PROTEINS: Structure, Function, and Genetics 3:71–84 (1988)

Identification of Structural Motifs From Protein Coordinate Data: Secondary Structure and First-Level Supersecondary Structure*

Frederic M. Richards and Craig E. Kundrot
Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06511

ABSTRACT **A computer program is described that produces a description of the secondary structure and supersecondary structure of a polypeptide chain using the list of alpha carbon coordinates as input. Restricting the term "secondary structure" to the conformation of *contiguous segments* of the chain, the program determines the initial and final residues in helices, extended strands, sharp turns, and omega loops. This is accomplished through the use of difference distance matrices. The distances in idealized models of the segments are compared with the actual structure, and the differences are evaluated for agreement within preset limits. The program assigns 90–95% of the residues in most proteins to at least one type of secondary element.**

In a second step the now-defined helices and strands are idealized as straight line segments, and the axial directions and locations are compiled from the input Cα coordinate list. These data are used to check for moderate curvature in strands and helices, and the secondary structure list is corrected where necessary. The geometric relations between these line segments are then calculated and output as the first level of supersecondary structure. A maximum of six parameters are required for a complete description of the relations between each pair. Frequently a less complete description will suffice, for example just the interaxial separation and angle. Both the secondary structure and one aspect of the supersecondary structure can be displayed in a character matrix analogous to the distance matrix format. This allows a quite accurate two-dimensional display of the three-dimensional structure, and several examples are presented.

A procedure for searching for arbitrary substructures in proteins using distance matrices is also described. A search for the DNA binding helix-turn-helix motif in the Protein Data Bank serves as an example.

A further abstraction of the above data can be made in the form of a metamatrix where each diagonal element represents an entire secondary segment rather than a single atom, and the off-diagonal elements contain all the parameters describing their interrelations. Such matrices can be used in a straightforward search for higher levels of supersecondary structure or used in toto as a representation of the entire tertiary structure of the polypeptide chain.

Key words: Cα coordinates, distance matrix, difference distance matrix, helix axes, strand axes, interaxial angles, turns

INTRODUCTION

Based on packing considerations Ponder and Richards[1] have suggested that each class of tertiary structure of proteins can be represented by a tertiary template, a list of sequences compatible with a given structure. A basic assumption in the template approach is that the classes are clearly separable and do not represent simply points in a continuous distribution of structures. There are, of course, variations within each class, and these atomic shifts present a major difficulty in calculating such templates. Objective and quantitative estimates of such variations in the secondary and tertiary structure of known proteins are necessary. This need led to the study reported here.

The direct result of an X-ray crystallographic structure determination is a list of the three-dimensional coordinates of each of the nonhydrogen atoms and an estimate of their positional variation as reflected in the temperature factors. While this list represents the full information available and is appropriate for computer queries, it is otherwise indigestible and is reduced to other forms for the purposes of description, discussion, and comparison. One can construct a physical model or, more commonly today, a computer graphics display of the model. From either form, the first stage of the description is the enumeration of the elements of secondary structure. This is normally done by visual inspection.

A number of authors have proposed algorithms for the automatic listing of secondary structure. A partial list is given in Table I. Idealized structures for strands, helices, and turns have been described in great detail. The conformational angles, interatomic distances, and specific interactions such as hydrogen bonds are defined precisely by the models. Difficulties

*A FORTRAN version of the program DEFINE_STRUCTURE, which defines the secondary structure and characterizes the first level of supersecondary structure, is available from the authors.

Received December 18, 1987; accepted December 30, 1987.

Address reprint requests to Dr. Frederic M. Richards, Department of Molecular Biophysics and Biochemistry, Yale University, P.O. Box 6666, 260 Whitney Ave., New Haven, CT 06511.

Craig E. Kundrot is now at MRC Laboratory of Molecular Biology, Hills Road, Cambridge, CB2 2QH, UK.

TABLE I. Partial List of Methods for Identifying Secondary Structure

	Method								
	Structure identified								
Reference	α helix	3_{10} helix	β strand	Turn	Cα distance	Cα torsion angle	Phi psi angles	Hydrogen bonds	Other
Lewis et al.[25]	–	–	–	+	+	–	–	–	
Kuntz[26]	–	–	–	+	+	+	–	–	
Crawford et al.[27]	–	–	–	+	+	–	+	+	
Levitt and Greer[28]	+	–	+	+	+	+	–	–	Pseudo HB from Cα coordinates
Rose and Seltzer[29]	+	–	+	+	–	–	–	–	Cα radius of curvature
Chou and Fasman[30]	+	–	–	+	+	–	+	+	HB not strictly required
Kolaskar et al.[31]	+	–	–	+	+	–	–	–	
Ramakrishman and Soman[32]	+	+	+	+	+	+	–	–	Cα virtual angle
Kabsch and Sander[33]	+	+	+	+	–	–	–	+	Cα radius of curvature, chirality
Hohne and Kretschmer[34]	+	–	+	–	–	–	+	+	

arise when the models are compared to actual structures. For example, the general appearance of a helix is clear, but this structure can be preserved with substantial changes in the parameters normally used in a definition. Rotation of the main chain angles psi_n, phi_{n+1} of a particular peptide bond "dog bone" through a large angular distance can be carried out with only marginal perturbation of the rest of the structure. Such action changes the position of the H and O atoms and thus the geometry of at least two hydrogen bonds. Alternatively, the pitch of the helix may change, and thus the H bond lengths, with only minor changes in the phi-psi angles. While the α helix is characterized by 1–4 H bonds and the 3_{10} helix by 1–3 bonds, it is not uncommon to find bifurcated 1–3, 1–4 bonds, which make the assignment of the helix type ambiguous. In loops, turns, and extended chains such variations can be even more pronounced.

