Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features

WOLFGANG KABSCH and CHRISTIAN SANDER, *Biophysics Department, Max Planck Institute of Medical Research, 6900 Heidelberg, Federal Republic of Germany*

Synopsis

For a successful analysis of the relation between amino acid sequence and protein structure, an unambiguous and physically meaningful definition of secondary structure is essential. We have developed a set of simple and physically motivated criteria for secondary structure, programmed as a pattern-recognition process of hydrogen-bonded and geometrical features extracted from x-ray coordinates. Cooperative secondary structure is recognized as repeats of the elementary hydrogen-bonding patterns "turn" and "bridge." Repeating turns are "helices," repeating bridges are "ladders," connected ladders are "sheets." Geometric structure is defined in terms of the concepts torsion and curvature of differential geometry. Local chain "chirality" is the torsional handedness of four consecutive C^α positions and is positive for right-handed helices and negative for ideal twisted β-sheets. Curved pieces are defined as "bends." Solvent "exposure" is given as the number of water molecules in possible contact with a residue. The end result is a compilation of the primary structure, including SS bonds, secondary structure, and solvent exposure of 62 different globular proteins. The presentation is in linear form: strip graphs for an overall view and strip tables for the details of each of 10,925 residues. The dictionary is also available in computer-readable form for protein structure prediction work.

INTRODUCTION

Background

α-Helices and pleated β-sheets were predicted in 1951 by Linus Pauling and Robert Corey[1] on the basis of hydrogen-bonding and cooperativity criteria. They were seen later, and beautifully, in the first structures shown in atomic detail by x-ray crystallography. Since then, the number of known protein structures has risen to over 100 and comprehensive analysis of secondary structure requires a computerized compilation of structure assignments, especially in the context of structure prediction methods. Existing compilations have various shortcomings. The crystallographers' assignments of secondary structure in the Brookhaven Protein Data Bank[2] are often subjective and sometimes incomplete. Objective algorithms exist, e.g., for defining turns[3–6] (reviewed in Refs. 7, 8), β-sheets,[9] and solvent accessibility,[10] but only Levitt and Greer[11] have published an extensive compilation of automatic assignments of helices and sheets. Their ap-

Biopolymers, Vol. 22, 2577–2637 (1983)

CCC 0006-3525/83/122577-61$07.10

proach has the advantage of giving assignments when only backbone C^α coordinates are known; the price paid is loss of accuracy when all-atom coordinates are known. Solvent exposure has been published for no more than a few proteins, and chirality only on microfiche.[12] We are thus motivated to make available an accurate, exhaustive, and up-to-date compilation.

The Main Ideas

Our goal is to approximate the intuitive notion of secondary structure by an objective algorithm. An algorithm for extracting structural features from the atomic coordinates is obviously a pattern-recognition process. The elementary patterns on which this process is based should be as simple as possible yet capable of discriminating among the main types of secondary structure. To discriminate whether a pattern is present or not in a continuum of possible atomic configurations, continuous decision parameters must be fixed. Using backbone φ,ψ angles or C^α positions requires the adjustment of several parameters, e.g., four angles for a rectangle in the φ,ψ plane for each type of secondary structure. In contrast, the presence or absence of an H bond can be characterized by a single decision parameter, a cutoff in the bond energy. Therefore, we base our secondary structure recognition algorithm mainly on H-bonding patterns: "n-turns" with an H-bond between the CO of residue i and the NH of residue $i + n$, where $n = 3,4,5$, and "bridges" with H bonds between residues not near each other in sequence. These two types of pattern essentially exhaust all backbone–backbone H bonds. Repeating 4-turns define α-helices, and repeating bridges define β-structure, in good agreement with intuitive assignments. All other occurrences of the basic patterns provide an interesting survey of 3_{10}-helices, π-helices, single turns, and single β-bridges.

The results are presented in short form as strip maps of secondary structure (Fig. A1), and in long form, together with the amino acid sequence as an easy-to-use dictionary (Table AIII). The computer program DSSP (Define Secondary Structure of Proteins) written in standard PASCAL will be available from the Protein Data Bank, Chemistry Dept., Brookhaven National Laboratory, Upton, N.Y. 11973. Publication of an update of this compilation is planned as more protein structures are solved.

DEFINITIONS

The definitions of H-bonded features form a hierarchy: first H bonds are defined; based on them, turns and bridges; and, based on them, α-helices and β-ladders, including common imperfections such as helical kinks and β-bulges. Features defined geometrically are bends, chirality, SS bonds, and solvent exposure. Each structural feature is defined independently of the others and structural overlaps are resolved by defining a secondary structure summary that assigns a single state to each residue. For brevity we express the pattern definitions in the form of equations. For example,

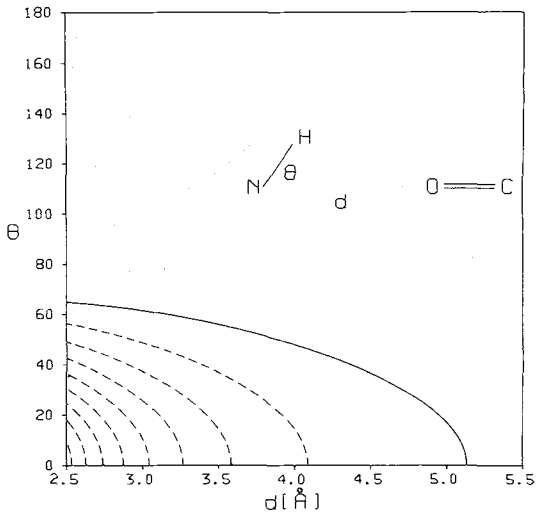


Fig. 1. H bond between peptide units is described here by the dominant electrostatic part E (see text) of the H-bond energy, drawn in contours of constant E at 0.5 kcal/mol intervals as a function of the distance, d, and the alignment angle θ. Dotted lines, E positive or zero; broken lines, E negative. An ideal H bond has $d = 2.9$ Å, $\theta = 0$, and $E = -3.0$ kcal/mol. We assume an H bond for E up to -0.5 kcal/mol (solid line). Thus, misalignment of up to 63° is allowed at the ideal length; an N-O distance of up to $d = 5.2$ Å is allowed for perfect alignment. This definition of H bonds is particularly simple and physically meaningful. It is more general than the historical definition of hydrogen "bond" and could be called polar interaction.

"Hbond(i,j)=: [E < −0.5kcal/mole]" means: there is an H bond (i,j) if E is less than −0.5 kcal/mol.

Hydrogen-Bonded Structure

Hydrogen Bonds

Hydrogen bonds in proteins have little wave-function overlap and are well described by an electrostatic model.[13] We calculate the electrostatic interaction energy between two H-bonding groups by placing partial charges on the C,O $(+q_1, -q_1)$ and N,H $(-q_2, +q_2)$ atoms, i.e.,

$$E = q_1q_2(1/r(\mathrm{ON}) + 1/r(\mathrm{CH}) - 1/r(\mathrm{OH}) - 1/r(\mathrm{CN}))*f$$

with $q_1 = 0.42e$ and $q_2 = 0.20e$, e being the unit electron charge and r(AB) the interatomic distance from A to B. In chemical units, r is in angstroms, the dimensional factor $f = 332$, and E is in kcal/mol. A good H bond has about −3 kcal/mol binding energy. We choose a generous cutoff to allow for bifurcated H bonds and errors in coordinates and assign an H bond between C=O of residue i and N-H of residue j if E is less than the cutoff, i.e., "Hbond(i,j)=: [E < −0.5kcal/mole]."

