J. Mol. Biol. (1983) **168**, 143–162

Differential Geometry of Proteins

Helical Approximations

A. H. Louie and R. L. Somorjai†

Division of Chemistry
National Research Council
Ottawa, Ontario, Canada K1A 0R6

(Received 16 September 1982, and in revised form 24 March 1983)

We regard a protein molecule as a geometric object, and in a first approximation represent it as a regular parametrized space curve passing through its α-carbon atoms (the backbone). In an earlier paper we argued that the regular patterns of secondary structures of proteins (morphons) correspond to geodesics on minimal surfaces. In this paper we discuss methods of recognizing these morphons on space curves that represent the protein backbone conformation. The mathematical tool we employ is the differential geometry of curves and surfaces.

We introduce a natural approximation of backbone space curves in terms of helical approximating elements and present a computer algorithm to implement the approximation. Simple recognition criteria are given for the various morphons of proteins. These are incorporated into our helical approximation algorithm, together with more non-local criteria for the recognition of β-sheet topologies. The method and the algorithm are illustrated with several examples of representative proteins. Generalizations of the helical approximation method are considered and their possible implications for protein energetics are sketched.

1. Introduction

Representations of protein structures

The complete three-dimensional structure of a protein is generally represented and documented as a list of atom identities and co-ordinates. Such lists are compiled at the Protein Data Bank, Brookhaven.

The complete structure of a protein, however, contains information on thousands of atoms and so remains incomprehensible unless a more intelligible representation is constructed. This involves a selective reduction of complexity. Such representations should ideally display only those morphological aspects that are of relevance to the particular question under study. Different requirements of structural details thus give rise to representations ranging from three-dimensional

† Author to whom all correspondence should be addressed.

0022-2836/83/210143-20 $03.00/0

space-filling, ball-and-stick, and wire models, to structural cartoons (e.g. see Dickerson & Geis, 1969; Richardson, 1981) showing only the conformation of the α-carbon backbones.

In addition to these "physical" models of protein structures, there are more abstract representations. The latter include plots of all dihedral backbone angles ϕ and ψ at each α-carbon atom (Balasubramanian, 1977), and a variety of α-carbon distance plots: all C_α–C_α distances (e.g. see Kuntz *et al.*, 1979; Dunn & Klotz, 1975; Sippl, 1982), C_α^i to C_α^{i+3} distances (Rose & Seltzer, 1977) and C_α^i to C_α^{i+1}, C_α^{i+2}, C_α^{i+3}, C_α^{i+4} distances (Goel & Ycas, 1979).

We regard a protein molecule as a geometric object, and in a first approximation represent it as a regular parametrized space curve passing through its α-carbon atoms. The requirement that the space curve be regular (continuously differentiable with a non-vanishing derivative) is not overly idealized and restrictive, because the 2 to 3 Å effective resolution of X-ray crystallography provides enough error margin for an analytically representable curve to be constructed. One may impose the smoothness condition that the arc-length between successive α-carbon atoms be within a prescribed upper limit and 3·8 Å (the established "average" C_α^i to C_α^{i+1} distance) so that the curve behaves in a physically meaningful way (e.g. does not have "loops") between consecutive points. The curvature and torsion of such space curves representing proteins are used to characterize structural patterns (morphons) of the backbones. (The reader is referred to any standard text on differential geometry, e.g. see Carmo (1976) for a review of the subject.) Some differential geometry-inspired concepts have been presented in the literature. Thus, Rose & Seltzer (1977) estimate radii of curvature along the backbone in their chain turn identifying algorithm. In a series of papers, Rackovsky & Scheraga (1978,1980,1981) represent the backbone of a protein molecule as a polygonal arc through the α-carbon atoms. Using four adjacent C_α atoms, they obtain a discrete representation of the backbone and can compute local versions of the curvature and torsion. Thus, they can operate on the well-defined length scale covering four adjacent C_α atoms. However, this "difference geometric" approach has its inherent limitations and non-physical peculiarities (e.g. negative curvature for a general space curve) and does not seem to be readily generalizable for different length scales.

We feel that the representation of a protein molecule by a regular space curve provides a natural, mathematically well-defined yet simple model that is readily generalizable. It provides us with a program for studying protein structure and function. In our earlier paper (Louie & Somorjai, 1982), we considered proteins in differential geometric terms, but on a higher hierarchical level: as curves lying on surfaces embedded in three-dimensional space $\mathbb{R}^3$. In fact, the protein curves could be well-represented as geodesics on minimal surfaces. This dual description of proteins as curves and surfaces provides a distinctive framework in which to discuss the structural and dynamical representations of protein patterns.

This paper addresses the problem of recognition of protein morphological patterns or morphons†. Given the representation of a protein molecule as a space

† From μορφή (shape, form) and the common scientific suffix -on to indicate a well-defined, stable identity (compare exciton, polaron, soliton).

curve, what is the most natural way of analyzing it to obtain specific structural information? The answer turns out to be: in terms of helical approximations; i.e. approximation of different sections of the space curve by helices of different radii and pitches. This method is most reasonable on physical grounds because, typically, large portions of a protein backbone are already helices, and it is most logical on mathematical grounds because helices correspond to the simplest class of space curves: those with constant non-vanishing curvature and torsion.

2. Theory and Methods

(a) *Helical approximations*

Let $\kappa(s) \geqslant 0$ and $\tau(s)$ be real-valued continuous functions of s on a real interval $I=[a, b]$. Then the fundamental theorem for space curves states that, except for position in space, there exists a unique space curve for which $\kappa(s)$ is the curvature, $\tau(s)$ is the torsion, and s is the arc-length parameter. Conversely, a curve (*modulo* a Euclidean transformation) is defined uniquely by its curvature and torsion. Thus, there is a bijection (one-to-one correspondence) between the class of all space curves and the class $C(I, \mathbb{R}^2)$ of continuous mappings from I to $\mathbb{R}^2$; i.e. of pairs of continuous functions on real intervals. In particular, the protein conformations form a subset of $C(I, \mathbb{R}^2)$.

The problem of recognizing morphons in proteins becomes this: how can one extract the information presented in a pair of continuous functions (κ, τ), and describe the associated space curve in morphological terms? In this section we give the mathematical background of an approach to this problem. A more general, abstract treatment of the theory has been described by Brown & Page (1970).

A mapping f of $I=[a, b]$ into $\mathbb{R}^2$ is called a stepped-mapping if there exists a finite sequence of real numbers $a=a_0<a_1<\ldots<a_n=b$, such that the restrictions of f to each of the open intervals (a_{i-1}, a_i), $i=1, 2, \ldots, n$, are constant. The ordered set $\{a_0, a_1 \ldots, a_n\}$ is an f-partition. Note that a stepped-mapping f has many different f-partitions. In particular, any finite subset of I that contains an f-partition is an f-partition. The set of all stepped-mappings from I to $\mathbb{R}^2$ is denoted by $S(I, \mathbb{R}^2)$.

The mathematical result that is of interest to us is that any given mapping $g=(\kappa, \tau)$ in $C(I, \mathbb{R}^2)$ can be arbitrarily closely approximated by mapping in $S(I, \mathbb{R}^2)$: for each $\varepsilon>0$, there exists $f \in S(I, \mathbb{R}^2)$, such that the maximal difference between f and g is less than ε, i.e.:

$$\sup\{\|f(s)-g(s)\|: a \leqslant s \leqslant b\}<\varepsilon, \tag{1}$$

where $\|\cdot\|$ is the standard norm in $\mathbb{R}^2$, $\|(x, y)\|=(x^2+y^2)^{\frac{1}{2}}$. (The mathematical statement is that the uniform closure of $S(I, \mathbb{R}^2)$ contains $C(I, \mathbb{R}^2)$. Subsets with this property are called approximating basis sets.)

