PROTEINS: Structure, Function, and Genetics 6:46–60 (1989)

Describing Protein Structure: A General Algorithm Yielding Complete Helicoidal Parameters and a Unique Overall Axis

Heinz Sklenar,[1] Catherine Etchebest,[2] and Richard Lavery[2]
[1]Central Institute of Molecular Biology, Academy of Sciences of the GDR, Robert Rössle Strasse 10, DDR-1115 Berlin Buch, German Democratic Republic, and [2]Institut de Biologie Physico-Chimique, 13 rue Pierre et Marie Curie, Paris 75005, France

ABSTRACT **We present a general and mathematically rigorous algorithm which allows the helicoidal structure of a protein to be calculated starting from the atomic coordinates of its peptide backbone. This algorithm yields a unique curved axis which quantifies the folding of the backbone and a full set of helicoidal parameters describing the location of each peptide unit. The parameters obtained form a complete and independent set and can therefore be used for analyzing, comparing, or reconstructing protein backbone geometry. This algorithm has been implemented in a computer program named P-Curve. Several examples of its possible applications are discussed.**

Key words: macromolecular conformation, protein folding, helical axis, secondary structure

INTRODUCTION

The increasing number of well-resolved protein structures available today poses the problem of how the conformations of these often very complex macromolecules can best be described. The simplest and most common solution to this problem is based on the calculation of the backbone and side chain torsion angles. In the case of the backbone, a Ramachandran plot[1] of ϕ/ψ torsions can subsequently indicate roughly the zones involved in recognizable secondary structure motifs such as α-helices or β-sheets. However, this approach cannot easily describe the folding of the protein backbone and is not very useful for finer studies such as the comparison of homologous structures, the description of turns, or the exact delimitation of secondary structures and detection of their internal distortions.

A number of partial solutions to these different problems have been proposed,[2–6] but no completely satisfactory description of protein backbone structure has yet been put forward. One attempt at an overall description has been made by Rackovsky and Scheraga[7–9] using differential geometry to obtain a continuous space curve based on the positions of successive C-α backbone atoms. This description, however, cannot be considered an ideal representation of folding since it remains curved and twisted even within secondary structure zones and thus renders the identification of real backbone kinks or turns difficult. Moreover, the resolution of the method for detecting secondary structures is limited by the fact that a minimum of four α carbons is necessary to obtain the parameters describing the form of the space curve.

The only way to overcome these difficulties appears to be the description of the protein backbone in terms of a rigorous definition of a generalized helical axis. This solution has the advantage of leading to a very simple and clear description of folding, since the backbones of all secondary structure zones will be reduced to more or less straight lines and true kinks or turns will be clearly visualized. Moreover, this approach, which must be based on the spacial location of successive peptides in the protein backbone, will enable a complete parameter set to be obtained for each monomeric unit.

Two attempts to obtain at least approximate helical axes for proteins or protein fragments have already been made.[10–12] These approaches are both based on the least-squares fitting of short probe helices to successive groups of backbone atoms within the proteins studied. In the work of Barlow and Thornton[12] an approximate overall helical axis is then defined by linking together successive locations on the probe axes. The disadvantage of this work is first, its approximate nature, the results obtained depending on both the length and the conformation of the probe helix employed. In addition, while regular secondary structure zones can be treated with this technique, extension of the analysis to irregular coil regions or sharp turns is much less obvious.

This situation clearly hinders a deeper under-

Received February 2, 1989; revision accepted May 30, 1989.

Address reprint requests to R. Lavery, Institut de Biologie Physico-Chimique, 13 rue Pierre et Marie Curie, Paris 75005, France.

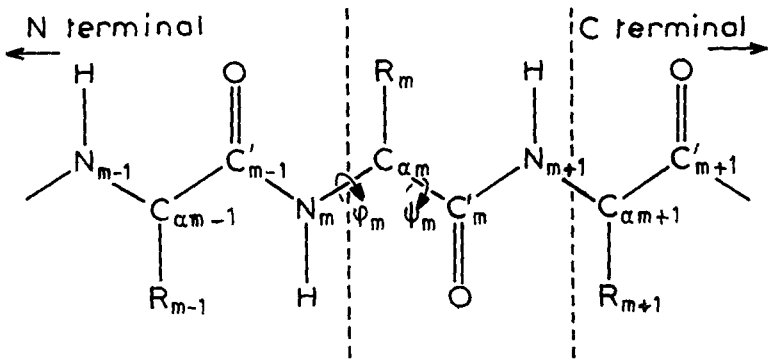

Fig. 1. Division of the polypeptide backbone for helical analysis. The *m*th unit is indicated by the dotted lines and comprises the peptide plane defined by C'_m and N_{m+1}.

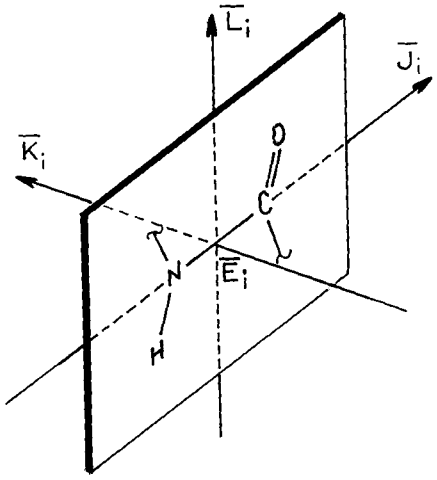

Fig. 2. Definition of the peptide fixed axis system $\bar{J}\bar{K}\bar{L}$. $\bar{E}$ is the mid-point of the peptide bond and the plane shown corresponds to the mean plane of the peptide group.

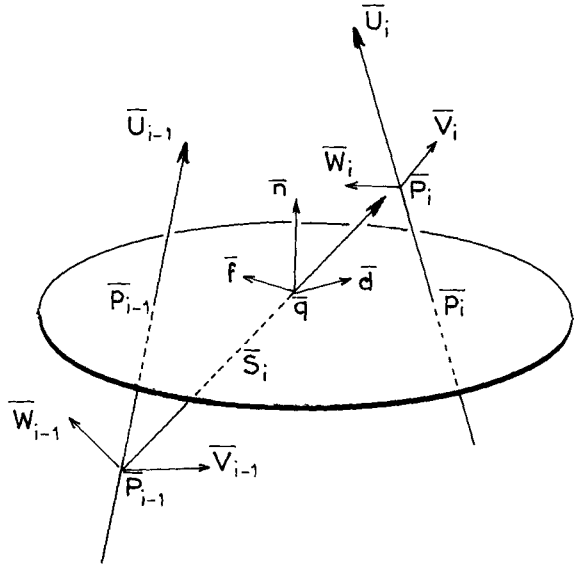

Fig. 4. Construction of the mean plane used for calculating interpeptide parameters.

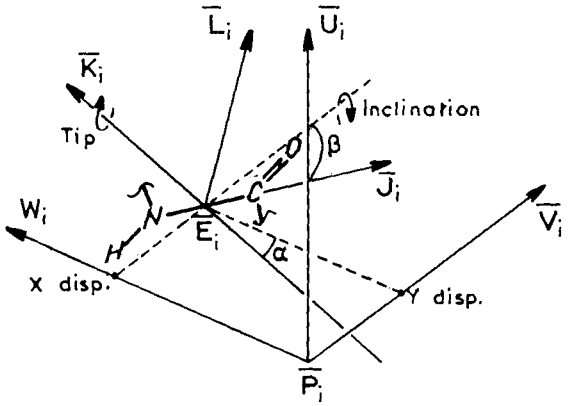

Fig. 3. Definition of the helicoidal parameters (X displacement, Y displacement, inclination, and tip) which relate the peptide axis system *JKL* to the local helical axis system $\bar{V}\bar{W}\bar{U}$.

standing of protein folding and limits the amount of data which can easily be extracted from data banks of protein structure. We would like to propose a possible solution to this problem by describing a general algorithm which can be used to obtain a complete and unique helicoidal description of any protein backbone. The algorithm we will describe is an adaption of the approach we have recently developed for describing nucleic acid structure.[13,14] It leads to a unique curved helicoidal axis and a full set of helicoidal parameters which locate each peptide unit with respect to this axis and with respect to its neighbors.