Ultimately, the precise delineation of secondary structures in a real molecule depends on which parameters are chosen for the definitions and how closely each segment is required to match the ideal definition. Algorithms based on phi-psi angles, on H bond lengths, or on Cα distances may be expected to give different answers, even though each, within its context, is correct. For each approach the match of the ideal and actual structures will have to allow for some disagreement or mismatch. If the amount of acceptable mismatch is set too small, the real structures will not be recognized; if too large, too many structures will "fit." The sensitivity of the results to the level of acceptable mismatch is important for each method. Although interatomic distances are used in a number of the reported algorithms, no one appears to have systematically used the difference distance matrix approach for structure analysis. In this work, the results obtained using interatomic distances normally agree very well with an intuitive visual examination of the structure. The method possesses the additional advantage of producing a quantitative measure of the quality of the fit to idealized models. This fit can be specified easily either through overall averages or down to the level of individual atoms.

The matrix of distances between all pairs of atomic centers provides an alternative to the Cartesian coordinate representation of a structure. With the exception of the handedness of the molecule, all of the structural information is contained in such a distance matrix. In contrast to the coordinate list, this representation is invariant to rotation or translation of the molecule. Nuclear magnetic resonance (NMR) techniques, which provide interatomic distance information, lead directly to a distance matrix representation of the protein. In this formulation, a structure composed of N atoms produces an N × N matrix whose elements (i,j) are the distances between atoms i and j. The matrix is easily calculated from a coordinate list, and the latter can be regenerated from a distance matrix with the techniques of distance geometry.[2] To maintain manageable size and yet retain the necessary data to define the main chain conformation, the subset of the matrix containing only the distances between Cα positions is usually used. Thus the method is useful in the early stages of a structure analysis where the Cα positions may be the only reasonably reliable molecular model data available.

Since the Cα distance matrix (DM) was introduced by Phillips[3] the most common graphic presentation has been the display of only those elements corresponding to distances shorter than some preset value. Thus, only atom pairs that are close together in space are represented in the diagram as nonblank ele-

ments. This leads to a characteristic thickening along the diagonal for α helices and to somewhat irregular lines parallel or perpendicular to the diagonal for parallel or antiparallel β strands, respectively. While qualitatively useful, this presentation does not do justice to the amount of information contained in the distance matrix.

Nishikawa et al.[4] pointed out some years ago that difference distance matrices (DDM) represent a powerful approach in comparing two structures. Element (i,j) in a DDM is the difference in distance between atoms i and j in two structures (e.g., the two structures might be the conformation of a protein with and without a bound ligand). Those parts of the structures that are identical will have identical distances between corresponding atoms and will produce values of zero in the appropriate DDM elements. Such difference matrices were used by Padlan and Davies[5] in comparisons of immunoglobulins and more recently by Kundrot and Richards[6] to show the effect of hydrostatic pressure on lysozyme, by Frauenfelder et al.[7] to show the effect of temperature on myoglobin, and by Elber and Karplus[8] to examine the changes occurring during molecular dynamics simulations. This type of matrix approach has also been implemented in the crystallographic modeling program FRODO to match putative Cα positions with Cα positions in a reference library of known structures.[9] While the application to chains of identical length is straightforward, Liebman[10,11] has pointed out how to handle chain length variation produced by insertions and deletions.

For many purposes the DDM approach is preferable to comparisons based on rms fitting procedures. The DDM procedure locates in detail the areas of identity, quantitates differences in structure down to the level of atom pairs, can accomodate insertions and deletions in the structure, and is very fast computationally. As two different descriptions of molecular structure, the DM and the Cartesian coordinate list are complementary, and the choice of representation will depend on the application.

In this paper, methods based both on matrix representations and on Cartesian coordinates are used to define secondary and supersecondary structure in proteins. The procedures are used sequentially. A DDM method is described for identifying secondary structure elements. Then a method based on the Cartesian coordinates is presented that describes the geometrical relationships of the previously defined secondary elements, i.e., the first level of supersecondary structure. The methods for searching for particular supersecondary structures follow. One is based on the detailed DDM approach, while the other matrix-based procedure builds on the previous Cartesian characterization of supersecondary relations. Sample applications of these methods are described, and possible extensions are discussed.

PROCEDURES

The distance matrices in this paper refer only to Cα atoms, and are all indexed with the N terminus of the chain(s) in the upper left corner. Since the matrices are symmetrical about the diagonal, only the upper triangle is shown. Distances between a set of contiguous Cα atoms appear in an upper triangular submatrix that has diagonal elements of the matrix for the hypoteneuse; such submatrices will be referred to as "triangles." Distances between two separate stretches of contiguous residues appear in an off-diagonal, rectangular submatrix, and these submatrices are referred to as "boxes." The distance between $C\alpha_i$ and Ca_j is abbreviated (i,j). Submatrices whose elements are the distances between atoms in an idealized structure are referred to as "masks."

Secondary Structure

A DM-based appoach is used to define elements of secondary structure. The distance between two consecutive Cα atoms (i,i+1) depends on the two intervening main chain conformational angles. However, the planarity of the peptide bond constrains the conformation enough to make the (i,i+1) elements adopt a narrow range of values. The (i,i+2) elements also adopt a narrow range of values since the (i−1, i,i+1) angle is essentially fixed. Neglecting experimental error in the coordinates, for any Cα distance matrix the diagonal elements are always zero, the (i,i+1) elements are 3.75 $\pm$ 0.02 Å and the (i,i+2) elements are 5.9 $\pm$ 0.6 Å for transpeptide bonds. Thus, distances characteristic of chain conformation begin to appear only at the third position (i,i+3), where the distance in an α helix is 5.0 Å and in an extended strand is about 9.9 Å.

For the reasons mentioned below, the secondary structure assignments are made in the order: 1) cis peptide bonds, 2) α helices, 3) 3_{10} extensions to α helices, 4) sharp turns, 5) extended strands, and 6) loops.

