J. theor. Biol. (1982) **98,** 189–209

Differential Geometry of Proteins: A Structural and Dynamical Representation of Patterns

A. H. LOUIE AND R. L. SOMORJAI†

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

(*Received* 9 *January* 1982)

This is an essay on the applications of the differential geometry of curves and surfaces to the analysis of the spatial patterns of proteins. Differential geometry is the natural mathematical tool when a protein molecule is represented as a space curve passing through its α-carbons.

We suggest a unifying and natural description of the three-dimensional conformation of proteins. In particular, we argue that the regular secondary structures correspond to geodesics on minimal surfaces, and that the tertiary structures result from the energetically best packing of these minimal surfaces. The α-helices and the strands of the β-barrels lie, respectively, on the conjugate minimal surfaces of the helicoid and the catenoid. These two surfaces can be transformed into each other isometrically, and the intermediate stages in the transformation model the various β-twisted sheets found in proteins.

The geometry of these protein curves and surfaces is studied in detail, and biological interpretations are given along with analytic expressions. Attention is paid to the relationship between local and global properties in both mathematical and biological terms.

Implications for the morphogenetic process of protein folding are outlined. The problem of the prediction of three-dimensional structures from amino acid sequence is also addressed. These turn out to be best formulated in terms of a dynamical analysis of vector fields on proteins.

Modern differential geometry centres its attention on manifolds, which are generalizations of curves and surfaces, and which behave locally like Euclidean spaces. From the theory of manifolds arises the notion of a fibre space. Fibre spaces are the mathematical objects used in Robert Rosen's unified approach to pattern generation. Since the spontaneous folding of proteins into native conformations is a prototype biological example of pattern generation, the abstract formalism of fibre spaces provides an appropriate general setting in which to further the study of the mathematical biology of proteins.

† To whom correspondence should be addressed.

0022-5193/82/180189+21 $03.00/0

1. Prologue

A protein molecule is a complex polymeric chain made up of α-amino acid residues linked by peptide bonds in a definite sequence. It is generally accepted that the sequence (the primary structure of the protein) determines all that is unique about a particular protein, including its three-dimensional structure and its biological function. To reduce complexity we idealize a protein molecule as a (regular parametrized) space curve passing through its α-carbons. Such a curve is determined uniquely except for position in space, i.e. up to a rigid motion. It then follows from the fundamental theorem of space curves that any curve, and in particular the backbone conformation (our "primitive protein"), can be completely characterized by two continuous functions of the arc-length $s \in [0, l]$ along the curve, the curvature κ and the torsion τ. Native conformations of all proteins contain distinctive, stable structural patterns and this is reflected in the spatial regularities present in our "primitive protein".

This paper suggests a unifying and natural description of the most important of the spatial patterns. It also addresses the dynamical problem of pattern generation and interconversion in the spontaneous folding of a protein molecule into its native form. Differential geometry is used as the mathematical tool, and the reader is referred to any one of the standard texts (e.g. Eisenhart, 1909; Weatherburn, 1927; Carmo, 1976) for a review of the subject.

The paper is structured such that each section introduces some relevant mathematical background which is then specialized to the protein structure-dynamics problem. A close parallel is maintained between the development of mathematical concepts and the description of protein morphology, starting with simple structural characterization, proceeding to pattern interconversion and concluding with some aspects of the dynamics of folding.

2. Geometry of the Helix

It is probably not too much of a hyperbole to say that we are made of helices. The helix is the most frequently encountered pattern in living organisms. There are numerous examples: the double helix of DNA, the α-helix in proteins, the triplet superhelix of collagen—the most abundant protein in mammals, the coiled-coil α-helix of myosin and the long helix of F-actin in the cytoskeleton and muscles, the helical capillaries which supply the contracting cardiac cells, and the helical tendrils of climbing plants, just to name a few. In fact the helix is such a common form of mollusk shells that there is a genus of pulmonate land mollusks named *Helix*.

Mathematically, a curve is called a *general helix* if its tangent vectors make a constant angle with a fixed direction, the *axis* of the helix. It is uniquely characterized as the space curve for which the ratio κ/τ is a constant ($\tau \neq 0$). The helix is right-handed if $\tau > 0$ and left-handed if $\tau < 0$. It has the canonical equation

$$\boldsymbol{\alpha}(s) = \left(\frac{a}{c}\int \sin\theta(s)\,ds, \frac{a}{c}\int \cos\theta(s)\,ds, \frac{b}{c}s\right), \tag{1}$$

where $c^2 = a^2 + b^2$, with axis $\mathbf{e}_3 = (0, 0, 1)$ and $\kappa/\tau = a/b$. In particular, $\theta(s) = s/c$ yields the (circular) helix, which is characterized as the unique curve for which $\kappa(= a/c^2)$ and $\tau(= b/c^2)$ are both non-zero constants.

The helix is usually represented as lying on the surface of a circular cylinder, but for our purpose it is both more convenient and mathematically more compelling (as we shall see later) to have it lying on a helicoid. A *helicoid* is the regular parametrized surface

$$\mathbf{x}(u, v) = (-b \sinh v \sin u, b \sinh v \cos u, bu), \tag{2}$$

$-\infty < u < +\infty$ and $-\infty < v < +\infty$ (Fig. 1(a)). An interval of the parameter u of length 2π corresponds to one turn of the helicoid with pitch $2\pi|b|$.

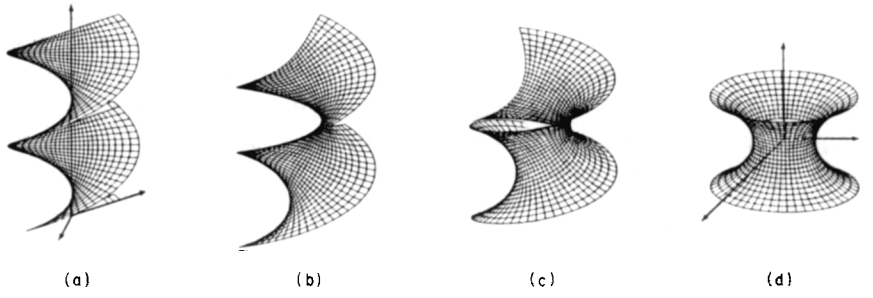

FIG. 1. Isometric transformation of helicoid to catenoid, given by $\mathbf{z}(w) = (\cos w)\mathbf{x} + (\sin w)\mathbf{y}$ with b = constant. (a) Helicoid $\mathbf{x} = \mathbf{z}(0)$. (b) & (c) Intermediate stages $\mathbf{z}(\pi/6)$ and $\mathbf{z}(\pi/3)$. (d) Catenoid $\mathbf{y} = \mathbf{z}(\pi/2)$.