Our method is a natural extension of the definitions used in helical descriptions of regular polymers to the case of irregular systems. The basis of this approach is the definition of a function which

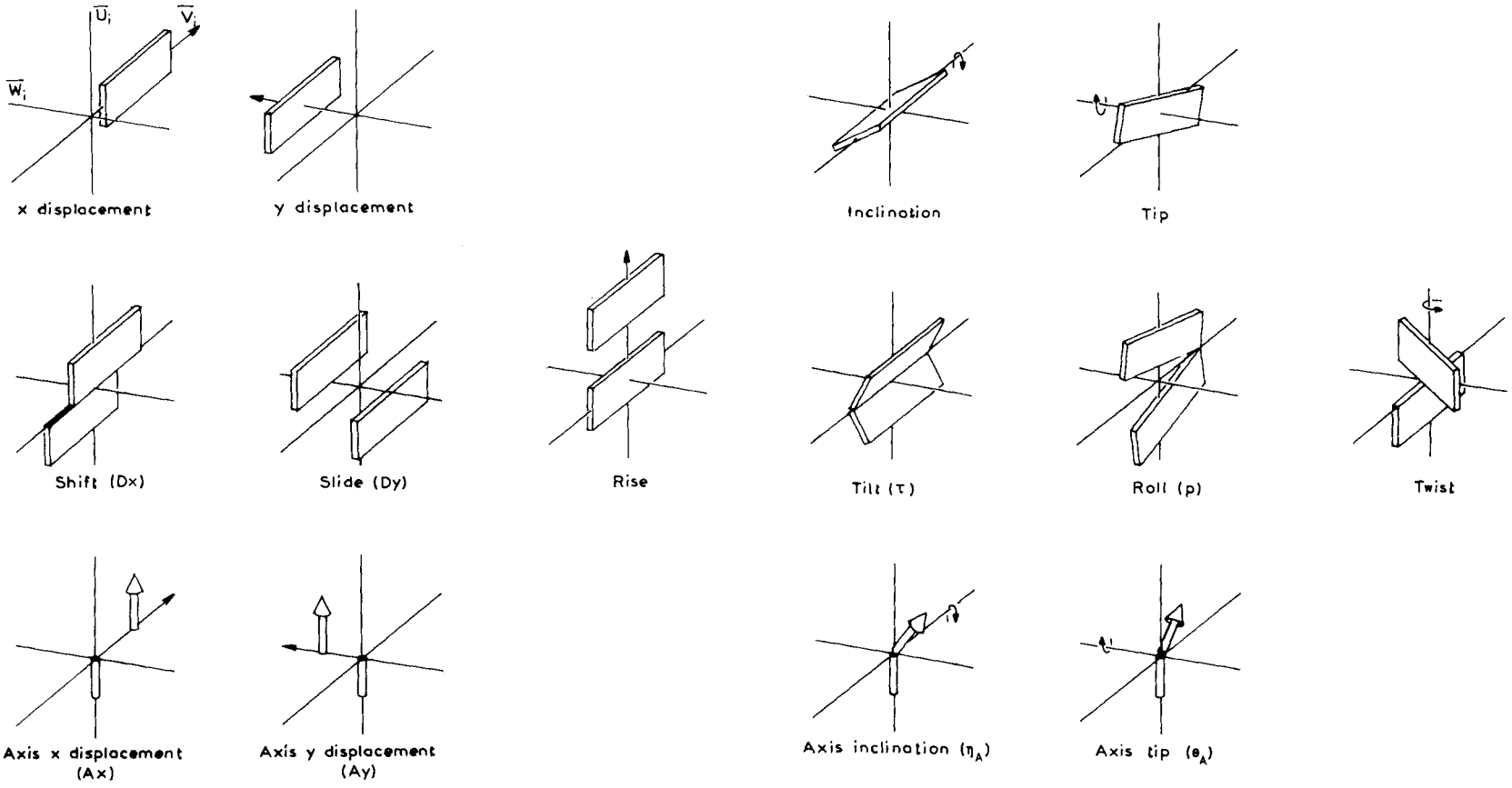

Fig. 5. Schematic illustration of the global helicoidal parameters defined by the P-Curve algorithm.

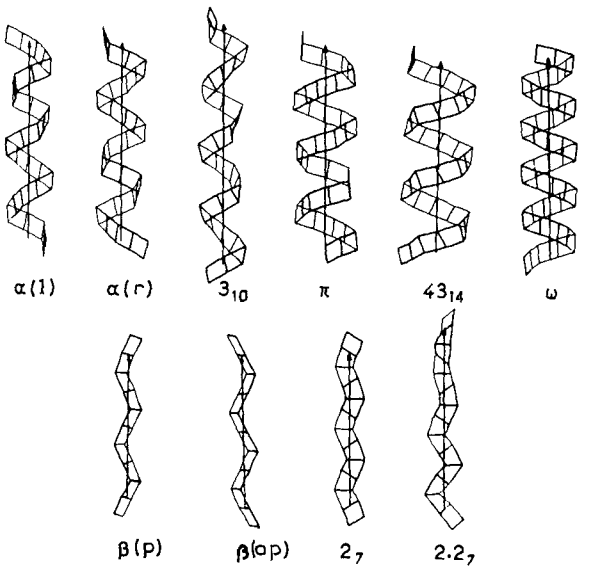

Fig. 6. Ribbon diagrams of symmetric polypeptide structures. In each case the helical axis of the structure is shown with an arrowhead indicating the C-terminal end (for references to data used in construction see Table II).

describes departure from perfect helical symmetry in terms of the curvature of the axis describing the polymer and in terms of changes in the position of successive monomers with respect to this axis. Minimization of this geometric function yields the generally curved axis of the polymer and provides a unique helical description where both types of irregularity have been "smoothed" in an optimal least-squares sense. Since the function is constructed so as to take into account simultaneously the position of all the monomeric units making up the polymer, the final description of any one of these units thus depends on the position of its neighbors. This leads to a much more coherent view of the overall conformation than that obtained with any purely local parameters such as the backbone torsion angles.

Once the analysis has been performed for a given protein, a great deal of information on its conformation can be obtained. It now becomes possible to rigorously define the location of secondary structures and to detect any deformations or anomalies that they may contain. One example of this type of application is described for the case of a small protein. Using the overall protein axis, it is also possible to rapidly compare related protein structures either graphically or numerically as part of a data base

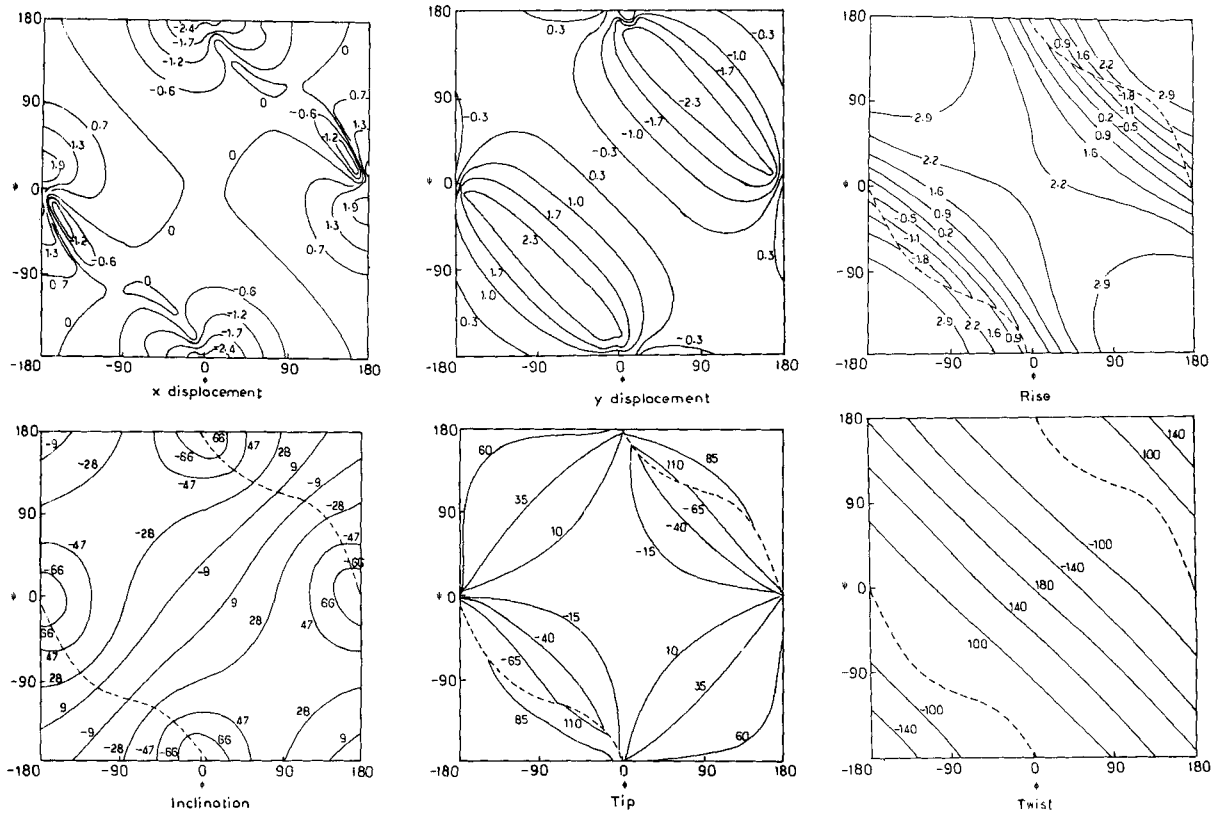

Fig. 7. Phi/psi plots of the helicoidal parameters of symmetric polypeptide structures. Values are in angstroms for translational parameters and in degrees for rotational parameters. Dotted lines indicate discontinuities.

