Biophysical Reviews and Letters, Vol. 3, Nos. 1 & 2 (2008) 77–101

World Scientific
www.worldscientific.com

THE DIFFERENTIAL GEOMETRY OF PROTEINS AND ITS APPLICATIONS TO STRUCTURE DETERMINATION

ALAIN GORIELY
Program in Applied Mathematics and Department of Mathematics, University of Arizona, Tucson, AZ 85721, USA
E-mail: goriely@math.arizona.edu

ANDREW HAUSRATH
Department of Biochemistry and Molecular Biophysics, University of Arizona, Tucson, AZ 85721, USA

SÉBASTIEN NEUKIRCH
Laboratoire de Modélisation en Mécanique, UMR 7607 CNRS & Université Pierre et Marie Curie, Paris, France

Understanding the three-dimensional structure of proteins is critical to understand their function. While great progress is being made in understanding the structures of soluble proteins, large classes of proteins such as membrane proteins, large macromolecular assemblies, and partially organized or heterogeneous structures are being comparatively neglected. Part of the difficulty is that the coordinate models we use to represent protein structure are discrete and static, whereas the molecules themselves are flexible and dynamic. In this article, we review methods to develop a continuous description of proteins more general than the traditional coordinate models and which can describe smooth changes in form. This description can be shown to be strictly equivalent to the traditional atomic coordinate description.

Keywords: Globular proteins; protein backbone; polyhelices.

1. Introduction

The accelerating growth of structural biology has created an enormous amount of information which we are only beginning to interpret. Today structural biology is approaching a comprehensive taxonomy of soluble globular proteins, and to make further progress it is imperative to address frontier areas which include membrane proteins, intrinsically disordered proteins, large macromolecular assemblies, and partially organized or heterogeneous structures such as cytoskeletal assemblies and amyloid fibers. These types of problems lie at the experimental frontier, as they severely tax the capabilities of existing physical techiques. Less widely appreciated is that these problems also lie at an intellectual frontier, as we currently lack a language to rigorously describe continuously variable or inexactly defined structures.

This review presents some recent efforts to use a *continuous* representation of protein structure to model and gain some insight into protein structure. The formalism employed is based on the differential geometry of continuous curves- the fold of a protein is considered as a curve which follows the path of the protein backbone. We have developed methods to interconvert continuous curves and discrete atomic coordinates as a way to create and manipulate protein models.[1,2] In this review we present new tools based on continuous methods to investigate and utilize experimental structures and for examining relationships between the known structures and fold families. The review addresses three main objectives:

(1) **Structure determination:** How to devise methods for obtaining the geometric parameters needed to specify a curve (and its associated atomic model) from experimental data for the purpose of structure determination by electron microscopy, NMR, and crystallography.
(2) **Study of Fold Continua for Structural Prediction and Comparison:** How to explore the possibilities simple protein architectures offer by variation of the underlying geometrical parameters to investigate the relations between known folds and as a novel means of predicting molecular structure.
(3) **Continuum Mechanics of Biological Structure:** How to model conformational changes associated with biological function in terms of deformations of elastic bodies.

The mathematical methods developed for creating smooth deformations of protein models can be used to investigate intrinsically disordered regions, evolutionary structural change, distributed conformational changes associated with binding, and accommodation of structures to quaternary structure rearrangements.

The starting point is to represent and model the protein backbone by continuous curves or surfaces. A basic mean of comparison between mathematical structures is to create transformations or mappings between them and then to investigate the properties of the mappings. In particular, properties of continuity and differentiability are crucial for such investigations. The natural representation of molecular structure with atomic coordinates is discrete and so precludes the application of continuous methods. However, continuous supersets which contain the discrete points corresponding to atomic positions can be operated on in this manner. Use of this alternative continuous representation allows the application of powerful analytical and geometric methods to answer questions about protein folds and conformations and their relationships through study of their corresponding space curves and surfaces.

It is critical not to lose sight of the atomic details, which are the foundation for understanding the properties of proteins and provide much of the underpinning for our current mechanistic view of biological processes. The challenge is to integrate both continuous and discrete representations to best understand the mathematical, physical, chemical, and biological properties of proteins and other macromolecules, using the appropriate viewpoint and theoretical tools for a particular question.

In Section 2 the relevant mathematics for protein curve construction and extraction of curve parameters are described, and various avenues for further mathematical developments will be indicated. Section 3 describes how the small number of geometric parameters needed to describe simple protein architectures allows the systematic search of an entire fold space. Section 5 applies the capabilities of continuous deformation of folds from Section 3 and the optimization of curve parameters from Section 2 to the problem of protein design.

2. Methods

The path of a protein backbone is often represented as a series of line segments connecting the alpha-carbon atoms. It could also be represented as a smooth curve passing through the same points. In general, regular three-dimensional curves can be completely specified by their *curvatures* which describe the local bending and twisting of the curve along its length. The local description in terms of curvatures and the global description in terms of spatial coordinates are entirely equivalent. For the particular case of proteins, we have developed specific methods to construct the curvatures of curves that follow the path of protein backbones, and to construct coordinate models of proteins from such curves. The interplay between the local and global descriptions and in particular the modulation of curvatures to control the three-dimensional shape of protein models plays a central role for the work described here. The rest of this section gives some technical details on how to construct curves from data sets and, conversely, how to build protein models with idealized geometry from curvatures. The next sections explore various applications of this formalism to structure determination, protein fold exploration, and protein design.

Let $\mathbf{r} = \mathbf{r}(s)$ be a curve in $\mathbb{R}^3$ parameterized by its arc-length s. At each point s on the curve, one can define (assuming sufficient regularity of the curve) a local general orthonormal basis $\{\mathbf{d}_1(s), \mathbf{d}_2(s), \mathbf{d}_3(s)\}$ by defining the orientation of the vector $\mathbf{d}_3$ with respect to the tangent vector $\mathbf{r}' = v_1\mathbf{d}_1 + v_2\mathbf{d}_2 + v_3\mathbf{d}_3$ to the curve. Since the vectors $\{\mathbf{d}_1(s), \mathbf{d}_2(s), \mathbf{d}_3(s)\}$ are orthonormal, their evolution is governed by

$$\frac{\partial \mathbf{d}_1}{\partial s} = k_3\mathbf{d}_2 - k_2\mathbf{d}_3, \quad \frac{\partial \mathbf{d}_2}{\partial s} = k_1\mathbf{d}_3 - k_3\mathbf{d}_1, \quad \frac{\partial \mathbf{d}_3}{\partial s} = k_2\mathbf{d}_1 - k_1\mathbf{d}_2. \tag{1}$$

That is, $D' = DK$, where D is the matrix whose columns are the basis vectors, $(\,)'$ denotes the derivative with respect to s, and K is the skew-symmetric matrix

$$K = \begin{bmatrix} 0 & -k_3 & k_2 \\ k_3 & 0 & -k_1 \\ -k_2 & k_1 & 0 \end{bmatrix}. \tag{2}$$

This general description of local bases for curves allows to define both the shape of the curve but also the evolution of a triad of orthonormal vectors attached to it through the specification of a vector of *curvatures* $\mathbf{k} = (k_1, k_2, k_3)$ and a vector of *basis orientations* $\mathbf{v} = (v_1, v_2, v_3)$. This description becomes more familiar

if we specialize the general basis by defining $\mathbf{d_3}$ as the tangent vector (that is, $(v_1, v_2, v_3) = (0, 0, 1)$ and $\mathbf{d}_1$ as the normal vector). In which case, Eq. (1) becomes the Frenet equations and the curvatures are $(k_1, k_2, k_3) = (0, \kappa, \tau)$ where κ and τ are, respectively, the curvature and torsion of the curve at the point s. Curvature and torsion can be determined from a given C^3 curve $\mathbf{r}(s)$ by standard differential geometry identities. The *curvatures* of $\mathbf{r}(s)$ are the entries of the Darboux vector $\mathbf{k} = (k_1, k_2, k_3)$ and they can only be determined up to a phase factor φ that describes the phase difference between the normal vector and the first basis vector $\mathbf{d}_1(s)$. The general basis has many advantages due to its compact form, the possibility to define local bases for less regular curves or to assign to the phase factor additional information (such as the twist and shear of a ribbon, or the material property of a tube surrounding the central curve). For clarity sake, in this review, we mostly restrict our analysis to the Frenet frame and define $D = [\mathbf{n}(s), \mathbf{b}(s), \mathbf{t}(s)]$ as the normal, binormal and tangent vectors and $\mathbf{k} = (0, \kappa, \tau)$ where κ is the signed curvature (defined on the real rather than strictly positive) and τ is the usual torsion. The exact description of protein backbones with piecewise helical curves in the next section will require the use of the general basis.