The relevant result of the analysis is that a pair of continuous functions (κ, τ), representing the curvature and torsion of a protein backbone, can be approximated by functions from I to $\mathbb{R}^2$ that are sectionally constant. (An example, which we discuss in Results and Discussion, section (d), is shown in Fig. 6.)

Now what is the shape of a space curve that corresponds to a stepped-mapping f in $S(I, \mathbb{R}^2)$? Between any 2 consecutive points a_{i-1} and a_i in an f-partition, f is constant; i.e. the corresponding space curve on $[a_{i-1}, a_i]$ has constant curvature and torsion. But a space curve has constant curvature and torsion if and only if it is a (circular) helix (or a degenerate helix). Indeed, if the constant curvature is κ_i and constant torsion τ_i, then for $\kappa_i \geqslant 0$ and $\tau_i \neq 0$, the space curve is a helix $\bar{\mathbf{X}}_i(s)$ with radius:

$$r_i=\frac{\kappa_i}{\kappa_i^2+\tau_i^2} \tag{2}$$

and pitch $2\pi p_i$, where:

$$p_i = \frac{\tau_i}{\kappa_i^2 + \tau_i^2}. \tag{3}$$

The helix is right-handed if $\tau_i > 0$ and left-handed if $\tau_i < 0$. If $\kappa_i \neq 0$ but $\tau_i = 0$, then the space curve is a (planar) circle with radius $1/\kappa_i$; if $\kappa_i = 0$ (implying $\tau_i = 0$), then the helix degenerates into a straight line.

Thus a stepped-mapping in $S(I, \mathbb{R}^2)$ gives a space curve that is a stepped-helix: on each subinterval $[a_{i-1}, a_i]$, the curve is a helix $\vec{\mathbf{X}}_i(s)$ with radius r_i and pitch $2\pi p_i$. Since the helices are determined only up to a Euclidean transformation (a rigid motion), they can be appropriately "patched" together so that:

$$\vec{\mathbf{X}}_{i-1}(a_i) = \vec{\mathbf{X}}_i(a_i) \tag{4}$$

(i.e. end points meet), and that the Frenet trihedron $(\vec{\mathbf{t}}, \vec{\mathbf{n}}, \vec{\mathbf{b}})$ moves smoothly along the entire curve (i.e. at the end points a_i, the trihedra from consecutive helices match). See Fig. 1.

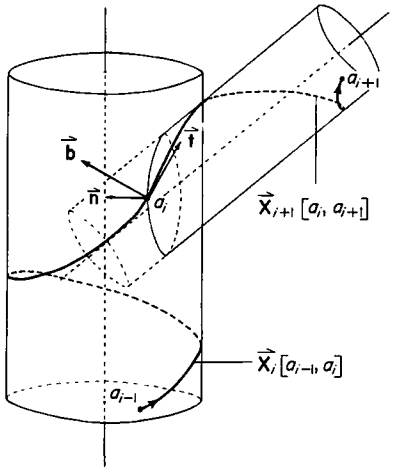

FIG. 1. Joining helical steps in a stepped-helix.

Thus we see that any given space curve, in particular any protein backbone, can be arbitrarily closely approximated by stepped-helices. It is through this helical approximation of proteins that the relevant structural patterns are recognized.

(b) *Line-groups*

The backbone of a protein forms what a crystallographer calls a line-group (a 1-dimensional repetition of a unit motif) if the regularity is such that all the dihedral angles (ϕ, ψ) are the same; i.e. if successive units have identical relative orientations. Every line-group in proteins is a helix, and this is another reason why helical approximations are natural mathematical tools in the analysis of protein morphons.

Helices are conveniently described by the radius r (of the cylinders on which the helices lie) and the pitch $2\pi p$, the canonical equation (with the z-axis as the axis of the helix) being:

$$\vec{\mathbf{X}}(s) = (r \cos s, r \sin s, ps). \tag{5}$$

Note that here the arc-length is actually $(r^2+p^2)^{\frac{1}{2}}\, s$; but for simplicity we shall ignore this scale factor. The inverse to equations (2) and (3) is:

$$\kappa = \frac{r}{r^2+p^2}, \tag{6}$$

$$\tau = \frac{p}{r^2+p^2}. \tag{7}$$

An additional parameter (residues per turn, rise per residue, or phase separation per residue) may be used to describe the regularity of the α-carbon distributions along the helix. The numerical values of these parameters are used to identify different portions of the protein backbone, in an algorithm we present in section (c), below.

The α-helix is the most abundant secondary structure in proteins. This line-group is a right-handed helix with an average radius $r=2{\cdot}3$ Å and pitch $2\pi p=5{\cdot}4$ Å ($p=0{\cdot}86$). The α-carbons are arranged in a helical coil having 3·6 residues per turn, hence the rise per residue is 1·5 Å and the phase separation per residue is 100° (=1·75 rad). Thus an α-carbon atom appears on the curve (5) for every s-interval of length 1·75. This helix has constant curvature $\kappa=0{\cdot}38$ and torsion $\tau=0{\cdot}14$.

A single strand of a (parallel or antiparallel) β-twisted sheet, to a good approximation, forms a line-group that is an extended left-handed helix. The average radius is about 1 Å and the average pitch is about −7 to −8 Å, although these values vary considerably among (and within) strands.

(c) *Approximation algorithm*

The α-carbon co-ordinates of many proteins are deposited at the Protein Data Bank, Brookhaven. Our computer algorithm transforms such discrete sets of information on protein backbone conformations *via* helical approximations so that various morphons can be recognized.

The theory of approximation of space curves by stepped-helices is slightly modified for fitting a discrete set of data as follows. The fitting is done sequentially and in an overlapping fashion such that at the ith step, the co-ordinates of α-carbon atoms i to $i+M-1$ are fitted onto the best approximating helix, where M is the "length" of each helical segment. We find that for an initial run, M is best taken to be 5, although 4 is the minimum for the number of parameters we wish to estimate. M can be increased appropriately after the different portions of the backbone are identified (see Results and Discussion, section (d)). In fact, M is the length scale on which one "filters" the information and is arbitrarily adjustable, $M \geqslant 4$.

Each step of fitting M adjacent α-carbon co-ordinates is divided into 2 consecutive stages. The first is a finite difference Levenberg–Marquardt minimization algorithm (from the IMSL library of Fortran subroutines; see Brown & Dennis, 1972) used to find the best fitting circular cylinder through these M points. From this stage we obtain the radius r_i of the cylinder (hence of the helix that lies on this cylinder), as well as the direction $\bar{\mathbf{V}}_i$ of the axis of the cylinder. This fitting is equivalent to the usual method of the crystallographer (e.g. see Rajan & Srinivasan, 1977), in which the α-carbon co-ordinates are projected onto a plane perpendicular to the axis and then fitted onto a circle. The second stage uses a linear regression algorithm, which gives the pitch p_i of the best fitting helix on the cylinder through the M points, and the phase separation Δs_i of consecutive points.