Figure 1 illustrates the relation of this one-parameter definition to the

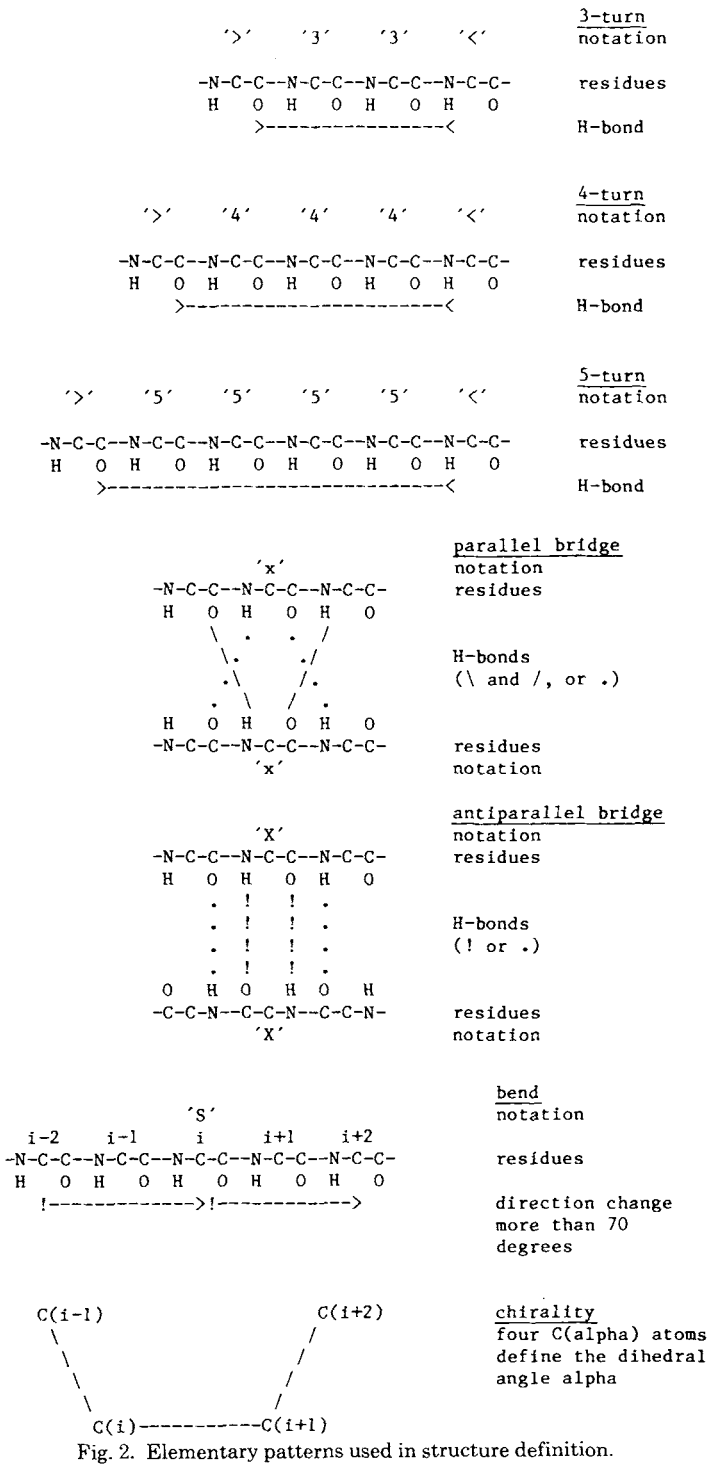


Fig. 2. Elementary patterns used in structure definition.

more complicated description of H bonds in terms of one distance and one angle. There is no generally correct H-bond definition, as there is no sharp border between the quantum-mechanical (wave-function overlap dominates at short distances) and electrostatic (electrostatic interaction dominates at larger distances) regimes and no discontinuity of the interaction energy as a function of distance or alignment. Thus, any H-bond definition is empirically tailored to a particular purpose. Our definition, well tested by trial and error, reflects a compromise suitable for the purpose of secondary structure definition. The cutoff chosen, which allows for an N-O distance up to 2.2 Å larger than the optimal value at perfect alignment or a misalignment of maximally 60° is similar to the tolerances used by Levitt and Greer[11] (1.8 Å excess and 60°) and was found to be sufficient to average over coordinate errors without leading to spurious secondary structure assignments. Were it not for historical reasons, we would use the term "polar interaction" rather than "hydrogen bond."

Elementary H-Bond Pattern: n-*Turn*

The basic turn pattern (Fig. 2) is a single H bond of type $(i,i+n)$. We assign an n-turn at residue i if there is an H bond from CO(i) to NH($i+n$), i.e., "n-turn(i)=: Hbond(i,i + n), n = 3,4,5."

When the pattern is found, the ends of the H bond are indicated by using "⟩" at i and "⟨" at $i+n$ in line 3-TURN, 4-TURN, or 5-TURN of Table AIII; the residues bracketed by the H bond are noted "3," "4," or "5" unless they are also the end points of other H bonds. Coincidence of "⟩" and "⟨" at one residue is indicated by "X." In line SUMMARY of Table AIII, residues bracketed by the hydrogen bond of an n-turn are marked "T," unless they are part of an n-helix (defined below).

Elementary H-Bond Pattern: Bridge

Two nonoverlapping stretches of three residues each, $i-1,i,i+1$ and $j-1,j,j+1$, form either a parallel or antiparallel bridge, depending on which of two basic patterns (Fig. 2) is matched. We assign a bridge between residues i and j if there are two H bonds characteristic of β-structure; in particular,

Parallel Bridge(i,j)=: [Hbond(i − 1,j) and Hbond(j,i + 1)] or
[Hbond(j − 1,i) and Hbond(i,j + 1)]

Antiparallel Bridge(i,j)=: [Hbond(i,j) and Hbond(j,i)] or
[Hbond(i − 1,j + 1) and Hbond(j − 1,i + 1)]

Parallel bridges are marked at i and j by lower-case letters, antiparallel ones by upper-case letters.

Cooperative H-Bond Pattern: Helices

A minimal helix is defined by two consecutive n-turns. For example, a 4-helix, of minimal length 4 from residues i to $i + 3$, requires 4-turns at residues $i - 1$ and i,

$$\text{4-helix(i,i + 3)=:} \quad [\text{4-turn(i} - 1) \text{ and 4-turn(i)}]$$

i.e., an H bond ($i - 1$,i +3) and an H bond (i,$i + 4$). Note that nothing is required about the H-bond state of residues $i + 1$ and $i + 2$. Similarly, two consecutive turns are required and a 3-helix of minimal length 3 from residue i to $i + 2$ and a 5-helix of minimal length 5 from residue i to $i + 5$:

$$\text{3-helix(i,i + 2)=:} \quad [\text{3-turn(i} - 1) \text{ and 3-turn(i)}]$$
$$\text{5-helix(i,i + 5)=:} \quad [\text{5-turn(i} - 1) \text{ and 5-turn(i)}]$$

Longer helices are defined as overlaps of minimal helices. Conventionally, these structures are called α-helix, 3_{10}-helix, and π-helix. In Table AIII, a 3-helix can be recognized by the pattern ⟩⟩3⟨⟨, a 4-helix by ⟩⟩44⟨⟨, and a 5-helix by ⟩⟩555⟨⟨. In the line SUMMARY, the residues bracketed by H bonds are labeled G, H, I, e.g.,

5-TURN			⟩⟩555⟨⟨
4-TURN		⟩⟩44⟨⟨	
3-TURN	⟩⟩3⟨⟨		
SUMMARY	GGG	HHHH	IIIII

These helices are one residue shorter at each end than they would be according to rule 6.3 of IUPAC-IUB.[14] Examples of a 3-helix and a 5-helix are shown in Fig. 3.

