Protein ⟨S⟩ Science

Repeat protein architectures predicted by a continuum representation of fold space

Andrew C. Hausrath and Alain Goriely

P<P	Published online March 7, 2006 in advance of the print journal.	
Email alerting service	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here**	

Notes

To subscribe to _Protein Science_ go to:
http://www.proteinscience.org/subscriptions/

Repeat protein architectures predicted by a continuum representation of fold space

ANDREW C. HAUSRATH[1] AND ALAIN GORIELY[2]

[1]Department of Biochemistry and Molecular Biophysics, and [2]Program in Applied Mathematics and Department of Mathematics, University of Arizona, Tucson, Arizona 85721, USA

Abstract

It is an open question whether nature has utilized all possible protein folds. For a simple protein architecture, the helical repeats, we report a method to address this question based on a mapping between the set of repetitive curves and a space of parameters specifying the curve. The exploration of the parameter space for a particular architecture enables a systematic exploration of the fold space for that protein architecture. In a planar subspace of the parameter space of helical repeats we have identified points corresponding to both naturally occurring folds and potential folds not observed so far.

Keywords: protein fold space; repeat proteins; fold evolution; polyhelix

Proteins with repetitive sequences have repetitive structures. A particularly diverse class of repeat proteins, including the ankyrin, tetratricopeptide, and HEAT repeats, are those with the general architecture (helix$_1$–turn$_1$–helix$_2$–turn$_2$)$_N$ (Groves and Barford 1999). Recent work shows that in some cases molecules whose sequence is the consensus for the repeat motif can fold to highly regular repeat molecules (Mosavi et al. 2002; Binz et al. 2003; Kohl et al. 2003; Main et al. 2003; Stumpp et al. 2003). These synthetic proteins represent the idealized form of the repeat motif. In contrast the naturally occurring examples of these same motifs display variations in structure around this basic form. Fundamental problems in the description of protein structures are the definition of a canonical form of a structural motif and the nature of relationships between these idealized forms.

To investigate these problems, we introduce a representation of the fold of a protein as a continuous space curve that follows the path of the protein backbone in three dimensions. The fold is considered as a geometric object that is distinct from the atomic model of the protein that displays that fold. Differential geometry is capable of describing general curves and so represents a natural language for the exploration of the possible forms that protein folds may take, and the variation about such forms.

Continuous representations have played an important role in the study of nucleic acid structure and dynamics (Marko and Siggia 1994; Manning et al. 1996; Goriely and Tabor 1997; Klapper and Qian 1998). Although there are some notable exceptions (Maritan et al. 2000; Banavar et al. 2002; Trovato et al. 2005), in general the greater conformational diversity of proteins has limited the use of continuous models. However, the rich repertoire of tertiary forms attained by proteins invites a geometric inquiry. Despite the apparent spatial complexity of protein folds, a considerable simplification is possible using a *local* geometrical description.

In general, sufficiently smooth three-dimensional space curves are completely specified up to a rotation and translation by their curvature and torsion (referred to together as curvatures), which are the local properties that describe the twisting and bending of the curve at each point along its length. The local description in terms of curvatures and the global description in terms of spatial coordinates are equivalent: Given any curvature profile, a corresponding space curve can be constructed, and conversely, given any curve, the corresponding curvature profile can be obtained. We have developed methods to construct

curvature profiles for curves that follow the path of protein backbones and to construct atomic coordinate models from such curves (A.C. Hausrath and A. Goriely, in prep.). In this report these methods are used to examine properties of the fold space of helical repeat proteins.

Results

We consider a space curve $r = r(s)$ parameterized by its arc length s. On every point, we define a local coordinate system, the Frenet frame, with orthonormal vectors consisting of the tangent $t(s)$, normal $n(s)$, and binormal $b(s)$ vectors. The changes in the orientation of this frame along the curve are specified by the Frenet equations in terms of its curvatures, that is, two local quantities curvature κ and torsion τ:

$$
\begin{aligned}
r' &= t \\
t' &= \kappa n \\
n' &= -\kappa t + \tau b \\
b' &= -\tau n
\end{aligned}
\tag{1}
$$

Here $(\;)'$ denotes differentiation with respect to s. The curvature and torsion can be extracted from the curve by repeated differentiation, and conversely, the curve can be constructed from the curvature and torsion profiles by integration of the Frenet equations. A convenient way to perform this integration is to introduce a 12-dimensional vector

$$
Y = \{t_1, n_1, b_1, t_2, n_2, b_2, t_3, n_3, b_3, r_1, r_2, r_3\}
\tag{2}
$$

whose entries are the nine components of the three basis vectors in the Frenet frame as well as the three coordinates of a point on the curve.

Then, the Frenet equations can be written as a differential matrix equation

$$
Y' = M(s) \cdot Y \qquad M = \begin{bmatrix} F & 0 & 0 & 0 \\ 0 & F & 0 & 0 \\ 0 & 0 & F & 0 \\ V_1 & V_2 & V_3 & 0 \end{bmatrix},
$$

$$
F = \begin{bmatrix} 0 & \kappa & 0 \\ -\kappa & 0 & \tau \\ 0 & -\tau & 0 \end{bmatrix}
\tag{3}
$$

where V_i is the 3×3 matrix whose single nonvanishing entry is a 1 in row i, column 1.