The algorithm is summarized succinctly in Fig. 2 (the Fortran program is available from us upon request). Note that our helical approximation algorithm has the distinct advantage that all portions of the protein backbone are fitted, not just the regular secondary structures. The regular structures merely correspond to sections where the running segment length M can be increased (i.e. longer helices) while the "random" portions are where the "helices" may contain less than a complete turn, but are helices nevertheless. With $M=5$, the average goodness-of-fit per residue (C_α: eqn (12)) for the 13

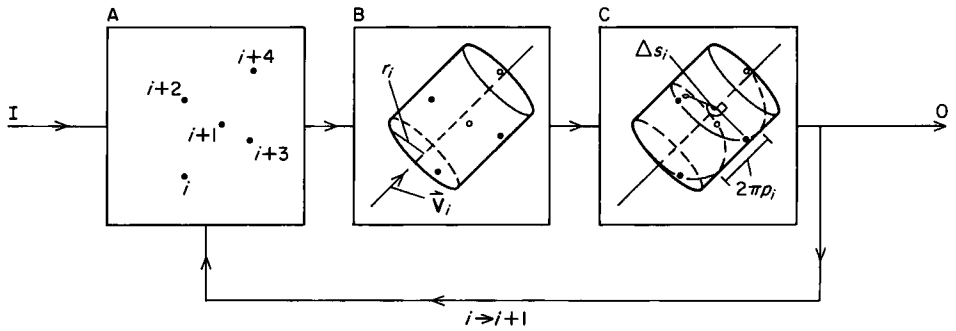

FIG. 2. The helical approximation alogorithm. I, begin for $i=1$. A, $M(=5)$ points (α-carbon atoms i to $i+M-1$) to be fitted. B, fit cylinder; $\hat{V}_i$, axial direction; r_i, radius. C, fit pitch; $2\pi p_i$, pitch; Δs_i, phase separation. O, end if $i+M-1=$number of residues of protein.

proteins we fitted is 0·55 Å, with a standard deviation of 0·16 Å. The best fit is for the haemoglobin α-subunit (0·228) and the worst is for the southern bean mosaic virus protein (0·711: see Table 1). These are exceptionally good fits, lending support to the validity of the helical approximation.

(d) *Recognition criteria*

The best-fitting helix through the M points is recognized as part of an α-helix if the parameters obtained from the algorithm do not differ from the expected values by more than about 10%. That is, if:

$$|r_i - 2{\cdot}3| < 0{\cdot}23, \tag{8}$$

$$|p_i - 0{\cdot}86| < 0{\cdot}086, \tag{9}$$

and

$$|\Delta s_i - 100°| < 10°, \tag{10}$$

then the M points of α-carbon atoms i to $i+M-1$ are considered as part of an α-helix. These and the other criteria that follow can be relaxed or tightened depending on how good the X-ray data are and on how "pure" a helix we require; hence, in particular, the algorithm may provide a valuable tool for analyzing preliminary X-ray data of proteins.

Because of the great variability of pitches of β-strands, the pitch parameter from the fitting turns out to be useless for recognition. We find that the single parameter, the radius r_i, is sufficient to identify strands, with the criterion:

$$r_i < 1{\cdot}5 \tag{11}$$

(identifying strands as "very thin helices"). This criterion, however, cannot distinguish isolated strands from those which are genuinely part of a β-sheet. This differentiation is to be carried out after the complete fitting of the protein, when all the "strands" are identified. A pairwise (both parallel and antiparallel) matching of strands is performed and the average distance between matched α-carbon atoms from the 2 strands is calculated. The calculation of these inter-strand distances is done automatically by the algorithm, once the number of strands and their C_α ranges are provided. When the distance between 2 strands is less than about 8 Å and the variance is small, the pair of strands is considered as neighbouring and as part of a β-sheet. This β-structure analysis also determines the β-sheet topologies (i.e. the orderings and orientations of β-strands) with the information obtained on strand distances and parallel/antiparallel pairings. See Richardson (1977) for a study of β-sheet topologies; we give examples of our β-structure analysis in the following sections.

Bends on the protein backbone are recognized as changes in axial angles of the fitting helices. The angles between consecutive helical axes, $\vec{V}_i$ and $\vec{V}_{i+1}$ can be used to indicate the curving of the backbone in general, and the "straightness" of strands and helices in particular.

The protein backbone also forms (reverse) turns with or without hydrogen bonds. These are recognized when 3 or more consecutive changes of axial angles are each greater than 40° (and add up to about 180°). Rose & Seltzer (1977) present an alternate approach to peptide chain turns. They consider the protein backbone as an ensemble of segments (LINCs in their terminology) and hinges, and the hinges (i.e. turns) are the sites where their radius of curvature is small.

It should be emphasized that none of our recognition criteria is based on the presence or absence of stabilizing hydrogen bonds, since only C_α information is used. This is a feasible explanation of why certain morphons we identify are not recorded in the official version stored in the Protein Data Bank. Furthermore, all segments of the complete backbone curve are represented on an equal footing: no distinction between regular and irregular regions is made.

The recognition algorithm contains another option that can compare the relative spatial arrangements of all α-helices found. This is carried out by computing the distances and angles between the helix axes $\vec{V}_i$, in a pairwise matching just as is done for the β-sheet topologies.

Furthermore, the input data sets of atomic co-ordinates are not restricted to α-carbon atoms. One can use the entire N–C_α–C backbone, the β-carbon atoms, the nitrogen atoms, the carboxyl carbon atoms, and so on. These other input data sets provide additional, more detailed information on the structures of proteins.

3. Results and Discussion

We have analyzed 13 representative proteins from the Data Bank, using our approximating algorithm and the recognition criteria given. The regular regions (both α-helices, and β-strands in the order they occupy the sheets) are listed in Table 1 together with the accepted ranges, as listed in the Protein Data Bank. The goodness of overall fit *GOF* (average metrical deviation/residue) is given by:

$$GOF = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{M} \sum_{j=i}^{i+M-1} [e_j^2(\text{cylinder}) + e_j^2(\text{pitch})]^{\frac{1}{2}}, \tag{12}$$

where the e_j terms are the errors (deviations) from the two sub-stages of helical fittings, M is the running segment length, and N is the total number of C_α residues of the particular protein. Note that e_j(cylinder) measures how well M points fit onto the best cylindrical surface, while e_j(pitch) determines how regularly one can distribute these points along the best helical segment on this cylinder. We discuss three examples in detail to show how an analysis is carried out, and to point out interesting features.

(a) *An example: ferredoxin*

The ferredoxins are small iron-sulphur proteins that function as electron transport agents. The structure of the bacterial ferredoxin from *Peptococcus aerogenes* has been determined at 2 Å resolution (Adman *et al.*, 1976), and we use this protein containing 54 amino acid residues as our first example.

TABLE 1

Comparison of sites of regular secondary structures of proteins

Ferredoxin 54

	L&S	Adman *et al.* (1976)
α1	14–18	
α2	39–44	
β2	22–26	Secondary
β3	33–27	structures
β1	1–6	not identified
β4	53–49	

GOF = 0·630

Staphylococcal nuclease 149

	L&S	Arnone *et al.* (1971)
α1	54–69	54–67
α2	98–105	99–106
α3	124–134	122–134
β3	31–37	30–36
β2	28–22	27–21
β1	6–19	12–19
β5	79–71	
β6	86–95	
β4	38–45	
β7	112–108	

GOF = 0·586

Triosephosphate isomerase 247

	L&S	Banner *et al.* (1975)
α1	16–30	17–31
α2	46–54	44–55
α3	81–86	79–87
α4	95–100	95–102
α5	104–119	105–120
α6	129–133	130–137
α7	137–153	138–154
α8	179–195	177–196
α9	196–201	197–204
α10	212–221	213–223
α11	237–244	237–246
3_{10}1	232–236	232–236
β1	3–12	6–12
β2	35–44	36–42
β3	57–68	60–63
β4	85–93	89–93
β5	122–130	122–129
β6	158–167	159–167
β7	202–209	205–209
β8	226–232	227–231
(β1)		

GOF = 0·501

Haemoglobin α-subunit 141

	L&S	Ladner *et al.* (1977)
α1	3–18	3–18
α2	20–36	20–35
		36–42
α3	52–73	52–71
α4	75–80	
α5	80–89	80–88
		87–92
α6	94–112	
α7	118–138	118–138