Cis peptide bonds

Cis peptide bonds in proteins are not common, but they do occur, most often involving the amide group of a proline residue. The (i,i+1) distance in a cis bond is about 2.9 Å, and a distance of 2.9 $\pm$ 0.2 Å is therefore flagged as a cis bond. On the same pass through the coordinate list all other (i,i+1) distances are checked to make sure they are in the expected range for a transpeptide bond: 3.8 $\pm$ 0.2 Å.[12]

The α helix and 3_{10} helical extensions

After the check for cis peptide bonds, the protein distance matrix is surveyed for matches to the α-helical mask. The helix is the most rigid and dimensionally well defined of the secondary structural elements. Once tested and accepted, therefore, the α-

helical residues are not checked against subsequent masks for other secondary structural elements. The distance submatrix for an ideal α helix of L residues will be a triangle with elements containing the distances given in Table II. A helix with $C\alpha_i$ at the N terminus produces a triangle whose first row, (i,i) to (i,i+L-1), contains distances to all other atoms in the helix. The elements in the next row will be identical but shifted to the right and truncated by one matrix element per line [i.e. (i+1,i+1) to (i+1, i+L-2)]. The distances in any line of elements parallel to the principal diagonal are identical. There is no limit in principal to the length of the ideal helix, but in our current algorithm it is limited to 50 residues. This is longer than most of the α helices in globular proteins, and, given the tendency of many helices to bend or kink, it is probably longer than any stretch of an actual α helix that is straight enough to fit the mask for an ideal straight helix.

To check for the location of an α helix, each $C\alpha$ in turn is considered as a possible N terminal atom. A helix is "grown" from a given atom, $C\alpha_i$, by moving one step at a time along the horizontal line i in the DM for positions, j=i+1,i+2, etc. The atom $C\alpha_j$ is the C terminus of this growing helix. All distances within the potential helix are given in the triangle (i,i), (i,j) (j,j). The distances along the right hand vertical line (i,j), (j,j) represent the distances from $C\alpha_j$ to all the previous atoms in the helix. These distances are compared to those in the equivalent position of the ideal α helix mask. The rms difference between the mask and protein distances in this subcolumn is checked

TABLE II. Reference Structure Distances or Coordinates*

```
DATA ALPHA/
  1   0.00,  3.75,  5.36,  5.02,  6.11,  8.53,  9.75, 10.43, 12.18, 14.09,
  2  15.15, 16.32, 18.19, 19.77, 20.85, 22.33, 24.13, 25.49, 26.70, 28.36,
  3  30.01, 31.27, 32.65, 34.35, 35.84, 37.11, 38.64, 40.30, 41.67, 43.06,
  4  44.64, 46.20, 47.52, 48.97, 50.61, 52.07, 53.41, 54.95, 56.55, 57.93,
  5  59.33, 60.93, 62.45, 63.81, 65.29, 66.89, 68.33, 69.72, 71.27, 72.82/

DATA BETA/      0.00,  3.75,  6.47,  9.89, 12.94,
  1            16.28, 19.40, 22.72, 25.87, 29.17,
  2            32.34, 35.62, 38.81, 42.09, 45.28,
  3            48.55, 51.74, 55.01, 58.21, 61.48/

DATA TURN/      0.00,  0.00,  0.00,  3.70,  0.00,
  1             0.00,  5.60,  3.70,  0.00,  5.00,
  2             5.10,  3.70,  5.40,  6.90,  6.10/

DATA BB/        4.95,  6.20,  8.45, 11.45, 14.55,
  1            17.65, 20.90, 24.15, 27.45, 30.65,
  2            34.00, 37.25, 40.55/

DATA A310/
  1   -0.042,  -1.388,   1.834,   1.472,  -1.616,  -1.673,   3.013,   1.872,  -1.334,
  2    4.464,   1.004,   2.097,   5.939,  -2.228,   0.679,   7.427,  -0.325,  -2.287,
  3    8.917,   2.295,   0.069,  10.402,  -0.459,   2.268,  11.884,  -2.174,  -0.812,
  4   13.360,   1.146,  -2.011,  14.844,   1.792,   1.456,  16.334,  -1.730,   1.525,
  5   17.815,  -1.221,  -1.965,  19.289,   2.143,  -0.899,  20.777,   0.548,   2.244,
  6   22.266,  -2.289,   0.155,  23.743,   0.238,  -2.309,  25.222,   2.260,   0.581,
  7   26.712,  -0.916,   2.104,  28.197,  -1.898,  -1.281,  29.673,   1.605,  -1.704,
  8   31.158,   1.456,   1.818,  32.647,  -1.992,   1.096,  34.126,  -0.717,  -2.193,
  9   35.607,   2.325,  -0.409,  37.096,   0.079,   2.304,  38.572,  -2.247,  -0.349,
  1   40.053,   0.737,  -2.226,  41.558,   2.124,   1.009,  43.051,  -1.299,   1.839,
  2   44.986,  -0.755,  -1.696,  46.921,   1.843,   0.220,  48.857,  -1.132,   1.471,
  3   50.792,  -0.684,  -1.726,  52.728,   1.833,   0.295,  54.663,  -1.192,   1.424,
  4   56.598,  -0.613,  -1.752,  58.534,   1.819,   0.370,  60.469,  -1.249,   1.373,
  5   62.404,  -0.541,  -1.776,  64.340,   1.803,   0.444,  66.275,  -1.304,   1.321,
  6   68.210,  -0.467,  -1.797,  70.146,   1.783,   0.518,  72.081,  -1.357,   1.266,
  7   74.017,  -0.393,  -1.814,  75.952,   1.760,   0.591,  77.887,  -1.408,   1.210,
  8   79.823,  -0.318,  -1.829,  81.758,   1.734,   0.662/
```

*ALPHA = linear mask for a 50-residue ideal α helix; BETA = linear mask for a 20-residue ideal strand from a parallel β sheet; TURN = triangular mask (3,5) for sharp beta turn; BB = distances between one $C\alpha$ atom and the atom in a neighbor strand in a beta sheet; A310 = X Y Z coordinates for an α helix (residues 1–30) followed by a 3_{10} helix (residues 31–50). The axes are colinear with each other and with the X coordinate axis. On execution the program computes a reference distance matrix from this coordinate set.

against the allowed mismatch limit. When this value is exceeded, the helix terminus is taken to be the residue corresponding to the previous column. In practice the initial and final positions of the helix normally are well defined, the rms difference increasing sharply for the first nonhelical atom. A helix is not recorded as such unless the match to the mask extends for at least five residues.