Cooperative H-Bond Patterns: β-Ladders and β-Sheets

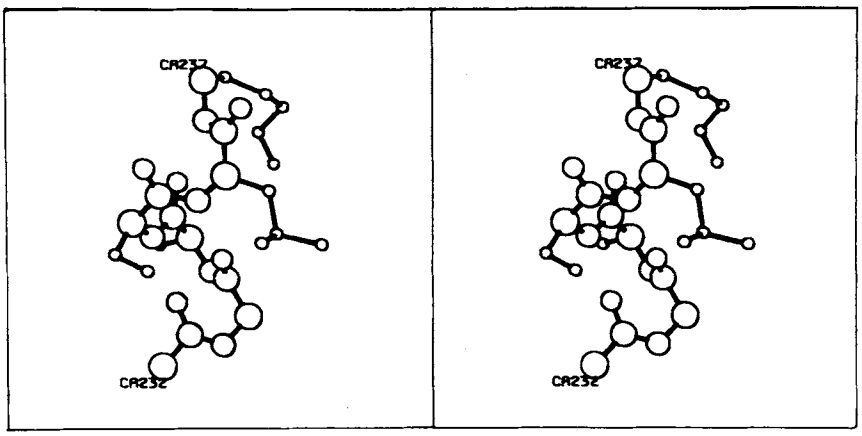
We coin the term "ladder" and define

ladder=: set of one or more consecutive bridges of identical type
sheet=: set of one or more ladders connected by shared residues

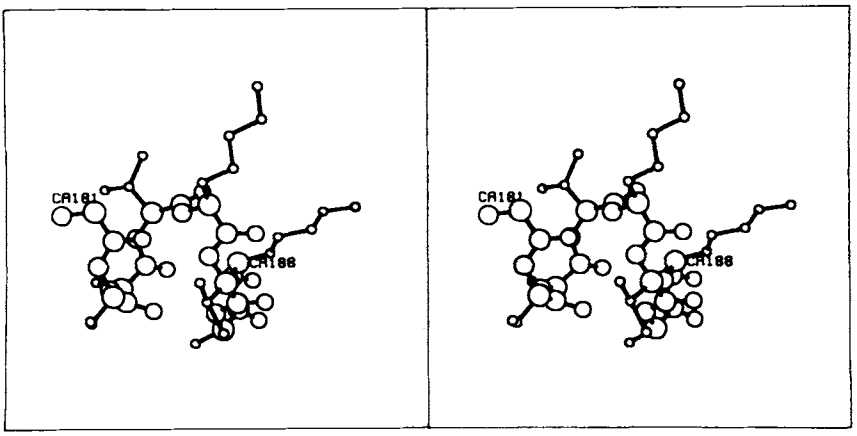
Ladders are given letter names, where a,b,c, . . . is for parallel, A,B,C . . . for antiparallel arrangement. Along the sequence, the first ladder is named "a" or "A," the second "b" or "B," etc. Sheets are also given letter names A,B,C . . . When the alphabet is exhausted, names restart at "a" or "A." In Table AIII, each residue is labeled in line SHEET by the sheet name and in lines BRIDGE by the names of the ladders in which it participates (at most two, one on each side). In line SUMMARY, residues in single bridges (ladders of length 1) are marked "B," all other ladder residues "E" (extended). Thus, continuous stretches of "E" are β-strands. The β-sheet notation is illustrated in Fig. 4.

Secondary Structure Irregularities

Long helices can deviate from regularity in that not all possible H bonds



(a)



(b)

Fig. 3. Stereoviews of secondary structure: (a) 3-helix (3_{10}-helix) and (b) 5-helix (π-helix). (a) 3-Helix Gly232–Lys237 from triose phosphate isomerase (1TIM). In Table AIII, it appears as the H-bond pattern

3-TURN	⟩⟩⟩⟨⟨⟨
SUMMARY	GGGG
SEQUENCE	GGASLK

3-Helices are not uncommon, but have only two or three weak H bonds with E about −1 kcal/mol and the C=O direction tilted away from the helix axis typically by 30°. (b) 5-Helix Gly181–Lys188 from alcohol dehydrogenase (4ADH), at the C-terminal end of a 4-helix. In Table AIII, it appears as the H-bond pattern

5-TURN	⟩⟩⟩55⟨⟨⟨
SEQUENCE	GSAVKVAK

5-Helices are extremely rare; the longest one, shown here, has three H bonds. All stereoviews are by PLUTO (Sam Motherwell, unpublished). In Figs. 3 and 5, the larger atoms are backbone atoms with ¼ their hard-sphere radius (C^{α}, 0.47; C of CO, 0.44; O, 0.35; N, 0.41 Å) and in Fig. 4 with twice these values; side-chain atoms are small, with 0.20-Å radius.

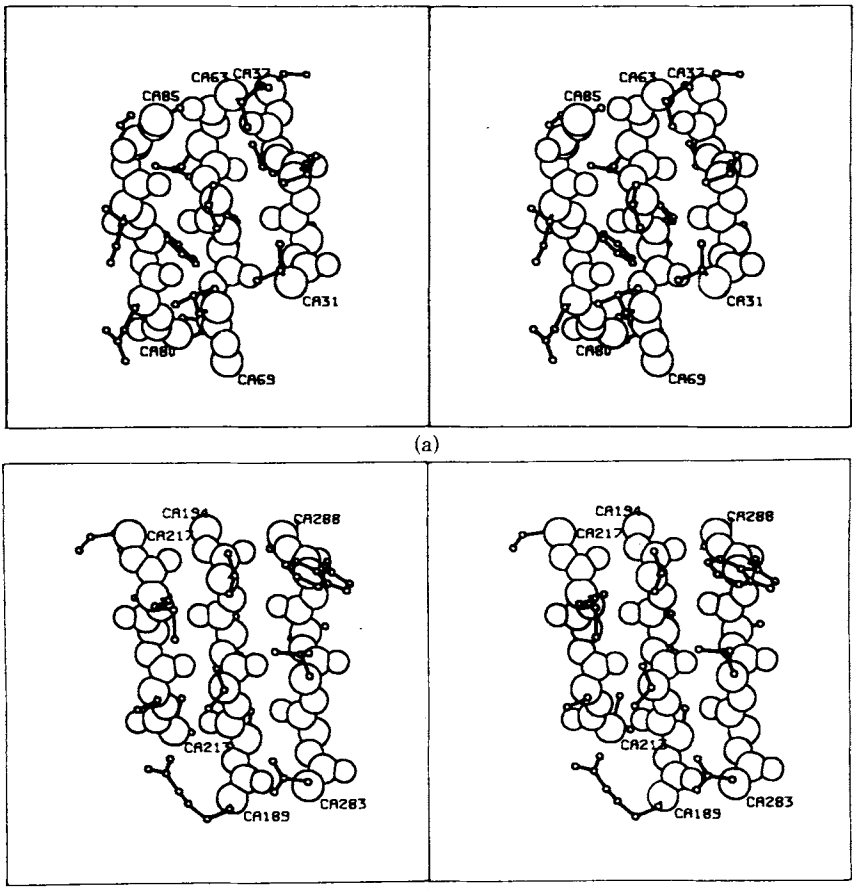


Fig. 4. Stereoviews of secondary structure: (a) antiparallel and parallel β-sheets with two ladders (three strands) each. (a) Two connected antiparallel β-ladders from trypsin (1PTN). The three participating strands are Val16(31)–Ser20(37), Ile46(63)–Gly51(69), and Glue62(80)–Ala67(85), where the first number is the sequential residue number from Table AIII and the number in parentheses the authors' residue identifier. The corresponding H-bond notation (Table AIII) is

```
SHEET..... CCC..... CCCC ..... CCCC....
BRIDGE2..... NNNN .....
BRIDGE1 ..... KKK ..... KKK..... NNNN....
SEQUENCE ..... VSLNS..... IQVRLG..... EQFISA ..
```