GOF = 0·228

Haemoglobin β-subunit 146

	L&S	Ladner *et al.* (1977)
α1	6–18	4–18
α2	19–35	19–34
α3	38–42	35–41
α4	50–56	50–56
α5	57–78	57–76
α6	80–84	
α7	85–95	85–93
		92–97
α8	100–119	99–117
α9	123–143	123–143

GOF = 0·251

TABLE 1 *(continued)*

Cu, Zn superoxide dismutase 151		
	L&S	Richardson *et al.* (1975)
α1	130–134	131–135
β1	2–11	4–9
β2	24–12	21–15
β3	25–36	27–34
β6	100–93	99–93
β5	80–90	80–87
β4	48–39	46–39
β7	113–117	113–118
β8	151–142	148–144
(β1)		
GOF = 0·705		

Phospholipase 123		
	L&S	Dijkstra *et al.* (1981)
α1	1–13	1–13
		17–22
α2	39–58	39–58
α3	58–65	58–66
α4	89–108	89–108
β1	72–79	74–78
β2	86–79	85–81
GOF = 0·421		

Elastase 240		
	L&S	Sawyer *et al.* (1978)
α1	154–159	154–160
α2	229–240	229–240
β1	15–23	14–22
β2	36–24	35–25
β3	38–45	38–44
β6	107–92	101–93
β5	67–86	69–80
β4	61–50	58–53
(β1)		
β7	125–132	124–135
β8	155–138	153–139
β9	167–179	171–180
β12	226–219	226–215
β11	198–212	203–209
β10	197–189	194–188
(β7)		
GOF = 0·667		

Papain 212		
	L&S	Drenth *et al.* (1971)
α1	24–40	24–43
α2	49–56	50–58
α3	67–78	67–78
α4	117–128	117–128
α5	139–144	137–143
β2	109–113	111–112
β7	212–204	208–206
β3	127–135	130–131
β4	168–164	167–162
β5	169–177	169–175
β6	192–185	191–185
β1	3–7	5–7
β4	168–164	167–162
β5	169–177	169–175
β6	192–185	191–185
GOF = 0·579		

Ribonuclease 124		
	L&S	Wyckoff *et al.* (1970)
α1	3–12	3–13
α2	24–34	24–34
α3	50–59	50–60
β1	42–49	41–48
β4	87–80	87–79
β6	93–103	94–104
β1	42–49	41–48
β5	92–88	91–80
β6	93–103	94–104

Southern bean mosaic virus protein 219		
	L&S	Abad-Zapatero *et al.* (1980)
α1	1–7	1–7
α2	79–84	78–86
α3	124–130	123–130
α4	163–172	160–172
α5	174–179	174–181
β1	18–23	18–21
β2	23–30	24–29

TABLE 1 *(continued)*

Ribonuclease 124 *(continued)*

β2	60–66	61–64
β3	77–71	75–71
β7	104–113	105–113
		119–114
β2	60–66	61–64
β3	77–71	75–71
β7	104–113	105–113
β8	124–119	124–121

GOF = 0·606

Southern bean mosaic virus protein 219 *(continued)*

β3	34–39	36–40
β4	55–41	56–41
β12	197–217	199–214
β6	102–87	101–87
β9	142–147	139–147
β10	155–160	155–158
β5	59–67	58–67
β11	192–181	193–183
β7	104–116	106–115
β8	138–133	138–132

GOF = 0·711

Immunoglobulin Fc fragment 206

	L&S	Diesenhofer (1981)
α1	10–15	
α2	72–79	
α3	117–122	
α4	176–183	
β1	1–11	2–7
β2	28–18	27–19
β6	61–73	62–72
β5	60–50	59–53
β4	44–48	45–47
β3	43–34	42–37
β7	78–90	81–88
β8	101–92	99–95
β9	104–118	105–114
β10	134–124	136–125
β14	165–178	166–176
β13	161–153	161–153
β12	148–152	149–151
β11	146–139	145–140
β15	182–194	185–192
β16	206–197	206–197

GOF = 0·709

Rhodanese 293

	L&S	Ploegman *et al.* (1978)
α1	12–22	11–22
α2	42–50	42–50
α3	76–88	76–87
α4	107–119	107–119
α5	129–137	129–137
α6	163–173	163–174
α7	183–189	183–189
α8	224–236	224–235
α9	253–264	251–264
α10	275–282	274–282
β1	8–12	8–11
β6	121–128	122–127
β5	92–101	94–98
β2	28–38	27–33
β4	55–61	56–58
β8	160–164	160–162
β12	267–272	269–271
β11	241–248	242–246
β9	175–181	177–181
β10	205–213	208–210
β3	49–53	
β7	136–159	

GOF = 0·549

The name of each protein is followed by its number of amino acid residues. L&S = sites recognized by our helical approximations, with *GOF* = goodness of fit (average metric error per residue) in ångström units. αi = the ith α-helix. $3_{10}j$ = the jth 3_{10}-helix. βk = the kth β-strand (the β-strands are arranged in the order of their sheet topologies with the sense of pairing; e.g. in ferredoxin, β2 22–26 is anti-parallel to its neighbour β3 33–27, but is parallel to β1 1–6). The right-hand columns are the "official" sites as recorded in the Protein Data Bank and denoted by their original references.

A helical approximation with step size $M = 5$ and the above recognition criteria identifies four strands (α-carbon atoms 1 to 6, 22 to 26, 27 to 33 and 49 to 53) and two α-helices (14 to 18, 39 to 44). Bends are identified within the following sections: 5 to 9, 16 to 21, 23 to 29, 31 to 37, 38 to 43 and 44 to 50.

A pairwise matching of strands results in the following minimal distances:

$$\begin{aligned} &d(\text{strand 1, strand 2}) = \text{parallel}, && 11{\cdot}2 \text{ Å} \\ &d(\text{strand 1, strand 3}) = \text{antiparallel}, && 7{\cdot}2 \text{ Å} \\ &d(\text{strand 1, strand 4}) = \text{antiparallel}, && 4{\cdot}8 \text{ Å} \\ &d(\text{strand 2, strand 3}) = \text{antiparallel}, && 4{\cdot}8 \text{ Å} \\ &d(\text{strand 2, strand 4}) = \text{antiparallel}, && 14{\cdot}5 \text{ Å} \\ &d(\text{strand 3, strand 4}) = \text{parallel}, && 10{\cdot}4 \text{ Å}. \end{aligned} \tag{13}$$

We therefore conclude that all four strands are parts of an antiparallel β-sheet with strands 1 and 3, 1 and 4, and 2 and 3 neighbouring. The distances also determine the relative positions of the β-strands. The β-sheet of ferredoxin is shown in Figure 3(a) and (b).

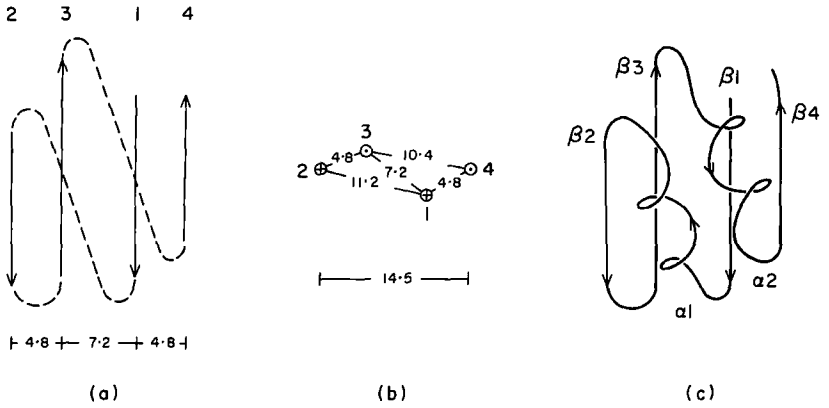

FIG. 3. Ferrodoxin. (a) Side view of β-sheet. (b) Top view of β-sheet. (c) Schematic native structure. All distances shown are in ångström units.