The co-ordinate curve $v = v_0$ (constant) corresponds to a helix with radius $|b \sinh v_0|$. (The "curved" grid lines in Fig. 1(a) are helices.) More generally, the image on the helicoid of the straight line $v = Au + v_0$ in the (u, v)-parameter space is a *conical helix*, the radius of which increases or decreases along the curve depending on the values and the relative signs of the parameters.

The first regular pattern of protein structure we consider is the right-handed α-helix, the most abundant secondary structure in proteins. In this

structure the α-carbons are arranged in a helical coil having 3·6 amino acid residues per turn, and the "rise" per residue is 1·5 Å i.e. the pitch of the helix is $1{\cdot}5\ \text{Å} \times 3{\cdot}6 = 5{\cdot}4$ Å. The radius of the α-helix is 2·3 Å. Thus the α-helix corresponds to the co-ordinate curve

$$v = 1{\cdot}7 \tag{3}$$

on the helicoid with $b = 0{\cdot}86$, hence it has the parametrization

$$\boldsymbol{\alpha}(u) = (-2{\cdot}3 \sin u, 2{\cdot}3 \cos u, 0{\cdot}86\, u). \tag{4}$$

It has a constant curvature $\kappa = 0{\cdot}38$ and a constant torsion $\tau = 0{\cdot}14$. An α-carbon appears on the curve for every u-interval of length $2\pi/3{\cdot}6 = 1{\cdot}75$. The average length of α-helices in proteins is 17 Å, corresponding to 11 residues or 3 turns (a u-interval of length 6π), and lengths of 7, 11 and 15 residues (2, 3 and 4 turns, respectively) seem to be preferred (Schulz & Schirmer, 1979).

3. Twisted Sheets and Barrels

The other major secondary structural pattern found in proteins is the β-sheet, formed from strands. Most observed sheets are nonplanar, with a left-handed twist when viewed along the sheet surface perpendicular to the strands. To a good approximation a single strand of a twisted sheet forms an extended left-handed helix with $b = -1{\cdot}2$ and $v = 0{\cdot}76$ (corresponding to a pitch of 7·6 Å and a radius of 1·0 Å). The average strand separation is 5 Å and neighbouring strands make an angle of about 25° with each other. The average strand length in a β-structure is about six residues (20 Å) and most sheets contain six or fewer strands (width 25 Å or less). These dimensions are those of a so-called "domain" in proteins. With these data, one can fit the average twisted β-sheet on a helicoid with $b = -11{\cdot}5$. A strand appears for every u-interval of length 0·44, and each strand can be parametrized by $u = \text{constant}$, $-1{\cdot}3 < v < 1{\cdot}3$ (as depicted by the "straight" grid lines in Fig. 1(a). (A better description of the twisted β-sheet, however, will be given later.)

The twisted sheet of the β-structure occasionally rolls up into a "barrel", the best example of which is that in triosephosphate isomerase (E.C. 5.3.1.1). β-barrels may consist of anywhere between 5 and 13 strands, but the circular cross-sections of all the barrels are nearly constant in appearance: the diameter is about 10 Å in the middle and slightly wider at the ends, and the barrels average 12 Å in height. The almost constant barrel size, independent of the number of strands, is due to the varying degree of strand twist around the barrel: the more strands there are, the less twist

each strand has. Twist can be measured by the angle which strands make with the axis of the barrel. That angle decreases from 50° for 5-stranded barrels to 15° for 9- to 13-stranded ones. (Beyond 8 or 9 strands the twist cannot decrease any further and the barrel cross-section "flattens out" and becomes elliptical. See Richardson, 1981.)

The surface of the β-barrel can be represented by a *catenoid* (Fig. 1(d)) with the regular parametrization

$$\mathbf{y}(u, v) = (-b \cosh v \cos u, -b \cosh v \sin u, bv), \tag{5}$$

$0 < u < 2\pi$, $-\infty < v < +\infty$. Sometimes it may be convenient to let $-\infty < u < +\infty$ (and then consider the u values modulo 2π), which can be thought of as a "spiral copy" of the catenoid, as long as it is understood that it does not "spiral in" or "spiral out": each u-interval of length 2π corresponds to one revolution of the surface, having the same geometry as the catenoid given by equation (5). The value of the parameter b dictates the "curvature" of the side of the barrel: the larger the magnitude $|b|$ the "straighter" the side. It also controls the "size" of the barrel, the diameter of the "waist" being $2|b|$.

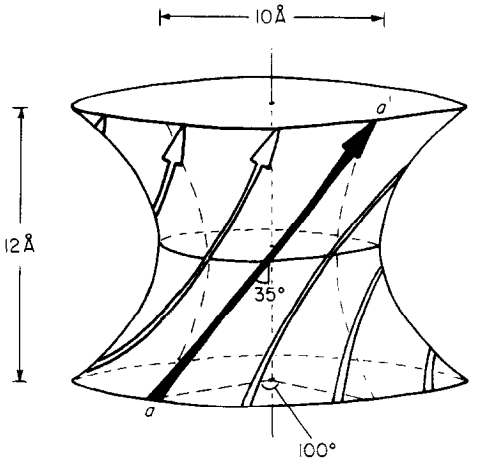

FIG. 2. Eight-stranded parallel β-barrel of triosephosphate isomerase.

Let us consider the β-barrel of triosephosphate isomerase as an example (Fig. 2). It is an eight-stranded parallel barrel with a twist angle of about 35°. A waist diameter of 10 Å and a height of 12 Å set the parameter values at $b = 5{\cdot}0$ and $-1{\cdot}2 < v < 1{\cdot}2$. The strand aa' in Fig. 2 is the image of the straight line segment

$$v = 1{\cdot}4u, \tag{6}$$

$-1{\cdot}2 < v < 1{\cdot}2$, in the (u, v)-parameter space. The eight parallel strands in fact correspond to

$$v = 1{\cdot}4\left(u - \frac{k\pi}{4}\right) \tag{7}$$

for $k = 0, 1, \ldots, 7$, where the u values are taken modulo 2π. Furthermore, as is shown for the strand aa', each strand spans an angle of about 100° on the barrel. This is in fact the observed value (Richardson, 1981). The neighbouring strand distance ranges from 3·2 Å at the waist to 5·8 Å at the ends of the barrel, and these again are approximately correct values.