The middle strand participates in two ladders. Both ladders belong to sheet C. (b) Two connected parallel β-ladders, Arg172(189)–Gly177(194), Thr196(213)–Ile200(217), Asp266(283)–Ala271(288) from glutathione reductase (2GRS). The corresponding H-bond notation (Table AIII) is

```
SHEET ..... EEEE ..... EEE..... EEEE ....
BRIDGE2..... lll .....
BRIDGE1..... kkkk ..... lll..... kkkk.....
SEQUENCE..... RSVIVG..... TSLMI ..... DCLLWA ...
```

The first strand has two ladder partners. The three strands are part of sheet E.

are formed. This possibility is implicit in the above helix definition, e.g., two overlapping minimal helices offset by two or three residues are joined into one helix:

```
    ))44((                          ))44((
+     ))44((                    +     ))44((
=   ))4)X(4((   irregular       =   ))44XX44((   irregular
=    HHHHHHH                    =    HHHHHHHH
    )))))X((((   perfect             ))))XX((((   perfect
```

even though the third and/or fourth H bond is missing, compared to a perfect seven- or eight-residue helix. Such imperfections are often associated with a kink in the helix, e.g., due to a proline residue.

For β-structure, we define explicitly: a bulge-linked ladder consists of two (perfect) ladders or bridges of the same type connected by at most one extra residue on one strand and at most four extra residues on the other strand. This definition follows Richardson's[8] observation of β-bulges, a frequent lattice fault in β-sheets, but includes more general bulges than her main types. In naming ladders, a bulge-linked ladder is treated as one ladder (lines BRIDGE). In line SUMMARY, all residues in bulge-linked ladders are marked "E," including the extra residues.

Geometrical Structure

Bend

Bends are regions with high curvature. We quantify chain curvature at the central residue i of five residues as the angle between the backbone direction of the first three and the last three residues. This definition of curvature is identical to that of Rose and Seltzer[5] but slightly different from that of Rackovsky and Scheraga.[15] For a bend at i, we require a curvature of at least 70°. The cutoff value was chosen by visual inspection of three-dimensional traces. With $\mathbf{C}^\alpha$ the position vector of C^α, we define

$$\text{Bend(i)} =: [\text{angle}\{(\mathbf{C}^\alpha(\text{i}) - \mathbf{C}^\alpha(\text{i} - 2)),(\mathbf{C}^\alpha(\text{i} + 2) - \mathbf{C}^\alpha(\text{i}))\} > 70°]$$

and assign "S" for a bend at residue i.

Chirality

We define chirality at each residue (except at the ends of the chain) as (Fig. 2)

$$\alpha(\text{i}) = \text{dihedral angle}(\mathbf{C}^\alpha(\text{i} - 1),\mathbf{C}^\alpha(\text{i}),\mathbf{C}^\alpha(\text{i} + 1),\mathbf{C}^\alpha(\text{i} + 2))$$

but report only the sign of α in Table AIII: "+" if $0° < \alpha < 180°$ and "−" if $-180° < \alpha < 0°$. Note that most helices have positive, most twisted β-ladders negative, chirality. We have found only one left-handed helix, in thermolysin. This rare specimen is shown in Fig. 5.

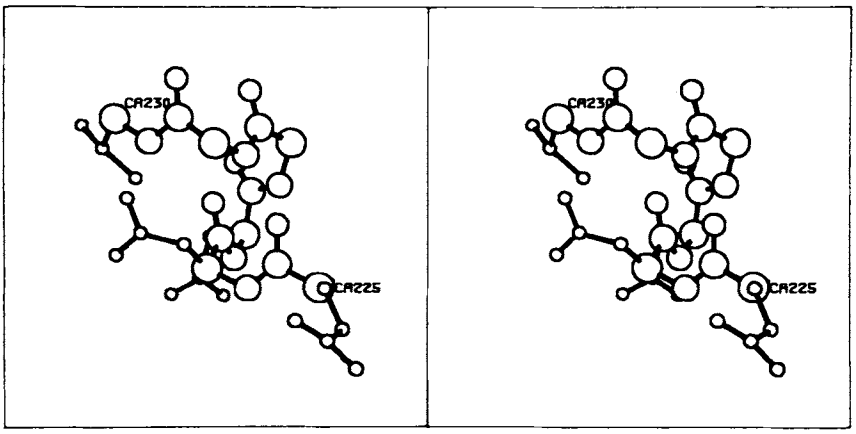


Fig. 5. Stereoviews of secondary structure: illustration of chirality. This short left-handed α-helix, Gln225–Val230 from thermolysin (2TLN) is the only one known to us. In Table AIII (note that chirality is entered at the second residue of each quartet) it appears as:

CHIRALITY	− − −
4-TURN	)) 4 4 ((
SUMMARY	H H H H
SEQUENCE	Q D N G G V

SS Bonds

SS bonds, i.e., covalent links between the S^{γ} atoms of two Cys residues, are taken directly from the Data Bank SSBOND records, as they can be considered part of the amino acid sequence (primary structure). For the coordinate data sets used here, an S-S distance of less than 3.0 Å can also serve as a definition. The SS bonds are given names a,b,c . . . , and the participating residues noted by this name in the line SEQUENCE in Table AIII. Thus, Cys appears in the amino sequence either as C or as a lowercase letter.

Chain Breaks

Chain breaks are assumed if the peptide bond length (distance C′-N) exceeds 2.5 Å. They are labeled "!" and counted as a break residue. Thus, "!" may reflect the absence of a chemical peptide bond, missing density in the crystallography map, or coordinate errors. The residues for which there are coordinates in the data set are numbered sequentially, including break residues. The resulting residue numbers often agree with the authors' except for proteins numbered according to sequence homology or those with missing density or chain breaks. In any case, inspection of the amino acid sequence in Table AIII always allows unambiguous identification of a residue.

Structure Summary

To make contact with the usual notation of secondary structure and to facilitate comparison with intuitive assignments, we summarize secondary structure in a single line (SUMMARY in Table AIII). Structural overlaps are eliminated in this line by giving priority to H,B,E,G,I,T,S in this order, i.e., when several symbols coincide, the first one in this list is written. For example, a helix is also a series of bends, but the state helix is given higher priority. Pieces of 3- or 5-helix, reduced to less than minimal size due to overlaps, are labeled "T." A blank, by implication, means a piece of low curvature not in H-bonded structure.

Static Solvent Exposure

Physically, we are interested in the number of water molecules in direct contact with the protein or with a particular part of the protein.