Incorporating the information we obtained on α-helices and bends into the above, we derive the three-dimensional native structure of ferredoxin (Fig. 3(c)). This, of course, agrees with the accepted structure.

(b) *A second example: triosephosphate isomerase*

The enzyme triosephosphate isomerase (EC 5.3.1.1) catalyzes the interconversion of dihydroxyacetone phosphate and D-glyceraldehyde-3-phosphate. The atomic co-ordinates of this protein from the chicken breast muscle have been determined crystallographically at 2·5 Å resolution (Banner *et al.*, 1975). A monomer of the triosephosphate isomerase molecule contains 247 amino acid residues.

Our helix-fitting algorithm with $M = 5$ identifies 12 strands. Among these 12 candidates, there are eight that have an average distance less than 8 Å from another strand in a pairwise matching. From this result (shown in Fig. 4(a)), we can readily deduce that the β-structure of triosephosphate isomerase is a barrel consisting of eight parallel β-strands.

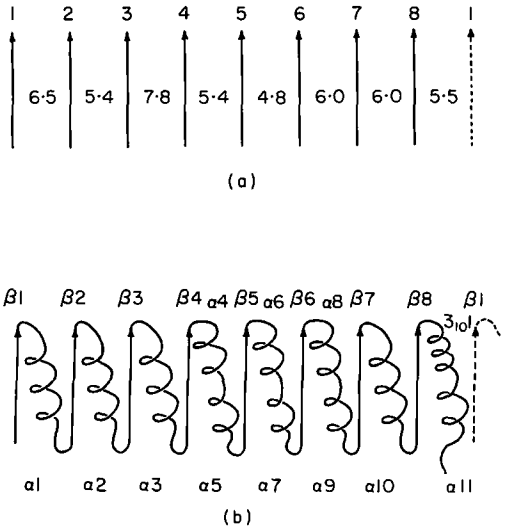

FIG. 4. Triosephosphate isomerase. (a) Parallel β-barrel; strand 1 (3–12), strand 2 (35–44), strand 3 (57–68), strand 4 (85–93), strand 5 (122–130), strand 6 (158–167), strand 7 (202–209), strand 8 (226–232). All distances shown are in ångström units. (b) Schematic native structure.

Eleven α-helices are recognized (α-carbon atoms 16 to 30, 46 to 54, 81 to 86, 95 to 100, 104 to 119, 129 to 133, 137 to 153, 179 to 195, 196 to 201, 212 to 221 and 237 to 244). Note that the α-helices lie between β-strands, forming a sequence of the so-called βαβ-units (Schulz & Schirmer, 1979).

Of special interest is the section of the protein backbone from α-carbon atoms 232 to 236. This is a helix of radius $r=1{\cdot}95$ and pitch $p=1{\cdot}1$ (corresponding to $\kappa=0{\cdot}39$ and $\tau=0{\cdot}22$). This is a 3_{10}-helix (the theoretical values of the parameters being $r=1{\cdot}9$, $p=0{\cdot}95$, $\kappa=0{\cdot}42$ and $\tau=0{\cdot}21$). The native structure of triosephosphate isomerase according to our helical approximation is shown in Figure 4(b), again agreeing well with the accepted structure.

(c) *A third example: staphylococcal nuclease*

The phosphoric diester hydrolase staphylococcal nuclease (EC 3.1.4.7; Arnone *et al.*, 1971) provides us with an example in which we can analyze the bending and relative displacement of β-strands. When this protein, which contains 149 amino acid residues, is approximated by helices with $M=5$, seven strands are recognized. They are the sections with α-carbon atoms 6 to 19, 22 to 28, 31 to 37, 38 to 45, 71 to 79, 86 to 95 and 108 to 112. The neighbouring relations and distances are as follows:

$$\begin{aligned} &d(\text{strand 1, strand 2}) = \text{antiparallel, } 5{\cdot}9\ \text{Å} \\ &d(\text{strand 1, strand 5}) = \text{antiparallel, } 6{\cdot}3\ \text{Å} \\ &d(\text{strand 2, strand 3}) = \text{antiparallel, } 5{\cdot}4\ \text{Å} \\ &d(\text{strand 4, strand 7}) = \text{antiparallel, } 5{\cdot}1\ \text{Å} \\ &d(\text{strand 5, strand 6}) = \text{antiparallel, } 4{\cdot}8\ \text{Å}. \end{aligned} \tag{14}$$

Since there is considerable variation in strand lengths (e.g. strand 1 has 14 residues, while strand 2 has 7), the relative displacements of strands (i.e. the pairwise matchings that give the smallest distances) are also important factors in determining the β-sheet topology. Furthermore, for the long strand 1 (α-carbon atoms 6 to 19), successive increases of the angle between $\vec{V}_i$ and $\vec{V}_6$ along the strand (Fig. 5(a)) show that it is bent. The large change in axis angle ($\sim 110°$) at α-carbon atoms 37–38 shows that there is a bend between strands 3 and 4. The β-structure of staphylococcal nuclease according to our algorithm is shown in Figure 5(b).

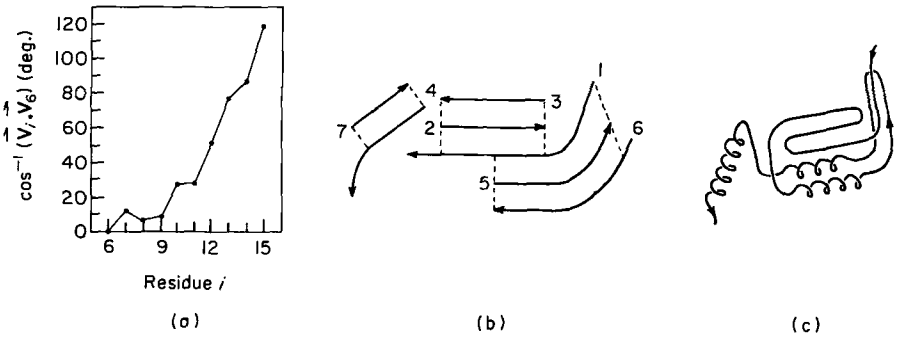

FIG. 5. Staphylococcal nuclease. (a) Bending of β-strand 1: residue i *versus* angle between axes $\vec{V}_i$ and $\vec{V}_6$. (b) β-structure with optimal matching of strands. (c) Schematic native structure.

Three α-helices are identified: α-carbon atoms 54 to 59, 98 to 105 and 124 to 134. These lie between strands 4 and 5, between 6 and 7, and after 7, respectively.

It is important to note that although Table 1 lists only the regular regions of the proteins and our analysis of proteins is based on the locations of these regular α and β-regions, our algorithm does treat all regions of the proteins equally in the fittings. We just saw how the irregular bends provide information on the structure of the β-sheet of staphylococcal nuclease. For completeness and for a comparison, Table 2 shows the bends (and turns) identified on this protein by our algorithm, *versus* those identified by Rose & Seltzer (1977). Bends at residues 81–82 and 90–91 are not identified by us because their bend angles (20° and 35°, respectively) do not exceed our 40° threshold. In other words, these are slight bends.

The complete native structure of staphylococcal nuclease by helical approximation is as shown in Figure 5(c).

