no helical residues among the four. The (θ, μ) values were then recorded for points i+1 and i+2 in the turns.

We are indebted to Pete von Hippel for his warm hospitality and to him, Jeremy Berg, and Brian Matthews for stimulating comments. This project received funding from the European Union's Horizon 2020 research and innovation program under Marie Skłodowska-Curie Grant Agreement No. 894784. The contents reflect only the authors' view and not the views of the European Commission. Support from the University of Oregon (through a Knight Chair to J.R.B.), NSF (G.D.R.), University of Padova through "Excellence Project 2018" of the Cariparo foundation (A.M.), MIUR PRIN-COFIN2017 *Soft Adaptive Networks* Grant No. 2017Z55KCW, and COST action Grant No. CA17139 (A.G.) is gratefully acknowledged. The computer calculations were performed on the Talapas cluster at the University of Oregon.

J.R.B., A.M., and T.Š. developed the ideas for the calculations, and all the authors participated in exploring the consequences of the theory. T.Š. carried out the calculations under the guidance of J.R.B. J.R.B. and G.D.R. wrote the paper. All the authors reviewed the manuscript.

[1] J. D. Bernal, Structure of Proteins, Nature **143**, 663 (1939).

[2] L. Pauling, R. B. Corey, and H. R. Branson, The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain, Proc. Natl. Acad. Sci. USA **37**, 205 (1951).

[3] L. Pauling and R. B. Corey, The pleated sheet, a new layer configuration of polypeptide chains, Proc. Natl. Acad. Sci. USA **37**, 251 (1951).

[4] G. N. Ramachandran and V. Sasisekharan, Conformation of polypeptides and proteins, Adv. Prot. Chem. **23**, 283 (1968).

[5] C. B. Anfinsen, Principles that govern the folding of protein chains, Science **181**, 223 (1973).

[6] F. M. Richards, The interpretation of protein structures: Total volume, group volume distributions and packing density, J. Mol. Biol. **82**, 1 (1974).

[7] J. L. Finney, Volume occupation, environment and accessibility in proteins. The problem of the protein surface, J. Mol. Biol. **96**, 721 (1975).

[8] M. Levitt and C. Chothia, Structural patterns in globular proteins, Nature **261**, 552 (1976).

[9] F. M. Richards, Areas, volumes, packing, and protein structure, Annu. Rev. Biophys. Bioeng. **6**, 151 (1977).

[10] P. S. Kim and R. L. Baldwin, Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding, Annu. Rev. Biochem. **51**, 459 (1982).

[11] W. Kabsch and C. Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, Biopolymers **22**, 2577 (1983).

[12] N. Gō, The consistency principle in protein structure and pathways of folding, Adv. Biophys. **18**, 149 (1984).

[13] G. D. Rose, L. M. Gierasch, and J. A. Smith, Turns in peptides and proteins, Adv. Protein Chem. **37**, 1 (1985).

[14] C. Chothia, One thousand families for the molecular biologist, Nature **357**, 543 (1992).

[15] D. T. Jones, W. R. Taylor, and J. M. Thornton, A new approach to protein fold recognition, Nature **358**, 86 (1992).

[16] B. W. Matthews, Structural and genetic analysis of protein stability, Annu. Rev. Biochem. **62**, 139 (1993).

[17] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Funnels, pathways, and the energy landscape of protein folding: A synthesis, Proteins **21**, 167 (1995).

[18] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, Navigating the folding routes, Science **267**, 1619 (1995).

[19] K. A. Dill and H. S. Chan, From Levinthal to pathways to funnels, Nat. Struct. Biol. **4**, 10 (1997).

[20] T. Przytycka, R. Aurora, and G. D. Rose, A protein taxonomy based on secondary structure, Nat. Struct. Biol. **6**, 672 (1999).

[21] C. Tanford and J. Reynolds, *Nature's Robots: A History of Proteins* (Oxford University Press, Oxford, 2001).

[22] M. Denton and C. Marshall, Laws of form revisited, Nature **410**, 417 (2001).

[23] W. Taylor, A 'periodic table' for protein structures, Nature **416**, 657 (2002).

[24] G. Wang and R. L. Dunbrack, Jr., PISCES: A protein sequence culling server, Bioinformatics **19**, 1589 (2003).

[25] J. R. Banavar, T. X. Hoang, A. Maritan, F. Seno, and A. Trovato, Unified perspective on proteins: A physics approach, Phys. Rev. E **70**, 041905 (2004).

[26] T. X. Hoang, A. Trovato, F. Seno, J. R. Banavar, and A. Maritan, Geometry and symmetry presculpt the free-energy landscape of proteins, Proc. Natl. Acad. Sci. USA **101**, 7960 (2004).

[27] A. M. Lesk, *Introduction to Protein Science: Architecture, Function and Genomics* (Oxford University Press, Oxford, 2004).