Geometrically, a very useful representation of a monomolecular layer of water is the surface described by all possible positions of a water molecule in touching contact with protein atoms. That was the idea of Lee and Richards'[10] water sphere rolling around the protein surface. Note that the surface associated with holes in the protein interior is very small, e.g., a hole that accommodates just one water molecule has zero area. For most of the protein exterior, however, the surface is proportional to the number of water molecules in the first hydration shell.

Mathematically, one calculates the surface by integrating a step function f over all points x on the surface of a sphere of radius r(atom) + r(water) around atom i. $f = 1$ if a water sphere centered at x (by definition in contact with atom i) does not intersect with any other protein atom; otherwise, $f = 0$.

Algorithmically, we integrate by summing over a polyhedron made of 20, 80, 320, or more approximately equal triangles. The integration points are the triangle centers, the weights are the triangle area. The polyhedron is generated starting from an icosahedron; a recursive procedure then divides each triangle into four by connecting the midpoints of the sides and projects the three new vertices onto the surface of the sphere, ready for the next level of recursion. The final polyhedron is reminiscent of the shells of certain viruses and of Buckminster Fuller's architecture of geodesic domes. Hence, we call the algorithm "geodesic sphere integration." It is similar to the algorithm of Shrake and Rupley[16] and conceptually simpler than z-layer integration.

With 320 integration points, the surface area of a residue is accurate to within 1 Å^2; with 80 points, to within 4 Å^2. For myoglobin, the numerical values agree with those of Lee and Richards,[10] using their parameters. The numbers given here are based on slightly different values of atomic radii: 1.40 for O, 1.65 for N, 1.87 for C$^\alpha$, 1.76 for C of CO in the backbone, 1.80 for

all side-chain atoms,[17] and 1.40 for a water molecule following observed water–protein distances (Ref. 18 as cited in Ref. 19).

In Table AIII, we report the average number, W, of water molecules in contact with each residue. W can be estimated from the surface area by

$$W = \frac{\text{Area}}{V(\text{water molecule})^{2/3}} \approx \frac{\text{Area}}{10}$$

since the surface is proportional to the volume of the monolayer, which, in turn, is proportional to the average number of molecules in the monolayer. For a water molecule volume of 30 Å^3 and area in Å^2, the conversion factor is $9.65 \approx 10$. Note that solvent exposure differs for a monomer and a dimer: here, it is calculated in the presence of all monomers in the data set (Table AI) but omitting HETATOMs (substrates, ligands, heme, etc.). The sum over all residues is the total solvent exposure of the protein.

RESULTS AND DISCUSSION

Choice of Proteins

Of the more than 100 coordinate data sets in the Protein Data Bank,[2] about 75 have complete backbone coordinates and a known amino acid sequence. When two protein data sets had more than a 50% sequence homology, i.e., identical amino acids in equivalent positions, the one with higher resolution, better refinement, or more secondary structure was chosen as representative, e.g., the first one was chosen of these pairs: serine proteinase 1SGA=1SGB by 61%; lactate dehydrogenases 4LHD=1LDX by 63%; carbonic anhydrase 1CAC=1CAB by 60%; chymotrypsin 2GCH=2CHA by 98%. Both were chosen of the following pairs: sulfhydryl proteinases actinidin/papain 2ACT=8PAP by 47%; immunoglobulins 1FAB=1REI by 47%; cytochrome c550/c2 155C=1C2C by 43%; chymotrypsin/trypsin 2GHA=1PTN by 42%; elastase/trypsin 1EST=1PTN by 38%; acid protease/penicillopepsin 1APR=1APP by 43%; α/β subunit of hemoglobin 2MHB(α)=2MHB(β) by 44%. The final 62 data sets thus cover essentially all known different protein structures, except those not deposited with the protein data bank (Table AI).

H-Bonded Structure

Backbone–backbone H bonds can be simply classified by the number of residues they bracket or, in our notation, by n of $(i,i+n) = (\text{CO}(i),\text{NH}(i+n))$. Let us discuss the structural role of H bonds for each n.

H bonds $n = 0$ and $n = 1$ are sterically disallowed. A hydrogen bond $(i,i+2)$ can be formed between two consecutive peptide units for certain ϕ,ψ values of residue $i + 1$. This local conformation is known as C_7 and leads to an extended strand roughly similar to a β-strand if it repeats. When it occurs as part of a tight turn, that turn is sometimes called a γ-turn.

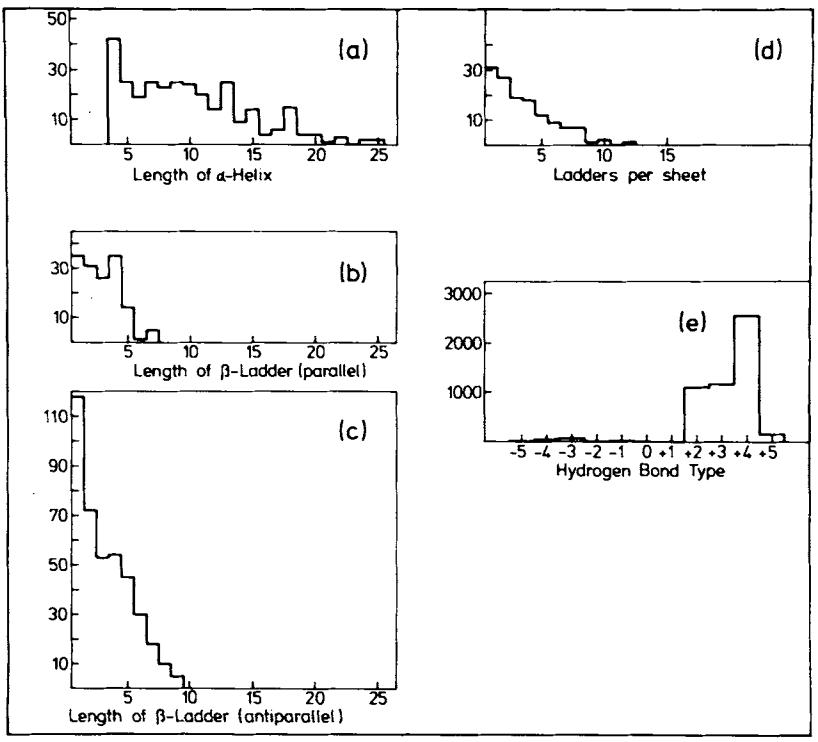


Fig. 6. The common feature of the size distribution of secondary structure segments is the gradual fall-off: larger sizes are less probable than smaller ones. Note that we give (b,c) the length of β-ladders (strand pairs) rather than the length of β-strands. A strand is often longer than the ladders in which it participates, since sheets tend to be trapezoidal rather than rectangular in shape. The number of bulge-linked ladders per sheet (d) is given as an indication of the width of the sheet. The width of a ladder is about 5 Å. In an ideal sheet, center strands take part in two ladders, edge strands in one: the number of ladders is equal to the number of strands minus one. In general, however, one strand can participate in more than one ladder on each side and the width of the sheet less than the number of ladders times 5 Å. Note: sheets consisting of a single bridge are not included in the histogram of ladders per sheet. (e) Number of H-bonds of type (CO(i),NH($i + n$)). Due to the nature of L-amino acids, positive n are heavily favored. The dominant peak at $n = 4$ represents α-helices and 4-turns. We find that H bonds (i,$i + 2$) and (i,$i + 3$) are surprisingly common, though generally weak.