2.1. *Curvatures to curve*

If the curvatures $\mathbf{k}$ and basis orientation $\mathbf{v}$ are given, the curve can be readily obtained by integrating Eq. (1) together with $\mathbf{r}' = \sum a_i \mathbf{d}_i$. These equations form a system of 12 linear non-autonomous equations for the basis vectors D and curve $\mathbf{r}$ that can be written in a compact form by introducing the vector $\mathbf{Z} = (d_{1,x}, d_{2,x}, d_{3,x}, d_{1,y}, d_{2,y}, d_{3,y}, d_{1,z}, d_{2,z}, d_{3,z}, x, y, z)^T$ in which case, the linear system reads

$$Z' = MZ, \quad \text{with} \quad M = \begin{bmatrix} K^T & 0 & 0 & 0 \\ 0 & K^T & 0 & 0 \\ 0 & 0 & K^T & 0 \\ V_1 & V_2 & V_3 & 0 \end{bmatrix}, \tag{3}$$

where K is defined above and V_i is the 3-matrix whose only non-zero entry is row i with value $\mathbf{v}$. The integration of Eq. (3) as a function of s provides both the curve position but also the evolution of the basis. In the particular case where Frenet frame is used $\mathbf{v} = (0, 0, 1)$ and for given curvature κ and torsion τ, the curve is reconstructed. A theoretical example of such a construction is given in Figure 1 and a more detailed fit of an actual protein is shown in Fig. 2.

2.2. *Curves from atomic models*

The first problem is to obtain curves describing the protein backbone from a set of C_α coordinates. One of such constructions can be found in Richardson ribbon diagrams, where curves, ribbons and helices are used to build a three-dimensional picture of the protein backbone. However, this remarkable construction, obtained

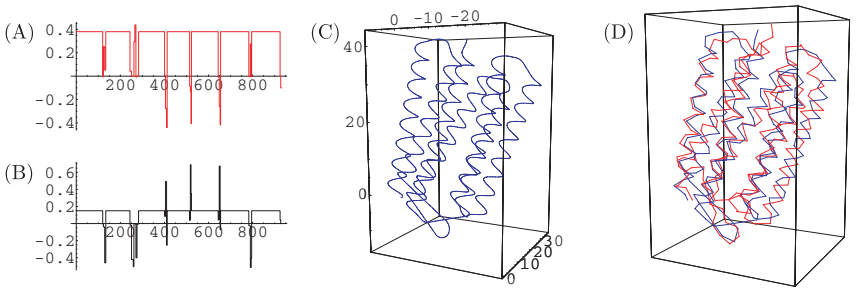


Fig. 1. Construction of curves from curvatures. A curve (C.) is constructed from its curvature-torsion profile (A. and B.). Note that the *signed* curvature (real) is used here rather than the curvature itself (assumed positive). D. The position of the C_α atoms obtained from the curve and the experimentally determined C_α coordinates of bacteriorhodopsin are superimposed. This example utilized 96 curve parameters to specify the curve, where as 3 coordinates are required for each of the 228 atoms in the C_α trace.

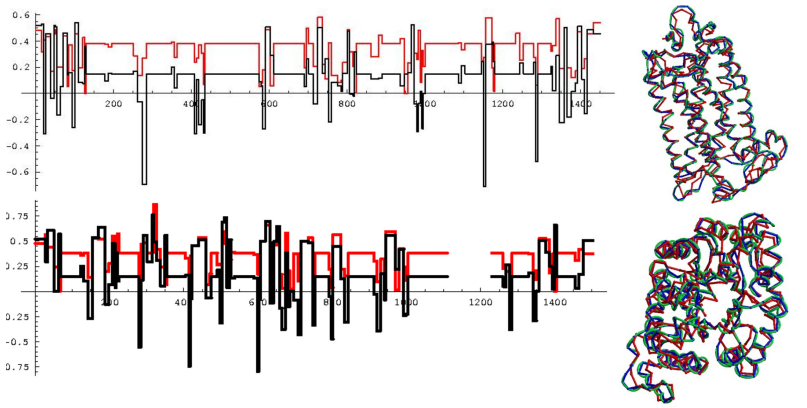
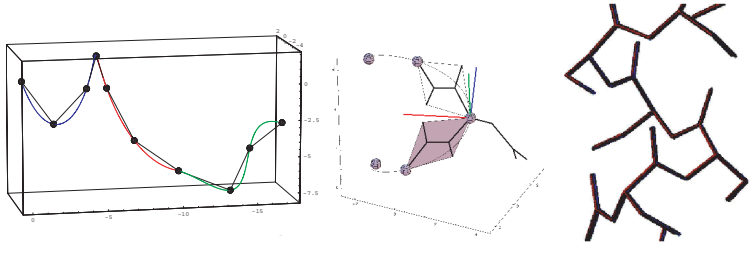


Fig. 2. (Left) curvature profiles, and (Right) Polyhelix models of the primarily α-helical protein bovine rhodopsin (1U19.pdb above) and the mixed α/β protein bacterial luciferase (1BRL.pdb, below). A disordered section in the latter is missing from the model, resulting in a gap in the curvature profile.

by spline fitting, is mostly used for visualization purposes and its mathematical formulation is not a faithful representation of the protein that can be used for other purposes.[3,4] Therefore, one first needs to obtain an exact curve representation of the protein backbone, that is a curve that passes exactly through each C_α atom. Obviously, this can be done in many different ways since there are infinitely many choices of curves passing through a given set of points. A possible choice is through the use of polynomial or rational splines by considering the set of N first atoms and find the spline of degree d through this set, then sequentially build another spline for the next N atoms (with possible overlap). Depending on the degree of smoothness required on the curve the parameters N and d can be adjusted within the proper algorithm.[5] This representation offers an exact description of the protein

backbone but does not carry much information on its global structure. In effect, it replaces one discrete set of points by a discrete set of curves through subsets of points.

A more useful way to represent the C_α trace is to find a piecewise helical curve (referred to hereafter as a *polyhelix*) through the points, that is a continuous curve built out of connected helices. Many authors have considered the problem of fitting helices though sets of points in space. This problem arises in protein structure,[6] engineering design of cables and springs,[7] and nuclear and particle physics for particle tracking.[8] To obtain an exact representation by a polyhelix, 4 consecutive C_α atoms are considered and a unique helix can be constructed (see for instance,[9]) the helix is characterized by its first C_α, curvature and torsion, and axis. To pinpoint the position of the other C_α, three arclengths are required. Together, it amounts to 12 data points corresponding to the 12 atomic coordinates, providing a 1-1 map between atomic data and helical data. The construction proceeds by considering the fourth to seventh C_α's for the next helical piece and so on (See Figure 3). This construction does not provide a purely local representation of the curve since extrinsic data (position of the axis in space) is required. However, using the general basis described above, a complete local representation of the curve can be obtained by specifying for each local piece the following data for the ith helix starting at the atom number $3i+1$: a constant curvature vector $\mathbf{k}^{(i)}$, the orientation of the basis given by a constant vector $\mathbf{v}^{(i)}$ and the arclength positions of the atoms on the helix $S^i = \{s_{3i+2}, s_{3i+3}, s_{3i+4}\}$. This represents 9 data for each successive 3 atoms. The change of orientation of the axis is characterized by the change in the vector $\mathbf{v}$ between the ith and $(i+1)$th helices. Computationally, the protein backbone and the position of the C_α is fully characterized by a list of triplets $H^{(i)} = \{\mathbf{v}^{(i)}, \mathbf{k}^{(i)}, S^{(i)}\}$ and the positions are recovered in extrinsic coordinates by integrating Eq. (3). While this operation seems to be a daunting task, it actually amounts to straightforward matrix algebra due to the exact analytical solution of these differential equations in the cases where curvatures and orientations are constant (See Section 2.4 on



(A) (B) (C)

Fig. 3. (A) Construction of curves from coordinates: A continuous polyhelix curve of 4 segments constructed from points. (B) Construction of coordinates from curves: The local basis on a curve centered on the 3rd C_α and selected atoms expressed in that local coordinate system. (C) Section of an idealized polyserine helix constructed with EDPDB compared with the model constructed in the local coordinates as in Section 2.3.

polyhelices). It is important to note that this representation is general and not restricted to the analysis of proteins with alpha-helices. A representation of atomic data in terms of polyhelices has many advantages: it is purely local in nature and so exploits the natural geometry of the protein, the curvatures carry global information on the curve through curvatures and torsion and therefore allows for direct identification of regions of interest (for instance alpha-helices or different types of turns), and modulation of these curvatures over long distances identifies long-range structure (*e.g.* bending and curving of helices, twist of beta-sheets, etc, ...). To a certain extent, different authors have explored the local geometry of existing proteins using similar approaches with alternative formulations.[10] This construction is the starting point of our analysis. The relevant aspect of local representation is that it helps connect experimental data to structure determination and modeling performed through the use of curve geometry as presented in the next Sections. There, idealized geometries based on polyhelices are used to explore possible folds in parameter space.