For arbitrary curvatures, Equation 3 cannot be solved exactly and numerical integration is required. However,

if the curvatures are piecewise constant, Equation 3 is piecewise linear with constant coefficients and an exact analytical solution can be obtained. Such a curvature profile is specified by a list of triples $P = \{(\kappa^{(i)}, \tau^{(i)}; L^{(i)}), i = 1...N\}$, with each triple corresponding to a segment. A curve with constant curvature and torsion is a helix. Therefore, a curve constructed from a piecewise constant curvature profile consists of a series of connected helical arcs and will be referred to as a polyhelix. The following polyhelix construction avoids the need for numerical integration techniques of differential equations normally required for obtaining solutions to the Frenet equations. It is computationally efficient and enables both curve and coordinate model construction utilizing only straightforward linear algebra well known in the structural biology community.

For a single segment with curvature κ and torsion τ starting at $s = 0$ and ending at $s = L$, the solution to Equation 3 is given by

$$
Y(s) = A(\kappa, \tau; s) \cdot Y(0), \qquad 0 \le s \le L,
\tag{4}
$$

where $Y(0)$ defines the initial position and the orientation of the Frenet basis at $s = 0$, and $A(\kappa, \tau; s) = e^{sM}$ is the matrix exponential which can be written

$$
A(\kappa, \tau; s) = \begin{bmatrix} a & 0 & 0 & 0 \\ 0 & a & 0 & 0 \\ 0 & 0 & a & 0 \\ b_1 & b_2 & b_3 & I_3 \end{bmatrix},
$$

$$
a = \begin{bmatrix} \frac{1}{\alpha^2}(\tau^2 + \kappa^2 \cos(\alpha s)) & \frac{\kappa}{\alpha}\sin(\alpha s) & \frac{\kappa\tau}{\alpha^2}(1 - \cos(\alpha s)) \\ -\frac{\kappa}{\alpha}\sin(\alpha s) & \cos(\alpha s) & \frac{\tau}{\alpha}\sin(\alpha s) \\ \frac{\kappa\tau}{\alpha^2}(1 - \cos(\alpha s)) & -\frac{\tau}{\alpha}\sin(\alpha s) & \frac{1}{\alpha^2}(\kappa^2 + \tau^2 \cos(\alpha s)) \end{bmatrix}
\tag{5}
$$

where $\alpha = \sqrt{\kappa^2 + \tau^2}$ and the 3×3 submatrices b_i have the single nonzero row i with entries

$$
\begin{aligned}
(b_i)_{i1} &= \frac{\alpha s \tau^2 + \kappa^2 \sin(\alpha s)}{\alpha^3}, \\
(b_i)_{i2} &= \frac{\kappa}{\alpha^2}(1 - \cos(\alpha s)), \\
(b_i)_{i3} &= \frac{\kappa\tau}{\alpha^3}(\alpha s - \sin(\alpha s)), \quad i = 1, 2, 3
\end{aligned}
\tag{6}
$$

A polyhelix with N segments is completely characterized by the list P and an initial position and basis orientation $Y(0)$. A parametric expression in arc length for the j th segment of the curve $r(s)$ is given by the last three components of the vector $Y^{(j)}(s)$:
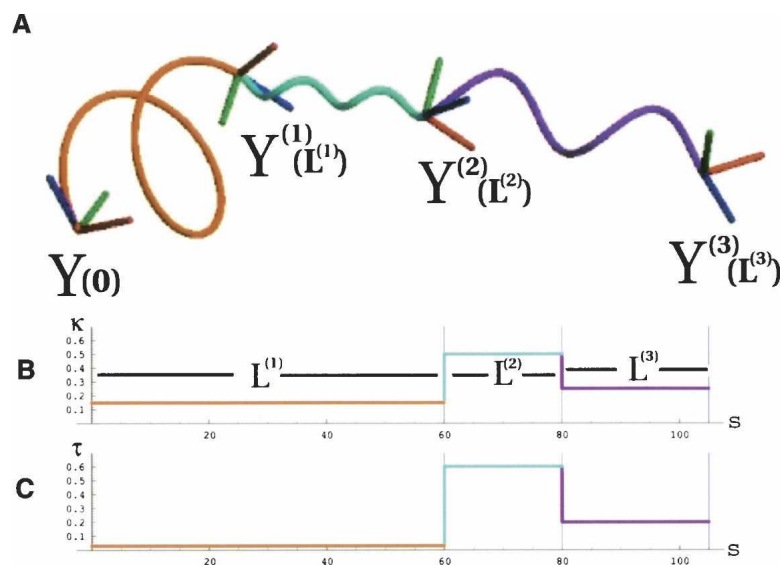
Figure 1. Polyhelical curve. An example three-segment polyhelix (*A*) and its corresponding curvature profile. The three successive segments are colored orange, cyan, and purple. In the Frenet frames at the endpoints of the helical segments, blue, green, and red vectors represent the directions of the tangent, normal, and binormal vectors, respectively. (*B*, *C*) Profiles of curvature $\kappa(s)$ (*B*) and torsion $\tau(s)$ (*C*) vs. arc length s, used to construct the curve in *A*.