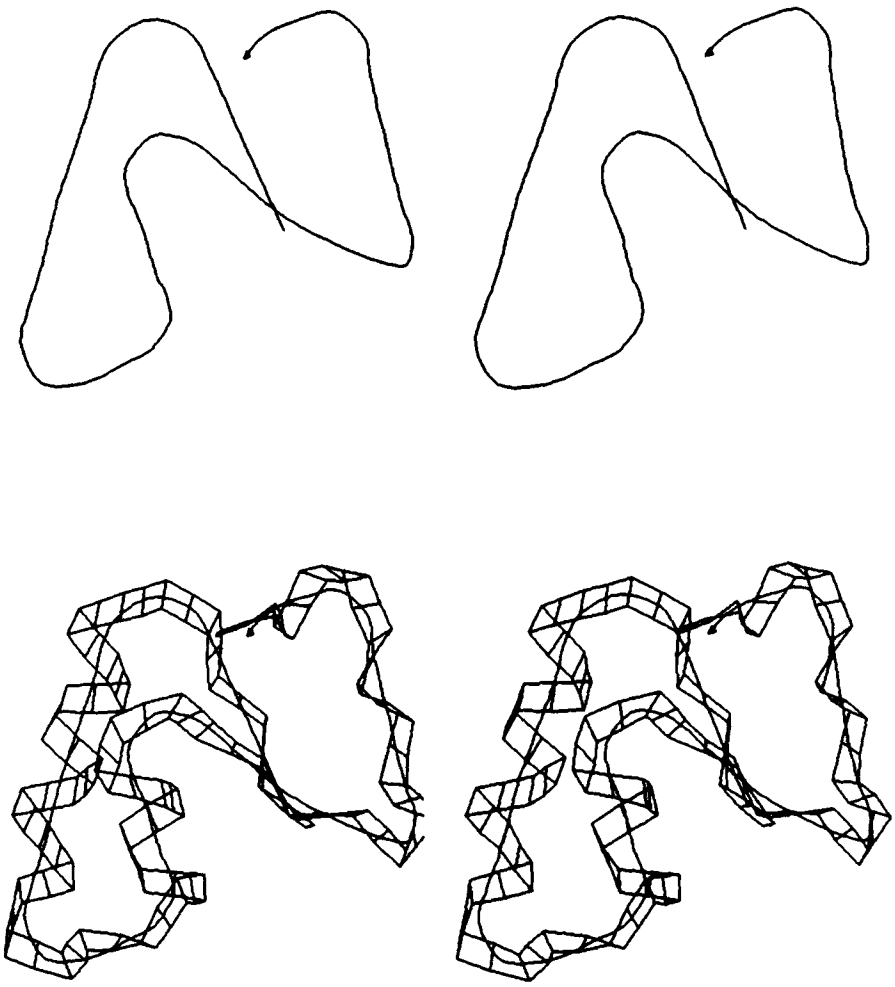

Fig. 8. Stereodiagrams of the helical axis alone (above) and of the backbone ribbon and the helical axis (below) resulting from the analysis of crambin (coordinates from ref. 19).

search program. Moreover, since this axis is a well-defined and concise alternative to the C-α backbone curve for describing the path of the polymer chain, the apparatus of differential geometry can be used to describe its shape in the same way as proposed by Rackovsky and Scheraga.[7] We will not make such an analysis presently, but will consider this possibility in our forthcoming studies.

MATERIALS AND METHODS

The first step in defining the helicoidal structure of a polymer is to choose the structural repeating element to be used. For a protein the most natural choice seems to be the successive peptide groups. For the purposes of our analysis we will consequently divide the protein chain as shown in Figure 1. We must subsequently associate a fixed axis system with each repeating element, which will serve to define its position in space. This axis system is shown in Figure 2. It is centered on the middle of the peptide bond ($\bar{E}$) and defined by three mutually perpendicular unit vectors ($\bar{J},\bar{K},\bar{L}$). The first of these vectors, $\bar{J}$, is simply the peptide bond vector in the direction N–C'. $\bar{L}$ lies in the mean plane of the peptide group and points toward the side carrying the carbonyl group. $\bar{K}$ is perpendicular to the mean peptide plane, and is defined by the vector product $\bar{L} \times \bar{J}$.

Next, it is necessary to define the position of each repeating element with respect to a local helical axis system. This requires four variables, two translations and two rotations. These variables are shown in Figure 3. The helical axis system is centered at point

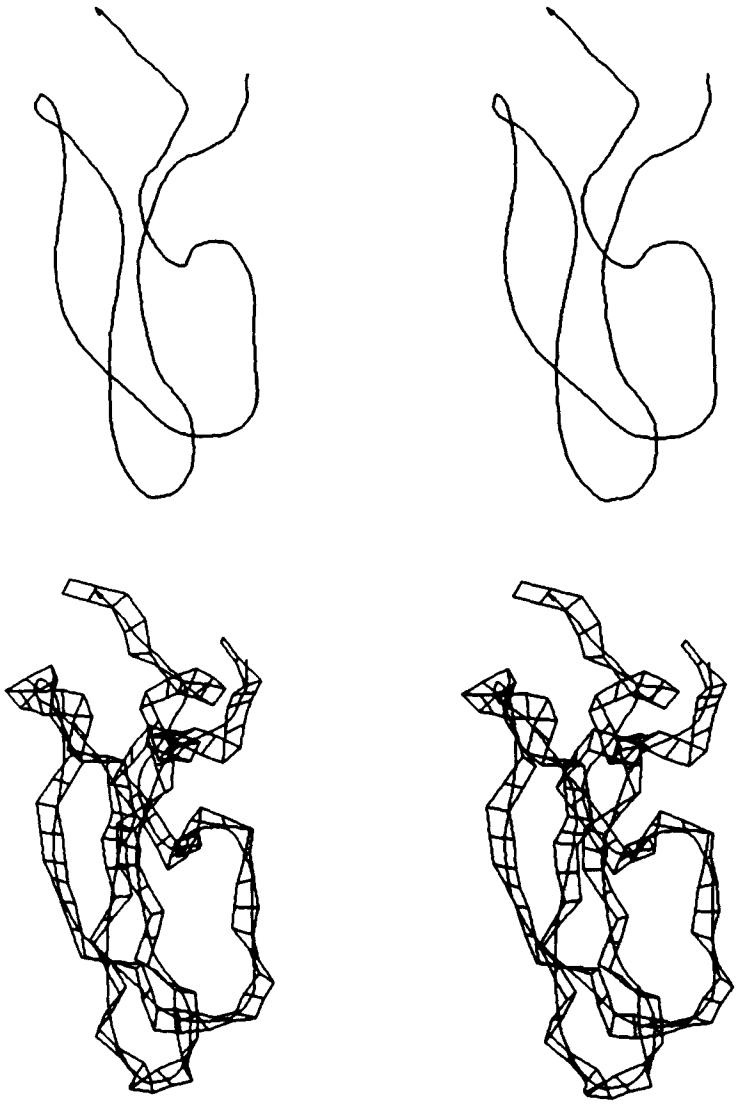

Fig. 9.—Stereodiagrams of the helical axis alone (above) and of the backbone ribbon and the helical axis (below) resulting from the analysis of BPTI (coordinates from ref. 20).

$\bar{P}$ and defined by the vectors $\bar{U}$ (the helical axis), $\bar{V}$, and $\bar{W}$. The two translations which link the helical axis system to the peptide fixed axes are then made along the axes $\bar{V}$ and $\bar{W}$ and termed, respectively, X displacement and Y displacement. The rotational position of the peptide system is obtained by a right-handed rotation termed "inclination" (angle α in Fig. 3) around a vector parallel to $\bar{V}$ passing through the point $\bar{E}$ and by a right-handed rotation termed "tip" (angle β in Figure 3) around the resulting position of the peptide fixed axis $\bar{K}$. It is remarked that the direction of the translations and the signs of the rotations used here have been chosen to be consistent with the those used in our algorithm for determining the structures of nucleic acids[13,14] and the conventions recently proposed for this problem during the EMBO Workshop at Cambridge in 1988.[15]

In order to describe a regular helicoidal conformation (such as an α-helix or a β-sheet) it is necessary to add only one further translation and rotation to the four variables described above. These additional variables correspond to the separation of successive repeating elements along the helical axis $\bar{U}$ (termed "rise") and their relative right-handed rotation around this axis (termed "twist").