Polyhelices provide an exact local representation of the protein backbone. However, in many studies, one may be interested in describing nonlocal properties of these structures. For instance, one may be interested in representing an alpha-helix that may not be strictly helical (due to bending or super-twisting) by one single helix. The problem is then to fit a given structure (helix, plane, ...) through a set of points. This can be done through some averaging process on the exact local representation or directly by fitting through least-square computations. An example of helix fitting is shown in Figure 1 where the curvature-torsion profile was optimized by fitting helical segments to the C_α coordinates from the protein bacteriorhodopsin. There are many different outstanding mathematical and computational issues associated with the problem of fitting a helix,[11] or a cylinder[12] through which will are addressed in separate papers.

2.3. *Atomic models from curves*

Curves can be used to describe proteins and construct models in many different ways. At the basic level, the curve can represent the backbone and C_α can be superimposed by imposing that they are located on the curves at determined position s. In particular, at suitable values of s, the local coordinate system has its origin at the C_α positions, which is a natural choice for the coordinate system in which to express the atomic coordinates for the remaining atoms in the residue. A set of local coordinate system $\mathbf{a} = \{a_1, a_2, a_3\}$ represents the point $\mathbf{p}_a = \mathbf{r}(s) + \sum_{i=1}^{3} a_i \mathbf{d}_i(s)$ in the external coordinates. Conversely, any point $\mathbf{p}_a(s)$ has local coordinates $a_i = (\mathbf{p}_a - \mathbf{r}(s)) \cdot \mathbf{d}_i(s),\ i = 1, 2, 3$. We have converted and tabulated the local coordinates for all the rotamers from[13] which allows the construction of the alpha-helical regions of protein models from curves with idealized geometry, and which can also be used, albeit with some distortion of the backbone atomic arrangements, in the rest of the protein.

2.4. *Polyhelices*

A particularly simple choice of curves is obtained by choosing the curvatures to be piecewise constant so that the curves are piecewise helical. These polyhelices can be used to map precisely atomic coordinates to continuous curves. Conversely, they can also be used to study and classify large families of proteins with idealized geometries. The advantages of this representation are threefold; first, it is consistent with the representation from the atomic coordinates making the comparison with experimental data straightforward; second, large families of proteins can be represented by few parameters and the exploration of fold spaces can be achieved with minimal effort; third, the computation of polyhelices can be reduced to simple linear algebra, making it computationally exact and reliable. The computation of polyhelices is achieved by integrating Eq. (3). Since M is now a constant matrix, the solution of this system is given by

$$Z(s) = A(\kappa, \tau; s) Z(s_0) \tag{4}$$

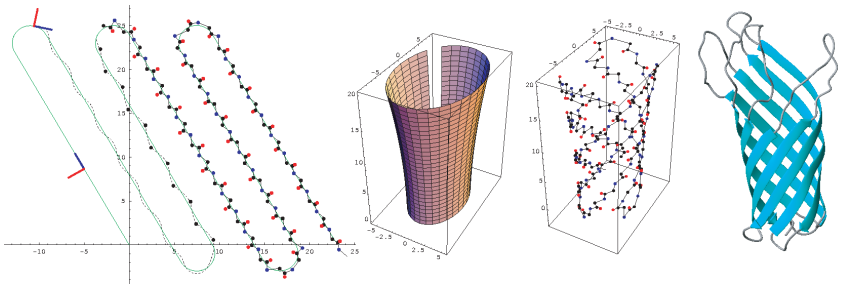
where $A(\kappa, \tau; s) = e^{(s-s_0)M}$ is the matrix exponential. Matrix A can be computed exactly and its entries are linear combinations of trigonometric and polynomial function of s with coefficients depending on κ and τ. A polyhelix with N helices is completely characterized by a list of curvatures and length: $P = \{(\kappa^{(i)}, \tau^{(i)}, L^{(i)}), i = 1..N\}$ and an initial position and basis orientation $Z^{(0)}$. The jth helix on the curve is given by the last three components of the vector Z^j:

$$Z^{(j)} = A(\kappa^{(j)}, \tau^{(j)}; s) Z^{(j-1)} = \prod_{k=1}^{j} A(\kappa^{(j)}, \tau^{(j)}; s) Z^{(0)}, \quad L_{j-1} \le s \le L_j, \tag{5}$$

where $L_j = \sum_{k=1}^{j} L^{(k)}$. Examples of such computations are given in Figure 1. This analytic expression for the curves provides an efficient way to explore fold space as shown in the next Sections.

2.5. *Embedding methods*

While α-helices lend themselves to a linear description, beta-sheets are inherently a linear but 2-dimensional structure. It is therefore natural to use a 2-dimensional embedding of a curve to describe them. The construction proceeds in 2 stages as illustrated in Figure 4. The linear character of the chain is accommodated by a description of the backbone as a plane curve specified by its curvature. The second stage is to map the curve into a two-dimensional surface plane in a three-dimensional space. The particular embedding chosen defines the surface characteristics of the beta-sheet model and can be described by the classical Darboux frame field representation. Here again, the atoms can be represented in the local Darboux frame. The choice of mapping is restricted by the constraints on bond angles and distances and possible beta-architecture with such constraint can be systematically explored within this framework (this and related ideas on possible architectures are discussed in Section 4).



(A) (B) (C) (D)

Fig. 4. (A) A plane curve and its local coordinate system: the side-to-side alternation of the beta-strands is accommodated by expression of the backbone plane atomic coordinates as a function of a sine wave expressed in that coordinate system. (B) A surface in three dimensions, schematically representing the form of a flared, asymmetric beta-barrel. (C) Atomic model resulting from embedding the plane from (A) onto the surface from (B). (D) Ribbon diagram of ompA (PDB code 1BXW). Figure and fit by Katie White.

More explicitly, the construction proceeds in the following steps.

(1) **Polyarc plane curves:**

A plane curve $c_p(s) = \{u(s), v(s)\}$ is described up to rotations and translations in terms of its plane curvature profile $\kappa_p(s)$. Any plane curve with constant curvature κ_p is a circular arc with radius $1/\kappa_p$, and a plane curve with zero curvature is a straight line. The Frenet equations in the plane are described in terms of the tangent and normal vectors $\mathbf{t}_p(s)$ and $\mathbf{n}_p(s)$ as the ODE system $\mathbf{c}_p' = \mathbf{t}$, $\mathbf{t}_p' = \kappa_p \mathbf{n}_p$, and $\mathbf{n}_p' = -\kappa_p \mathbf{t}_p$. Denoting the initial basis as $Z_0 = \{t_u, n_u, t_v, n_v, c_u, c_v\}$, a plane curve and coordinate system may be cast as a matrix equation and its solution obtained in closed form as a matrix exponential.

$$Z' = NZ \quad Z(s) = B(\kappa_p, s)Z_0 \quad \text{where} \quad B(\kappa, s) = e^{Ns}. \tag{6}$$

Explicitly

$$N = \begin{pmatrix} 0 & \kappa & 0 & 0 & 0 & 0 \\ -\kappa & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \kappa & 0 & 0 \\ 0 & 0 & -\kappa & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \tag{7}$$

$$B(\kappa, s) = \begin{pmatrix} \cos \kappa s & \sin \kappa s & 0 & 0 & 0 & 0 \\ -\sin \kappa s & \cos \kappa s & 0 & 0 & 0 & 0 \\ 0 & 0 & \cos \kappa s & \sin \kappa s & 0 & 0 \\ 0 & 0 & -\sin \kappa s & \cos \kappa s & 0 & 0 \\ \frac{\sin \kappa s}{\kappa} & \frac{1-\cos \kappa s}{\kappa} & 0 & 0 & 1 & 0 \\ 0 & 0 & \frac{\sin \kappa s}{\kappa} & \frac{1-\cos \kappa s}{\kappa} & 0 & 1 \end{pmatrix}. \tag{8}$$

If the curvature profile is specified with a list of pairs $Q = \{(\kappa^{(i)}, L^{(i)}), i = 1..N\}$, then the vector Z is obtained with

$$Z^{(j)}(s) = B(\kappa^{(j)}, s - s_0^{(j)}) \cdot \prod_{k=1}^{j-1} B(\kappa^{(k)}, L^{(k)}) \cdot Z^{(0)}, \quad s_0^{(j)} \leq s \leq s_0^{(j-1)} \tag{9}$$

where $s_0^{(j)} = \sum_{k=1}^{j-1} L^{(k)}$. This expression for polyarc curves has a similar structure to the polyhelix construction in that it provides a recursive parametric expression for the plane curve and its local coordinate system.