4. First and Second Fundamental Forms

The fundamental theorem of space curves states that a curve is uniquely determined by two local invariants, curvature and torsion, as functions of arc length. There is similarly a fundamental theorem for surfaces embedded in R^3, which states that a surface is uniquely determined also by two local invariants, the first and second fundamental forms, as quadratic forms defined on vectors in the tangent planes.

Let $\mathbf{x}(u, v)$ be a regular parametrized surface in R^3. At each point on the surface there is a unit normal

$$\mathbf{N} = \frac{\mathbf{x}_u \times \mathbf{x}_v}{|\mathbf{x}_u \times \mathbf{x}_v|}. \tag{8}$$

The *first fundamental form*, I, of $\mathbf{x}$ is the quadratic form defined by

$$\mathrm{I}(\mathrm{d}u, \mathrm{d}v) = \langle \mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{x} \rangle = (\mathrm{d}u\ \mathrm{d}v)\begin{pmatrix} E & F \\ F & G \end{pmatrix}\begin{pmatrix} \mathrm{d}u \\ \mathrm{d}v \end{pmatrix} \tag{9}$$

(where $\langle .\,, . \rangle$ denotes the inner product on the tangent planes of the surface), and the *first fundamental coefficients* are

$$\begin{aligned} E &= \langle \mathbf{x}_u, \mathbf{x}_u \rangle, \\ F &= \langle \mathbf{x}_u, \mathbf{x}_v \rangle, \\ G &= \langle \mathbf{x}_v, \mathbf{x}_v \rangle. \end{aligned} \tag{10}$$

Note that

$$|\mathrm{I}| = EG - F^2 = |\mathbf{x}_u \times \mathbf{x}_v|^2 > 0, \tag{11}$$

thus the first fundamental form is positive definite. The *second fundamental form*, II, of $\mathbf{x}$ is the quadratic form

$$\mathrm{II}(\mathrm{d}u, \mathrm{d}v) = -\langle \mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{N} \rangle = (\mathrm{d}u\ \mathrm{d}v)\begin{pmatrix} L & M \\ M & N \end{pmatrix}\begin{pmatrix} \mathrm{d}u \\ \mathrm{d}v \end{pmatrix} \tag{12}$$

and the *second fundamental coefficients* are

$$\begin{aligned} L &= -\langle \mathbf{x}_u, \mathbf{N}_u \rangle, \\ M &= -\tfrac{1}{2}(\langle \mathbf{x}_u, \mathbf{N}_v \rangle + \langle \mathbf{x}_v, \mathbf{N}_u \rangle), \\ N &= -\langle \mathbf{x}_v, \mathbf{N}_v \rangle. \end{aligned} \tag{13}$$

Since $\mathbf{x}_u$ and $\mathbf{x}_v$ are perpendicular to $\mathbf{N}$, one has the alternate expressions

$$\begin{aligned} L &= \langle \mathbf{x}_{uu}, \mathbf{N} \rangle, \\ M &= \langle \mathbf{x}_{uv}, \mathbf{N} \rangle, \\ N &= \langle \mathbf{x}_{vv}, \mathbf{N} \rangle, \end{aligned} \tag{14}$$

hence

$$\mathrm{II}(\mathrm{d}u, \mathrm{d}v) = \langle \mathrm{d}^2\mathbf{x}, \mathbf{N} \rangle. \tag{15}$$

(Definitions (9) and (12) suggest naturally the *third fundamental form*

$$\mathrm{III}(\mathrm{d}u, \mathrm{d}v) = \langle \mathrm{d}\mathbf{N}, \mathrm{d}\mathbf{N} \rangle. \tag{16}$$

But we shall not be concerned with III here.)

The first fundamental form plays a basic geometric role in the expression of lengths, angles, and surface area. A regular parametrized curve $\mathbf{x}(u(t), v(t))$, $a \le t \le b$, on the surface $\mathbf{x}$ has arc length

$$s = \int_a^b \mathrm{I}\left(\frac{\mathrm{d}u}{\mathrm{d}t}, \frac{\mathrm{d}v}{\mathrm{d}t}\right)^{1/2} \mathrm{d}t. \tag{17}$$

The angle between two vectors $d_1\mathbf{x} = \mathbf{x}_u d_1u + \mathbf{x}_v d_1v$ and $d_2\mathbf{x} = \mathbf{x}_u d_2u + \mathbf{x}_v d_2v$ on the tangent plane at a point on $\mathbf{x}$ is

$$\theta = \cos^{-1}\left[\frac{(\mathrm{d}_1u \ \mathrm{d}_1v)\begin{pmatrix} E & F \\ F & G \end{pmatrix}\begin{pmatrix} \mathrm{d}_2u \\ \mathrm{d}_2v \end{pmatrix}}{\mathrm{I}(\mathrm{d}_1u, \mathrm{d}_1v)^{1/2}\mathrm{I}(\mathrm{d}_2u, \mathrm{d}_2v)^{1/2}}\right]. \tag{18}$$

The surface area of a region $\mathbf{x}(R)$ on $\mathbf{x}$ is

$$A = \iint_R |\mathrm{I}|^{1/2} \, \mathrm{d}u \, \mathrm{d}v. \tag{19}$$

Because of these "metric" properties associated with I, it is often called the *metric tensor* of the surface.

The first fundamental forms of the helicoid and the catenoid are identical, with coefficients

$$E = G = b^2 \cosh^2 v,$$
$$F = 0. \tag{20}$$

As we shall see in section 6, this is not a mere concidence.

The second fundamental form determines qualitatively the nature ("shape") of the surface in neighbourhoods of points, depending on the determinant

$$|\mathrm{II}| = LN - M^2. \tag{21}$$

If $LN - M^2 > 0$ at a point p on $\mathbf{x}$, the point p is called an *elliptic point.* In the neighbourhood of an elliptic point the surface lies on one side of the tangent plane. If $LN - M^2 < 0$, the point is called a *hyperbolic point,* around which the surface lies on both sides of the tangent plane. The point p is a *parabolic point* if $LN - M^2 = 0$ and $\mathrm{II} \neq 0$, in which case a single line in the tangent plane at p lies on the surface in a neighbourhood of p. The point is a *planar point* if $\mathrm{II} = 0$. In this final case the degree of contact between the surface and the tangent plane is of higher order than in the preceding cases.