Although substantial isolated 3_{10} helices have not yet been found in the known protein structures, it is not uncommon for several residues at the C-terminal end of an α helix to adopt the 3_{10} conformation. When the end of an α-helical section has been located, the program checks for a 3_{10} extension. A single subcolumn as used in the above mask is not sufficient for the identification of 3_{10}-like extensions on α helices since the distances observed in the region of the extension will depend on the length of the α-helical segment. Rather, an upper triangular matrix mask is computed from the coordinates of a 30-residue α helix followed by 20 residues of a colinear 3_{10} helix. The appropriate section of this mask is then selected to check for the 3_{10} extension. There is an interaction between the "appearance" of a 3_{10} segment and the allowed mismatch level, which is used to terminate both the α and 3_{10} matches. This problem is taken up in the discussion after the description of supersecondary relations.

Sharp turns

A turn provides a characteristic set of distances between a very limited set of Cα atoms rather than the indefinite sets of the helix or strand. The distance between Cα atoms in the various types of sharp turns are so similar that useful distinctions cannot be made with so few distances. The mask used at the present time consists of distances between (i,i+1,i+2,i+3,i+4) as listed in Table II. There are ten off-diagonal distances in this set, but only three of these are significantly conformationally dependent. These distances are very similar to those found in an α helix and any reasonable allowed mismatch level would produce an acceptable match to a single turn of a helix. Thus, it is essential to check for helices first and to check for turns only in the regions not assigned to helices.

The extended strand

The check for strands is made after the helix and turn checks. Extended strands are intrinsically more variable in conformation than an α helix. The wide range of phi-psi pairs permit variable degrees of curvature even within a single strand. Strands in antiparallel sheets tend to be the most fully extended, while those in parallel sheets are slightly shorter. We have used distances based on the latter as our best compromise for a "generalized strand" (see Table II). This distance set is used exactly as described above for the helix. Strand lengths of four residues or greater are noted.

Loops

Loops that are longer than tight turns have been discussed by Sibanda and Thornton,[13] and the type described as omega loops has beeen discussed by Leszczynski and Rose.[14] These structures are intrinsically less regular and more variable than any of the elements described above. The criteria specified by Leszczynski and Rose[14] for omega loops is easily implemented and only requires examination of specific elements in the distance matrix. The loop length may vary from six to 16 residues. For any given length, an omega loop must have an end-to-end distance that is less than some predetermined value, i.e., 10 Å, and that is also less than two-thirds of the maximum distance within the loop. Such loops can often be fitted around sharp turns, and loops of differing length can frequently be constructed from a common starting point, but they are not allowed to override the prior specification of helices, sharp turns, or strands and are listed separately in the secondary structure output.

Establishing helix and strand axes and a rough check for moderate curvature

At this point in the program all the elements of secondary structure have been located. This first pass has assumed that all helices and strands are straight, and the masks used for comparison have been constructed on that basis. However, it is well known that strands and even helices can be uniformly curved or bent sharply. A sharp bend is not a problem. The straight axis assumption will lead to recognition of the bend and the identification of the separate leading and following helices or strands. However, for a curved element the program will frequently insert a "break" and produce two helices or strands where it should not. The curvature leads to failure of the match to the "straight" mask before the end of the actual element is reached. In such a break a particular residue is assigned as the C terminus of the first element and the N terminus of the second and flags the need for a curvature check.

The program now returns to the Cartesian representation. The axes of the secondary segments can be determined from the coordinate list once the starting and ending residue numbers are determined most easily from the distance matrix as described above. The N to C direction is taken as positive. The vector $V_i = [(C\alpha_{i+1} - C\alpha_i) + (C\alpha_{i-1} - C\alpha_i)]$ starts at $C\alpha_i$, is perpendicular to the segment axis, and passes through, or close to, the mean axis. For strands, the axis point, A_i, is placed one-quarter of the way along V_i from $C\alpha_i$. For helices, the axis point A_i, is taken as the intersection of three planes defined by ($C\alpha_{i-1}$, $C\alpha_i$, $C\alpha_{i+1}$), (V_i, $V_i \times V_{i+1}$), (V_{i+1}, $V_i \times V_{i+1}$). For a secondary element (i, i+n) axis points are thus defined for positions i+1 to i+n−1. A segment of the axis is defined as the vector between axis points (j)

and (j+1). The direction of the axis is taken as the mean of the directions of the consecutive segments. A matrix is produced from the angles between all segment pairs. If the rms value of the matrix elements is 10° or less, the secondary unit is essentially straight. If the rms is greater than 10°, then the secondary element is curved.

In testing for curvature across an apparent break the program computes the angle matrix for the joined secondary elements. If the rms value of the terms in the matrix is less than 25°, the pair is accepted as a single curved segment, and the secondary structure list is modified accordingly. If one of the pair is short, minimum length 4 α carbon atoms, then it is "peeled" back one atom at a time to get the longest acceptable element from the pair. The deleted atoms appear in the unassigned category in the secondary structure list. If the rms value is greater than about 25°C, then the element is so distorted that it cannot be effectively approximated by a straight line and should be left in two pieces for the purpose of the current description.

Following the establishment of the final secondary structure list, the mean repeat distance between axis points and the mean radius of the Cα atom from the axis are recorded. Even though an element is moderately curved, it is still approximated by a single straight axis for the purposes described later. The axis points A_{i+1} to A_{i+n-1} are calculated from the coordinates as described bove, but the points A_i and A_{i+n}, the projections of $C\alpha_i$ and $C\alpha_{i+n}$ on the axis, can be obtained only by extrapolation of the axis in the N- and C-terminal directions. The mean repeat distance between the rest of the points is used to place A_i and A_{i+n} on the axis as measured from A_{i+1} and A_{i+n-1}. These special points are referred to subsequently as points N_k and C_k of the k_{th} secondary element. N_k is used as the origin, and the length of the element k is taken as the distance between N_k and C_k. The vector representing the axis direction and length and the point N_k together define the segment axis.