Using our H bond definition, we find that many β-strands have, in addition to the main interstrand H bonds, minor (i,$i + 2$) intrastrand H bonds [see peak in Fig. 6(e)]. These reflect part of the electrostatic stabilization of extended conformations due to the polar interaction of the C-O and N-H groups of adjacent peptide units, first shown by Flory's group[20] to be essential in stabilizing the C_7 conformation in solution. We speculate that β-strands originate as extended C_7 strands as the protein folds up. Outside of β-strands, we typically find one or two weak ($E < -1.0$ kcal/mol) (i,$i + 2$) H bonds per 100 residues, but most of them are neither repeating nor part of a tight turn.

H bonds with $n = +3,+4,+5$ are reported as turns or helices. Most $(i,i + n)$ hydrogen bonds for $n > 5$ or $n < -5$ are part of a bridge or ladder. Interestingly, H bonds $(i,i - 2)$, $(i,i - 3) \ldots (i,i - 5)$ are also rare. There is steric hindrance, e.g., in an $(i,i - 4)$ helix between the backbone oxygen and the first side-chain atom C^{β}.

3-Helices are more frequent than previously believed, although they are usually short and have mediocre hydrogen bonds. α-Helices are rarely entirely pure: numerous H bonds in them are bifurcated, i.e., $(i,i + 4)$ and $(i,i + 3)$ or sometimes $(i,i + 5)$. The ends of α-helices often are overwound, ending in a 3-turn or 3-helix, or underwound, ending in a 5-turn. Some of these cases were already noted and generalized by Schellman[21] and Richardson.[8] We even find a few 5-helices (π-helices)—see Fig. 3.

Tabulation of the relative number of H bonds in Table AI may be useful in calibrating spectroscopic determination (CD, laser Raman) of the percentage of secondary structure (e.g., by the algorithm of Provencher and Gloeckner[22]). In particular, we suggest that the distinction between parallel and antiparallel β-structure[23,24] in the reference spectra will improve the overall accuracy of these experiments.

Accuracy of H-Bond and Secondary Structure Assignments

At best, secondary structure assignments can only be as accurate as the coordinates on which they are based. In using this dictionary, it is therefore very important to be aware of the state of resolution and refinement of each structure indicated in Table AI. The coordinate data sets range from refined structures at better than 1.5-Å resolution, where individual side chains can clearly be seen, to unrefined structures at a resolution just sufficient to trace the protein chain. As a test, we compare our assignments with those of the crystallographers and of Levitt and Greer[11] for three proteins of 1.5, 2.5, and 3.0 Å resolution (Table I).

For the *higher-resolution* structure of trypsin inhibitor (3PTI), Deisenhofer and Steigemann[25] assign an H bond when the N-O distance d is no greater than 3.1 Å and list 18 backbone–backbone H bonds. Of these, we find all except Tyr35(CO)-Ala16(NH), which has $d = 3.1$; instead, we have Gly36(CO)-Ala16(NH), which has $E = -2.2$. In addition, we assign 11 others, due to the rather generous energy cutoff in our definition. One, Tyr35(CO)-Ile18(NH) is quite strong, with $E = -2.0$, consistent with the slow hydrogen-exchange rate of 2.6×10^{-5} min^{-1} measured by nmr.[26] Three others of type $(i,i + 3)$, with $E = -1.3, -1.7, -0.9$, form the well-known[8] 3-helix Asp3-Leu6. One $(i,i + 5)$ H bond, Asn24(CO)-Leu29(NH), is part of the β-hairpin. Six are of type $(i,i + 2)$, characteristic of the C_7 configuration: five weak ones and one stronger one ($E = -1.8$) in a γ-turn at Asn43. The additional H bonds assigned by us lead to identification of two unambiguous segments of secondary structure not cited by the authors but also assigned by Levitt and Greer.[11]

For the *medium-resolution* structure of cytochrome c550, Timkovich and Dickerson[27] use a conservative interpretation of hydrogen bonds and

TABLE I
Comparison of Secondary Structure Assignments for Three Proteins of Higher, Medium, and Lower Resolution

Structure[a]	Original Authors (AU)	Levitt & Greer (LG)	This Work (KS)	
3PTI				
G1	—[b]	2–7	3–6	Clearly 3_{10}; LG have α
E1	16–25	14–25	18–24	
E2	28–36	28–37	29–35	
E3	—[b]	43–46	45–45	β-Bridge, 2 H bonds
H1	47–56	47–55	48–55	
155C				
H1	6–11	4–16[b]	6–12	4-Turn 13–16
G1	11–13	—	11–13	Overlaps with H1
E1	—	17–23[b]	19–20	AU have 2 H bonds; KS, 4
E2	—	26–31[b]	—	Discontinuity at Asp28–Ile29
E3	—	33–39	35–37	AU have 2 H bonds; KS have 4
H1	—	40–44[b]	—	KS have 3-Turn
H2	56–63	55–65	56–64	
H3	73–79	71–80	73–80	
H4	—	81–90[b]	—	Pro at 82, 84; possible helix
H5	107–118	106–118	107–117	
2ADK				
H1	1–8	1–7	2–7	
E1	10–14	8–15	10–14	
H2	23–30	21–31	23–31	
E2	35–38	34–38	35–38	
H3	41–48	39–49	39–48	
H4	53–62	52–61	52–62	
H5	69–84	68–83	69–83	
E3	90–94	88–95	90–93	
H6	100–107	100–109	101–108	
E4	114–118	113–120	114–118	
H7	123–133	121–136[b]	122–132	α-Helix ends in 3-turn
H8	144–158	141–157[b]	143–157	No $(i,i+4)$ H bond at Asp 141
H9	160–164	159–166	160–167[b]	Two weak H bonds at 167,168
E5	169–173	169–175	170–173	
H10	179–194	179–192	179–193	

[a] H = α-helix, G = 3_{10}-helix, E = β-strand. 3PTI = pancreatic trypsin inhibitor, 1.5-Å resolution, Diamond real-space refinement (Ref. 25). 155C = cytochrome c550, 2.5-Å resolution, Diamond model building to guide coordinates, assignments derived from the H-bonding diagram of Ref. 27. 2ADK = adenylate kinase, 3.0-Å resolution, unrefined (Ref. 28).

[b] Serious discrepancy (segment missing or boundary different by three or more residues).

give a minimal set of 41 backbone–backbone H bonds. We assign all of these, except Ala115(CO)-Gln119(NH), at the end of an α-helix; instead, we see the helix end with the $(i,i+3)$ H bond Ala115(CO)-Asp118(NH). We assign an additional 24 H bonds, of which 7 are the secondary partners of a bifurcated H bond, which is common in helices, and 8 others are marginal, with $E > -1.0$ kcal/mol. Of the remaining 9, four are of type $(i,i+$

2) in approximate γ-turns at Glu2, Gly40, Lys 53, and Lys88; two are (i,i + 4) H bonds at the end of α-helices; two are (i,i − 3) and (i,i − 6) in the loop region Gln22-Asp28; and one is involved in forming the heme pocket by a tertiary contact between Thr80(CO) at a helix end and Met103(NH) in an extended strand. All of these have a meaningful structural interpretation. The resulting secondary structure assignments are consistent with the authors' H-bond list, except for the additional short parallel bulged β-strand pair, 19–20/35–37, which is due to two additional weak H bonds. Levitt and Greer[11] assign considerably more secondary structure (Table I), including a much longer parallel β-sheet 17–23/33–39 (probably too long), a β-strand 26–31 (roughly antiparallel to 17–23), a helix 40–44 (we assign a 3-turn), and a longer helix 81–90 (which has only two of the seven possible H bonds but looks very much like a helix in a C^{α} chain tracing and therefore may be seen to be a helix at higher resolution).