For the helicoid, the second fundamental form has

$$L = N = 0,$$
$$M = b, \tag{22}$$

hence $|\mathrm{II}| = LN - M^2 = -b^2 < 0 (b \neq 0)$. Thus all points on the helicoid are hyperbolic. The catenoid has

$$L = -b,$$
$$M = 0, \tag{23}$$
$$N = b,$$

hence $|\mathrm{II}| = -b^2 < 0$ also. Thus all points on the catenoid are hyperbolic as well.

Let p be a point on the surface $\mathbf{x}$ and let $\mathrm{d}\mathbf{x} = \mathbf{x}_u\, \mathrm{d}u + \mathbf{x}_v\, \mathrm{d}v$ be a (nonzero) vector in the tangent plane at p. Then the *normal curvature* at p in the direction $\mathrm{d}u : \mathrm{d}v$ is the number

$$\kappa_n = \frac{\mathrm{II}(\mathrm{d}u, \mathrm{d}v)}{\mathrm{I}(\mathrm{d}u, \mathrm{d}v)}. \tag{24}$$

It turns out that κ_n takes on distinct maximum and minimum values in two *perpendicular* directions. These are called the *principal directions,* and the

corresponding normal curvatures, κ_1 and κ_2, are called the *principal curvatures*. Their average

$$H = \tfrac{1}{2}(\kappa_1 + \kappa_2) \tag{25}$$

is the *mean curvature* at p, and their product

$$K = \kappa_1 \kappa_2 \tag{26}$$

is the *Gaussian curvature* at p. These have the alternate expressions

$$H = \frac{1}{2}\frac{EN - 2FM + GL}{EG - F^2}, \tag{27}$$

$$K = \frac{LN - M^2}{EG - F^2} = \frac{|\mathrm{II}|}{|\mathrm{I}|}. \tag{28}$$

It then then follows that a point on a surface is elliptic, hyperbolic or parabolic/planar according to $K > 0, < 0$, or $= 0$ at the point. We shall return to discussing the relevance of these concepts later.

5. Minimal Surfaces

A *minimal* surface is one for which the mean curvature H vanishes everywhere. A more physically intuitive characterization of a minimal surface is that it is a surface of *minimal area* bounded by a given closed curve in space; i.e. it is a solution of the classical Plateau's problem in the calculus of variations.

A regular parametrized surface $\mathbf{x}(u, v)$ is called *isothermal* if $\langle \mathbf{x}_u, \mathbf{x}_u \rangle = \langle \mathbf{x}_v, \mathbf{x}_v \rangle$ and $\langle \mathbf{x}_u, \mathbf{x}_v \rangle = 0$; i.e., a surface is isothermal if its first fundamental form is such that $E = G$ and $F = 0$. An isothermally parametrized surface is minimal if and only if its co-ordinate functions are harmonic, i.e. $\mathbf{x}_{uu} + \mathbf{x}_{vv} = \mathbf{0}$. One can easily check, then, that both the helicoid (2) and the catenoid (5) are isothermally parametrized and minimal. (Alternatively, one can verify directly that $H = 0$.) In fact the helicoid is the *only* ruled minimal surface other than the plane (a ruled surface is one which is generated by a family of straight lines; here, by the grid lines u = constant), while the catenoid is the *only* nonplanar minimal surface of revolution. Thus one could speculate that the uniqueness of the α- and β-structures of proteins might be a direct consequence of (or at least related to) the uniqueness of these two minimal surfaces. In any case, it is interesting to note that these regular protein structures lie on minimal surfaces.

The most frequently encountered minimal surfaces are probably soap films. The geometry of soap films is governed by the physical requirement

that the surface (potential) energy be minimal at equilibrium. The surface energy arises as a result of unbalanced attractive forces among molecules at the surface of the film. These unbalanced forces tend to minimize surface area, and hence lead to minimal surface energy. Since the three-dimensional structure of a protein is determined by the non-covalent forces among the amino acid residues of the sequence and between these residues and the solvent, an analogous physical principle may apply to proteins. Nonpolar groups tend to aggregate in order to reduce the interface area between them and the solvent (i.e. water), hence a soap-film type of surface minimization procedure (although in reverse "polarity") may be in operation here. Alternatively, the minimality could be due to the molecular forces that try to relieve stress in the regular structures (Salemme, 1981) while satisfying H-bond and backbone constraints. In any case, we have physically plausible explanations for the claim that at least the local regular structures of proteins lie on minimal surfaces. The native conformation of a protein molecule can then be considered as a collection of minimal surfaces, linked by turns and "random" coils which contain mainly polar groups and are exposed to the solvent. These "irregular" coils and turns also provide the necessary flexibility for the energetically best packing of the whole ensemble.

6. Conjugacy and Isometry

When two differentiable real-valued functions f and g, defined on an open subset of the plane, satisfy the Cauchy–Riemann equations $f_u = g_v$, $f_v = -g_u$, they are harmonic; hence f and g are called harmonic conjugate, and the complex function defined by $h(z) = h(u+iv) = f(u, v) + ig(u, v)$ is analytic. If the component functions of two isothermal parametrizations $\mathbf{x}$ and $\mathbf{y}$ of minimal surfaces are pairwise harmonic conjugate, i.e.

$$\mathbf{x}_u = \mathbf{y}_v, \qquad \mathbf{x}_v = -\mathbf{y}_u, \tag{29}$$

then $\mathbf{x}$ and $\mathbf{y}$ are called *conjugate minimal surfaces*, and the components of $\mathbf{x}+i\mathbf{y}$ are analytic functions. Furthermore, just as every harmonic function has a conjugate, every isothermal minimal surface has a conjugate, also obtained by solving the Cauchy–Riemann equations (29).

Given two conjugate minimal surfaces $\mathbf{x}$ and $\mathbf{y}$, the surface

$$\mathbf{z}(w) = (\cos w)\mathbf{x} + (\sin w)\mathbf{y} \tag{30}$$

is again minimal for all $w \in R$. Note that $\mathbf{z}(0) = \mathbf{x}$ and $\mathbf{z}(\pi/2) = \mathbf{y}$. All surfaces

of the one-parameter family (30) have the same first fundamental form with

$$E = G = \langle \mathbf{x}_u, \mathbf{x}_u \rangle (= \langle \mathbf{y}_u, \mathbf{y}_u \rangle = \langle \mathbf{x}_v, \mathbf{x}_v \rangle = \langle \mathbf{y}_v, \mathbf{y}_v \rangle),$$
$$F = 0. \tag{31}$$

Thus any two conjugate minimal surfaces can be joined through a one-parameter family of minimal surfaces, and the first fundamental form (hence the intrinsic geometry) of this family is independent of w. In other words, two conjugate minimal surfaces can be interconverted via an *isometry*.