$$
Y^{(j)}(s) = A\left(\kappa^{(j)}, \tau^{(j)}; s - s_o^{(j)}\right)
$$
$$
\cdot \left[\prod_{k=j-1}^{1} A\left(\kappa^{(k)}, \tau^{(k)}; L^{(k)}\right)\right] \cdot Y(0),
$$
$$
s_o^{(j)} \le s \le s_o^{(j-1)} \tag{7}
$$

where $s_o^{(j)} = \sum_{k=1}^{j-1} L^{(k)}$. The matrix A propagates both its corresponding helix and the associated Frenet frame from the initial frame to the frame at its endpoint, thereby supplying the initial frame for the subsequent segment. Joining these recursively defined parametric representations for each segment creates a parametric representation of the entire curve.

The nature of the polyhelix construction can be more clearly seen in the explicit expressions for matrix products specifying the first three segments:

$$
Y^{(1)}(s) = A\left(\kappa^{(1)}, \tau^{(1)}; s\right) \cdot Y(0)
$$
$$
Y^{(2)}(s) = A\left(\kappa^{(2)}, \tau^{(2)}; s - L^{(1)}\right) \cdot A\left(\kappa^{(1)}, \tau^{(1)}; L^{(1)}\right) \cdot Y(0)
$$
$$
= A\left(\kappa^{(2)}, \tau^{(2)}; s - L^{(1)}\right) \cdot Y^{(1)}\left(L^{(1)}\right)
$$
$$
Y^{(3)}(s) = A\left(\kappa^{(3)}, \tau^{(3)}; s - \left(L^{(1)} + L^{(2)}\right)\right) \cdot
$$
$$
A\left(\kappa^{(2)}, \tau^{(2)}; L^{(2)}\right) \cdot A\left(\kappa^{(1)}, \tau^{(1)}; L^{(1)}\right) \cdot Y(0)
$$
$$
= A\left(\kappa^{(3)}, \tau^{(3)}; s - \left(L^{(1)} + L^{(2)}\right)\right) \cdot Y^{(2)}\left(L^{(2)}\right) \tag{8}
$$

Note that only the first matrix in the product contains the parametrization in *s*. The rest of the product amounts to a constant vector. This constant vector supplies the initial basis for that segment and is obtained as the endpoint of the previous segment. As an example, a three-segment polyhelix is shown in Figure 1.

The first nine components of $Y^{(j)}(s)$ give the components of the basis vectors for the local Frenet frame at *s*. These coordinate systems are particularly convenient for constructing atomic coordinate models from the curve. Selecting points on the curve separated by the length of a peptide plane gives the positions of the atoms in the model and defines a discrete set of corresponding arc-length values. The Frenet frames at positions $\{s_1, ..., s_N\}$ are the natural coordinate systems in which to express the atoms of the corresponding residue in the atomic model. A set of local coordinates $a = (a_1, a_2, a_3)$ in the Frenet frame at *s* represents the point $p_{ext} = r(s) + a_1 t(s) + a_2 n(s) + a_3 b(s)$ in the external coordinates. Conversely, any external point p_{ext} has local coordinates $a = \{(p_{ext} - r(s)) \cdot t(s), (p_{ext} - r(s)) \cdot n(s), (p_{ext} - r(s)) \cdot b(s)\}$ in the Frenet frame at *s*. The closed-form expression for $Y^{(j)}(s)$ allows the coordinates of the atoms of the model, and therefore derived quantities such as energies, geometric quantities including bond geometries or buried surface areas, and also agreement with experimental data, to be the subject of *analytical* studies. Local coordinates used for construction of backbone atomic models (e.g., see Figure 4, below) are shown in Table 1. The backbone atoms are specified by the curvature profile. The

Table 1. *Local coordinates for backbone atoms*

Atom	a_1	a_2	a_3
N	−0.9854	0.8980	−0.5914
CA	0.0	0.0	0.0
C	1.0894	0.7819	0.7276
O	1.4468	0.4591	1.8605

construction of side chains can be accomplished in a similar manner although information in addition to the curvature profile must be supplied (A.C. Hausrath and A. Goriely, in prep.).

Discussion

In principle, a protein fold containing any type of secondary structure can be described with an appropriate curvature profile. Piecewise constant curvature profiles are especially well suited to the description of α-helical proteins. General helical protein folds can be described with such curvature profiles, although the difficulty of obtaining an accurate representation increases rapidly for more complex structures. In this article we focus on a simple family of such structures—the helical repeat proteins. As the fold of a repeat protein is a repetitive curve, its curvature profiles are periodic. The periodicity greatly simplifies the curvature profile, as only one period need be specified to generate an extended regular structure.

Using the polyhelix construction, a two-helix repeat protein fold can be specified with six $\{\kappa, \tau, L\}$ triples with each turn represented by two such triples. We use two helical arcs to connect successive helices. In general, six parameters are required to specify the relative orientation of two rigid bodies, and two $\{\kappa, \tau, L\}$ triples therefore supply the necessary parameters.