These examples illustrate that the algorithm is simple, yet powerful enough to identify complicated topological features, as well as providing new assignments of regular regions. Table 1 contains some discrepancies between morphons we identify and the official versions. For example, for staphylococcal nuclease the Protein Data Bank only records the three α-helices with β-strands 3, 2 and half of 1. Thus only the "front half" of the β-sheet (β3, 2, 1, 5, 6) is identified there. Also, β4 and β7 are identified by us and not by them, and this is probably due to the criterion of hydrogen bonding that we mentioned before. There are morphons

TABLE 2

Comparison of bends (and turns) in staphylococcal nuclease

L&S†	Rose & Seltzer (1977)
5–7	5–6
15	15
20–21	20
28–30	28–29
38–39	38–40
45–49	45
	48
53–54	53–57
69–70	70
78–80	79
	81–82
85–87	85
	90–91
95–96	95
100	100
106–108	107
117–118	116
120	120–124
135–136	135–136
138–140	

† Sites recognized by our helical approximations.

picked out by the Data Bank (e.g. α-helices 36 to 42 and 87 to 92 on the haemoglobin α-subunit) but not recognized by our algorithm because the parameter values fall outside the range specified by the inequalities (8) to (10). Thus, these are probably rather "irregular" structures. Other similar discrepancies can be likewise explained.

In the next section, we proceed with the detailed illustration of the next phase of the algorithm.

(d) *Stepped-helix with minimal partition*

After an initial helix-fitting with $M=5$ and with the different morphons of the protein backbone identified, we can partition the protein curve into structural sections. Let us use the 141-residue α-subunit of haemoglobin (Ladner *et al.*, 1977) as an illustration.

The fitting algorithm identifies seven α-helices at α-carbon positions 3 to 18, 20 to 36, 52 to 73, 75 to 80, 80 to 89, 94 to 112 and 118 to 138, a single strand (hence not part of a β-sheet) at 45 to 50, and bends in between the regular segments of this antiparallel α-domain protein. Thus the backbone of haemoglobin α-subunit has the minimal partition (in terms of the α-carbon residue number instead of the arc-length):

$$\{a_i\}=\{1, 3, 18, 20, 36, 40, 45, 52, 73, 75, 80, 89, 94, 112, 118, 138, 141\}. \quad (15)$$

The partition point $a_5=40$ is added to subdivide the long loop 36 to 45 into

shorter intervals for better helix-fitting. This partition is termed minimal to denote the minimal number of structural sections. This number is commensurate with the requirement that the accuracy of sectional fitting be reasonably uniform over the whole length of the curve.

The same two-stage cylinder-pitch fitting algorithm is then applied to each of the intervals $[a_{i-1}, a_i]$; i.e. α-carbon atoms a_{i-1} to a_i. The resulting κ_i and τ_i values constitute the step-constants in the approximation of the pair (κ, τ) in $C(I, \mathbb{R}^2)$, which represents the protein backbone. Figure 6 shows the curvature

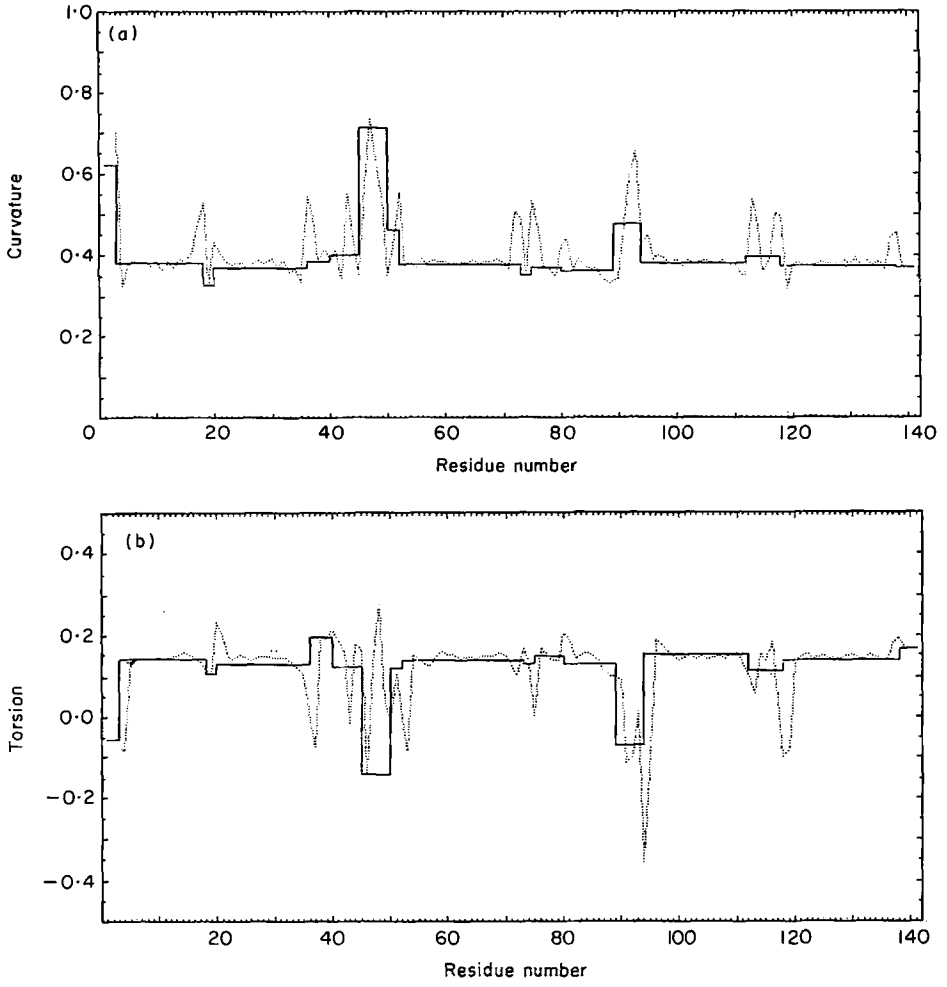

FIG. 6. (a) Curvature and (b) torsion of the space curve representing haemoglobin α-subunit (.....) and of the approximating step helix with minimal partition (————).

and torsion of the space curve of haemoglobin α-subunit, with the sequence of (κ_i, τ_i) values of the best-approximating stepped-helix.

It is interesting to compare the goodness-of-fit (per residue) from different fittings. With the first sequential fitting using a sliding sectional length of $M=5$,

$GOF = 0{\cdot}228$. The error increases as M increases: for $M = 6$, 7, 8, 9 and 10, $GOF = 0{\cdot}332$, $0{\cdot}443$, $0{\cdot}505$, $0{\cdot}595$ and $0{\cdot}715$, respectively. With the minimal partition (14), since longer sections are involved, we obtain a GOF value of $0{\cdot}441$, which is, incidentally, a better fit than the GOF value of $0{\cdot}477$ obtained from the partition using the official Protein Data Bank sections.

This secondary fitting by stepped-helix with a partition also provides information on the bending of the path of the protein backbone. The direction $\vec{\mathbf{V}}_i$ of the helical axis on $[a_{i-1}, a_i]$ gives the direction of that section of the space curve. The sequence of triples $(r_i, p_i, \vec{\mathbf{V}}_i)$ can be used to construct the stepped-helix: on the interval from α-carbon a_{i-1} to α-carbon a_i, the helix has radius r_i, pitch $2\pi p_i$, and axial direction $\vec{\mathbf{V}}_i$.