For the unrefined, *lower-resolution* structure of adenylate kinase (2ADK[28]), all secondary structure assignments (ours, the original authors',[28] and Levitt and Greer's[11]) are similar. Other lower-resolution coordinate data sets show more discrepancies, depending on the quality of the H bonds.

This detailed comparison shows that our H-bond energy cutoff, chosen out of necessity to allow for coordinate errors in lower-resolution data, typically leads to 50% more H bonds than conservative assignments in higher-resolution data (example, 3PTI). All these have a physical meaning in terms of electrostatic interaction energy and nearly all have an interpretation in terms of canonical secondary structure; and, most importantly, the increased number of H bonds does not give rise to spurious secondary structure assignments.

H-bond assignments become less certain for some lower-resolution data. For example, in the data sets 1APR, 3PGM, and 1ABP, Richardson[8] sees a number of β-strands, which, in Table AIII, do appear as uncurved (non-"S") strands but with relatively few H-bonded bridges between them. At least for 1APR, only partially refined at 2.5-Å resolution with tentative amino acid sequence, one may expect that more H bonds will form in the β-sheets on further refinement.

We conclude that our criteria for H-bonded secondary structure are relatively strict, in spite of a generous cutoff in the H-bonding energy. For higher-resolution data sets, our assignments are more accurate than those of Levitt and Greer,[11] and for lower-resolution data, they are conservative compared with both Levitt and Greer's program and Richardon's[8] visual processing.

Secondary Structure Size

What is the extent of secondary structure cooperativity? Are there any preferred lengths of secondary structure segments? The length distributions [Fig. 6(a–c)] fall off almost monotonically with increasing length up

to a maximum segment length of about 30 Å, with parallel β-ladders slightly shorter. There appear to be no statistically significant peaks, either for an integral number of helical repeats or for typical domain sizes, with the possible exception of four-residue parallel β-ladders characteristic of the $\alpha/\beta/\alpha$ folding unit and, perhaps, 13- and 18-residue α-helices. We speculate that protein folding, although cooperative, follows random polymer statistics approximately in that long segments are statistically less likely than short ones. The apparent maximum size of 30 Å perhaps reflects the maximum size of globular domains.

OUTLOOK

The structure of influenza virus hemagglutinin,[29] with its 50-residue helix, shows that our data base certainly does not exhaust all possible variations in protein architecture. In spite of this limitation, this compilation will be used in the ongoing development of protein structure prediction methods.

APPENDIX: DICTIONARY OF PROTEIN SECONDARY STRUCTURE

Notes to Table AI

Proteins are ordered by function and can be found in the strip tables (Table AIII) and strip maps (Fig. AI) by their running number. % α-helix, % β-antiparallel, % β-parallel = number of H bonds per 100 residues of type 4-turn, parallel and antiparallel bridge; these percentages can be compared with results from spectroscopy (CD, Raman, ir). Exposure = estimated number of water molecules in contact with protein surface (first hydration shell); it can also be read as the static exposed surface area in units of 10 Å^2. Exposure is calculated for the entire data set and then divided by the multiplicity of sequence-unique molecules, e.g., the data set 1INS has two copies each of the insulin A- and B-chain (multiplicity 2). Exposure given is that of the A- and B-chain in the tetramer. Number of residues is also for the sequence-unique molecule. Crystallographic resolution (Å) and refinement give some indication of the quality of the coordinates; both are taken from the Data Bank without further checking. In case of doubt, consult the original papers. Refinement code: D1 = Diamond model building to guide coordinates (Ref. 30); D2 = Diamond real-space refinement (Ref. 31); HK = Hendrickson-Konnert (Ref. 32); DO = Dodson, Isaacs, and Rollett (Ref. 33); JL = Jack and Levitt (Ref. 34); DS = Deisenhofer and Steigemann (Ref. 25); DF = difference Fourier; DC = difference Fourier with constraints; FD = difference Fourier and D1; LS = least squares; RL = restrained least squares; CL = constrained least squares; SD = steepest descent; LL = energy minimization of Levitt and Lifson (Ref. 35); HH = D2 and Hermans' REFINE2 and HK; DD = DS and D2; DL = DF and LS; DJ = D2 and JL; AD = Agarwal least squares (Ref. 36) and DO; DH = D2 and HK; DE = D2 and LL; MD = energy minimization of McQueen and DO; CS = constrained difference Fourier of Chambers and Stroud (Ref. 37); RE = real space and energy minimization; CC = constrained crystallographic refinement; CD = D2 and CORELS (Ref. 38).