In the α-helical segments, curvature and torsion are fixed. The construction of a two-helix repeat therefore requires 14 parameters. The canonical form for each fold of this type corresponds to a single point in this 14-dimensional space. The space of parameters for a given architecture will be referred to as the curvature space. There is an exact correspondence between points in the curvature space and individual three-dimensional curves obtained using the polyhelix construction. Further, using the local coordinate systems appropriately placed along the curve, backbone atomic models may be constructed using these curvature parameters. The polyhelix construction represents a mathematically well-defined mapping between the Euclidean curvature space and the complicated fold space of helical repeat proteins.

As an example, Figure 2 shows curvature profiles for three different helical repeat proteins: yeast vesicular transport protein sec17 (PDB code 1qqe), human protein phosphatase 2A (1b3u), and bacterial transcription factor MalT (1hz4), which will be subsequently referred to by their PDB codes for brevity (Groves et al. 1999; Rice and Brunger 1999; Steegborn et al. 2001). The parameters defining the curvature profiles are provided in Table 2. Figure 3 shows stereo views of the three repeats of the repetitive curves specified by these profiles and the Cα trace to which they were fitted. The profiles were also used to construct the backbone models represented by ribbon diagrams in Figure 4B.

Protein folds represent a subset of the possible space curves. Our strategy to predict new folds is to search within a curvature space specifying a family of space curves and to identify instances that are compatible with protein geometry. Mathematically, this is accomplished by devising a function to score the potential of a curve to be realized as a protein. While it is unrealistic to expect that a single function could reliably confirm the existence of a protein whose fold conforms to a given curve, it is not difficult to create functions that can eliminate curves that are incompatible with realization as a protein. We refer to these functions as protein quality functions. Given an appropriately constructed protein quality function, a contour plot over a curvature space will have islands in regions that correspond to protein-like curves. These islands may correspond to either: (1) curves that resemble known folds; (2) "false-positives," curves that are not realizable as protein
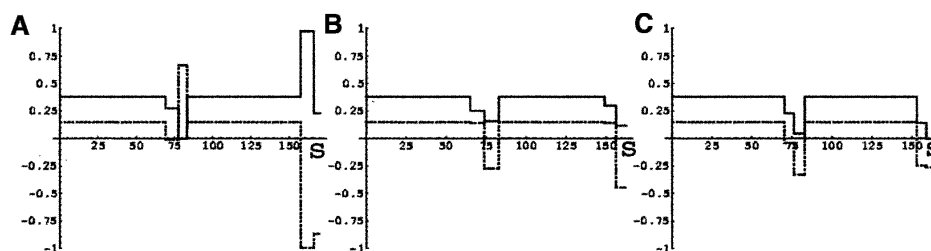
Figure 2. Curvature profiles of repetitive curves representing repeat protein folds. Curvature $\kappa(s)$ (solid line) and torsion $\tau(s)$ (dotted line) profiles for a single repeat of the curves fitted to 1qqe (*A*), 1b3u (*B*), and 1hz4 (*C*). The plateaus in the profiles at 0.38 (curvature) and 0.15 (torsion) specify the α-helical segments.

Table 2. *Curvature parameters for 1qqe, 1b3u, and 1hz4*

| PDB	1qqe			1b3u			1hz4		
segment	κ	τ	L	κ	τ	L	κ	τ	L
1	0.38	0.15	68.5665	0.38	0.15	65.1111	0.38	0.15	70.1281
2	0.2733	−0.0082	8.5389	0.2490	0.1380	8.7975	0.2258	−0.0432	6.0429
3	0.0	0.6629	5.6701	0.1569	−0.2731	8.9004	0.0425	−0.03288	6.3368
4	0.38	0.15	74.6944	0.38	0.15	66.5759	0.38	0.15	70.2449
5	0.9781	−1.000	8.2056	0.3019	0.1381	6.7962	0.1382	−0.2512	6.0209
6	0.2317	−0.8655	4.3622	0.1151	−0.4492	6.4457	0.0	−0.2688	4.0167

folds but that have not been eliminated from the search because of the limitations of the quality function; (3) folds that exist in nature but have not been experimentally observed; or (4) folds that could be realized but have not been utilized in nature. Once such a function has been constructed for a particular architecture, it is possible to seek new folds within this architecture and to examine relationships between the individual instances of folds having this architecture. The computational efficiency of the polyhelix construction allows the exploration of the entire curvature space. Therefore it is possible to examine properties of the continuum of possible forms, such as the density of folds in fold space or the connectedness of fold space (Shindyalov and Bourne 2000; Harrison et al. 2002; Hou et al. 2003).

Remarkably, simple, well-chosen quality functions are adequate to resolve coarse features of the fold space and to identify regions of interest. Energetic calculations on complete models built from the curve-derived scaffolds could then be applied to promising candidates in these regions. Actual proof of existence requires constructing a physical protein molecule with the correct fold and the experimental determination of its three-dimensional structure (Kuhlman et al. 2003). Modern protein design methods are capable of designing sequences compatible with a backbone scaffold model (Dahiyat and Mayo 1997; Dwyer and Hellinga 2004).