The helicoid (2) *and the catenoid* (5), with $0 < u < 2\pi$, $-\infty < v < +\infty$, *are conjugate minimal surfaces.* The analytic manifold is

$$\mathbf{x} + i\mathbf{y} = (-ib \cos(u + iv), b \sin(u + iv), b(u + iv)). \tag{32}$$

The first fundamental form of the corresponding one-parameter family (30) has (recalling equation (20))

$$E = G = b^2 \cosh^2 v,$$
$$F = 0. \tag{33}$$

As the parameter w increases from 0 to $\pi/2$, while the parameter b remains constant, the helicoid is transformed isometrically into the catenoid (Fig. 1(a)–(d)). Furthermore, with a change in the value of the parameter b from 0·86 to 5·0, the helicoid on which the α-helix lies can be transformed into the catenoid which forms the β-barrel.

In section 3 we fitted the "average" β-sheet on a helicoid. But a quick glance through the sketches in Richardson's (1981) survey will convince the reader that the β-sheet is not a very "regular" structure, assuming different degrees of twists and sizes in different proteins. The β-sheet is better described by an intermediate $\mathbf{z}(w)$ (e.g. Fig. 1(b), (c)) than by the "pure" helicoid $\mathbf{x} = \mathbf{z}(0)$, with the parameter w controlling the "twist" and the parameter b controlling the "size". Note that as the α-helicoid is transformed into the β-catenoid, the value of b does not necessarily increase monotonically with w. In fact b may assume some negative intermediate values to represent the "left-handed" β-sheets. The family of minimal surfaces (30), when $\mathbf{x}$ is the helicoid and $\mathbf{y}$ the catenoid, then provides a universal description of the regular secondary structures of proteins.

The second fundamental form of this family depends on w:

$$L = -b \sin w,$$
$$M = b \cos w, \tag{34}$$
$$N = b \sin w,$$

but

$$|\mathrm{II}| = LN - M^2 = -b^2 < 0 \tag{35}$$

for all values of w. Thus all points on each $\mathbf{z}(w)$ are hyperbolic points. The principal, mean, and Gaussian curvatures are all independent of w:

$$\kappa_1, \kappa_2 = \frac{\pm 1}{b \cosh^2 v}, \tag{36}$$

$$H = 0 \text{ (hence minimal surfaces)}, \tag{37}$$

$$K = \frac{-1}{b^2 \cosh^4 v} < 0. \tag{38}$$

Thus the isometric helicoid-to-catenoid transformation does not alter the nature of the surfaces.

7. Geodesics

On any surface there are intrinsic curves, called *geodesics*, which have zero geodesic curvature κ_g everywhere. κ_g can be defined by the equation

$$\kappa^2 = \kappa_n^2 + \kappa_g^2, \tag{39}$$

hence a geodesic has

$$\kappa = \pm \kappa_n. \tag{40}$$

(Note that κ_n at a point of a curve on a surface is κ_n evaluated at that point in the direction of the tangent to the curve.) Geodesics are generalizations of straight lines in Euclidean space because they are "shortest paths", since an arc of minimum length on a surface joining two points is necessarily a geodesic. A more illuminating physical characterization is that a geodesic on a surface is an energy-critical curve. The "energy" of a smooth curve $\boldsymbol{\gamma}:[a, b] \to \mathbf{x}$ on the surface $\mathbf{x}$ is the quantity

$$\varepsilon(\boldsymbol{\gamma}) = \tfrac{1}{2} \int_a^b \langle \boldsymbol{\gamma}'(s), \boldsymbol{\gamma}'(s) \rangle \, ds \tag{41}$$

which is analogous to the expression of kinetic energy in mechanics. By "energy-critical" we mean that the first variation of $\varepsilon(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$ is zero. Thus a geodesic can be considered as a curve with minimal energy. This makes geodesics on minimal surfaces doubly interesting because they are minimal kinetic energy curves on minimal potential energy surfaces.

One can calculate what the geodesics on a surface are by solving the so-called geodesic equations. Since these are a set of differential equations

in which the coefficients depend only on the first fundamental form, geodesics are a part of the intrinsic geometry of surfaces. In particular, the helicoid (2) and the catenoid (5) (and the whole family (30)) have the *same* set of geodesics. Their equations turn out to be

$$\begin{aligned} \frac{d^2u}{ds^2} &= 0, \\ \frac{d^2v}{ds^2} + \tanh v\left(\left(\frac{dv}{ds}\right)^2 - \left(\frac{du}{ds}\right)^2\right) &= 0, \end{aligned} \tag{42}$$

(where s is the arc-length parameter). The solutions are

$$\begin{aligned} \sinh v(s) &= A \sinh (u(s) + B), \\ u(s) &= Cs + D, \end{aligned} \tag{43}$$

for constants A, B, C, D. For small $|u|$ and $|v|$ values, to a first approximation the geodesics are

$$v = A(u + B). \tag{44}$$

In other words, the helicoid–catenoid family has images of straight line segments from the (u, v)-parameter space as (approximate) geodesics. Recalling the descriptions of the α-helices and β-strands, we then have the remarkable conclusion that these *regular backbone conformations of proteins are geodesics on minimal surfaces*, and that *these surfaces can be transformed into one another in a natural way*.

It is important to note that under this α-to-β transformation the parameter b varies along with the increasing parameter w. Also, a β-strand, say $\boldsymbol{\beta}$: $v = 1{\cdot}4u$ (equation (6)), is not the direct image of an α-helix, $\boldsymbol{\alpha}$: $v = 1{\cdot}7$ (equation (3)). The situation is best represented by a continuous deformation from the latter curve to the fomer via a one-parameter family of geodesic curves

$$\{\boldsymbol{\gamma}(w) \subset \mathbf{z}(w) : 0 \leq w \leq \pi/2\}, \tag{45}$$

such that

$$\boldsymbol{\gamma}(0) = \boldsymbol{\alpha} \subset \mathbf{z}(0) = \mathbf{x} \tag{46}$$

and

$$\boldsymbol{\gamma}(\pi/2) = \boldsymbol{\beta} \subset \mathbf{z}(\pi/2) = \mathbf{y}, \tag{47}$$

where $\mathbf{x}$ is the helicoid with $b = 0{\cdot}86$ and $\mathbf{y}$ is the catenoid with $b = 5{\cdot}0$. For intermediate values of w, $\boldsymbol{\gamma}(w)$ could represent a β-strand on a twisted sheet with an appropriate value for b.