We now consider how to define a function that will enable us to obtain the optimal description of the helical structure of irregular conformations such as those found in native proteins. The system we have defined above consists of an axis system $\bar{J}\bar{K}\bar{L}$ and a reference point $\bar{E}$ attached to each peptide plane and a local helical axis system $\bar{V}\bar{W}\bar{U}$ with a reference point $\bar{P}$ again for each peptide. These two axis systems are related by the helicoidal variables X dis-

TABLE I. Global Parameter Definitions

Name	Class	Code	Symbol	Definition
X displacement	A	XDP	dx	$\bar{V}_i^T(\bar{E}_i-\bar{P}_i)$
Y displacement	A	YDP	dy	$\bar{W}_i^T(\bar{E}_i-\bar{P}_i)$
Inclination	A	INC	η	$\pm\cos^{-1}(\bar{K}_i^T\bar{W}_i)$
Tip	A	TIP	θ	$\pm\cos^{-1}(\bar{J}_i^T\bar{V}_i)$
Shift	C	SHF	Dx	$dx(i)+Ax-dx(i-1)$
Slide	C	SLD	Dy	$dy(i)+Ay-dy(i-1)$
Rise	C	RIS	Dz	$\lvert\bar{p}_{i-1}-\bar{P}_{i-1}\rvert + \lvert\bar{P}_i-p_i\rvert$
Tilt	C	TLT	τ	$\eta(i)+\eta_A-\eta(i-1)$
Roll	C	ROL	ρ	$\theta(i)+\theta_A-\theta(i-1)$
Twist	C	TWT	Ω	$\pm\cos^{-1}(\bar{W}_i^T\bar{f}^+)\pm\cos^{-1}(\bar{W}_{i-1}^T\bar{f}^-)$
Axis X displacement	D	AXD	Ax	$\bar{d}^T(\bar{p}_i-\bar{p}_{i-1})$
Axis Y displacement	D	AYD	Ay	$\bar{f}^T(\bar{p}_i-\bar{p}_{i-1})$
Axis inclination	D	AIN	η_A	$\pm2\cos^{-1}(\bar{f}^T\bar{t})$
Axis tip	D	ATP	θ_A	$\pm2\cos^{-1}(\bar{r}^T\bar{U}_i)$

TABLE II. Helicoidal Parameters for Standard Secondary Structure Motifs*

Name†	ϕ	ψ	X displacement	Y displacement	Inclination	Tip	Rise	Twist
α(r)	−57.0	−47.0	0.1	1.5	−6.1	−20.6	1.5	100.2
α(l)	57.0	47.0	0.1	−1.5	6.1	−20.6	1.5	−100.2
3_{10}	−71.0	−18.0	0.3	1.1	−20.1	−13.3	1.8	111.9
π	−57.1	−69.7	0.0	2.1	2.0	−26.6	1.1	81.8
43_{14}	88.1	91.7	0.1	−2.6	−2.1	118.3	1.2	70.3
ω	64.4	55.4	0.1	−1.8	6.3	−28.5	1.0	−92.7
β(p)	−119.0	113.0	0.1	0.0	−31.3	45.4	3.3	178.0
β(ap)	−139.0	135.0	0.1	0.0	−23.2	55.4	3.5	178.9
2_7	−75.0	70.0	0.1	0.0	−33.4	16.2	2.8	179.0
2.2_7	−78.1	59.2	0.2	0.1	−33.0	13.6	2.7	167.6

*In Tables II–V, translational parameters are in angstroms and rotational parameters are in degrees.
†r, right handed; l, left handed; p, parallel; ap, antiparallel. ϕ/ψ values for α and β structures from ref. 17, for the 3_{10} helix from ref. 6, and for π, 43_{14}, ω, 2_7, and 2.2_7 from ref. 18.

placement, Y displacement, inclination, and tip. In practice, we know the atomic coordinates of the protein so the $\bar{J}\bar{K}\bar{L}$ axes are fixed in space and the solution to our problem consists of finding the optimal positions and orientations of the local helical axis systems, $\bar{V}\bar{W}\bar{U}$.

This aim is achieved by formulating a function which, first, quantifies the irregularity in the helicoidal parameters between successive peptide planes and, second, quantifies the disruption between successive local helical axes (this disruption taking the form of a change of direction of the $\bar{U}$ vectors or a lateral shift between successive $\bar{P}$ points). The first aim is easily satisfied by summing terms (over the N peptides in the protein backbone) which represent the change in position of successive peptides with respect to their local helical axis systems. These terms involve calculating the differences (as sums of squares to avoid the influence of signs) between projections of the local helical axes $\bar{U}$ and the vectors $\bar{P}$–$\bar{E}$ onto the local axis systems of successive peptides. These projections are defined respectively by,

$$D_i = \Sigma_{X\epsilon J,K,L}\,(\bar{U}_i^T\bar{X}_i - \bar{U}_{i-1}^T\bar{X}_{i-1})^2$$

and

$$C_i = \Sigma_{X\epsilon J,K,L}\,[(\bar{P}_i-\bar{E}_i)^T\bar{X}_i - (\bar{P}_{i-1}-\bar{E}_{i-1})^T\bar{X}_{i-1})]^2$$

(note $\bar{U}_i^T\bar{X}_i$ is the scalar product between the vectors $\bar{U}_i$ and $\bar{X}_i$). We thus arrive at the first two terms of our function,

$$A1 = \Sigma_{i=2,N}\,D_i$$

and

$$A2 = \Sigma_{i=2,N}\,C_i$$

To deal with deformations between successive local helical axes we require one term to compare their vectorial directions which can be formulated as shown below.

$$B1 = \Sigma_{i=2,N}\,(\bar{U}_i - \bar{U}_{i-1})^2$$

If we now define the mean unit vector between successive helical axes as

$$\langle\bar{U}_i\rangle = (\bar{U}_i + \bar{U}_{i-1}) / \lvert\bar{U}_i + \bar{U}_{i-1}\rvert$$

and the vector between successive $\bar{P}$ points as

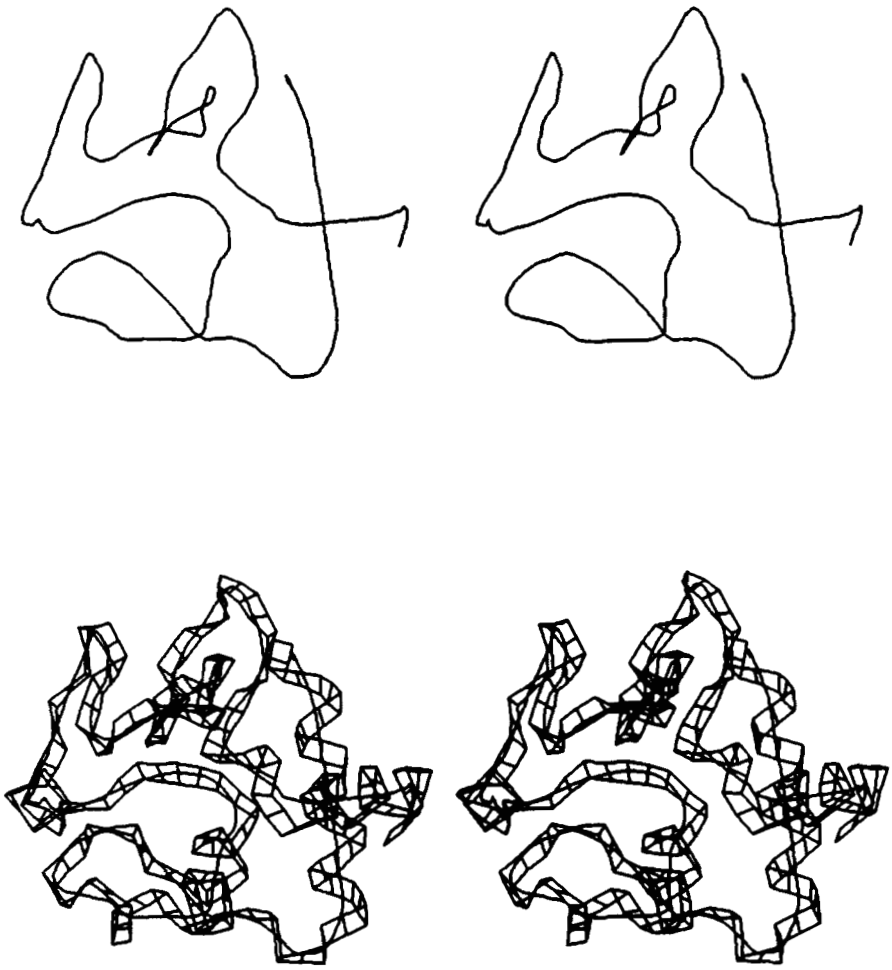

Fig. 10. Stereodiagrams of the helical axis alone (above) and of the backbone ribbon and the helical axis (below) resulting from the analysis of myoglobin (coordinates from ref. 21).