(2) **Plane Curves on Surfaces:**

Given a plane curve, it can be embedded into three dimensions by mapping the region of the plane which it occupies onto a surface. Any combination of a plane curve and a surface will generate a space curve. This construction is the strategy proposed here for modeling beta-sheet structure.

A general plane curve which can be represented as a map $c : \Re \to \Re^2$, whereas a surface is $S : \Re^2 \to \Re^3$. Therefore the space curve $C : \Re \to \Re^3$ can be written as the composition $C = S \circ c$. We use polyarcs as the explicit map from $\Re \to \Re^2$. Choice of the map $\Re^2 \to \Re^3$ depends on the particular type of structure being modeled. A general means to describe a surface is a 2-dimensional parametrization. $S_3(u, v) = \{X(u, v), X(u, v), Z(u, v)\}$. Therefore the space curve obtained by mapping a plane curve in 2 dimensions to its corresponding space curve in three dimensions can be written $C_3(s) = \{X(C_2(s)), Y(C_2(s)), Z(C_2(s))\}$. One well-understood class, surfaces of revolution (,[14] Chapter 20) is especially applicable to beta barrels. More generally, the flexible and compact representation of curves as polyhelices can be utilized to specify a class of surfaces (termed polysheets) useful for representing β-sheets. Because this type of surface is constructed in terms of polyhelices, it is easily integrated into a polyhelix description to allow for models of mixed alpha and beta architecture. (Although beta strands can be represented within the polyhelix construction as seen in Figure 2, the geometry of the sheet is not explicitly used.)

3. Fold Space Exploration

The set of protein folds is a subset of the set of all possible space curves that can be constructed by standard differential geometry tools. By investigating the set of possible curves, we can find within it the possible protein folds. The key problem is to identify among all curves the ones that may constitute the path of a protein backbone. Mathematically, the problem amounts to finding functions to score the potential of a curve to take the shape of a protein. The main idea is then to explore continuous families of curves defined by a set of parameters and isolate good protein candidates that correspond to points within that parameter space. Small parameter spaces which describe simple protein architectures can be exhaustively sampled. The ability to explore the entire realm of possibilities inherent

in a particular architecture makes it possible to see relations between folds that may not be apparent. The method has applications to protein structure prediction, genome interpretation, and homology modeling. Sequences can be threaded onto curves obtained by a systematic or guided parameter space search. Postulated folds may also serve as protein design targets. Finally, the ability to "interpolate" or "extrapolate" from existing folds may allow prediction of new folds (and explicit construction of their coordinate models) before they are experimentally observed.

3.1. *Protein quality functions*

Most curves in space could not be realized as paths of protein backbones, because they have impossibly tight bends, unrealistically straight segments, have regions that approach too close to other regions of the curve, or are too loosely packed to have sufficient interactions to remain folded. However, there are some curves that satisfy all those criteria. A fundamental question is to identify simple criteria (geometric and physical) to quantify whether a given curve might be realizable as protein backbone conformations if the right amino acid sequence could be found. To address this question, we introduce the idea of *protein quality functions*, to quantify the potential of a curve to be realized as a protein fold. The *curvature space* is the space of parameters defining a a family of curves (for instance, the family of helices is a three-dimensional space defined by curvature, torsion, and length) and a protein quality function is defined at each point of the curvature space and takes real values. A contour plot of a quality function over the curvature space would have islands in regions that correspond to protein-like curves (for instance, α-helices would score very high in the family of helices and a small island would be centered around the ideal value of curvature and torsion for α-helices). Once such a function has been identified, it is possible to investigate questions about the density of folds in fold space, or the connectedness of fold space (*i.e.* are regions of protein folds connected or widely separated?). What are the possible choices for quality functions? Clearly, to conduct a search over large regions of the fold space, the quality functions should be easy to evaluate.

A simple protein quality function can be expressed as a ratio of a term that expresses curve compactness and a term that penalizes a curve which approaches itself too closely. To quantify compactness, the notion of contact order[15] serves nicely. Given a set of points $\{P(s_k)\}$ on a curve, two points $P(s_i)$ and $P(s_j)$ form a contact when within a prescribed contact distance in space. The contact order is the sequence distance $|s_j - s_i|$ averaged over all contacts. Contact order is large for curves in which many pairs of points distant on the curve are close in space and so serves as a simple quantitative measure of compactness.

$$C_O = \frac{1}{LN} \sum^{N} |s_j - s_i| \tag{10}$$

where N is the number of contacts and L the number of points. However, a curve that is too compact will approach too close to itself. Defining a clash as a pair

of points which are closer than a prescribed clash distance in space, curves with self-intersections are severely penalized by using the quality function:

$$Q_1 = \frac{C_O}{2^M} \tag{11}$$

where C_O is the contact order of the curve and M the number of clashes. Examination of the distances between C_α coordinates of several repeat proteins suggested the use of 9 Angstroms as the contact distance and 4 Angstroms as the clash distance. An advantage of this quality function is that it can be used both on C_α coordinates from real proteins and on points obtained from curves.

There many different choices for quality functions emphasizing different features of a candidate curve. Such functions could be evaluated on smooth curves which may have advantages for theoretical investigations or on discrete sets of points obtained from the curve, which have the advantage of ease of comparison with experimentally obtained coordinate sets. Other interesting possibilities include the use of the global radius of curvature[16] providing both a local estimate of curvature and a global estimate of self-contact; simplified versions of energy functionals as used in homology modeling and structure prediction[17,18] ; family of Vasiliev knots invariants for protein which have already shown great promise for classification purposes;[19,20] or statistics of distributions of curve parameters in curves fitted to experimentally determined coordinates.

3.2. *The fold spaces of polyhelices*

As an example of fold space exploration we use the quality function Q_1 defined above to study the fold space of some polyhelical families. As discussed in Section 2, a curve consisting of N helical segments can be specified by a list of N {curvature,torsion,length} triples and the fold space is a $3N$-dimensional which can be systematically explored with the function Q_1. This space is very convenient to model proteins with α-helices since a few parameters are necessary to describe the main building blocks of the proteins.

First, a simple application of the use of the quality function to rank the turns connecting two α helices is shown in Figure 5. The curvature/torsion profile and its corresponding helical hairpin are shown. The turn is parameterized by the two triples $\{\{\kappa_1, \tau_1, l_1\}, \{\kappa_2, \tau_2, l_2\}\}$. This 6-dimensional fold space can be searched exhaustively. Contour plots (in which light colors indicate high (favorable) values of Q_1 function) as a function of κ_2 and τ_2, for different values of l_2 are shown for 2 different choices of κ_1, τ_1, and l_1 in Figure 5.

The presence of islands and plateaus indicates that only certain combinations of curvature and torsion gives rise to reasonable turns. By selecting points within the white regions, a list of turns which specify high-scoring protein-like helical hairpin curves can be collected in a library of candidates for connecting turns. Once this library is built, one can proceed hierarchically with the search for protein candidate

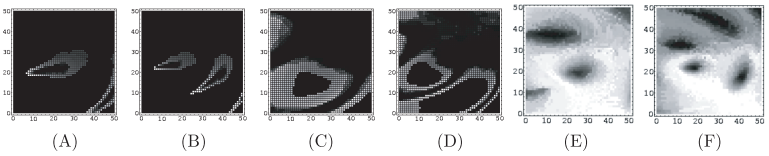


Fig. 5. (A and B) Contour plot of the PQ function (κ_2 vs. τ_2) for $l_2 = 3.0$ and $l_3 = 5.0$. (C and D) A similar plot using different κ_1, τ_1, and l_1 values. (E and F) Contour plot of the best value of the quality function for any choice of κ_2, τ_2, and l_2 plotted as a function of κ_1, and τ_1. Here l_2 is 3.0 (E) and 5.0 in (F).

by finding helical repeats where the connecting turns are given by helical hairpin from the library.