8. Asymptotic Curves

When $\kappa_g = 0$ in equation (39), one obtains the class of geodesics on a surface. It is then natural to investigate the situation when the κ_n "component" is zero, i.e. when

$$\kappa = \pm\kappa_g. \tag{48}$$

Since $\kappa_n = \mathrm{II}(\mathrm{d}u, \mathrm{d}v)/\mathrm{I}(\mathrm{d}u, dv)$ (equation (24)) and I is positive definite, $\kappa_n = 0$ precisely in the direction for which $\mathrm{II} = 0$. This is called an *asymptotic direction*. At an elliptic point there are no asymptotic directions; at a hyperbolic point there are two distinct asymptotic directions; at a parabolic point there is one asymptotic direction; and at a planar point every direction is asymptotic.

A curve on a surface is an *asymptotic curve* if its tangent at every point is an asymptotic direction at that point. In particular, in a neighbourhood of a hyperbolic point on a surface there exist two distinct families of asymptotic curves. Since our family (30) of minimal surfaces from the helicoid to the catenoid contains only hyperbolic points, this is the only case we consider here. The asymptotic directions of this family are

$$\mathrm{d}u : \mathrm{d}v = \cos\frac{w}{2} : \sin\frac{w}{2}, \; -\sin\frac{w}{2} : \cos\frac{w}{2}. \tag{49}$$

Thus for the helicoid ($w = 0$) they are in the direction of the u- and v-coordinate curves, while for the catenoid ($w = \pi/2$) the asymptotic directions make 45° angles with the u- and v-parameter axes. Asymptotic curves on the helicoid are, then, the helices and straight lines (i.e. the grid lines in Fig. 1(a)). It is interesting that they represent the α-helices and β-strands-on-twisted-sheets, respectively. Asymptotic curves on the catenoid make angles of ±45° with the barrel axis, and, significantly, this is about the correct twist angle value for the strands in an "average" β-barrel.

Asymptotic curves have a physical significance as well. While $\kappa_g = 0$ gives the "space-critical" curves of geodesics, $\kappa_n = 0$ corresponds to the "time-critical" curves. This is why both geodesics and asymptotic curves are considered as generalizations of straight lines: a straight line, for which $\kappa = 0$, is both space-critical and time-critical. Thus equation (39), $\kappa^2 = \kappa_n^2 + \kappa_g^2$, can be interpreted as a decomposition of the curvature into its "time and space components". Since the regular protein conformations turn out to be asymptotic curves (in addition to being approximate geodesics), we suggest that this results from a "generalized Fermat's principle", that "energy propagation along protein molecules takes paths of critical time". Many biological phenomena can be associated with energy

transfer on protein molecules, notably that of the excitation of active sites on enzymes (Somorjai, 1978). This problem in bioenergetics is related to the motion of solitons (Lamb, 1975; Davydov, 1977). We will discuss elsewhere the relations between asymptotic curves and soliton motions, and their implications for enzyme functions.

9. Implications for Protein Folding

The morphogenetic process of protein folding, which generates the specific three-dimensional geometric form (i.e. the tertiary structure), is regarded most often in physicochemical terms as a transition to a state of minimal free energy, because of the spontaneity of the process and the stability of the native conformation. A basic dogma of molecular biology is that the tertiary structure is completely determined by the primary structure, the amino acid sequence. Thus in principle one should be able to compute the tertiary structure of a protein when the primary sequence is known. Despite intensive effort, however, the free energy minimization approach to the protein folding problem has not proven fruitful, mainly because of the enormous complexity of the total potential function involving the interactions of thousands of atoms in space, and the technical difficulties attending the subsequent minimization of such a function (the "curse" of getting trapped in local minima in the high-dimensional energy hyperspace). Furthermore, even the more basic question of whether the native structure constitutes a local or a global energy minimum (or neither) has not received a satisfactory answer. This strongly suggests that free energy minimization is not the correct practical approach for elucidating the folding pathway.

The most intricate problem in folding dynamics is perhaps the characterization of intermediate states. Much effort has been invested to find the so-called nuclei and domains which serve as scaffolds for further folding. It has commonly been postulated that early in the folding process, among the fluctuating conformational states local elements of secondary structure form and survive for a significant length of time (Yaron *et al.*, 1971; Ralston & DeCoen, 1974; Chou & Fasman, 1974). We propose that after the creation of folding "nuclei", the process continues with the formation of several local minimal surfaces. (It is particularly tempting to paraphrase Lim, 1980, and assume that "initially there were only fluctuating α-helices".) Then the interactions between these helicoidal surfaces (and the changing protein environment) vary the size parameter b and the positions of the geodesics, and these in turn change the twist parameter w. Thus, the formation of β-sheets and eventually β-barrels do not require postulating alternative pathways as is now done (e.g. Finkelstein & Ptitsyn, 1980).

As an illustration, consider the parametrized geodesic, equation (44): when $A > 0$, it corresponds on the helicoid to a conical helix with increasing radius as it winds around the axis, and corresponds on the catenoid to a β-strand which has increasing "height" with twist (as is the case for the strands in Fig. 2); when $A < 0$, one has a conical helix with decreasing radius and a β-strand pointing in the opposite direction. Now consider an α-helix consisting of 12 residues. It can be parametrized on the helicoid as

$$\alpha: v = 1{\cdot}7, \qquad 0 \leq u \leq 19{\cdot}2 (= 1100°). \tag{50}$$

The α-carbons appear when

$$u = 1{\cdot}75j, \qquad j = 0, 1, \ldots, 11 \tag{51}$$

(Fig. 3(a)). Suppose the segments 1–4 and 9–12 interact. They would deform into a pair of conical helices in which the radius increases on one and decreases on the other (Fig. 3(b)). This pair, under a helicoid-to-catenoid transformation, would then result in an antiparallel pair of β-strand connected by a four-residue turn (Fig. 3(c)).

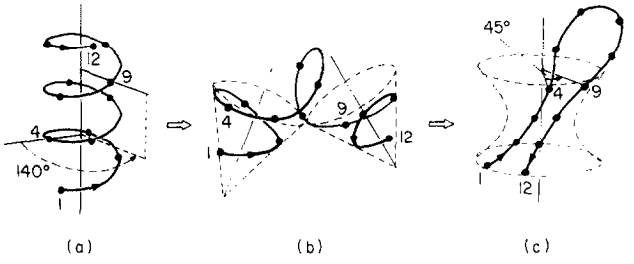

FIG. 3. Transformation of α-helices into antiparallel β-strands.