$$\bar{S}_i = \bar{P}_i - \bar{P}_{i-1}$$

then we can calculate the lateral dislocation between these points, perpendicular to the mean axis as

$$\bar{Q}_i = \bar{S}_i - \langle \bar{U}_i \rangle (\langle \bar{U}_i \rangle^{\mathrm{T}} \bar{S}_i)$$

from which the last term of our function is obtained as

$$\mathrm{B2} = \Sigma_{i=2,N} \bar{Q}_i^2$$

In order to obtain a balanced weighting between the rotational terms of the function (A1,B1) and the translational terms (A2,B2) it is necessary to multiply the rotation angles contained within the former terms by the average distance separating successive units in the polymer. This implies that A1 and B1 should be multiplied by the square of this distance. We use a value of 6 for this weighting, corresponding to an average separation of roughly 2.5 Å.

The full expression for the function to be minimized is then

$$F(h) = 6\,(\mathrm{A1} + \mathrm{B1}) + \mathrm{A2} + \mathrm{B2}$$

The variables of the function, denoted by the letter h, are simply the four helicoidal variables (X dis-

TABLE III. ϕ/ψ Torsions and Peptide-Axis (Class A) Helicoidal Parameters for the First 25 Residues of Crambin[14]*

Residue	ϕ	ψ	X displacement	Y displacement	Inclination	Tip
T 1		147.7	−0.13	0.04	−27.95	49.81
T 2	−107.8	144.3	−0.06	0.03	−27.41	55.62
C 3	−131.2	133.2	−0.15	0.00	−27.04	54.49
C 4	−118.9	151.2	−0.14	−0.05	−28.68	46.41
P 5	−76.2	19.0	0.00	0.21	−2.75	40.69
S 6	−157.9	166.0	−0.16	0.58	−15.16	28.18
I 7	−63.6	−42.1	−0.26	1.09	1.75	−3.67
V 8	−55.9	−44.6	−0.13	1.42	−0.90	−17.98
A 9	−61.4	−43.8	−0.01	1.57	−5.23	−20.76
R 10	−63.2	−43.3	0.05	1.56	−7.80	−22.36
S 11	−61.2	−42.4	0.13	1.52	−8.83	−21.01
N 12	−64.9	−39.5	0.13	1.52	−9.39	−21.11
F 13	−59.1	−47.2	0.12	1.49	−9.41	−20.20
N 14	−62.8	−35.2	0.19	1.49	−12.44	−20.94
V 15	−69.2	−41.2	0.17	1.44	−12.13	−20.37
C 16	−56.6	−36.0	0.35	1.36	−16.61	−12.42
R 17	−77.1	−16.1	0.48	1.18	−20.38	−12.19
L 18	−53.2	−46.2	0.53	0.82	−15.20	4.33
P 19	−77.2	−7.6	0.25	0.03	1.02	13.74
G 20	106.3	7.3	−0.34	−0.23	−18.25	9.95
T 21	−52.7	136.3	−0.33	−0.39	−31.95	29.89
P 22	−57.0	146.6	−0.38	0.21	−27.62	17.13
E 23	−56.4	−36.2	−0.24	1.04	−4.58	−3.71
A 24	−63.4	−34.9	−0.06	1.44	−6.91	−16.47
I 25	−74.8	−37.9	−0.02	1.60	−8.94	−23.23

*Divisions (indicated by space) in Tables III–V indicate secondary structure zones (see text).

TABLE IV. Interpeptide (class C) Helicoidal Parameters for the First 25 Residues of Crambin[19]

Junction	Shift	Slide	Rise	Tilt	Roll	Twist
T 1/T 2	0.10	0.05	3.32	−3.29	7.17	−146.60
T 2/C 3	−0.14	−0.24	3.41	5.50	−2.63	−176.25
C 3/C 4	0.13	0.25	3.40	7.15	−12.43	−154.26
C 4/P 5	0.31	0.77	2.99	76.04	−16.83	68.88
P 5/S 6	0.22	0.23	4.01	−70.20	−16.80	178.37
S 6/I 7	−0.29	0.32	2.09	51.85	−15.49	83.47
I 7/V 8	−0.09	0.28	1.58	11.82	−17.80	100.84
V 8/A 9	−0.01	0.22	1.39	−0.75	−6.30	99.06
A 9/R 10	−0.02	0.07	1.45	−2.73	−5.70	97.80
R10/S 11	0.14	−0.01	1.48	−0.80	1.37	101.15
S11/N12	0.02	0.01	1.50	−1.69	−2.04	99.97
N12/F 13	−0.05	−0.05	1.47	−0.78	2.63	98.48
F13/N14	0.14	−0.02	1.61	−4.81	1.46	103.26
N14/V15	−0.17	−0.07	1.36	6.02	0.55	96.59
V15/C16	0.27	−0.08	1.63	−2.95	11.61	108.84
C16/R17	0.01	−0.32	1.92	−8.11	−6.70	105.70
R17/L 18	−0.29	−0.18	1.98	20.26	29.28	99.02
L18/P 19	−0.54	−0.61	2.29	32.99	−20.57	107.55
P19/G20	−1.40	0.31	3.13	−31.57	−61.51	−66.88
G20/T21	−0.22	−0.20	3.14	13.32	34.55	−96.93
T21/P22	−0.70	0.88	2.92	19.19	12.60	−111.48
P22/E23	−0.22	0.90	2.13	43.59	1.71	93.56
E23/A24	−0.04	0.42	1.70	9.69	−11.11	99.98
A24/I 25	−0.17	0.28	1.31	5.48	−14.08	93.64

placement, Y displacement, inclination, tip) of each peptide in the strand. It should be noted that each term of $F(h)$ has been chosen so that an identical sum will be obtained whether the protein backbone is analyzed in the sense N-terminal to C-terminal or vice versa. In order to achieve a rapid convergence of the function $F(h)$ we also calculate the analytic first derivatives of F with respect to the helicoidal vari-

TABLE V. Interaxis (Class D) Helicoidal Parameters for the First 25 Residues of Crambin[19]

Junction	*Ax*	*Ay*	Ainc	Atip	Adis	Bend
T 1/T 2	0.03	0.06	−3.84	1.35	0.07	4.07
T 2/C 3	−0.04	−0.21	5.14	−1.49	0.22	5.35
C 3/C 4	0.12	0.29	8.79	−4.35	0.32	9.80
C 4/P 5	0.17	0.50	50.11	−11.11	0.53	51.25
P 5/S 6	0.38	−0.14	−57.79	−4.29	0.40	57.93
S 6/I 7	−0.20	−0.18	34.95	16.36	0.27	38.48
I 7/V 8	−0.21	−0.05	14.47	−3.49	0.22	14.88
V 8/A 9	−0.14	0.07	3.58	−3.52	0.16	5.02
A 9/R10	−0.07	0.08	−0.16	−4.10	0.11	4.10
R10/S 11	0.06	0.03	0.23	0.02	0.07	0.23
S11/N12	0.01	0.01	−1.13	−1.95	0.01	2.25
N12/F 13	−0.03	−0.02	−0.76	1.73	0.04	1.89
F13/N14	0.07	−0.02	−1.79	2.20	0.07	2.84
N14/V15	−0.14	−0.02	5.71	−0.02	0.15	5.72
V15/C16	0.09	0.00	1.52	3.66	0.09	3.97
C16/R17	−0.12	−0.14	−4.33	−6.93	0.18	8.17
R17/L18	−0.33	0.19	15.08	12.76	0.38	19.73
L18/P19	−0.26	0.18	16.77	−29.99	0.32	34.26
P19/G20	−0.80	0.57	−12.30	−57.71	0.98	58.90
G20/T21	−0.23	−0.04	27.02	14.61	0.23	30.66
T21/P22	−0.65	0.28	14.86	25.36	0.71	29.33
P22/E23	−0.37	0.07	20.55	22.55	0.37	30.42
E23/A24	−0.21	0.01	12.02	1.66	0.21	12.13
A24/I 25	−0.22	0.12	7.51	−7.32	0.25	10.48

ables of each peptide. The development of these derivatives has been fully described in our previous publication[13] and is not repeated here.