(e) *General stepped-helices*

The basic principle underlying the approximation of protein backbone by stepped-helices (recall Theory and Methods, section (a)) is that any given mapping (κ, τ) in $C(I, \mathbb{R}^2)$ can be arbitrarily closely approximated by mappings in $S(I, \mathbb{R}^2)$, because the uniform closure of the subset $S(I, \mathbb{R}^2)$ contains $C(I, \mathbb{R}^2)$. Any other subset of $C(I, \mathbb{R}^2)$ that has this property will of course fulfil the role of an approximating basis set as well. What is special about the stepped-mappings $S(I, \mathbb{R}^2)$ is that the corresponding space curves, the stepped-helices, are easy to describe and analyze as geometric objects. A subset of $C(I, \mathbb{R}^2)$ that contains $S(I, \mathbb{R}^2)$ in turn as a subset would be an approximating basis set, and it would also retain the special role of stepped-helices without being limited to helices with circular cross-sections. A natural such subset is the set $H(I, \mathbb{R}^2)$ of all pairs (κ, τ), for which partitions $\{a_i\}$ exist with the property that on each subinterval (a_{i-1}, a_i), the ratio τ/κ is constant (where $\kappa \neq 0$, and $\tau = 0$ whenever $\kappa = 0$), without either $\tau(s)$ or $\kappa(s)$ being independent of s.

A space curve that has a constant torsion-to-curvature ratio is a general helix. A space curve is a general helix if and only if there is a fixed vector in space, called the axis of the helix, such that the angle θ between the tangent vectors and the axis is constant. In fact in this case:

$$\tau/\kappa = \cot\theta, \tag{16}$$

and the general helix has a canonical representation of the form:

$$\vec{\mathbf{X}}(s) = (x_1(s), x_2(s), s\cos\theta), \tag{17}$$

where the z-axis is the axis of the helix. Note that the circular helix (5) is a general helix with:

$$p = \cos\theta. \tag{18}$$

Thus a mapping in $H(I, \mathbb{R}^2)$ with sectionally constant τ/κ ratios gives a space curve that is a general stepped-helix. The axial directions of the sections are well-determined, so can again be used to describe the bending path of the protein backbone. Since on each subinterval the space curve is not confined to a circular cylinder, there is more flexibility and hence a better fit in the approximation of proteins by general stepped-helices.

(f) *Elliptic helices, helicoids and catenoids*

The simplest natural extension of the set of circular helices is the set of elliptic helices; i.e. helices that lie on cylinders with elliptic cross-sections. An elliptic helix has the equation:

$$\vec{\mathbf{X}}(s) = (a \cos s,\ b \sin s,\ ps), \tag{19}$$

where $a \geqslant b$ are the semi-major and semi-minor axes, and $2\pi p$ is the constant pitch. Associated with elliptic helices, there is a parameter that describes the twist of protein backbones. The twist along the curve is defined as the angle ϕ between the elliptic axes of consecutive helices.

Our helix-fitting algorithm has been modified to use elliptic helices as the approximating basis set. The first step (the Levenberg–Marquardt method) was changed to find the best-fitting elliptic, instead of circular, cylinder. Two additional parameters have to be estimated: the pair (a,b) (replacing the radius r), and the new twist parameter ϕ. Elliptic-helical approximation of a typical α-helix (e.g. α-carbon atoms 16 to 30 of triosephosphate isomerase) yields values for a and b between 2·1 and 2·5 Å (with a geometric mean $\sqrt{ab} \approx 2{\cdot}3$ Å) and a small twist value of less than $\pm 5°$. Other parts of the protein backbone give larger differences in the a and b values, and larger twist parameters. The *GOF* value for elliptic-helical approximations is typically about 10% smaller than that of circular-helical approximations, since more parameters lead to more degrees of freedom and hence better fits. In fact the *GOF* values shown in Table 1 correspond to elliptic-helical fits. We feel that the relative values of a, b and ϕ reflect the nature of the side-chains, and that the correlations between the ellipticity and the twist of the protein backbone and the bulk and packing properties of the amino acid residues are significant. These questions are under investigation.

Another approximating basis set can be formed by general helices-with-constant-pitch (eqn (17)). These general helices lie on the surface of a helicoid (Fig. 7(a)), and this fact establishes a connection between this paper and our

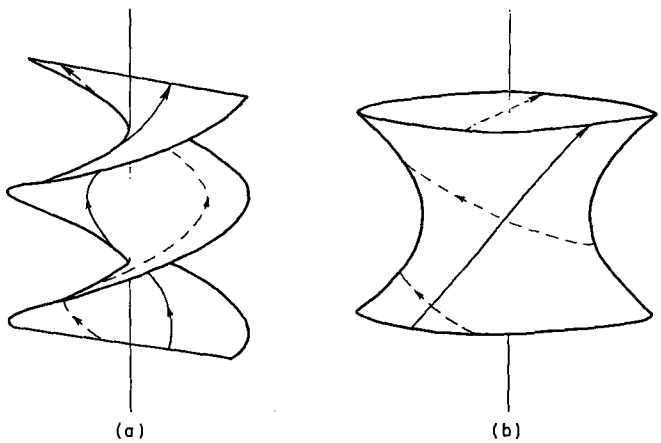

FIG. 7. (a) Helicoid with constant-pitch general helices. (b) Catenoid with catenoidal helices.

earlier paper (Louie & Somorjai, 1982) on proteins as geodesics on minimal surfaces, in which we represent (circular) helices as space curves on the surface of a helicoid.

General helices on the surface of a catenoid (Fig. 7(b)), the conjugate surface of the helicoid (Louie & Somorjai, 1982, section 6), can also be used as an approximating basis set. This basis set works especially well for β-barrels, since these can be represented well as catenoids. For example, the eight-strand parallel β-barrel of triosephosphate isomerase is a catenoid with waist diameter 10 Å, and the strands make the constant angle of $\theta=35°$ with the axis of the barrel. Thus the eight β-strands of this protein can be represented as "catenoidal helices".

An elliptic catenoidal helix has the equation:

$$\vec{\mathbf{X}}(s)=(a\cosh(qs)\cos s,\ b\cosh(qs)\sin s,\ ps), \tag{20}$$

which lies on the elliptic catenoid:

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=\cosh^2\frac{z}{c}, \tag{21}$$

where $c=p/q$. Thus, the eight-strand barrel can be described on two levels as follows. The eight strands can be fitted individually by the elliptic catenoidal helix (eqn (20)), resulting in eight sets of values of the parameters (a, b, p, q), from which a set of average values $(\bar{a}, \bar{b}, \bar{p}, \bar{q})$ can be calculated. On the other hand, the co-ordinates of all the C_α atoms on the eight strands can be fitted onto the surface described by equation (21), hence one obtains a set of values $(\hat{a}, \hat{b}, \hat{c})$. The accuracy of the equalities:

$$\bar{a}=\hat{a},\quad \bar{b}=\hat{b},\quad \bar{p}/\bar{q}=\hat{c}, \tag{22}$$

can then be used to determine how well the eight strands fit on a catenoidal barrel.

Equation (21) reduces to the equation of an elliptic cylinder:

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1, \tag{23}$$

when $c\to\infty$. Thus an α-helix, for example, can be described as a catenoidal helix with $a,b\approx 2{\cdot}3$ Å and $c\to\infty$. As an illustration, again consider the α-helix 16 to 30 of triosephosphate isomerase. The best-fitting elliptic catenoid has parameters $a=2{\cdot}40$, $b=2{\cdot}27$ and $c>10^6$. An advantage of using catenoidal helices as the approximating basis set is that the same set can be used to describe secondary structures and certain supersecondary structures, such as the supercoiling and bundles of α-helices (e.g. see Weber & Salemme, 1980) and the β-barrels mentioned above. In other words, catenoidal helices provide "prior indicators" for the type of surfaces onto which the protein curves are wound; i.e. they form a universal basis set for the description of morphological patterns of protein backbones at two hierarchical levels: space curves and surfaces embedded in $\mathbb{R}^3$.