The α-carbons 1 and 4 on the helicoid are separated by a "phase angle" of $u = 3 \times 1{\cdot}75 = 300°$, while on the catenoid they are separated by a "phase angle" of $u = 100°$ (see section 3). Thus there is a phase reduction by a factor of 1/3. Now α-carbons 4 and 9 are separated by $u = 5 \times 1{\cdot}75 = 500° \equiv 140°$ on the helicoid, hence they should be separated by $u = 140°/3 \approx 45°$ on the catenoid. This is indeed the phase value of strand separation obtained in section 3. So it appears that the α-to-β transformation is quantitatively correct as well.

Parallel β-strands can also be derived from interacting helices in a similar fashion (Fig. 4). Here we require an initial helix of 16 residues in length, and the interaction is between the 1–4 and 13–16 segments. The longer "linkage" 5–12 makes the resultant longer "loop" possible. Note that the

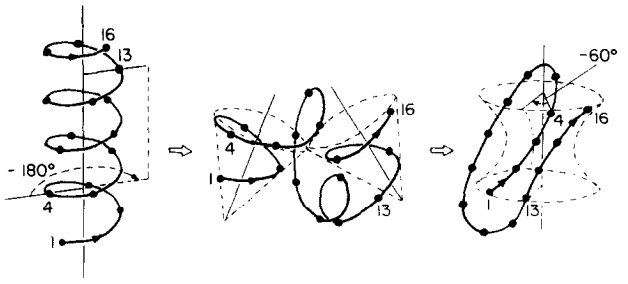

FIG. 4. Transformation of α-helices into parallel β-strands.

phase separation of α-carbons 4 and 13 is $u = 9 \times 1{\cdot}75 = 900° \equiv -180° \to -180°/3 = -60°$, again approximately correct on the catenoid.

Furthermore, the illustrative calculations presented, in which a four-residue turn was produced linking the antiparallel strands, and an eight-residue longer loop linking two parallel strands, can be readily and reasonably modified to include fewer or larger number of intervening "irregular" residues. This is most readily accomplished by changing the local characteristics of the helicoid; the important variable is the relative phase angle.

The above α-to-β transition can naturally be extended to more than two structures, and to a mixture of parallel and antiparallel strands. It is interesting to note that the surface area of the eight-stranded β-barrel (calculated by the surface integral

$$\iint \sqrt{EG - F^2}\, du\, dv) \tag{52}$$

which contains about 30 amino acid residues, is almost exactly the same as the cylindrical surface area determined by an eight-turn α-helix (both being about 600 Å^2). Thus one can conclude that under the α-to-β transition, the surface area "exposed to the outside" is approximately conserved.

Since the native conformation of a protein molecule is considered as an ensemble of linked minimal surfaces, and since the α-to-β (helicoid-to-catenoid) interconversion depends solely on the values of the parameters b and w, it follows that the overall scheme of protein folding can be formulated as a dynamical problem in terms of the two control parameters b and w, which can be treated as functions of time.

10. Vector Fields on Proteins

The parameters b and w can and should also be considered as functions of the arc-length parameter s along the curve of the protein backbone.

The "size" parameter b reflects the "bulkiness" of the amino acid residues while the parameter w could be a measure of the α- and β-forming tendencies of the amino acids (recalling that $w = 0$ gives a helicoid and $w = \pi/2$ a catenoid). Thus the problem of prediction of protein structure from the amino acid sequence can be formulated in terms of vector fields on curves: to each point $\boldsymbol{\gamma}(s)$ on the protein backbone attach the "vector" $(b(s), w(s))$, and from the vector field (b, w) derive the curvature and torsion of the space curve, which then determine the three-dimensional conformation of protein backbone. The implication $(b, w) \rightarrow (\kappa, \tau)$ is equivalent to the statement that from the *equation of state*

$$\Phi(b, w, \kappa, \tau) = 0 \tag{53}$$

(where the arguments are functions of both the arc-length s and time t), κ and τ can be solved in terms of b and w; i.e. b and w are the *fundamental variables* while κ and τ are the *derived variables*. The problem is then reduced to finding the exact form of the functional operator Φ (which may involve derivatives and integrals of the arguments), and is under investigation.

The (b, w) vector field on proteins can be generalized as follows. Let $U = U(s)$ be the *base curve* representing a protein backbone. By attaching to each point of $U(s)$ a vector space $X(s)$ we obtain what is known as a *fibre space* (in fact a *vector bundle*). If we choose from each $X(s)$ a vector $\mathbf{x}(s)$, we obtain a vector field on the curve U, i.e. a dynamics on U. We can represent U with the vector field $\mathbf{x}$ by a "ribbon" in space, the width of which at the point $U(s)$ is $|\mathbf{x}(s)|$. Actually this "ribbon" is nothing but a portion of the ruled surface

$$\mathbf{y}(u, v) = U(u) + v\mathbf{x}(u) \tag{54}$$

with straight line generators $v\mathbf{x}(u_0)$, $u_0 = \text{constant}$. In this formulation the vector field $\mathbf{x}$ can be chosen to reflect any of the known properties of the amino acids, such as bulkiness, polarity, hydrophobicity, pK, etc. These properties can then be used to determine the shape of the backbone U.

The relations between the curves

$$U(s) = \mathbf{y}(s, 0) \tag{55}$$

and

$$V(s) = \mathbf{y}(s, 1) \tag{56}$$

would also be of interest. When $\mathbf{x}$ is chosen to locate the β-carbons, for example, the *winding number* of V on U reflects the orientations of the side chains of the protein. When $\mathbf{x}$ is chosen to represent the different sizes

of the residues, the relative κ and τ of the two curves characterize the packing of the molecule. These are just a few of the many possible representations of protein patterns with a vector field approach.

11. Epilogue: Protein Patterns on Fibre Spaces

The ideas of fibre spaces apply to the study of proteins on a higher hierarchical level as well. We end this essay on differential geometry of proteins with a mathematical formulation of a unified approach to pattern generation, as suggested in Rosen (1981). We hope the powerful abstract formalism developed below will provide the appropriate setting in which to study the mathematical biology of proteins.

Let A be the "alphabet" of twenty letters representing the twenty amino acids found in proteins, together with an additional letter "0" representing an "empty" element. Then a finite sequence of amino acids is equivalent to an infinite sequence $u = \{u_j\}$ in A with "finite support"; i.e., $u_j = 0$ for all but finitely many j. Let U be the collection of all sequences in A with finite support; then the collection of all proteins forms a subset of U. U can be made a metric space by using any one of the sequence space metrics.