Finally, we must consider the definition of the interpeptide parameters in the general case where the local helical axes are not aligned. In this event, the simple definitions of rise and twist given earlier no longer apply and we must also be able to describe the relative position of the two helical axes in space. This is done using a mean axis system ($\bar{n},\bar{d},\bar{f}$ centered at point $\bar{q}$) as shown in Figure 4. This system is defined by the equations below.

$$\bar{q} = (\bar{P}_{i-1} + \bar{P}_i)/2$$

$$\bar{n} = (\bar{U}_i + \bar{U}_{i-1})/|\bar{U}_i + \bar{U}_{i-1}|$$

$$\bar{g} = (\bar{V}_{i-1} + \bar{V}_i)/|\bar{V}_{i-1} + \bar{V}_i|$$

$$\bar{d} = [\bar{g} - \bar{n}(\bar{n}^T\bar{g})]/|\bar{g} - \bar{n}(\bar{n}^T\bar{g})$$

$$\bar{f} = \bar{n} \times \bar{d}$$

The intersection of the $\bar{U}$ vectors with the mean plane (perpendicular to $\bar{n}$) are then

$$\bar{p}_{i-1} = \bar{P}_{i-1} + \bar{U}_{i-1}[n^T(\bar{q} - \bar{P}_{i-1})]/(\bar{n}^T\bar{U}_{i-1})$$

and

$$\bar{p}_i = \bar{P}_i - \bar{U}_i[\bar{n}^T(\bar{P}_i - \bar{q})]/(\bar{n}^T\bar{U}_i)$$

We can now derive expressions for parameters describing the relative position of the helical axis systems at this junction. Two translations along the $\bar{d}$ and $\bar{f}$ axes are defined as

$$\text{Axis X displacement} = \bar{d}^T(\bar{p}_i - \bar{p}_{i-1})$$

$$\text{Axis Y displacement} = \bar{f}^T(\bar{p}_i - \bar{p}_{i-1})$$

Similarly, two rotations, analogous to inclination and tip, are defined as

$$\text{axis inclination} = 2\cos^{-1}(\bar{f}^T\bar{t}),\ \phi \text{ positive if } \bar{d}^T(\bar{f} \times \bar{t}) > 0$$

$$\text{axis tip} = 2\cos^{-1}(\bar{r}^T\bar{U}_i),\ \text{positive if } \bar{t}^T(\bar{r} \times \bar{U}_i) > 0$$

where $\bar{t} = (\bar{U}_i \times \bar{d})/|\bar{U}_i \times \bar{d}|$ and $\bar{r} = (\bar{d} \times \bar{t})/|\bar{d} \times \bar{t}|$.

It is also possible to derive three subsidiary parameters which can be useful. These parameters measure the net angle formed between successive helical axis vectors [axis bend, $\text{Ab} = \cos^{-1}(\bar{U}_{i-1}{}^T\bar{U}_i)$], the net lateral dislocation between successive $\bar{P}$ points [axis dislocation, $\text{Ad} = \sqrt{(A\alpha^2 + Ay^2)}$, and the distance between successive P points (path length, path $= |\bar{P}_i - \bar{P}_{i-1}|$).

Lastly, we can define the general parameters for the interpeptide junction by three translations:

$$\text{shift} = dx(i) + Ax - dx(i-1)$$

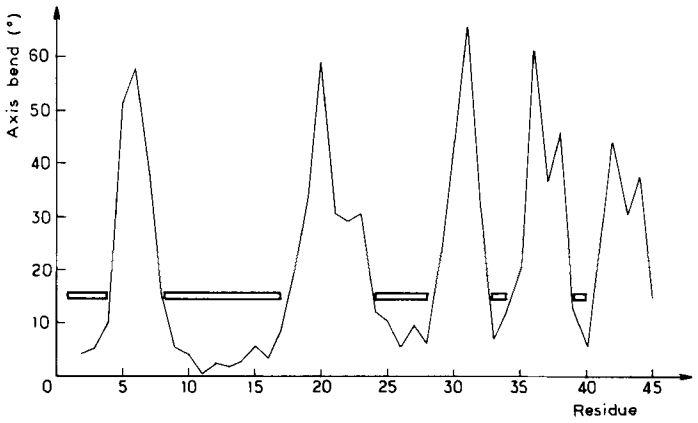

Fig. 11. Localization of secondary structure zones in crambin[19] (open rectangles) using axis bend angles.

$$\text{slide} = dy\,(i) + Ay - dy\,(i-1)$$

$$\text{rise} = |\,\bar{p}_{i-1} - \bar{P}_{i-1}\,| + |\,\bar{P}_i - \bar{p}_i\,|$$

and three rotations:

$$\text{tilt} = \eta(i) + \eta_A - \eta\,(i-1)$$

$$\text{roll} = \theta(i) + \theta_A - \theta(i-1)$$

$$\text{twist} = \pm\cos^{-1}(\bar{W}_i{}^T\bar{f}^+) \pm \cos^{-1}(\bar{W}_{i-1}{}^T\bar{f}^-)$$

Note that the $\bar{f}^+$ and $\bar{f}^-$ vectors are obtained by rotating the $\bar{f}$ vector by (axis incl.)/2 and $-$(axis incl.)/2, respectively, around $\bar{d}$. The first term of twist is positive if $\bar{U}_i{}^T(\bar{f}^+ \times \bar{W}_i) > 0$ and the second term is positive if $\bar{U}_{i-1}{}^T(\bar{f}^- \times \bar{W}_{i-1}) < 0$.

All of the parameters we have defined are summarized in Table I and illustrated schematically in Figure 5. For clarity they have been divided into three classes: A, peptide-axis parameters; C, interpeptide parameters; D, axis junction parameters. Note that these classes correspond to those we have previously defined for the nucleic acids[14] and that class B parameters (which apply to base pairs and thus exist only for double helices) are not defined for polypeptides.

Before turning to the applications of the P-Curve algorithm we should remark that the parameters we have discussed are all "global" parameters since they are defined with respect to a unique overall helical axis. It would also be possible to define a set of "local" positional parameters which would simply locate each peptide with respect to the preceding residue in the chain. This would require six parameters, three translations and three rotations, but would not generate, nor make any use of an overall helical axis. While such an approach is mathematically rigorous and, if properly formulated, leads to a complete and independent set of parameters,[14] it would not lead to much improvement over the simplest local view of protein conformation, namely, the ϕ/ψ angles. In particular, the possibility of measuring the curvature and the exact folding of the backbone would be lost and thus global comparisons of related proteins would also become very difficult.

The "P-Curve" algorithm described has been incorporated in a fortran program which is available on request (from R.L.). This program accepts protein atomic coordinates in a variety of formats including that of the protein data bank[16] and outputs both a helicoidal analysis and a full torsion angle analysis. Graphic output files can also be generated showing the protein helical axis and a ribbon representation of the backbone, with or without side chain atoms.

RESULTS AND DISCUSSION

We will now present some applications of the P-Curve approach to both regular polypeptides and native protein conformations and we will also discuss the study of the fine details of secondary structure.

Analysis of Symmetric Polypeptide Conformations

In order to gain some feeling for the new helicoidal parameters that we have defined, it is easiest to begin by applying the P-Curve algorithm to regular secondary structures. In these cases the helical axis is, by definition, a straight line and only the four peptide-axis parameters and interpeptide rise and twist can have nonzero values. The parameters obtained for a number of well-known structures are listed in Table II along with the corresponding ϕ/ψ values. The same structures are illustrated by backbone ribbons in Figure 6.