Secondary structure output

In our algorithm the results of the above searches are recorded sequentially in a *character matrix* equivalent to the distance matrix. When the character matrix is complete, it can easily be "read" to provide an output list of the secondary structure. All the matrix elements are blank except for those that correspond to matches, within the preset limits, between the observed distances in the protein and the idealized distances in the various masks. For each element of secondary structure, the upper triangle submatrix representing all Cα–Cα distances within the secondary segment is filled with characters identifying this type of secondary structure. The diagonal elements of the character matrix, however, are left blank for the following reason. It is perfectly possible for a residue to be part of two secondary structural units because the amide hydrogen and carbonyl oxygen of one residue may participate in different secondary structures. Thus, the end residue of a helix may simultaneously be the first residue of a following helix, turn, or strand and so forth. The diagonal elements of the character matrix are thus not identified as to secondary structural unit. If there is no "competition," the final residue of a segment of secondary structure is assigned uniquely to that segment in the output list. If there is competition, it is assigned to both the previous and following segments. Thus, the summed secondary structure segment lengths will normally be greater than the total number of residues in the peptide chain. This list is the full description of the conformation of consecutive segments of residues, and the match to the idealized structures is quantitatively specified through the mismatch parameters set at the time of initiation of the program. The secondary structure list is put in the output file twice, once after the first pass where all elements are compared solely to straight masks and again after the check and allowance for moderate curvature in strands and helices.

Supersecondary Structure

Although β sheets are commonly listed as part of the secondary structure, the term "secondary structure" logically should be restricted to the conformation of contiguous residues in the polypeptide chain. The first level of supersecondary structure thus involves the relationships of strand to strand, strand to helix, and helix to helix. The hydrogen bonding between strands makes β sheet structures a special class of strand–strand relations, and they are the first pairs to be considered at the supersecondary level. The distances between secondary segments appear in off-diagonal rectangular submatrices, or "boxes," in the distance matrix. In this section, methods are described for 1) identifying the extended strand–strand relationships found in β sheets using distance matrices, 2) characterizing the orientation of strand–helix and helix–helix interactions using Cartesian coordinate methods, and 3) identifying any specifiable supersecondary structure using distance matrices.

Strand–strand pairs in β sheets

The interstrand hydrogen bonding characteristic of β sheets imposes closely defined geometrical restrictions (whereas the nonbinding interactions characteristic of all other supersecondary relations impose much looser distance restraints). The search for strand–strand pairs is performed using distance matrices. When distance matrices were first introduced, the presentation of short distances displayed the parallel and antiparallel strand pairs as lines parallel or perpendicular to the matrix diagonal. Our mask for such a pair is based on the distance of closest approach of atom $C\alpha_i$ in one strand to $C\alpha_j$ in the

adjacent strand and the distances expected between $C\alpha_i$ and $C\alpha_{j\pm n}$, if the strands were parallel or antiparallel. The extent of the match between the mask distances and the actual distances is then a measure of the fit to a flat sheet. The variable twist that occurs in real sheets will only change these distances to a minor extent unless the strands are very long, an unusual situation in globular proteins. A perfect fit produces an off-diagonal rectangle in the character matrix (see "Secondary Structure Output" above) with the distances of closest approach represented as a line parallel or perpendicular to the principal diagonal. In practice, for a given $C\alpha_i$ there may be 1, 2, or even 3 $C\alpha_j$ atoms within the test distance, and the match to the neighboring strand will vary from one $C\alpha_i$ atom to the next. On the other hand, a β bulge shows up clearly as a gap in the strand pairs.

Helix and strand relations

The methods of characterizing helix and strand relations are based on Cartesian coordinates rather than distance matrices. For the purpose of examining the spatial relations between strands and helices and between helices, these elements are modeled as straight line segments. The setting up of the axes has been discussed above in "Establishing Helix and Strand Axes and a Rough Clock for Moderate Curvature."

The relation between any two pairs of segments can be specified completely by six parameters, Figure 1. The common perpendicular to the two segment axes provides the points P_1 and P_2, the interaxial distance D, and angle θ_A. Points P_1 and P_2 may be on the N-terminal side, the C-terminal side, or within the secondary segment. For easy characterization by a single number, the position of P is expressed as a signed fraction of the segment length. Thus P_k is negative for any position on the N-terminal side of N_k, has a value between O and +1 if between N_k and C_k, and a value greater than +1 if on the C-terminal side of C_k.

The only additional parameters needed for a full description are the rotational position of each segment about its own axis; the angle θ_k is measured between that direction and the vector from N_k to the $C\alpha$ atom for which it is the projection. Depending on the level of detail required, a sequence of 4 or 6 numbers is all that is necessary to describe the relation between two secondary segments: strand–strand in β sandwiches, helix–helix, helix–strand, or strand–helix. While the full set of parameters can easily be listed, only one or two can be usefully represented at any one time in the character matrix for two-dimensional presentation.

Searching for homologous substructures

The same distance matrix-based methods used to identify elements of secondary structure can be used to identify substructures that are composed of two or more noncontiguous segments of the polypeptide chain. The method described in this section is not confined to elements of defined secondary structure. The individual segments may be any part of the structure; part of a helix, part of an extended strand, and their connecting loop could be used as one segment. Since this method conducts a general search, it is implemented separately from the previously described methods in this section and constitutes a separate computer program.