As a more complex example, consider the various types of helical repeat proteins which share a common architecture of (helix$_1$-turn$_1$-helix$_2$-turn$_2$)N. Curves corresponding to this architecture can be specified with 14 parameters. (The curvature and torsion of an alpha helix are fixed, so only 2 parameters for the lengths of the helices are needed. Each turn is described by 2 $\{\kappa, \tau, l\}$ segments.) A curve is determined by a point the 14 dimensional fold space. A systematic search of this space is still a daunting undertaking and some simplifying assumptions are necessary. We assume that the sections helix$_1$-turn$_1$-helix$_2$ and helix$_2$-turn$_2$-helix$_1$ as helical hairpins. This reduces the search to a 4-dimensional space, over the two (continuous) helix lengths and two turns from the (discrete) helical hairpin list. For high-scoring curves, we construct polyalanine atomic models and overlay them on experimentally determined repeat protein coordinate sets (PDB entries 1i7x,1b89,and 1b3u. We will expand the comparison and use more sensitive methods of structure comparison[21]). An example of a "hit" from this search is shown in Figure 6, in which a curve close to the armadillo repeat protein β-catenin (PDB code 1i7x) was obtained. Quite remarkably, this result shows that a construction solely based on simple geometric principles can capture the form of existing proteins accurately, and that simple quality functions can be used to search rapidly through the curve specification parameter space and identify protein-like curves. Suggestively, some regions of

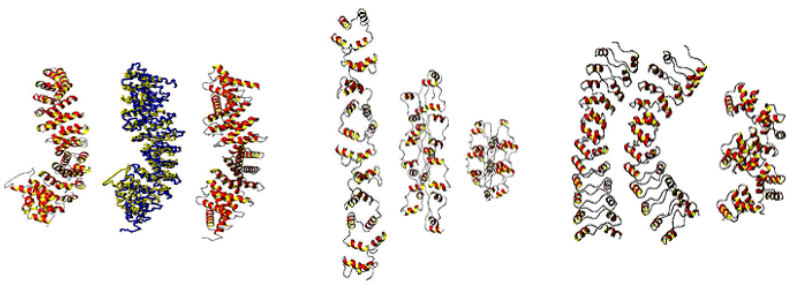


Fig. 6. On the left is a ribbon diagram of beta-catenin, and on the right is the ribbon diagram of coordinates constructed from the curve- in between is their superposition with beta-catenin and the curve-derived model.

the fold space have high quality values and yet describe curves that do not resemble any known proteins. Some appear to have plausible packing arrangements- examples are shown in Figure 6. The left three show examples of a family which can be described as a stack of antiparallel coiled-coils. The right three show examples of a family which most closely resembles leucine-rich repeats but which have a second helix in place of the beta-strand. Once interesting regions of a fold space have been identified with one quality functions, it can be further explored by using other quality functions providing, in effect, a series of filters for plausible proteins.

This methodology has been demonstrated with helical repeat proteins due to the small number of parameters needed to describe curves with such architecture. However, the same idea is applicable both to non-repeat proteins and also to beta- or mixed alpha-beta architectures by using different (or mixed-) representations of curves such as the one given in Section 2.5 for beta sheets.

3.3. *Modeling helical bundle and β-barrel membrane proteins*

Experimental structure determinations of membrane proteins are more difficult than those for soluble proteins, and there is a continuing need for improved modeling methods applicable to membrane proteins.[22] Most known membrane proteins fall into two structural categories: helical bundles and β-barrels. These classes of proteins are well suited to our continuous descriptions. The helical bundle category is conveniently represented with polyhelices (Fig. 2). The β-barrel models are constructed using the embedding procedure described in Section 1.3, and using the Darboux frame for atomic model construction (Fig. 7).

For these protein architectures, we will use these efficient parametrizations of structure to explore the classes of likely folds. Backbone coordinate models of the possible β-barrel structures with different numbers of strands and different sheet registers will be constructed.[23] Similarly, helical bundles with different numbers of helices and different patterns of membrane insertion and helical arrangement will be constructed (cf[24]). These will be scored by geometric criteria to devise concrete

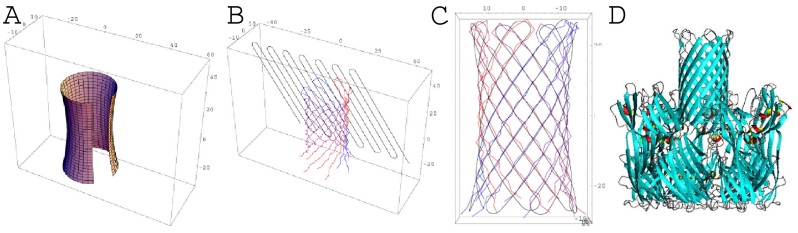


Fig. 7. Embedding of plane curve onto surface to model β-barrel proteins. (A) Surface of revolution. (B) An antiparallel plane curve in the x-z plane shown with C_α trace of the regular β-barrel section from α-hemolysin (7AHL.pdb Resulting fit to these C_α coordinates of the plane curve from (B) embedded into the surface from (A). (D) Ribbon diagram of α-hemolysin for comparison. (Note this molecule is trimeric, whereas only a single plane curve is used to model the barrel in this example. More complex curves and surfaces are used for modeling less regular β-barrels.)

examples of likely membrane protein architectures. Since membrane proteins tend to have a simpler form than globular proteins a coarse sampling of the possible structures can of provide useful starting models for other applications. For instance, plausible folds can be used for modeling purposes by threading protein sequences onto them. This may be useful in conjunction with different structural biology methods described in section 2, in that a small number of experimental constraints on the structure may be sufficient to distinguish between certain classes of models. If more detailed information becomes available, such generic structures will serve as structural templates for curvature-based optimization.

4. Protein Design

The goal of *de novo* design of proteins is to create sequences that fold into a desired three-dimensional structure.[25,26] The continuous representation of proteins is of great utility for design work as it allows the overall specification of protein architecture without requiring that all the atomic details be considered at the outset. It separates the geometric problem of finding a suitable backbone from the problem of consistent atomic interactions. Therefore, it provides tools to specify a protein scaffold that may not resemble natural protein folds.

The primary determinant of globular protein folds is encoded in the binary pattern of hydrophobic and polar residues which defines the hydrophobic core and hydrophilic surface of the protein.[27–29] Recognizing the fundamental character of the hydrophobic core for structure specification, the computational protein design field initially focused on means to design cores compatible with existing natural scaffolds.[30] In its simplest form, the problem of protein design is the problem of selecting the best sequence which fills the interior volume of a given scaffold. The need for criteria to distinguish among different sequences that satisfy the crude inside/outside constraint but which led to stable but non-unique structures[31] has led to the development of sophisticated energy functions that take into account the many physical interactions specifying a unique folded protein structure.[32] By incorporating these finely-tuned energy functions with the power and efficiency of new combinatorial algorithms,[33] core-packing algorithms have matured to the point where realizable designs have become routine. However, it has become increasingly clear that protein backbones are not rigid scaffolds but rather can move in unexpected ways in response to changes in core composition.[34,35] The problem of how to incorporate such backbone freedom into protein design algorithms, and to control the computational cost of exploring these extra degrees of freedom, is the main challenge to the field protein design field.[36]

The greatest appeal of a geometric approach to protein design is that it separates the specification of structure from the validation of structure. The geometric representation can be used to construct new plausible models, and different energy functionals can be used to rank them. By separating the two problems, the constraints of sequence and structure can be looked at as independent, and either can

be varied to best satisfy the other. In essence, one can ask the question "What is the best sequence for this particular curve?" as well as "What is the best curve for this particular sequence?" By iterating between the two problems, a solution for which both sequence and structure are mutually optimal can be determined .

The continuous representation provides a theoretical foundation for backbone design as well as natural and inexpensive computational methods especially compared to the computations in dihedral angle space.

Recently hybrid methods which alternate between sequence rearrangement and energy minimization to allow backbone relaxation have shown great promise[37] We propose to generalize these ideas by adapting the continuous description to allow simultaneous minimization of an externally specified energy function with respect to curve parameters and side-chain rotamers as a way to find mutually optimal sequence and structure solutions.

4.1. *Creation of structure specification and optimization tools*

An important part of protein design is the development of algorithms for optimization of models optimizing user-supplied design criteria (such as a the minimization of an energy function) by adjusting curve parameters, and the identification of an interior volume needed to construct a hydrophobic core. Considered in combination, these two constraints will identify suitable scaffolds for design targets. Section 5.2 will discuss how to create sequences that best conform to these scaffolds.

The optimization of curves and curve-derived coordinate models by variation of curve parameters is based on the the same underlying methods as the ones described of Section 3 where agreement with experimental data rather than with design constraints was required. From an atomic perspective, the design task is to create energy functions balancing the relative contributions of the different types of interactions (*e.g.* electrostatic, Vanderwaal's, hydrophobic) involved in stabilizing folded proteins. From the coarse-grained perspective given by the continuous description, simple geometric criteria based on protein quality functions have considerable discriminating power. The main idea is therefore to use the optimization algorithms of Section 3 with a protein quality function of Section 4 to identify suitable design targets which can then be used to construct detailed atomic models as in section 5.2.