Comparing Table II and Figure 6, it can be seen that, in addition to the easily understandable rise and twist values, the remaining parameters allow

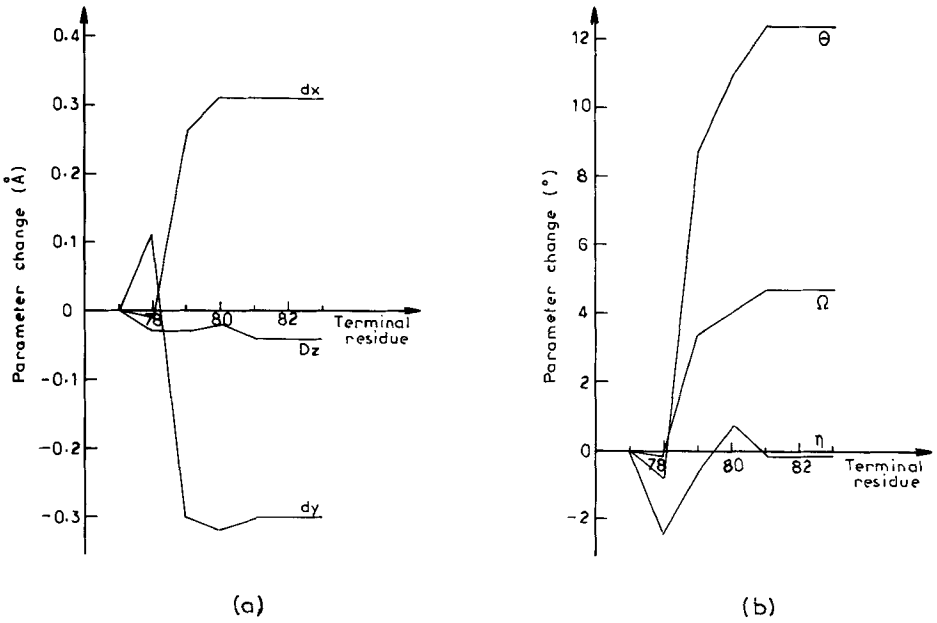

Fig. 12. Variation of the helicoidal parameters of the terminal residue of an α-helix (residues 58–77) extracted from oxymyoglobin[21] as a function of the limit of the P-Curve analysis in the C-terminal direction.

us to judge the radius of the helical structure (almost directly equal Y displacement, since X displacement values tend to be very small) and also of the orientation of the peptide plane with respect to the helical axis, through the inclination and tip values.

Since regular structures naturally lead to regular helicoidal parameters it is also possible to make ϕ/ψ plots of these parameters for the full range of the backbone torsion angles. The results of such a study are presented in Figure 7. From these data it is possible to see immediately what type of structure will result for any given ϕ/ψ combination and to determine the correlations which exist between the different helicoidal parameters.

It should be mentioned at this point that we have not assumed anything about the nature of the linkages within the peptide backbone and thus six parameters are necessary to define the position of successive peptides in space. In fact the chemical bonding within a polypeptide means that there are only two single bond torsions between any two peptides and thus we can expect quite strong correlations between our six helicoidal parameters. However, in any real protein structure, variations in the peptide bond torsion (ω) and in valence angles or bond lengths will mean that these correlations can be only approximate. Thus, a rigorous geometrical description of a protein must nevertheless conserve the full number of variables that we have defined.

It should further be stressed that the unique relationship between the backbone torsions ϕ/ψ and the helicoidal parameters illustrated by Figure 7 is true only for regular structures. In real proteins the conformational environment of any given peptide unit (that is to say, the structure adopted by the residues preceding and following the peptide unit considered) will influence the position of its local helical axis, during minimization of the function $F(h)$, and consequently its helicoidal parameters. Therefore it is clear that no simple relationship between these parameters and the ϕ/ψ values of the peptide can exist. We will return to this point shortly.

Analysis of Native Protein Conformations

We now apply the P-Curve algorithm to protein conformation analysis. Three proteins have been chosen as examples to illustrate the nature of the data which can be obtained: crambin,[19] bovine pancreatic trypsin inhibitor,[20] and sperm whale oxymyoglobin.[21] The graphic data resulting from the analysis of these proteins are presented in Figures 8–10 by stereodiagrams of the protein backbone ribbon with the calculated helical axis and of the helical axis alone. Note that the smooth curve presented for the helical axis is generated by a cubic spline fit to the $\bar{P}_i,\bar{U}_i$ data obtained from P-Curve.

Figures 8–10 show the visual nature of the results obtained from our algorithm. For any protein, regions with regular secondary structures (most commonly, α-helices and β-sheets) are easily visible as straight portions of the axis. These segments are linked by curved zones corresponding to irregular conformations of the polypeptide backbone. On a

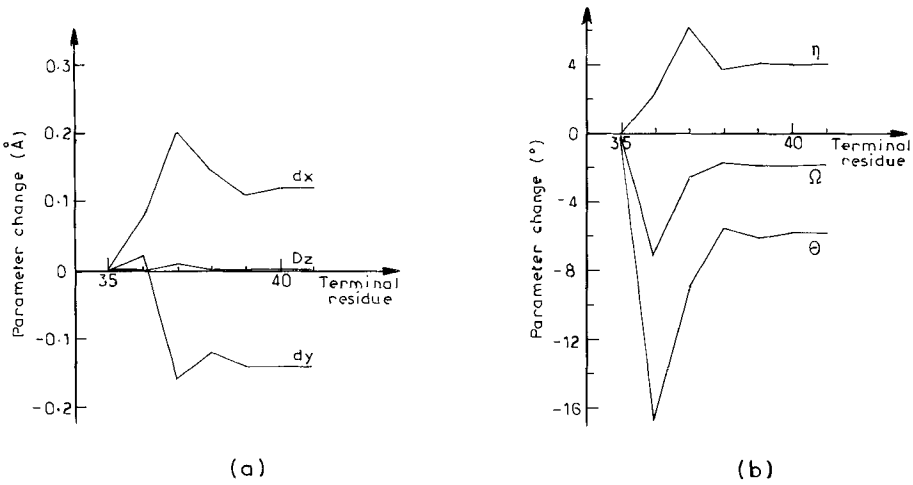

Fig. 13. Variation of the helicoidal parameters of the terminal residue of a β-sheet (residues 29–35) extracted from BPTI[20] as a function of the limit of the P-Curve analysis in the C-terminal direction.

graphics system it is possible to color the axis following the type of secondary structure detected and thus to obtain a very simple, but nevertheless rigorous, description of the folding of the polypeptide chain.

In addition to the graphical data, the P-Curve analysis also lists all the helicoidal parameters described previously. For reasons of space we can present only a part of these data here. We have chosen to discuss the first 25 residues of the smallest protein treated, crambin. The parameters obtained for these residues are listed in Tables III–V. These results will serve to illustrate how our analysis differs from simple ϕ/ψ torsion angle or hydrogen bonding data.

In order to get a quick idea of the localization of regular secondary structures, the three-dimensional helical axis of crambin shown in Figure 8 can be simplified to two dimensions by plotting the bending angle at each junction of the helical axis. This is shown in Figure 11. Secondary structures are now easily distinguished from irregular zones by the low values of their axis bends. If we place a bar at 15° bending, we immediately detect five relatively straight zones: 1–4, 8–17, 24–28, 33–34, and 39–40. (Note from Fig. 1 that the division of the peptide chain we have adopted implies that if an interpeptide junction i–j is bent, then the peptide j should be included in the irregular zone.)

Looking at the data in Tables III–V and comparing it with the standard structures in Table II, we can rapidly identify the first three zones as a β-sheet followed by two α-helices (the remaining two zones, not shown in the tables, are again β-sheets). The most useful values for this identification are Y displacement, tip, rise, and twist, which all distinguish clearly between α and β structures. Regular zones may also be detected through the small values of the interpeptide parameters shift, slide, tilt, and roll. However, it should be noted that none of the secondary structures is quantitatively regular and we will return to the description of distortion within these segments in the following section.

If we compare our findings with those listed by the author of the crystallographic study of crambin[19] within the corresponding protein data bank entry, several differences can be found. The original assignments for secondary structure zones were limited to four segments: 1–4 (β), 7–19 (α), 23–30 (α), and 32–35 (β). All but the first of these zones is wider than our findings and the last β-sheet we have located was not seen. It is interesting to look in detail at the assignment of the first α-helix with the help of the parameters in Tables III–V. From these data it would seem the residues 7, 18, and 19 cannot easily be classed as belonging to the helix. All these residues have tip values far from those of the α-helix and Y displacement for 18 and 19 is too small. Moreover, the junctions 6–7, 17–18, and 18–19 are all distinctly bent and associated with important tilt and roll values.

Looking at the ϕ/ψ values in Table III it is easy to see that, in the case of a folded polypeptide chain, there is no longer any precise correlation between the backbone torsions and our helicoidal parameters. When the $F(h)$ function of the P-Curve algorithm is minimized, the residues preceding and following any given peptide group influence its helicoidal parameters. Thus, while at least peptides 7 and 18 can be classed as α-helical on the basis of their ϕ/ψ values, they cannot in terms of their helicoidal parameters.