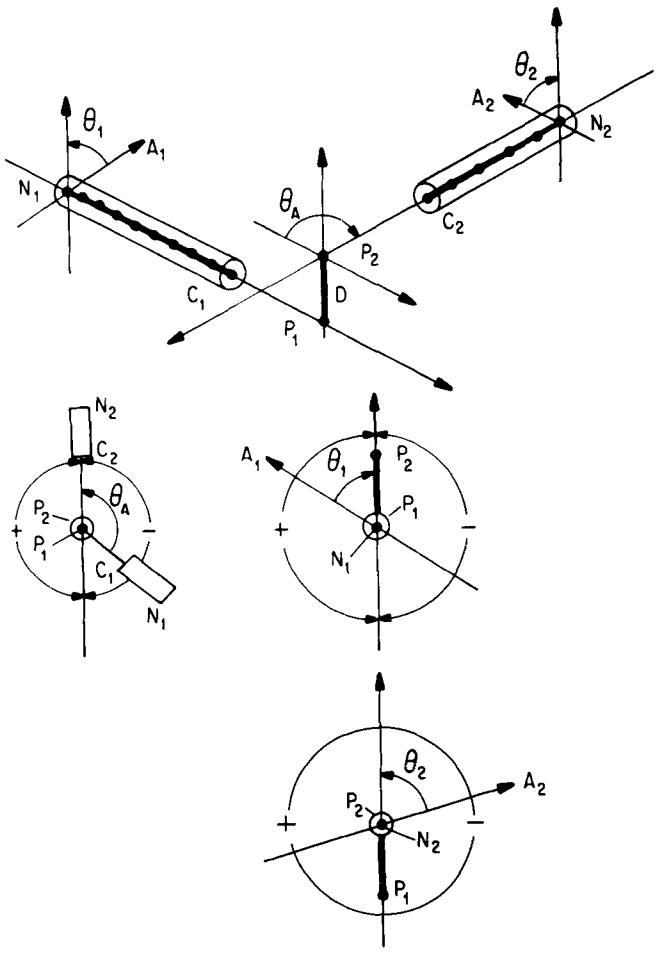

Fig. 1. Diagram of the relationships between any two straight secondary segments. The two thickened lines within the cyclinders represent the length of each segment, and the dots are the axis points, the $C\alpha$ positions projected onto the axes. These segments may be either helices or strands. The direction $N_i \rightarrow C_i$ is taken as positive. The distance of closest approach between the two axes, D, is shown on the common perpendicular between the axial intersection points P_1 and P_2. Either point may be in the region C terminal to the segment, as shown, in the N terminal region, or within the segment. The position of P_i is listed as a signed fraction of the real space segment length (C_i–N_i). The value is negative in the N-terminal region, 0 to +1 within the segment, and >1 in the C-terminal region. θ_A is the angle between the two axes. The rotation of each secondary segment about its own axis is shown by the angles θ_1 and θ_2. All angles are given in the range 0 to 180° with sign conventions shown in the lower Newman diagrams.

The Cα distances between atoms in two separate segments i to i+m and j to j+n (i<j) are contained in rectangular submatrices, or "boxes," of the DM. The upper left corner of one box is above the principal diagonal and has indices (i,j), and the lower right corner has the indices (i+m,j+n). The other submatrix has indices (j,i) and (j+n,i+m) at the upper left and lower right corners, respectively. Insertions or deletions in the polypeptide will change the values of i and j, but the DM will still contain the boxes characteristic of the substructure.

A weighted rms difference is used to compare the mask of a general substructure and the DM of the protein under examination. The weighted rms difference is used because the general substructure may be defined using several examples of the substructure occurring in actual structures rather than an idealized structure (as in the above methods). Each element of the mask, d_{ij}^m, is defined as the average value of d_{ij} in the set of substructures used for the definition. The variance of each element, σ_{ij}^2, is used to weight the difference to obtain the score

$$s = \sum_i \sum_j \frac{1}{\sigma_{ij}^2} (d_{ij}^m - d_{ij}^o)^2 \qquad (1)$$

where d_{ij}^o is the observed distance in a structure under examination, and the sums are over the atoms in two different contiguous segments as outlined in the previous paragraph. Thus, a search for a substructure is conducted by 1) calculating the mask and the variances of each element, 2) calculated the DM of the protein under examination, and 3) scoring the fit between the mask "box" and the protein DM at each possible position in the DM using equation 1. The output is a list of all positions in the protein that score less than a specified amount.

Searches for substructures composed of more than two segments are conducted by executing a series of searches for two-segment substructures. For example, a search for a substructure composed of segments A, B, and C begins with a search for the A B box. If an acceptable A B match is found, the A "row" is searched for the A C box. If an acceptable A C match is found, the B C box, whose position is now defined, is checked as the final step.

RESULTS

Standard Secondary and Supersecondary Structure

Experience to date indicates that roughly 90–95% of all residues in a protein will be assigned to at least one secondary structural segment as defined in the procedures described above. The bulk of the unassigned residues are in 1 or 2 residue connecting links that do not happen to fit the masks. The detailed output, useful for structural comparisons between proteins, is a file of numerical data listing all the parameters of the secondary elements and their pairwise relations. However, for visual examination and comparison, graphical output of the character matrix is useful. No single two-dimensional representation can conveniently present all of the parameters, so only a subset is shown in the following figures as examples.

The four panels of Figure 2 show ribonuclease A. Panel a shows a classic distance matrix where a matrix element contains an "X" if the distance is less than 10 Å. The helix and strand secondary structure is hidden along the diagonal with helices tending to be slightly thicker and strands slightly smaller than the average. The antiparallel β strands are visible in the off-diagonal region, but the starting and stopping points are not entirely clear. There are many "accidental" short distances whose interpretation is quite unclear. Figure 2b shows a graphical line version of the character matrix containing only the secondary structure estimates. The symbol key is given in the legend. The DDM procedure has dramatically sharpened the definition of the secondary elements and their sequence positions. Figure 2c includes the H-bonded β strands. Only those short distances in Figure 2a that truly correspond to β sheets appear, and the ends are clearly defined. In Figure 2d the angular relations between secondary elements other than the hydrogen-bonded strands are shown. For clarity the potential full output has been cut back by showing helix–helix pairs only if the closest Cα–Cα distance is less than 15 Å and by showing strand–helix–strand, or strand–strand pairs only if the minimum distance is less than 10 Å. This type of figure could have been arranged to show any of the other parameters such as interaxial distances, points 1 and 2, and so forth, but they cannot all be usefully shown at once. Therefore, angles are frequently the most obvious parameters in three-dimensional models.