Here we use a simple protein quality function expressing a balance between curve compactness and self-avoidance. It is formulated as the ratio of a term that quantifies curve compactness and a term that penalizes a curve that approaches itself too closely. Given a set of points $\{p(s_k)\}$ on a curve, two points $p(s_i)$ and $p(s_j)$ are said to form a contact when they are within a prescribed contact distance d in space. Applied to points on continuous curves, the contact order is the arc-length separation $|s_j - s_i|$ averaged over all contacts (Plaxco et al. 1998). Contact order is large for curves in which many pairs of points distant in arc length are close in space and so serves as a simple quantitative measure of compactness. Explicitly, the contact order is

$$CO(d) = \frac{1}{LN} \sum^{N} |s_j - s_i|, \qquad (9)$$

where N is the total number of contacts, L is the number of points, and the sum runs over all contacts. However, a curve that is too compact will approach too close to
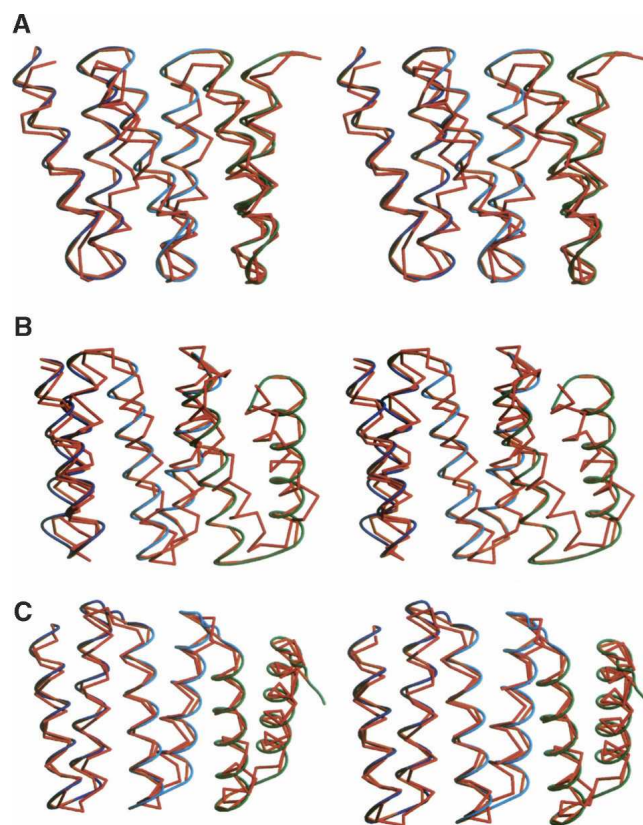
Figure 3. Continuous and discrete representations of example repeat protein folds. In each figure the three successive repeats of the curve are indicated in blue, cyan, and green. The experimentally determined Cα trace is shown in red, and the Cα model obtained from the curve is shown in orange. (*A*) Residues A120–236 of 1b3u (3.3 Å RMSD on Cα positions), (*B*) residues 47–165 of 1hz4. (4.8 Å RMSD), (*C*) residues 33–153 of 1qqe (2.0 Å RMSD).
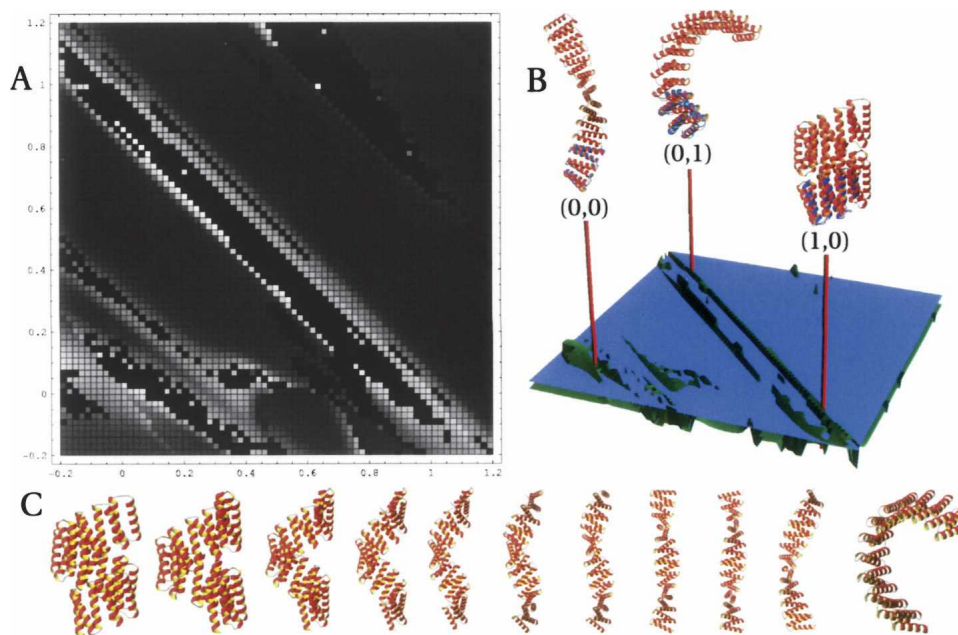
Figure 4. Visualization of curvature space. (*A*) Density representation of the quality function Q (1.5, 14.0) on a plane in the 14-dimensional curvature space containing three idealized natural protein forms. The white areas correspond to higher values, i.e., points that specify more compact, less-clashing curves. Contact order was calculated over two repeats, so that inter-repeat contacts tend to have greater weight than intra-repeat contacts. Clashes were calculated over six repeats to penalize long-range clashes. Contacts and clashes from points spaced five or fewer positions along the curve were excluded from consideration. (*B*) Surface representation (green) of the plot in *A* in which the points (0,0), (1,0), and (0,1) are marked with red ribbon diagrams of the models specified by these points. The superimposed blue ribbon diagrams represent three repeats each of the corresponding natural protein. A bridge linking (1,0) and (0,1) can be seen *above* the plane (blue), which represents the contour level of 2.8 standard deviations above the mean. (*C*) Ribbon diagrams of models constructed from points along the bridge between coordinates (1,0) and (0,1).