TABLE AI

List of 62 Different Globular Proteins

%AH	%BA	%BP	EXPO	M	LEN	RES	RF	PROTEIN IDENTIFIER, NAME
								% ALPHA HELICAL AND 4-TURN HYDROGEN BONDS
								% BETA ANTIPARALLEL HYDROGEN BONDS
								% BETA PARALLEL HYDROGEN BONDS
								WATER EXPOSURE
								MULTIPLICITY OF DATA SET
								NUMBER OF RESIDUES
								RESOLUTION
								REFINEMENT
binding proteins								
38	4	0	610		108	1.9	DF	1) 1CPV CALCIUM-BINDING PARVALBUMIN B
31	1	6	1423		306	2.4	--	2) 1ABP L-ARABINOSE-BINDING PROTEIN
electron transfer								
7	13	2	490		85	2.0	DF	3) 1HIP OXIDIZED HIGH POTENTIAL IRON PROTEIN (HIPIP).
electron transport								
19	15	7	566		85	2.8	D2	4) 2B5C CYTOCHROME B5 (OXIDIZED)
57	0	0	665		103	2.5	--	5) 156B CYTOCHROME B562 (E. COLI, OXIDIZED)
34	2	0	620	2	103	2.0	RL	6) 1CYT CYTOCHROME C (OXIDIZED).
23	2	2	642		112	2.0	DC	7) 1C2C CYTOCHROME C2 (FERRI)
23	0	3	781		134	2.5	D1	8) 155C CYTOCHROME C550
38	2	0	482		82	2.0	CS	9) 251C CYTOCHROME C551 (OXIDIZED)
9	9	0	316		54	2.0	DC	10) 1FDX FERREDOXIN (PEPTOCOCCUS AEROGENES)
0	0	0	623		98	2.8	--	11) 1FXC FERREDOXIN (SPIRULINA PLATENSIS)
30	1	18	715		138	1.9	DE	12) 3FXN FLAVODOXIN (OXIDIZED)
7	17	0	376		54	1.5	LS	13) 2RXN RUBREDOXIN (OXIDIZED, FE(III))
10	17	7	645		125	2.7	HK	14) 1AZU AZURIN
2	21	10	513		99	1.6	DH	15) 1PCY PLASTOCYANIN
hormones								
44	0	0	343		36	1.4	RL	16) 1PPT AVIAN PANCREATIC POLYPEPTIDE
38	0	0	354		29	3.0	RE	17) 1GCN GLUCAGON (PH 6-7)
29	12	0	301	2	51	1.5	DL	18) 1INS INSULIN (A AND B CHAIN)
hydrolase, phosphatide acyl								
37	7	0	712		123	1.7	AD	19) 1BP2 PHOSPHOLIPASE A2
hydrolases, O-glycosyl								
38	7	0	918		164	2.4	CL	20) 1LZM LYSOZYME (BACTERIOPHAGE T4)
24	8	2	665		129	2.5	CD	21) 7LYZ LYSOZYME (HEN EGG WHITE, TRICLINIC)
hydrolases, phosphoric diester								
19	20	3	842		142	<4	--	22) 1SNS STAPHYLOCOCCAL NUCLEASE (COMPLEX)
14	28	2	709		124	2.0	SD	23) 1RNS RIBONUCLEASE-S
hydrolases, proteinases								
28	5	9	1209		308	2.0	--	24) 1CPA CARBOXYPEPTIDASE A
3	13	3	1333		324	2.5	HK	25) 1APR ACID PROTEASE (RHIZOPUS CHINENSIS)
7	32	6	1272		323	2.8	D2	26) 1APP ACID PROTEINASE (PENICILLOPEPSIN, FUNGUS)
27	9	4	1266		316	2.3	D2	27) 2TLN THERMOLYSIN
6	31	1	1033		236	1.9	HH	28) 2GCH GAMMA CHYMOTRYPSIN A
3	37	3	821		198	2.8	D1	29) 1ALP ALPHA LYTIC PROTEASE
7	31	1	929		223	1.5	D2	30) 1PTN BETA-TRYPSIN (NATIVE AT PH 8)
4	34	2	745		181	2.8	D1	31) 1SGA PROTEINASE A FROM STREPTOMYCES GRISEUS (SGPA)
23	5	12	1058		275	2.5	FD	32) 1SBT SUBTILISIN BPN'
4	35	0	1089		240	2.5	--	33) 1EST TOSYL-ELASTASE
21	19	1	923		218	1.7	LS	34) 2ACT ACTINIDIN
19	15	1	968		212	2.8	D1	35) 8PAP PAPAIN
immunoglobulins								
1	34	2	2101		428	2.0	--	36) 1FAB LAMBDA IMMUNOGLOBULIN FAB
1	37	5	492	2	107	2.0	CC	37) 1REI BENCE-JONES IMMUNOGLOBULIN (VARIABLE PORTION)
isomerases								
23	2	5	1220		230	2.8	HK	38) 3PGM PHOSPHOGLYCERATE MUTASE (DE-PHOSPHO)
35	1	15	1026	2	246	2.5	DO	39) 1TIM TRIOSE PHOSPHATE ISOMERASE
lectin (agglutinin)								
2	35	0	1125		237	2.4	MD	40) 3CNA CONCANAVALIN A
lyase, carbon-oxygen								
4	20	5	1273		256	2.0	D1	41) 1CAC CARBONIC ANHYDRASE FORM C
oxidoreductases								
15	12	13	937		162	2.5	--	42) 1DFR DIHYDROFOLATE REDUCTASE (COMPLEX)
22	11	10	1505	2	333	2.9	D1	43) 1GPD D-GYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE
17	11	9	1639		374	2.4	DJ	44) 4ADH APO-LIVER ALCOHOL DEHYDROGENASE
27	6	7	1753		329	2.0	D2	45) 4LDH LACTATE DEHYDROGENASE, APO ENZYME M4
22	12	7	2354		461	2.0	--	46) 2GRS GLUTATHIONE REDUCTASE
1	33	1	686	4	151	2.0	HK	47) 2SOD CU,ZN SUPEROXIDE DISMUTASE
oxygen storage								
65	0	0	842		153	2.0	D2	48) 1MBN MYOGLOBIN (FERRIC IRON - METMYOGLOBIN)
oxygen transport								
62	0	0	706		136	1.4	DS	49) 1ECD HEMOGLOBIN (ERYTHROCRUORIN DEOXY)
58	0	0	1415		287	2.0	DD	50) 2MHB HEMOGLOBIN (HORSE, AQUO MET)
47	0	0	864		148	2.0	D1	51) 1LHB HEMOGLOBIN(MET)-CYANIDE V (SEA LAMPREY)
62	0	0	824		153	2.0	D2	52) 1HBL LEGHEMOGLOBIN (ACETATE,MET) (YELLOW LUPIN)
plant seed protein								
33	6	0	301		46	1.5	HK	53) 1CRN CRAMBIN
proteinase inhibitors								
13	14	2	351	4	56	1.9	DJ	54) 1OVO OVOMUCOID THIRD DOMAIN
13	23	2	632		107	2.6	DO	55) 2SSI STREPTOMYCES SUBTILISIN INHIBITOR
12	17	0	412		58	1.5	D2	56) 3PTI TRYPSIN INHIBITOR
toxins								
3	17	0	511		71	2.8	--	57) 1CTX ALPHA COBRATOXIN
71	0	0	222	2	26	2.0	HK	58) 1MLT MELITTIN
0	29	0	406		62	1.4	HK	59) 1NXB NEUROTOXIN B (PROBABLY IDENTICAL TO ERABUTOXIN B
transferases								
47	0	10	1251		194	3.0	--	60) 2ADK ADENYLATE KINASE
20	2	10	1456		293	2.5	D1	61) 1RHD RHODANESE
transport								
7	33	7	652		114	1.8	DO	62) 2PAB PREALBUMIN (HUMAN PLASMA)

TABLE AII
Structure Notation Used in Table AIII

First line: running number 1–62, data set identifier (3PTI,4LDH . . .), protein name, [function], {source}	
SHEET. . .	One-character name of β-sheet ("A," "B," "C" . . .) in which residue i participates.
BRIDGE2 . . .	One-character name of β-ladders in which residue i participates,
BRIDGE1 . . .	"A," "B," "C" . . . = antiparallel, "a," "b," "c" . . . = parallel. Ladders are named sequentially from N- to C-terminus. A β-strand can be part of two ladders, one to each side, so there are two lines for the possible ladder partners. Each ladder name appears twice, once for each participating strand. Partner strands can thus be easily identified by identical letters. The sheet topology can be reconstructed by starting from a β-strand and tracing all partners and their partners.
CHIRALITY	"+" or "−" Chirality at residue i is the sign of the dihedral angle defined by C^{α} $i-1$ to $i+2$. Thus, a right-handed α-helix has "+," an ideal twisted β-strand "−."
BEND . . .	"S" = five-residue bend centered at residue i.
5-TURN . . .	Hydrogen-bonding pattern for turns and helices:
4-TURN . . .	")" = backbone CO of this residue makes H bond $(i,i+n)$
3-TURN . . .	"(" = backbone NH of this residue makes H bond $(i-n,i)$ "X" = both CO and NH make H bond "3," "4," "5" = residues bracketed by H bond
SUMMARY . . .	Structure summary: "H" = 4-helix (α-helix) "B" = residue in isolated β-bridge "E" - extended strand, participates in β-ladder "G" = 3-helix (3_{10}-helix) "I" = 5-helix (π-helix) "T" = H-bonded turn "S" = bend In case of structural overlaps, priority is given to the structure first in this list.
EXPOSURE . . .	Solvent exposure is the estimated number of water molecules in contact with residue i. The scale is 0–9; "*" = more than 9 water molecules. Exposure can be read as solvated surface area in units of 10 Å^2.
SEQUENCE . . .	Amino acid sequence in one letter code: "a," "b," "c" . . . are Cys residues labeled by their SS-bond name. "!" = chain break (peptide bond length exceeds 2.5 Å). Residues including chain breaks are numbered sequentially within the coordinate data set, irrespective of the residue identifier given there. Thus, the total number of residues is equal to the total number of print positions minus the number of chain breaks.