Escherichia coli thioredoxin is shown in Figure 3. There is a single cis peptide whose significance to the folding kinetics of this protein has recently been described by Kelley and Richards.[15] This small protein has a five-stranded β sheet whose topology can be directly "read" from the position of the H bond strand pairs. The topology may be represented by the one-dimensional symbol 1p3p2a4a5. In this alphanumeric symbol the strands are labeled by their sequence position, and the relation between each strand pair is shown by "p" for parallel or "a" for antiparallel.

Two all-helical proteins are shown in Figure 4. Hemerythrin (Fig. 4a) is the paradigm of a four-helix bundle. The alternating parallel and antiparallel arrangement of the helices is clear from the off diagonal boxes. The interhelical angles are remarkably constant within the bundle. Figure 4b,c represent leghemoglobin and the globin fold. The substantial difference between a and b in the interaxial angle pattern is clear. The patterns for all the globins are very similar (not shown), but the angular variations show upon close inspection.

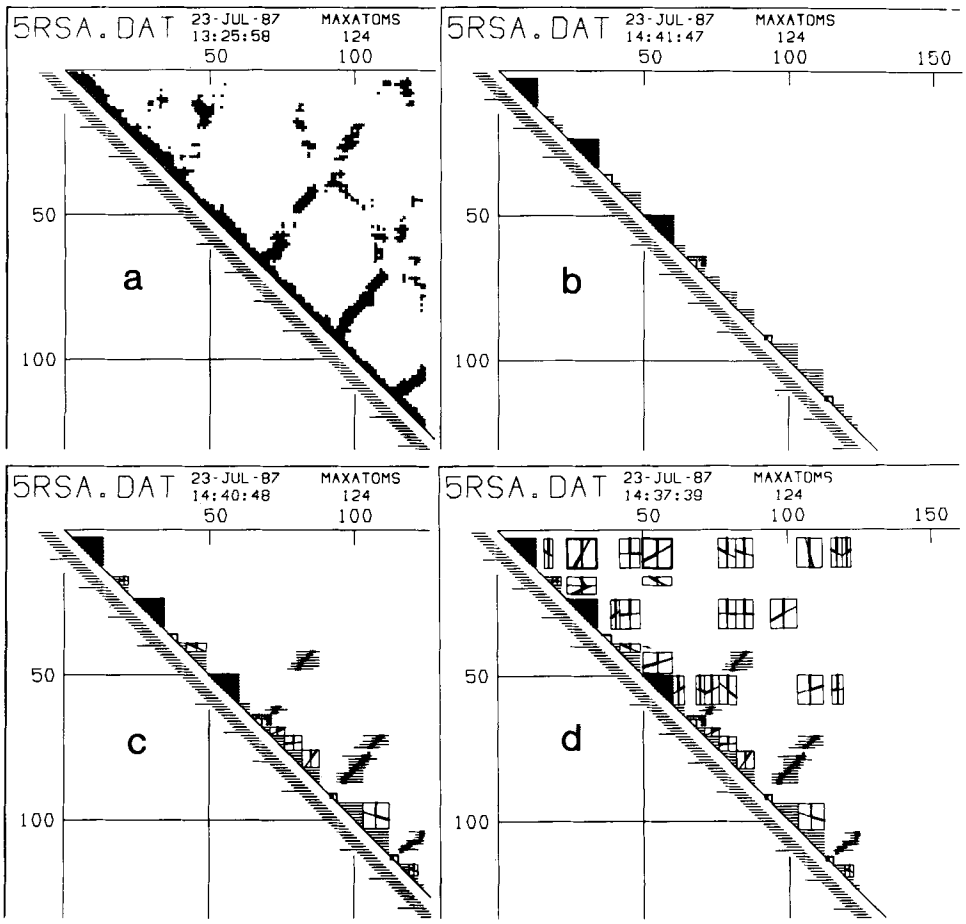

Fig. 2. Schematic distance matrices for the α carbon atoms of ribonuclease A (PDB file 5RSA AUTHOR: A. Wlodawer). In each panel the origin is at the upper left. The α carbon sequence numbers are listed on the upper abscissa and left ordinate. **a:** Shows a "classic" matrix in which the elements representing distances less than 10 Å are marked with an "x." **b:** Shows the graphic output for secondary structure only of the program DEFINE_STRUCTURE. The N-terminal residue of the two cis proline bonds in this molecule are identified by the square boxes on the diagonal at positions 92 and 113. The triangles on the diagonal made up of thick lines are α helices. The slightly thinner lines in the third helix represent a 3_{10} extension. The narrow line triangles are straight strand segments. The small open triangles are sharp β turns. The diamonds outline possible omega loops, and there are alternate possibilities in the regions in which they occur. The program does not overwrite any other symbol with the loop diamonds that may be considered to extend horizontally or vertically to the diagonal wherever they do not do so. **c:** Part of the supersecondary structure is shown. The Xs mark those atom pairs that are closest in a hydrogen-bonded sheet structure, either parallel or antiparallel. The lines extending to either side of the X show the range over which the distances match those expected for the strand adjacent to that atom. The boxes close to the diagonal show the angles between strands that are not H bonded to each other in sheets. The axis of the left strand is shown by the vertical line in the center of the rectangle. The thick N-terminal half is at the top of the thin C-terminal half below. The axis of the lower strand is shown by the other line at the calculated interaxial angle. Again the N-terminal half is thick, the C-terminal half thin. The angle convention for sign is shown in Figure 1. **d:** The full supersecondary relations, which now include the interaxial angles for helix–helix, strand–helix, and helix–strand pairs, are shown. The helix–helix pairs have a box with a thick outline; all others have a thin outline. The axis and angle conventions are as in c. For this particular example only those segment pairs are shown where the distance of closest approach is less than 15 Å for helix–helix pairs and 10 Å for all others.