The starting point for sequence design is the determination of the binary hydrophobic/hydrophilic pattern in the primary sequence[38] A design target conformation can be characterized by a curve. To determine a binary pattern along the curve, it is necessary to identify the interior volume enclosed by the curve. To do so, a simple method (See Figure 8) consists in locating C_α atom on the curve, and to construct a grid covering the extent of the C_α. For every point on the grid, the number of neighboring C_α atoms is determined. Choosing only the points which have many neighboring C_α atoms as the center of spheres, an approximate "interior" volume is obtained by taking the union of these spheres. Constructing the C_β

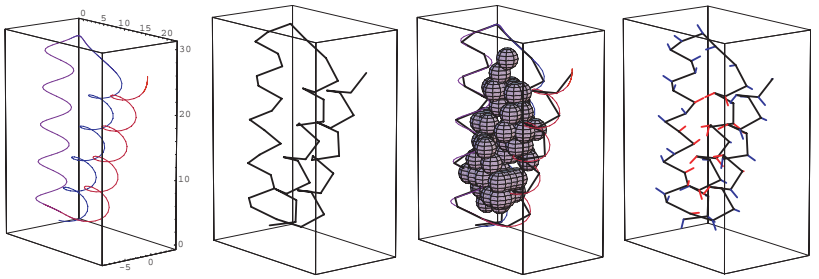


Fig. 8. Determination of binary patterning. (A) A three-helix bundle curve. (B) C_α atoms on the curve. (C) "Inside" volume as the union of spheres centered at points near (within 8 Angstroms) sufficiently many (more than 12) C_α atoms. (D) C_β atoms constructed from the curve located within the "inside" volume.

atoms from the curve, the atoms facing the "interior" volume can be assigned a hydrophobic character by standard core-packing algorithm like DEE.[33,39] Subsequently, the curve can be modified so as to bury more or fewer side-chains according to desired criterion such as buried hydrophobic surface or volume.[40,41]

5. Continuum Mechanics of Biological Structures

The methods described in the previous sections can describe protein models accurately but are of geometric nature and exist without reference to physical assumptions about molecules. In this section we build on this foundation, refining the models by incorporating physical considerations to relate structural and energetic properties of molecules and moreover make experimentally testable predictions. We will use classical elasticity theory for this purpose. Primarily we will seek analytical solutions by employing the semi-inverse method of Saint-Venant, although with modern computational resources, numerical solution of the elastic field equations can be applied when needed as an alternative.

For the purposes of structural modeling, elastic deformations can be used to describe large-scale, distributed conformational changes. But importantly, these responses are described in terms of changes in body coordinate systems, which can be made to coincide with the local coordinate frames we have used for construction of atomic models. Thus elasticity provides a natural formalism for devising physical theories which make structural predictions. For instance the elastic energy of filamentous structures can be described with the equation

$$E_{elastic} = \frac{1}{2}\int_0^L \left(B_1(s)\kappa_1^2(s) + B_2(s)\kappa_2^2(s) + B_3(s)\kappa_3^2(s)\right) ds \qquad (12)$$

where B_1, B_2, and B_3 are the elastic constants and the κ_i are the deviations from equilibrium values of the curvatures. Additional terms may be included, for instance to include nonlocal effects or to obtain force-extension curves for study of mechanical responses.

5.1. *Continuum elastic theory of coiled-coils*

Within this formalism, a coiled-coil can be modeled as conjoined elastic filaments with elastic constants $B_1 = B_2 = B$ which describe resistance to bending and $B_3 = C$ which describes resistance to twisting. In a coiled-coil configuration, the center line $\mathbf{r}(s)$ of each filament itself is a helix. The axis is along z, the radius is written R, the pitch is $2\pi R/\tan\theta$ and the super-helical angle θ is the complement of the pitch angle. We parametrise the (helical) center line as:

$$\mathbf{r}(s) = \begin{pmatrix} +R\ \sin\psi(s) \\ -R\ \cos\psi(s) \\ s\ \cos\theta + z_0 \end{pmatrix}, \qquad \psi(s) = \frac{\sin\theta}{R}s + \psi_0 \tag{13}$$

where $\psi(s)$ is the equatorial angle in the (x, y) plane. The (constant) curvature and torsion of the super-helical axis are $\kappa = \sin^2\theta/R$ and $\tau = \sin\theta\cos\theta/R$. The Frenet and Cosserat frame vectors are obtained from this parametrization by standard identities and can so be used to construct atomic models.

Coiled-coils are held together by interactions along one face of their constituent helices, and this interaction face can be parametrized in terms of the Cosserat directors and the parameter $\hat{\tau}$ which describes the twisting of the interaction face along the surface of the alpha helix. Constraining these interaction surfaces to be joined provides a structural constraint, and minimizing Equation 12 subject to this constraint yields the equilibrium conformation of the coiled-coil in terms of a relation between the elastic constants B and C, the equilibrium super-helical angle θ_0 and the interface parameter $\hat{\tau}$.

$$-\frac{2B\sin^3\theta_0\cos\theta_0}{C\cos 2\theta_0} = \sin\theta_0\cos\theta_0 - \hat{\tau}R. \tag{14}$$

This elastic theory[42] agrees well with structures of leucine zipper coiled-coils (see table below) and can be used to construct atomic models of large macromolecular complexes not easily accessible to experimental structural analysis.

	X-ray data				model	
GCN4	res./turn	rise/res.	R	2θ	$\hat{\tau}$ (rad/Å)	2θ
dimer	3.62	1.51 Å	4.9 Å	$-23.4°$	-0.039	$-22°$
trimer	3.60	1.53 Å	6.7 Å	$-26.8°$	-0.033	$-25°$
tetramer	3.59	1.52 Å	7.6 Å	$-26.0°$	-0.030	$-26°$

5.2. *Modeling the open and closed states of the CusCFBA bacterial efflux complex*

An application of the coiled-coil theory is to the bacterial metal efflux complex cusCFBA. This complex allows bacteria to survive high concentrations of toxic heavy metals such as copper by pumping them out of the cell.[43] Other members

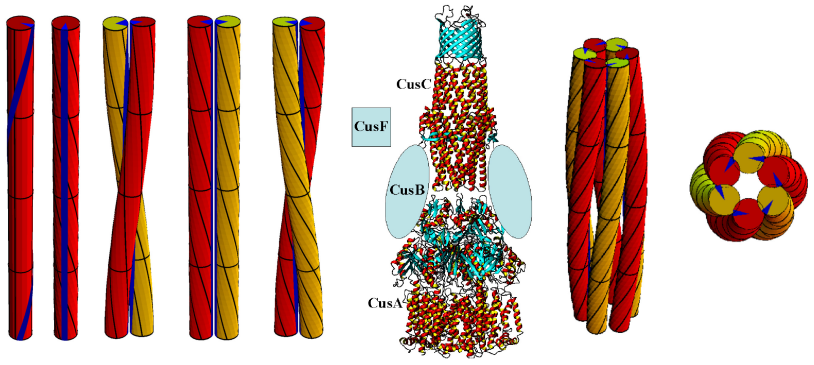


Fig. 9. From left: Individual filaments of a coiled-coil in the (A) unstressed state (B) twisted state. (C and D) Different coiled-coil configurations constrained by their interaction surfaces. The equilibrium state requires a balance between bending and twisting of the filaments. (F) Ribbon diagram representing the bacterial efflux complex. (Coordinates are from the closely related multidrug resistance system components TolC (1EK9.pdb). (G) Schematic models of an iris formed from a hexameric coiled-coil. (Note CusC is trimeric but its iris is formed by 12 helices.)

of this family of proteins are involved in bacterial drug resistance. The structure of this complex is shown schematically in Figure 9. CusC forms the channel, CusA is thought to function as a pump, and the peripheral subunits CusF and CusB are thought to effect the opening and closing of the channel in response to metal concentration.[44] A crystal structure of TolC[45] provides a structural model for the closed form of CusC in which the channel is blocked, and these authors postulated a distinct open conformation for this molecular complex. The particular bending and twisting which is needed to open and close an iris-like arrangement of helices is also well-described within our coiled-coil theory. This theory can also be used to devise atomic coordinate models of the states of the complex to investigate the mechanism of action of the cusCFBA complex, and in particular the role of the periplasmic subunits cusF and cusB.[44,46]

Our theory of coiled-coils[42] relies on a geometric relation between the interface residues by which the individual helices associate. A simple modification allows the constraint to be modified to devise barrel-like structures, which are open in the middle. But it is considerably more complicated to devise a constraint equally compatible with *two distinct states* which must be accommodated in an iris-like structure. In the simple coiled-coil theory an interaction surface on the individual helices (such as the stripe of leucine residues in a leucine zipper) defines the coiled-coil geometry. Our hypothesis for the general problem of forming an iris is that two distinct interaction surfaces (with different values of $\hat{\tau}$) must necessarily exist, which specify distinct coiled-coil geometries corresponding to the open and closed states. In the context of the cusCFBA system, the periplasmic subunits cusB and

cusF could bind to one or the other interaction surface so as to control the state of the channel subunit cusC.