The effect of neighboring residues in our analysis can be shown clearly if we look at other examples of

TABLE VI. Classification of the Deformation in the Structural Zones Found for Crambin[19]*

Residues	σ_1(Å)	σ_c(Å)	Radius(Å)	Path(Å)	Type	Class
1–4	0.11	0.05	59.29	10.13	β	Curved
5–7	0.47	0.00	3.43	5.52	—	Curved
8–17	0.12	0.10	145.89	13.85	α	Linear
18–23	1.14	0.35	6.38	13.13	—	Curved
24–28	0.12	0.07	19.39	5.72	α	Curved
29–32	0.95	0.18	3.80	8.48	—	Curved
33–34	0.00	—	—	3.28	β	Linear
35–38	0.94	0.15	4.06	8.79	—	Curved
39–40	0.00	—	—	3.08	β	Linear
41–46	1.02	0.17	3.71	9.11	—	Curved

*σ_1, rms deviation with respect to a line; σ_c, rms deviation with respect to a circle.

regular secondary structures which are followed by sharp bends. We have chosen two such cases from other proteins, first, an α-helix between residues 58 and 77 in oxymyoglobin[21] and, second, a β-sheet between residues 29 and 35 in BPTI.[20] In both cases we will concentrate our attention on one peptide at the C-terminal end of these secondary structures. Analyses have been made for these fragments alone and then repeated while adding residues one by one to the C-terminal end.

The effect of extending the fragment analyzed on the parameters of the terminal peptides (numbers 77 in myoglobin and 35 in BPTI) can be seen in Figures 12 and 13, respectively. Both graphics show that all parameters are indeed influenced by the change of the conformational environment. The change amounts to a maximum of roughly 0.5 Å for translational parameters and can exceed 10° for rotational parameters. Reference to the parameters obtained for standard conformations in Table II and the ϕ/ψ plots of the helicoidal parameters in Figure 7 shows that these changes are by no means negligible.

We can thus conclude that the P-Curve analysis differs from any description of protein structure based only on local data such as backbone torsion angles or hydrogen bonds. The data in Figures 12 and 13 show that, with our approach, the parameters describing any given peptide (with given ϕ/ψ angles) will depend on the position of at least four residues on either side of it in the polypeptide chain. This clearly leads to differences concerning positioning and the deformation of secondary structure motifs, but it also corresponds to a more global and coherent view of the overall protein conformation than can be obtained from data referring only to isolated peptide groups.

Analysis of Secondary Structure Deformation

We finally consider how the P-Curve analysis can be used to study fine deformations within secondary structures. If we return to the example of crambin, it is possible to make a detailed analysis of each zone of the protein that was detected by plotting the axis bend angles (see Figure 11). The analysis is performed by testing each segment (secondary structures and intermediate zones) through least squares fits to the P_i points of the constituent residues using both a straight line and a circle. Note that a similar analysis of protein α-helices has been presented by Barlow and Thornton[12] using as data points on an approximate helical axis determined by least-squares fitting of a "probe" helix[11,12] to successive residues of the segment.

The results are presented in Table VI which contains the standard deviation obtained with a straight line (σ_1), the standard deviation and the radius obtained with a circle (σ_cR) and the length of each segment (path, defined as the sum of the distances between successive P_i points). The distinction between the secondary structure segments and the intermediate zones of crambin now becomes clear. The longest α-helix (8–17) can effectively be classed as linear, while the second (24–28) is curved with a radius of 19 Å. The first β-sheet (1–4) is also curved, but with a larger radius (59 Å). (The remaining β-sheets are automatically classed as linear since they contain only two residues.) In contrast, the intermediate zones are all very strongly curved with radii varying between 3 and 6 Å. Within these zones the only notable departure from a circular pathway occurs for the six residue segment 18–23 ($\sigma_{c=0.35}$).

Work is now in progress to analyze a large number of well-resolved protein structures. Distribution plots of the helicoidal parameters obtained from the sum of these analyses will enable us to generate rigorous definitions for each type of secondary structure. In combination with the localization of secondary structure zones described above we will then hopefully be in a position to extract new and interesting information from protein crystallographic data.

CONCLUSIONS

We have described a rigorous algorithm for obtaining a helicoidal description of protein conformation. This method, termed P-Curve, yields a complete and independent set of helicoidal parameters

and a unique overall helical axis for any protein whose backbone atomic coordinates are known. The approach makes use of an extended least-squares minimization procedure to yield an optimal helical description where structural irregularities are distributed between changes in the orientation of successive peptide groups and curvature of the overall helical axis. Using the P-Curve algorithm has two fundamental advantages. First, the algorithm gives a coherent overall view of the entire protein conformation and also allows detailed information of the positioning of individual peptides to be extracted. Second, the location of secondary structures and measures of the deformation of these segments or intermediate zones of the backbone can be obtained automatically.

The P-Curve algorithm has obvious applications for describing protein folding patterns and for the automatic comparison of related proteins or searches of chosen conformational fragments within data banks of protein structure. It may also be of considerable interest in analyzing the data from molecular dynamics studies of proteins where the extraction of easily readable information is often a major problem (for a similar application to a nucleic acid oligomer see ref. 22).

REFERENCES

1. Ramakrishnan, C., Ramachandran, G.N. Stereochemical criteria for polypeptide and protein chain conformations II. Allowed conformations for a pair of peptide units. Biophys. J. 5:909–933, 1965.
2. Richardson, J.S. The anatomy and taxonomy of protein structure. Adv. Protein Chem. 34:167–339, 1981.
3. Lifson, S., Sander, C. Specific recognition in the tertiary structure of β-sheets of proteins. J. Mol. Biol. 139:627–639, 1980.
4. Levitt, M., Greer, J. Automatic identification of secondary structure in globular proteins. J. Mol. Biol. 114:181–239, 1977.
5. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
6. Rose, G.D., Gierasch, L.M., Smith, J.A. Turns in peptides and proteins. Adv. Protein Chem. 37:1–109, 1985.
7. Rackovsky, S., Scheraga, H.A. Differential geometry and polymer conformation. I. Comparison of protein conformations. Macromolecules 11:1168–1174, 1978.
8. Rackovsky, S., Scheraga, H.A. Differential geometry and polymer conformation. II. Development of a conformational distance function. Macromolecules 13:1440–1453, 1980.
9. Rackovsky, S., Scheraga, H.A. Differential geometry and polymer conformation. III. Single-site and nearest-neighbor distributions and nucleation of protein folding. Macromolecules 14:1259–1269, 1981.
10. Louie, A.H., Somorjai, R.L. Differential geometry of proteins. Helical approximations. J. Mol. Biol. 168:143–162, 1983.
11. Blundell, T., Barlow, D., Borakoti, N., Thornton, J. Solvent-induced distortions and the curvature of α-helices. Nature (London) 306:281–283, 1983.
12. Barlow, D.J., Thornton, J.M. Helix geometry in proteins. J. Mol. Biol. 201:601–619, 1988.
13. Lavery, R., Sklenar, H. The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. J. Biomol. Struct. Dynam. 6:63–91, 1988.
14. Lavery, R., Sklenar, H. Defining the structure of irregular nucleic acids: Conventions and principles. J. Biomol. Struct. Dynam. 6:655–667, 1989.
15. Dickerson, R.E., Bansal, M., Calladine, C.R., Diekmann, S., Hunter, W.N., Kennard, O., Lavery, R., Nelson, H.C.M., Olson, W.K., Saenger, W., Shakked, Z., Sklenar, H., Soumpasis, D.M., Tung, C-S., von Kitzing, E., Wang, A.H-J., Zhurkin, V.B. Definitions and nomenclature of nucleic acid structure parameters. EMBO J. 8:1–4, 1989; J. Biomol. Struct. Dynam. 6:627–634, 1989; J. Mol. Biol. 205: 787–791, 1989.
16. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rogers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The protein data bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
17. IUPAC-IUB Commission on biochemical nomenclature. Abbreviations and symbols for the description of the conformation of polypeptide chains. Biochemistry 9:3471–3479, 1970 and references therein.
18. Ramachandran, G.N., Sasisekharan, V. Conformation of polypeptides and proteins. Adv. Protein Chem. 23:283–437, 1983.
19. Teeter, M.M. Water structure of a hydrophobic protein at atomic resolution. Pentagon rings of water molecules in the crystals of crambin. Proc. Natl. Acad. Sci. U.S.A. 81: 6014–6018, 1984.
20. Walter, J., Huber, R. Pancreatic trypsin inhibitor. A new crystal form and its analysis. J. Mol. Biol. 167:911–917, 1983.
21. Phillips, S.E.V. Structure and refinement of oxymyoglobin at 1.6Å resolution. J. Mol. Biol. 142:531–554, 1980.
22. Ravishankar, G., Swaminathan, S., Beveridge, D.L., Lavery, R., Sklenar, H. Conformational and helicoidal analysis of 30ps of molecular dynamics on the d(CGCGAATTCGCG) double helix: "Curves", Dials and Windows. J. Biomol. Struct. Dynam. 6:669–699, 1989.