itself. Defining a clash as a pair of points that are closer than a prescribed clash distance c in space, curves with more than a very few close self-approaches are severely penalized by using the quality function

$$Q(c,d) = \frac{CO(d)}{2^{M(c)}}, \qquad (10)$$

where $CO(d)$ is the contact order of the curve and $M(c)$ the number of clashes. With this function, curves with no such self-approaches are not penalized and so are ranked only by their relative contact order. Other functions could certainly be devised, but as any effective functions should be in agreement about the "poor" regions of the curvature space, a very simple scheme such as this one is suitable for an initial survey.

The quality function Q was used to investigate the curves parameterized by points in a subspace of the larger 14-dimensional curvature space. The three points corresponding to the profiles in Figure 2 define a plane and a coordinate system on that plane. Defining $v_{0,1,2}$ as the vectors of curvature parameters for 1qqe, 1hz4, and 1b3u, any point on this plane can be represented by an ordered pair (a, b) corresponding to the position in the curvature space of $v_0 + a(v_1 - v_0) + b(v_2 - v_0)$, so that the parameter vector for 1qqe is at $(0,0)$, 1hz4 is at $(1,0)$, and 1b3u is at $(0,1)$. We have plotted the value of the function Q for a and b in the ranges –0.2 to 1.2 and displayed the results in Figure 4A.

In this plane, representing a small subset of the curvature space, a variety of distinct forms can be found, including some whose "curve quality" compares favorably with the fitted curves we have used to represent natural proteins and yet not corresponding to any known protein structure. (Ribbon diagrams of some examples are shown in Fig. 4C.) This result suggests that there are quite a large number of curves consistent with protein geometry, which could be found and constructed by a more systematic search of curvature spaces, either of the two-helix repeats or other protein architectures.

With the rapid increase in structural knowledge has come the realization that nature has made use of a limited set of protein folds (Chothia 1992; Zhang and DeLisi 1998). It is not clear to what extent the set of

folds used in nature represents the structural repertoire attainable by a polypeptide chain. The number of possible amino acid sequences clearly means that nature has only sampled a tiny fraction of possible polypeptides, but a given fold may be compatible with a large number of sequences. The numbers of sequences associated with stable folds has been investigated using lattice models (Li et al. 1996). This study suggested that protein folds may be those tertiary forms with energetically unique states that also have unusually large numbers of compatible sequences. Lattice models are currently the only types of models for which it is possible to enumerate exhaustively the structures associated with all possible sequences, and, despite their simplicity, appear to capture the essential characteristics of the folding problem. But lattice models are not intended to represent particular natural proteins, so they are not able to be predictive of natural protein folds.

Our method differs in that a continuum representation is used, and it can represent particular natural protein folds. It is not capable of considering all possible types of folds at once, but the method can be comprehensive within a given architecture. In contrast to the discrete lattice models, smooth and continuous deformations of the curve representation can be used to model both subtle and large-scale changes in such protein folds. Consideration of fitness or energetic properties could be superimposed on top of the representation, and the "quality function" Q is a first step in this direction. The key difference is that folds of particular natural proteins can be parameterized, and so in principle the method can be predictive. But the method does not address whether a sequence might exist that could confer such a fold.

It is interesting to consider the relationship between curvature space and sequence space. The evolutionary history of a protein sequence can often be reconstructed from phylogenies. Doing so creates the path through sequence space that the protein has followed during its evolution. The question arises as to whether it is possible to establish a correspondence between such paths in sequence space and paths in curvature space. John Maynard Smith (1970) articulated a model of protein evolution as a series of sitewise changes in sequence that ultimately led from one sequence to a completely different sequence, but all the while retaining the ability to fold and carry out a cellular function. (Larger scale modifications of sequence may happen that result in discontinuous changes in structure [Cui et al. 2002].) It is not known if the evolution of new folds can be accomplished by the stepwise process envisioned by Smith: Can the tertiary structure of proteins be changed by successive pointwise changes from one fold to another?