[28] N. C. Fitzkee and G. D. Rose, Steric restrictions in protein folding: An α-helix cannot be followed by a contiguous β-strand, Protein Sci. **13**, 633 (2004).

[29] G. D. Rose, P. J. Fleming, J. R. Banavar, and A. Maritan, A backbone-based theory of protein folding, Proc. Natl. Acad. Sci. USA **103**, 16623 (2006).

[30] K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl, The protein folding problem, Annu. Rev. Biophys. **37**, 289 (2008).

[31] D. E. Shaw, P. Maragakis, K. Lindorf-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, and W. Wriggers,Atomic-level characterization of the structural dynamics of proteins, Science **330**, 341 (2010).

[32] A.-F. Bitbol, R. S. Dwyer, L. J. Colwell, and N. S. Wingreen, Inferring interaction patterns from protein sequences, Proc. Natl. Acad. Sci. USA **113**, 12180 (2016).

[33] I. Bahar, R. L. Jernigan, and K. A. Dill, *Protein Actions* (Garland Science, New York, 2017).

[34] J. W. Rocks, N. Pashine, I. Bischofberger, C. P. Goodrich, A. J. Liu, and S. R. Nagel, Designing allostery-inspired response in mechanical networks, Proc. Natl. Acad. Sci. USA **114**, 2520 (2017).

[35] C. M. Runnels, K. A. Lanier, J. K. Williams, J. C. Bowman, A. S. Petrov, N. V. Hud, and L. D. Williams, Folding, assembly, and persistence: The essential nature and origins of biopolymers, J. Mol. Evol. **86**, 598 (2018).

[36] J. M. Berg., J. L. Tymoczko, G. J. Gatto, Jr., and L. Stryer, *Biochemistry*, 9th ed. (Macmillan Learning, New York, 2019).

[37] J. K. Leman *et al.*, Macromolecular modeling and design in ROSETTA: Recent methods and frameworks, Nat. Methods **17**, 665 (2020).

[38] C. M. Dobson, T. P. J. Knowles, and M Vendruscolo, The amyloid phenomenon and its significance in biology and medicine, Cold Spring Harbor Perspect. Biol. **12**, a033878 (2020).

[39] M. Fantini, S. Lisi, P. De Los Rios, A. Cattaneo, and A. Pastore, Protein structural information and evolutionary landscape by in vitro evolution, Mol. Biol. Evol. **37**, 1179 (2020).

[40] H. I. Merritt, N. Sawyer, and P. S. Arora, Bent into shape: Folded peptides to mimic protein structure and modulate protein function, Peptide Sci. **112**, e24145 (2020).

[41] J. C. Bowman, A. S. Petrov, M. Frenkel-Pinter, P. I. Penev, and L. D. Williams, Root of the tree: The significance, evolution, and origins of the ribosome, Chem. Rev. **120**, 4848 (2020).

[42] N. Goldenfeld and L. P. Kadanoff, Simple lessons from complexity, Science **284**, 87 (1999).

[43] A. Maritan, C. Micheletti, A. Trovato, and J. R. Banavar, Optimal shapes of compact strings, Nature **406**, 287 (2000).

[44] A. Stasiak and J. H. Maddocks, Best packing in proteins and DNA, Nature **406**, 251 (2000).

[45] S. Przybyl and P. Pieranski, Helical close packings of ideal ropes, Eur. Phys. J. E **4**, 445 (2001).

[46] Y. Snir and R. D. Kamien, Entropically driven helix formation, Science **307**, 1067 (2005).

[47] Y. Snir and R. D. Kamien, Helical tubes in crowded environments, Phys. Rev. E **75**, 051114 (2007).

[48] K. Olsen and J. Bohr, The generic geometry of helices and their close-packed structures, Theor. Chem. Acc. **125**, 207 (2010).

[49] E. P. Wigner, Unreasonable effectivness of mathematics in the natural sciences, Commun. Pure Appl. Math. **13**, 1 (1960).

[50] P. Chaikin and T. Lubensky, *Principles of Condensed Matter Physics* (Cambridge University Press, Cambridge, 2000).

[51] R. Dawkins, *The Blind Watchmaker* (W. W. Norton & Company, London, 1986).

[52] N. Goldenfeld and C. Woese, Biology's next revolution, Nature **445**, 369 (2007).

[53] P. Davies, *The Eerie Silence: Renewing Our Search for Alien Intelligence* (Mariner Books, Boston, 2011).

[54] S. Levy, *Artificial Life: The Quest for a New Creation* (Penguin Books, London, 1993).

[55] C. Darwin, *On the Origin of Species* (John Murray, London, 1859).

[56] http://kinemage.biochem.duke.edu/databases/top8000.php

[57] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.104.014402 for the list of the PDB names of 4416 protein structures used to validate the predictions of our theory.

Correction: The previously published Figure 5 contained an error in the angle given in panel (h) and has been replaced. Corresponding changes have been made to the caption of Figure 5, Table I and its caption, and the caption of Figure 2.