5.3. *Modeling oligomerization states of Adiponectin*

A second application of the coiled-coil theory is to the anti-diabetic signalling hormone adiponectin, which stimulates insulin sensitivity. This molecule circulates in the bloodstream in three distinct oligomerization states (trimer, hexamer and 18mer) but only the largest appears to be active in signalling.[47,48] The Acrp30 gene coding for adiponectin is organized into segments which code for a globular headpiece, a collagen-like domain, and a short tail region. The structure of the globular headpiece was determined by crystallography but the detailed structure of the remaining portion of the molecule in unknown. The collagen-like domain is necessary for the assembly of adiponectin monomers into trimers through formation of a triple-helical collagen I-type coiled-coil.[49,50]

We hypothesize that the organizing principle by which the coiled-coil coil trimers are assembled into the hexameric and high-molecular weight oligomerization states is a coiled-coiled-coil. Our elastic coiled-coil model is well suited to modeling not only the collagen triple helix, but also the higher-order forms by treating the collagen triple helix itself as a single elastic filament.

The first step in the mathematical approach consists in modeling adiponectin as different collagen strands twined together. This provides a family of possible models depending on the geometric parameters (radii, twist, etc.). Purely geometric requirements on the possible forms already strongly constrain the possible models. The second step is to identify within this large family of coiled structures, subfamilies for which collagen domains are complementary (see Figure 11). This is done by defining a suitable energy that can be minimized within the family of coils. The third and crucial step is to consider specific adiponectin molecules and use constraints from experimental data to refine the structure. The experimental data comes from various sources. The stoichiometry of the three states has been accurately determined by analytical ultracentrifugation and estimates of the radial and axial dimensions of the different species are available from electron micrographs. This information provides a basis for modeling the oligomers, but more subtle details such as the relative twists of the individual molecules in the oligomers or the juxtaposition of sidechains requires higher-resolution or site-specific information.

6. Conclusions

We have presented here various results regarding the use of simple differential geometry to model protein structures. The classical approach to model proteins is in term of discrete models based on atomic coordinates. These models have shown to be very successful for a variety of problems from identifying folds to protein functions.

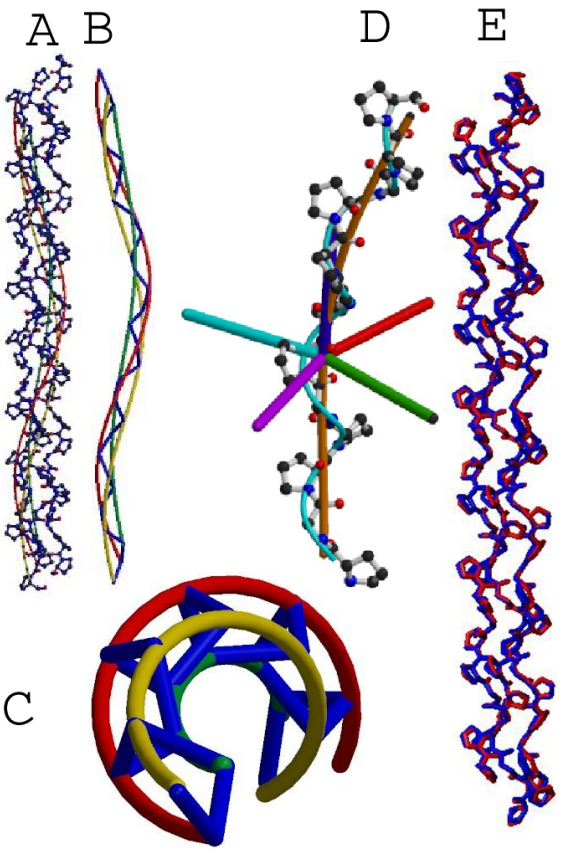


Fig. 10. Construction of coordinate models of collagen using Cosserat frames. (A) X-ray crystallographic model (blue) of $(Gly - Pro - Pro)_{10}$ (1K6f.pdb) with helical curves (red, green, and yellow) fitted to C_α coordinates of one chain. (B) C_α trace of chain from (A) with fitted helical curves. The green curve corresponds to the Gly positions, red to the X positions, and yellow to the Y positions in the Gly-X-Y pattern (C) Top view of (B). (D) The cyan curve is a coiled-coil curve passing through the $C\alpha$ positions, and the orange curve corresponds to the axis of the cyan curve. The Cosserat frame attached to the orange curve ($\mathbf{d}_1$, cyan; $\mathbf{d}_2$, magenta; $\mathbf{d}_3$, blue) is obtained via a rotation of the Frenet frame ($\mathbf{t}$, blue; $\mathbf{n}$, green; $\mathbf{b}$, red). The cyan vector twists about the orange curve so as to pass through each C_α position in turn. The particular Cosserat frames where the $\mathbf{d}_1$ vector passes through a C_α position can be used to construct coordinate models. (E) Complete coordinate model constructed using the Cosserat frames for each amino acid is shown in red, superimposed on the experimentally determined coordinates from 1K6F.pdb, in blue. The rmsd on all atoms is 0.3 $\AA$. (The above color figures can be found in the electronic version of this paper.)

However, we believe that they are limited in many respects. First, computationally, they require considerable effort and ingenuity and will always be limited by computer speed and memory. In this regard, it appears clearly that there is a need for hybrid methods that combine both discrete and continuous models. We expect that the lines of research presented here will be relevant to solve these problems. Second, from a conceptual standpoint, much can be gained from understanding

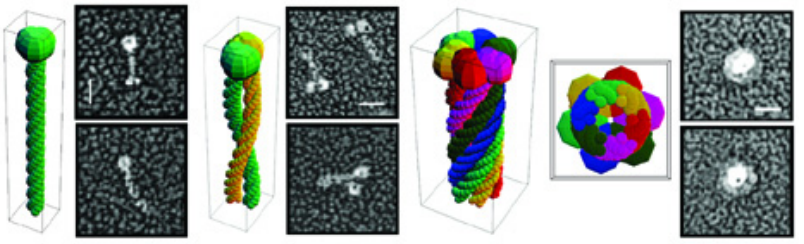


Fig. 11. Models of the oligomerization states of adiponectin compared to electron micrographs. (Left) trimer (a single coiled-coil). (Center) Hexamer (two trimeric coiled-coils wound around each other to form a coiled-coiled-coil). (Right) Octadecamer (a distinct coiled-coiled-coil composed of 6 trimeric coiled-coils)

proteins as continuous flexible objects; the full power of continuum mechanics will allow us to gain insight into some basic phenomena such as energy transfer in the ATP-synthase or the mechanical response of fibrous proteins.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation under grants No. DMS-0604704 and DMS-IGMS-0623989 to A. G. and a BIO5 Institute Grant to A. G. and A. H. We also acknowledge many discussions and ongoing collaborations with Tsushuen Tsao and Megan McEvoy. The beta barrel figures were produced by Katie White in collaboration with A. H.

References

1. A. C. Hausrath and A. Goriely. Repeat protein architectures predicted by a continuum representation of fold space. *Protein Science*, 15(4):753–760, 2006.
2. A. C. Hausrath and A. Goriely. Continuous representations of proteins: Construction of coordinate models from curvature profiles. *Journal of Structural Biology*, (In press), 2006.
3. R. Koradi, M. Billeter, and K. Wuthrich. Molmol: A program for display and analysis of macromolecular structures. *Journal of Molecular Graphics*, 14(1):51, 1996.
4. P. J. Kraulis. Molscript - a program to produce both detailed and schematic plots of protein structures. *Journal of Applied Crystallography*, 24:946–950, 1991.
5. S. Asaturyan, P. Costantini, and C. Manni. Local shape-preserving interpolation by space curves. *Ima Journal of Numerical Analysis*, 21(1):301–325, 2001.
6. J. A. Christopher, R. Swanson, and T. O. Baldwin. Algorithms for finding the axis of a helix: fast rotational and parametric least-squares methods. *Computers Chem.*, 20:339–345, 1996.
7. M. J. Keil and J. Rodriguez. Methods for generating compound spring element curves. *J. Geometry and Graphics*, 3:67–76, 1999.
8. R. Frühwirth, A. Strandlie, and W. Waltenberger. Helix fitting by an extended riemann fit. *Nuclear Instruments and Methods in Physics Research A*, 490:366–378, 2002.
9. H. Sugeta and T. Miyazawa. General methods for calculating helical parameters of polymer chains from bond lengths, bond angles, and internal rotation angles. *Biopolymers*, 5:673–679, 1967.