The helical repeat proteins provide an example where incremental tertiary structure changes *are* observed. The variation in sequence among individual repeats creates small differences in the relative orientations of successive repeat units in the structure. The overall superhelical character of the array of repeats is a consequence of such fine adjustments, especially when amplified by repetition. A dramatic example is provided by the two HEAT proteins importin-β and protein phosphatase 2A PR65/A subunit, which form right- and left-handed superhelical arrays, respectively; yet, the two proteins are derived from a common ancestral sequence (Cingolani et al. 1999; Groves et al. 1999; Andrade et al. 2001). Examination of the spatial relationships between the individual repeats of these and other HEAT repeat structures shows that there is considerable diversity. The large collection of these structures with small differences between them suggests that the helical repeats constitute a densely sampled continuum of tertiary forms.

A mathematically explicit example of a continuous change between natural forms within this continuum can be viewed in Figure 4B. The diagonal bridge in the quality function between coordinates (0,1) and (1,0) suggests that realizable protein forms parameterized by the points along this path through the curvature space may exist. More abstractly, two forms that are connected by a high-scoring path might be related in the sense that if an important cellular function originally resided on a polypeptide with a fold somewhere on this path, the process of evolution might conceivably allow continuous change in the tertiary structure of this protein to both points while retaining a specific folded form and thereby also retaining any capability dependent on that structure. Points that are not connected could not be related in this manner. For example, in Figure 4B, no path exists between (0,0) and (1,0), suggesting that these structures do not share a common precursor located in the portion of the curvature space sampled so far. Parametrization by curvatures is a means to investigate Smith's abstract protein space as an explicit mathematical object, and quality functions might be thought of as the fitness of forms inhabiting this landscape (Macken and Perelson 1989). Investigation of the connectivity of the level sets of quality functions on curvature spaces allows for insight into whether folds could be evolutionarily related or disjoint.

Materials and methods

Backbone coordinates (Table 1) were obtained by creating an idealized polyglycine α-helix using EDPDB (Zhang and Matthews 1995). A helix with curvature 0.3812 and torsion 0.1492 was created and the set of points with spatial separation of 3.8 Å, one peptide plane, along this curve were obtained. The polyglycine model was superimposed on the curve by overlaying its Cα

positions on the set of corresponding points obtained from the curve. The Frenet frames at each Cα position were then used to express the coordinates of the remaining backbone atoms. A helix of 100 residues was used to minimize any error introduced by the superposition procedure or conversion process, and the local coordinates reported in Table 1 are the average of these 100 instances.

For curve construction, curvature profiles were devised by a combination of manual adjustment and least-squares minimization of the sum of pairwise distances between point sets obtained from curves with the spatial separation of 3.8 Å and Cα traces of coordinate models. Initial orientation vectors $Y(0)$ were obtained by superposition of the Cα trace of the first α-helix from a coordinate set and a corresponding set of points from an α-helical curve. This transformation was then applied to the initial Frenet frame from the α-helical curve to obtain the initial basis for curve construction.

Each curvature profile was fitted to a three-repeat section from its corresponding coordinate set as follows. For each helix–turn–helix motif from the selected section of the coordinate model, an initial basis was created as above and a four-segment polyhelix, comprised of two α-helical curves (with curvature 0.38 and torsion 0.15) connected by two general helical arcs, was fitted to its Cα coordinates. An average of the values for the curvatures in the turn regions obtained from this procedure supplied the starting values for a periodic curvature profile. This periodic curvature profile was fitted first against the coordinates of the first two repeats in the selected section and then against the coordinates of the three repeats.

Calculations were carried out using Mathematica (Mathematica 5.2, Wolfram Research Inc.), Maple (Maple 10, Maplesoft), or with custom C programs. Figures were created using MOLSCRIPT (Kraulis 1991), Raster3D (Merritt and Bacon 1997), MOLMOL (Koradi et al. 1996), and Mathematica (Wolfram Research Inc.).

Acknowledgments

References

Andrade, M.A., Petosa, C., O'Donoghue, S.I., Muller, C.W., and Bork, P. 2001. Comparison of ARM and HEAT protein repeats. *J. Mol. Biol.* **309:** 1–18.

Banavar, J.R., Maritan, A., Micheletti, C., and Trovato, A. 2002. Geometry and physics of proteins. *Proteins* **47:** 315–322.

Binz, H.K., Stumpp, M.T., Forrer, P., Amstutz, P., and Pluckthun, A. 2003. Designing repeat proteins: Well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins. *J. Mol. Biol.* **332:** 489–503.

Chothia, C. 1992. Proteins—1000 families for the molecular biologist. *Nature* **357:** 543–544.

Cingolani, G., Petosa, C., Weis, K., and Muller, C.W. 1999. Structure of importin-β-bound to the IBB domain of importin-α. *Nature* **399:** 221–229.

Cui, Y., Wong, W.H., Bornberg-Bauer, E., and Chan, H.S. 2002. Recombinatoric exploration of novel folded structures: A heteropolymer-based model of protein evolutionary landscapes. *Proc. Natl. Acad. Sci.* **99:** 809–814.