(g) *Implications for protein energetics*

We suggested previously (Louie & Somorjai, 1982) a generalized Fermat's principle, that "energy propagation along protein molecules takes paths of critical

time". This principle has particular significance for helices, since a helix is a geodesic (shortest curve) on a cylinder as well as an asymptotic curve ("fastest" curve) on a helicoid. Thus helices are analogues of straight lines in the *surface* geometry of proteins, in that they provide both shortest and fastest paths for energy transfer. This dual description of helices again gives them a special role; indeed, a helix with radius r and pitch $2\pi p$ is the curve of intersection of a cylinder with radius r and a helicoid with pitch $2\pi p$. Approximations of proteins by stepped-helices, therefore, may provide the natural geometric tools to analyze proteins in dynamical terms as well. The reader is referred to Louie & Somorjai (1982) for a detailed discussion of the concepts presented in this paragraph.

The motion of helical curves is intimately related to the propagation of solitons. It has been shown (Lamb, 1976) that the intrinsic equations governing the curvature and torsion (i.e. the Frenet equations that establish the fundamental theorem of space curves) of a helix can be reduced to a non-linear Schrödinger equation. The single-soliton solution of this equation provides a description of the response to excitation of a helical curve. Furthermore, the sine-Gordon equation is association with curves of constant curvature, and the modified Korteweg–de Vries equation is associated with curves of constant torsion. These constants play the role of the eigenvalue parameters in the inverse-scattering method. (See Scott *et al.* (1973) for an introduction to solitons.)

The connections between solitons and the motion of helical curves have been studied by Davydov (1973,1977,1979) in the context of energy transfer along α-helices of proteins. It was found that α-helices contract under the excitation of their peptide groups. In the region of excitation, the pitch of the helix decreases and the contracted region moves along the direction of the axis at a rate proportional to the energy of interaction of the residues. A numerical analysis of Davydov solitons on α-helices has been presented by Scott (1982). A major reason why only α-helical proteins were investigated until now is that these provide the "obvious" helices; non-α-helical regions were not considered. Since in our approach all regions of a protein molecule are identified as (approximate) helices of various radii and pitches, the idea of soliton propagation along helices can be applied to the whole protein backbone. In fact, this extension takes on a new significance in terms of the excitation of active sites on enzymes. Since the energy of the excitation and the rate of propagation of the soliton are related to the curvature and torsion of the carrying helix, it is possible to consider concentration of energy at certain specific sites of proteins. In other words, the rate of energy transfer (i.e. soliton propagation) along a protein backbone is modulated by the non-constant character of its approximating helical curve, creating local maxima and minima in the energy distribution at specific (helical) sites, with possible trapping at appropriate regions. These ideas may provide the concretization of the qualitative concept of entatic states, according to which the active sites of enzymes are locally excited relative to the rest of the molecule, thus facilitating the creation of the productive transition state. We shall discuss this circle of ideas in more detail elsewhere.

This paper is recorded as NRCC no. 20967.

REFERENCES

Abad-Zapatero, C., Abdel-Meguid, S. S., Johnson, J. E., Leslie, A. G. W., Rayment, I., Rossmann, M. G., Suck, D. & Tsukihara, T. (1980). *Nature (London)*, **286**, 33–39.

Adman, E. T., Sieker, L. C. & Jensen, L. H. (1976). *J. Biol. Chem.* **251**, 3801–3806.

Arnone, A., Bier, C. J., Cotton, F. A., Day, V. W., Hazen, E. E., Richardson, D. C., Richardson, J. S. & Yonath, A. (1971). *J. Biol. Chem.* **246**, 2302–2316.

Balasubramanian, R. (1977). *Nature (London)*, **266**, 856–857.

Banner, D. W., Bloomer, A. C., Petsko, G. A., Phillips, D. C., Pogson, C. I. & Wilson, I. A. (1975). *Nature (London)*, **255**, 609–614.

Brown, A. L. & Page, A. (1970). In *Elements of Functional Analysis*, sections 3.7 to 3.8, Van Nostrand Reinhold, London.

Brown, K. M. & Dennis, J. E. (1972). *Numer. Math.* **18**, 289–297.

Carmo, M. P. do (1976). *Differential Geometry of Curves and Surfaces*, Prentice-Hall, London.

Davydov, A. S. (1973). *J. Theor. Biol.* **38**, 559–569.

Davydov, A. S. (1977). *J. Theor. Biol.* **66**, 379–387.

Davydov, A. S. (1979). *Int. J. Quantum Chem.* **16**, 5–17.

Dickerson, R. E. & Geis, I. (1969). *The Structure and Action of Proteins*, Harper & Row, New York.

Diesenhofer, J. (1981). *Biochemistry*, **20**, 2361–2370.

Dijkstra, B. W., Kalk, K. H., Hol, W. G. J. & Drenth, J. (1981). *J. Mol. Biol.* **147**, 97–123.

Drenth, J., Jansonius, J. N., Koekoek, R. & Wolthers, B. G. (1971). *Advan. Protein Chem.* **25**, 79–115.

Dunn, J. B. R. & Klotz, I. M. (1975). *Arch. Biochem. Biophys.* **167**, 615–626.

Goel, N. S. & Ycas, M. (1979). *J. Theor. Biol.* **77**, 253–305.

Kuntz, I. D., Crippen, G. M. & Kollman, P. A. (1979). *Biopolymers*, **18**, 939–957.

Ladner, R. C., Heidner, E. J. & Perutz, M. F. (1977). *J. Mol. Biol.* **114**, 385–414.

Lamb, G. L. (1976). *Phys. Rev. Letters*, **37**, 235–237.

Louie, A. H. & Somorjai, R. L. (1982). *J. Theor. Biol.* **98**, 189–209.

Ploegman, J. H., Drent, G., Kalk, K. H. & Hol, W. G. J. (1978). *J. Mol. Biol.* **123**, 557–594.

Rackovsky, S. & Scheraga, H. A. (1978). *Macromolecules*, **11**, 1168–1174.

Rackovsky, S. & Scheraga, H. A. (1980). *Macromolecules*, **13**, 1440–1453.

Rackovsky, S. & Scheraga, H. A. (1981). *Macromolecules*, **14**, 1259–1269.

Rajan, S. S. & Srinivasan, R. (1977). *Biopolymers*, **16**, 1617–1634.

Richardson, J. S. (1977). *Nature (London)*, **268**, 495–500.

Richardson, J. S. (1981). *Advan. Protein Chem.* **34**, 167–339.

Richardson, J. S., Thomas, K. E., Rubin, B. H. & Richardson, D. C. (1975). *Proc. Nat. Acad. Sci., U.S.A.* **72**, 1349–1353.

Rose, G. D. & Seltzer, J. P. (1977). *J. Mol. Biol.* **113**, 153–164.

Sawyer, L., Shottin, D. M., Campbell, J. W., Wendell, P. L., Muirhead, H., Watson, H. C., Diamond, R. & Ladner, R. C. (1978). *J. Mol. Biol.* **118**, 137–208.

Schulz, G. E. & Schirmer, R. H. (1979). In *Principles of Protein Structure*, p. 81. Springer-Verlag, New York.

Scott, A. C. (1982). *Phys. Rev. sect. A*, **26**, 578–595.

Scott, A. C., Chu, F. Y. F. & McLaughlin, D. W. (1973). *Proc. IEEE*, **61**, 1443–1483.

Sippl, M. J. (1982). *J. Mol. Biol.* **156**, 359–380.

Weber, P. C. & Salemme, F. R. (1980). *Nature (London)*, **287**, 82–84.

Wyckoff, H. W., Tsernoglou, D., Hanson, A. W., Knox, J. R., Lee, B. & Richards, F. M. (1970). *J. Biol. Chem.* **245**, 305–328.

Edited by A. Klug