10. M. Bansal, S. Kumar, and R. Velavan. Helanal: A program to characterize helix geometry in proteins. *Journal of Biomolecular Structure & Dynamics*, 17:811–820, 2000.
11. Y. Nievergelt. Fitting helices to data by total least squares. *Computer Aided Geometric Design*, 14:707–718, 1997.
12. O. Devillers, F. P. Mourrain, and P. Trebuchet. Circular cylinders through four or five points in space. *Discrete Comput Geom*, 29:83–104, 2003.
13. S. C. Lovell, J. M. Word, J. S. Richardson, and D. C. Richardson. The penultimate rotamer library. *Proteins-structure Function Genetics*, 40(3):389–408, 2000.
14. Alfred Gray. *Modern differential geometry of curves and surfaces with Mathematica.* CRC Press, Boca Raton, 2nd edition, 1998.
15. K. W. Plaxco, K. T. Simons, and D. Baker. Contact order, transition state placement and the refolding rates of single domain proteins. *Journal of Molecular Biology*, 277(4):985–994, 1998.
16. O. Gonzalez and J. H. Maddocks. Global curvature, thickness, and the ideal shapes of knots. *Proc. National Acad. Sciences United States Am.*, 96(9):4769–4773, 1999.
17. T. Lazaridis and M. Karplus. Effective energy functions for protein structure prediction. *Current Opinion in Structural Biology*, 10(2):139–145, 2000.
18. F. Melo, R. Sanchez, and A. Sali. Statistical potentials for fold assessment. *Protein Science*, 11(2):430–448, 2002.
19. P. Rogen and B. Fain. Automatic classification of protein structure by using Gauss integrals. *Proc. Natl. Acad. Sci. USA*, 100:119–124, 2003.
20. P. Rogen and H. Bohr. A new family of global protein shape descriptors. *Math. Biosc.*, 182:167–181, 2003.
21. L. Holm and C. Sander. Protein-structure comparison by alignment of distance matrices. *Journal of Molecular Biology*, 233(1):123–138, 1993.
22. A. Oberai, Y. Ihm, S. Kim, and J. U. Bowie. A limited universe of membrane protein families and folds. *Protein Science*, 15(7):1723–1734, 2006.
23. A. G. Murzin, A. M. Lesk, and C. Chothia. Principles determining the structure of beta-sheet barrels in proteins .2. the observed structures. *J. Mol. Biol.*, 236(5):1382–1400, 1994.
24. J. U. Bowie. Helix-bundle membrane protein fold templates. *Protein Science*, 8(12):2711–2719, 1999.
25. R. B. Hill, D. P. Raleigh, A. Lombardi, and N. F. Degrado. De novo design of helical bundles as models for understanding protein folding and function. *Accounts of Chemical Research*, 33(11):745–754, 2000.
26. N. Pokala and T. M. Handel. Review: Protein design - where we were, where we are, where we're going. *Journal of Structural Biology*, 134(2-3):269–281, 2001.
27. W. A. Lim and R. T. Sauer. Alternative packing arrangements in the hydrophobic core of lambda-repressor. *Nature*, 339(6219):31–36, 1989.
28. N. C. Gassner, W. A. Baase, and B. W. Matthews. A test of the "jigsaw puzzle" model for protein folding by multiple methionine substitutions within the core of t4 lysozyme. *Proceedings of the National Academy of Sciences of the United States of America*, 93(22):12155–12158, 1996.
29. S. Dalal, S. Balasubramanian, and L. Regan. Protein alchemy: Changing beta-sheet into alpha-helix. *Nature Struct. Biol.*, 4(7):548–552, 1997.
30. J. W. Ponder and F. M. Richards. Tertiary templates for proteins - use of packing criteria in the enumeration of allowed sequences for different structural classes. *Journal of Molecular Biology*, 193(4):775–791, 1987.

31. S. F. Betz, D. P. Raleigh, and W. F. Degrado. De-novo protein design - from molten globules to native-like states. *Current Opinion in Structural Biology*, 3(4):601–610, 1993.
32. D. B. Gordon, S. A. Marshall, and S. L. Mayo. Energy functions for protein design. *Current Opinion in Structural Biology*, 9(4):509–513, 1999.
33. J. Desmet, M. Demaeyer, B. Hazes, and I. Lasters. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, 356(6369):539–542, 1992.
34. E. P. Baldwin, O. Hajiseyedjavadi, W. A. Baase, and B. W. Matthews. The role of backbone flexibility in the accommodation of variants that repack the core of t4-lysozyme. *Science*, 262(5140):1715–1718, 1993.
35. B. H. M. Mooers, D. Datta, W. A. Baase, E. S. Zollars, S. L. Mayo, and B. W. Matthews. Repacking the core of t4 lysozyme by automated design. *Journal of Molecular Biology*, 332(3):741–756, 2003.
36. J. R. Desjarlais and T. M. Handel. Side-chain and backbone flexibility in protein core design. *Journal of Molecular Biology*, 290(1):305–318, 1999.
37. B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker. Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302(5649):1364–1368, 2003.
38. G. A. Lazar and T. M. Handel. Hydrophobic core packing and protein design. *Current Opinion in Chemical Biology*, 2(6):675–679, 1998.
39. L. L. Looger and H. W. Hellinga. Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: Implications for protein design and structural genomics. *Journal of Molecular Biology*, 307(1):429–445, 2001.
40. T. J. Richmond. Solvent accessible surface-area and excluded volume in proteins - analytical equations for overlapping spheres and implications for the hydrophobic effect. *Journal of Molecular Biology*, 178(1):63–89, 1984.
41. C. E. Kundrot, J. W. Ponder, and F. M. Richards. Algorithms for calculating excluded volume and its derivatives as a function of molecular-conformation and their use in energy minimization. *Journal of Computational Chemistry*, 12(3):402–409, 1991.
42. S. Neukirch, A. Goriely, and A. C. Hausrath. A continuum elastic theory of coiled-coils with applications to the mechanical properties of fibrous proteins and energy transduction by the atp synthase. *Biophysical Journal*, (Submitted), 2006.
43. G. Grass and C. Rensing. Genes involved in copper homeostasis in Escherichia coli. *J. Bacteriology*, 183(6):2145–2147, 2001.
44. I. R. Loftin, S. Franke, S. A. Roberts, A. Weichsel, A. Heroux, W. R. Montfort, C. Rensing, and M. M. McEvoy. A novel copper-binding fold for the periplasmic copper resistance protein cusf. *Biochem.*, 44(31):10533–10540, 2005.
45. V. Koronakis, A. Sharff, E. Koronakis, B. Luisi, and C. Hughes. Crystal structure of the bacterial membrane protein ToOC central to multidrug efflux and protein export. *Nature*, 405(6789):914–919, 2000.
46. J. T. Kittleson, I. R. Loftin, A. C. Hausrath, K. P. Engelhardt, C. Rensing, and M. M. McEvoy. Periplasmic metal-resistance protein CusF exhibits high affinity and specificity for both Cu-I and ag-i. *Biochem.*, 45(37):11096–11102, 2006.
47. T. S. Tsao, H. E. Murrey, C. Hug, D. H. Lee, and H. F. Lodish. Oligomerization state-dependent activation of NF-kappa B signaling pathway by adipocyte complement-related protein of 30 kDa (acrp30). *J. Biological Chem.*, 277(33):29359–29362, 2002.
48. T. S. Tsao, E. Tomas, H. E. Murrey, C. Hug, D. H. Lee, N. B. Ruderman, J. E. Heuser, and H. F. Lodish. Role of disulfide bonds in Acrp30/adiponectin structure

and signaling specificity - Different oligomers activate different signal transduction pathways. *J. Biological Chem.*, 278(50):50810–50817, 2003.

49. K. Kobayashi and T. Inoguchi. Adipokines: Therapeutic targets for metabolic syndrome. *Current Drug Targets*, 6(4):525–529, 2005.
50. G. W. Wong, J. Wang, C. Hug, T. S. Tsao, and H. F. Lodish. A family of Acrp30/adiponectin structural and functional paralogs. *Proc. National Acad. Sciences United States Am.*, 101(28):10302–10307, 2004.