Dahiyat, B.I. and Mayo, S.L. 1997. De novo protein design: Fully automated sequence selection. *Science* **278:** 82–87.

Dwyer, M.A. and Hellinga, H.W. 2004. Periplasmic binding proteins: A versatile superfamily for protein engineering. *Curr. Opin. Struct. Biol.* **14:** 495–504.

Goriely, A. and Tabor, M. 1997. Nonlinear dynamics of filaments. 1. Dynamical instabilities. *Physica D.* **105:** 20–44.

Groves, M.R. and Barford, D. 1999. Topological characteristics of helical repeat proteins. *Curr. Opin. Struct. Biol.* **9:** 383–389.

Groves, M.R., Hanlon, N., Turowski, P., Hemmings, B.A., and Barford, D. 1999. The structure of the protein phosphatase 2A PR65/A subunit reveals the conformation of its 15 tandemly repeated HEAT motifs. *Cell* **96:** 99–110.

Harrison, A., Pearl, F., Mott, R., Thornton, J., and Orengo, C. 2002. Quantifying the similarities within fold space. *J. Mol. Biol.* **323:** 909–926.

Hou, J.T., Sims, G.E., Zhang, C., and Kim, S.H. 2003. A global representation of the protein fold space. *Proc. Natl. Acad. Sci.* **100:** 2386–2390.

Klapper, I. and Qian, H. 1998. Remarks on discrete and continuous large-scale models of DNA dynamics. *Biophys. J.* **74:** 2504–2514.

Kohl, A., Binz, H.K., Forrer, P., Stumpp, M.T., Pluckthun, A., and Grutter, M.G. 2003. Designed to be stable: Crystal structure of a consensus ankyrin repeat protein. *Proc. Natl. Acad. Sci.* **100:** 1700–1705.

Koradi, R., Billeter, M., and Wuthrich, K. 1996. MOLMOL: A program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14:** 51–55.

Kraulis, P.J. 1991. MOLSCRIPT—A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24:** 946–950.

Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L., and Baker, D. 2003. Design of a novel globular protein fold with atomic-level accuracy. *Science* **302:** 1364–1368.

Li, H., Helling, R., Tang, C., and Wingreen, N. 1996. Emergence of preferred structures in a simple model of protein folding. *Science* **273:** 666–669.

Macken, C.A. and Perelson, A.S. 1989. Protein evolution on rugged landscapes. *Proc. Natl. Acad. Sci.* **86:** 6191–6195.

Main, E.R.G., Xiong, Y., Cocco, M.J., D'Andrea, L., and Regan, L. 2003. Design of stable α-helical arrays from an idealized TPR motif. *Structure* **11:** 497–508.

Manning, R.S., Maddocks, J.H., and Kahn, J.D. 1996. A continuum rod model of sequence-dependent DNA structure. *J. Chem. Phys.* **105:** 5626–5646.

Maritan, A., Micheletti, C., Trovato, A., and Banavar, J.R. 2000. Optimal shapes of compact strings. *Nature* **406:** 287–290.

Marko, J.F. and Siggia, E.D. 1994. Bending and twisting elasticity of DNA. *Macromolecules* **27:** 981–988.

Merritt, E.A. and Bacon, D.J. 1997. Raster3D: Photorealistic molecular graphics. *Methods Enzymol.* **277:** 505–524.

Mosavi, L.K., Minor, D.L., and Peng, Z.Y. 2002. Consensus-derived structural determinants of the ankyrin repeat motif. *Proc. Natl. Acad. Sci.* **99:** 16029–16034.

Plaxco, K.W., Simons, K.T., and Baker, D. 1998. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277:** 985–994.

Rice, L.M. and Brunger, A.T. 1999. Crystal structure of the vesicular transport protein Sec17: Implications for SNAP function in SNARE complex disassembly. *Mol. Cell* **4:** 85–95.

Shindyalov, I.N. and Bourne, P.E. 2000. An alternative view of protein fold space. *Proteins* **38:** 247–260.

Smith, J.M. 1970. Natural selection and concept of a protein space. *Nature* **225:** 563–564.

Steegborn, C., Danot, O., Huber, R., and Clausen, T. 2001. Crystal structure of transcription factor MaiT domain III: A novel helix repeat fold implicated in regulated oligomerization. *Structure* **9:** 1051–1060.

Stumpp, M.T., Forrer, P., Binz, H.K., and Pluckthun, A. 2003. Designing repeat proteins: Modular leucine-rich repeat protein libraries based on the mammalian ribonuclease inhibitor family. *J. Mol. Biol.* **332:** 471–487.

Trovato, A., Hoang, T.X., Banavar, J.R., Maritan, A., and Seno, F. 2005. What determines the structures of native folds of proteins? *J. Phys. Condens. Matter* **17:** S1515–S1522.

Zhang, C.O. and DeLisi, C. 1998. Estimating the number of protein folds. *J. Mol. Biol.* **284:** 1301–1305.

Zhang, X.J. and Matthews, B.W. 1995. EDPDB—A multifunctional tool for protein-structure analysis. *J. Appl. Crystallogr.* **28:** 624–630.