Let X be the collection of all space curves. Then associated with each protein $u \in U$ there is a set $X_u \subset X$ of all "physically accessible conformation" of U. The set $X \sim X_u$ would contain, for example, all the curves with lengths not equal to that of u, all the non-simple curves, and so on. One interesting problem in the study of proteins is to determine exactly what each X_u is; i.e. what the physically accessible conformations of a specific protein are under all possible environments. Each set X_u contains a special curve n_u which represents the native conformation of the protein u in physiological environment. The problem of protein folding is then equivalent to the convergence problem $x_u(k) \to n_u$ of a net $\{x_u(k)\}_k$ in X_u, where the index k can be taken as either continuous or discrete. The convergence here is with respect to the topology on $X_u \subset X$, which is probably best taken as the topology of uniform convergence on the function space of pairs of continuous functions (κ, τ) (recalling that the set X is equivalent to the set of all pairs (κ, τ)).

Now consider the cartesian product fibre space $B = X \times U$ equipped with the projection p onto its factor (base) space U. This projection has the property that for each $u \in U$, the inverse image $p^{-1}(u)$ is a copy of the fixed fibre X over u, which has a distinguished subset X_u, and which in turn contains a distinguished member n_u (Fig. 5).

One of the basic notions in the study of fibre spaces is that of the cross-section, which is simply a mapping $f: U \to B$ with the property that

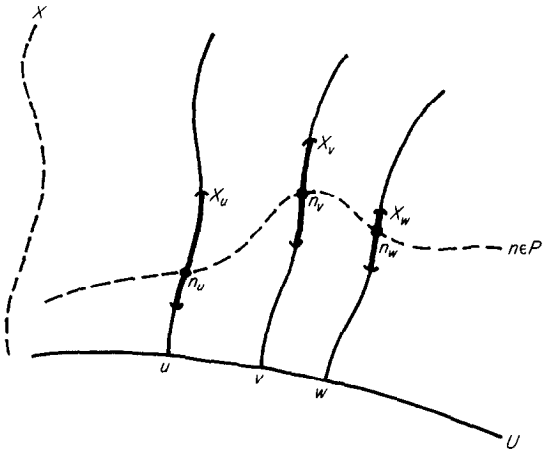

FIG. 5. Fibre space B.

$p \circ f = 1_U : U \to U$. In other words, a cross section selects for each $u \in U$ a point in the fibre X over u. For example, the cross-section "native conformation" is a mapping $n : U \to B$ sending u to $n_u \in X_u$. If we consider the base space U of all proteins as the domain over which patterns are formed, and the fibre X as the set of states which may be assigned to each point of that domain, then a pattern is obtained by assigning to each point of U a unique physically accessible state from X. This of course says, then, that pattern and cross-section are the same concepts. Let us denote the space of cross-sections (patterns) by P.

The problems of pattern recognition and pattern generation are then most appropriately formulated in the space P. The folding of proteins is a dynamics in P, the convergence of a net $x(k) \to n$. This is actually the approach we have taken because we considered the structure and transformation of protein patterns independent of any particular protein; i.e. we studied the space P instead of a specific X_u.

A member of P can also be interpreted as a "similarity deformation" among proteins. This is like playing a D'Arcy Thompson type of "deformation game" at the molecular level. We can also think of continuous deformation of binding sites of one protein to that of another under this interpretation, mapping "function" to "function", consistent with the postulates of Rashevsky's (1960) relational biology. This concept of "corresponding states" can be more precisely formulated in terms of the equation of state (53). Let the functional parameters (b, w) correspond to a protein u and (b', w') to a protein v. Then we say the curve defined by

(κ, τ) describing u is *similar* to the curve defined by (κ', τ') describing v if

$$\Phi(b, w, \kappa, \tau) = 0 \tag{57}$$

and

$$\Phi(b', w', \kappa', \tau') = 0. \tag{58}$$

Then the relation "similarity" in B is an equivalence, the equivalence classes of which are precisely what we call patterns.

The concept of a fibre space is one of the keystones of contemporary mathematics, appearing in connection with the theory of partial differential equations, algebraic topology, and especially differential geometry. These various areas of "pure" mathematics have proven invaluable in the development of the mathematical physics of quantum mechanics, relativity, and field theory. Fibre spaces seem to be the underlying foundation of them all. We saw in this paper how the simplest ideas of differential geometry apply in the structural and dynamical representation of protein patterns. The more general mathematical theory of fibre spaces could undoubtedly also be translated into biological terms.

REFERENCES

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

CHOU, P. Y. & FASMAN, G. D. (1974). *Biochemistry* **13,** 211.

DAVYDOV, A. S. (1977). *J. theor. Biol.* **66,** 379.

EISENHART, L. P. (1909). *A Treatise on the Differential Geometry of Curves and Surfaces.* London: Ginn.

FINKELSTEIN, A. V. & PTITSYN, O. B. (1976). *J. mol. Biol.* **103,** 15.

LAMB, G. L. (1976). *Phys. Rev. Lett.* **37,** 235.

LIM, V. I. (1980). In *Protein Folding* (R. Jaenicke, ed.), Amsterdam: Elsevier, p. 149.

PTITSYN, O. B. & FINKELSTEIN, A. V. (1980). In *Protein Folding* (R. Jaenicke, ed.), Amsterdam: Elsevier, p. 101.

RALSTON, E. & DECOEN, J. L. (1974). *J. mol. Biol.* **83,** 393.

RASHEVSKY, N. (1960). *Mathematical Biophysics* (3rd edition). New York: Dover.

RICHARDSON, J. S. (1981). *Adv. Protein Chem.* **34,** 167.

ROSEN, R. (1981). *Bull. math. Biol.* **43,** 513.

SALEMME, F. R. (1981). *J. mol. Biol.* **146,** 143.

SCHULZ, G. E. & SCHIRMER, R. H. (1979). *Principles of Protein Structure.* New York: Springer-Verlag.

SOMORJAI, R. L. (1978). In *Molecular Movements and Chemical Reactivity* (R. Lefever & A. Goldbeter, eds.) New York: Wiley.

WEATHERBURN, C. E. (1927). *Differential Geometry.* Cambridge: Cambridge University Press.

YARON, A., KATCHALSKI, E., BERGER, A., FASMAN, G. D. & SOBER, H. A. (1971). *Biopolymers* **10,** 1107.