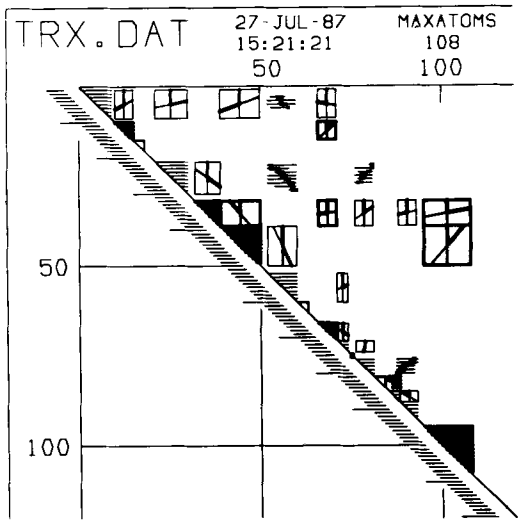

Fig. 3. Character matrix representation of the structure of *E. coli* thioredoxin from coordinates kindly supplied by H. Eklund. The symbol key is given in the legend to Figure 2.

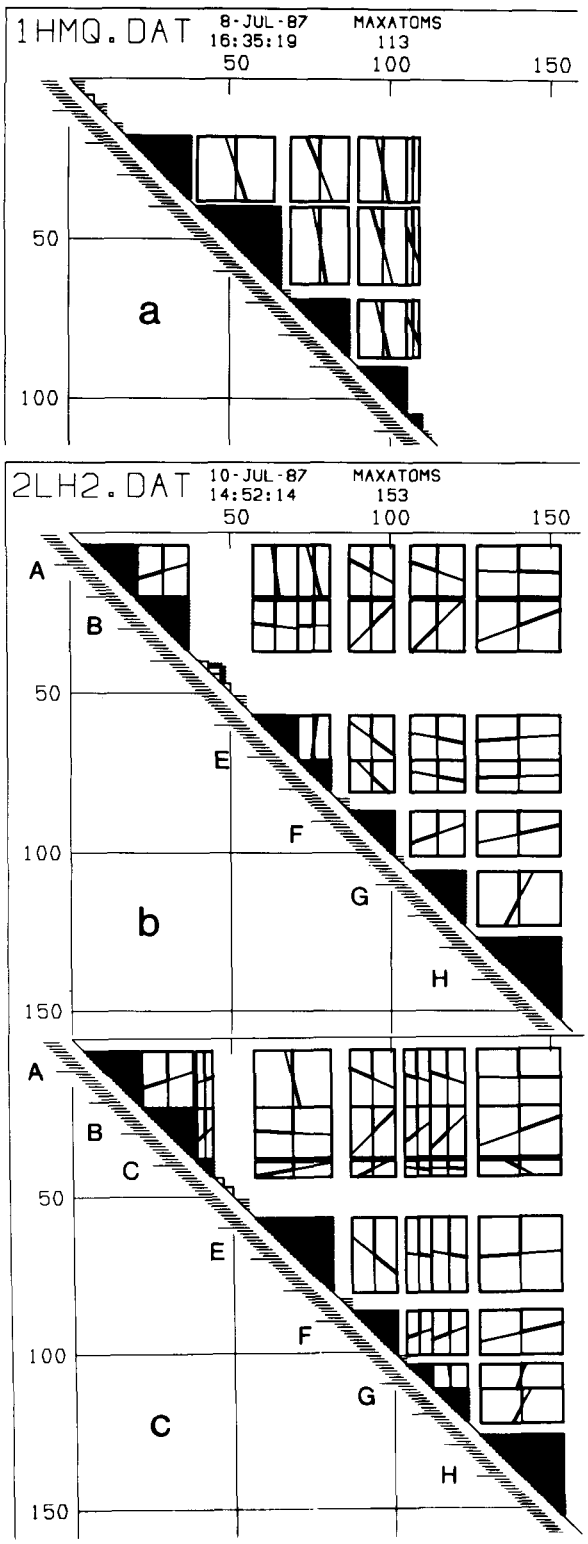

Fig. 4. Character matrix representation of two all helical proteins. **a:** Myohemerythrin, PDB file 1HMQ.[35] **b:** Leghemoglobin, PDB file 2LH2, AUTHORS: B.K. Vainshtein, E.H. Haruyunyan, I.P. Kuranova, V.V. Borisov, N.I. Sosfenov, A.G. Pavlovsky, A.I. Grebenko, N.V. Konareva, with a permitted mismatch to the AHELIX mask of 1.0 Å. **c:** Same as b with a cutoff of 1.5 Å. See discussion in text.

The only difference between b and c in Figure 5 is the mismatch permitted in the helix mask comparisons. In b the value is 1.0 Å, and in c it is 1.5 Å. The A, B, F, and H helices are unaffected by this choice. In b the G helix (106–123) is shown as a single segment, but the E helix is split (57–61, 61–71). A mask fit failure occurs in the middle of the helix, which is then terminated and is the starting point for a new helix. In general, this could reflect a kink in the helix, curvature, or just coordinate error. Since the interaxial angle is very small as seen in the box between the two segments, curvature is the likely explanation. In Figure 5c, where the allowed error limit is larger, the E helix appears as a single unit (57–71). Helix G behaves exactly the opposite, being a single helix (106–128) in Figure 5b and two segments (104–112, 112–123) in Figure 5c. Again the very small interaxial angle in Figure 5c shows that it really is a single helix. In this case the larger allowed error permits an acceptable mask fit to start two residues earlier in the sequence (104 instead of 106). The fit is not really good, so the helix terminates early and the program starts again with a new segment. This dependence on the allowed mismatch appears inevitable. The curve check on a second pass through the secondary structure list shows that in both cases a single helix is the appropriate designation. A true kink in a helix will show a significantly larger rms value from the axial segment angle matrix than is seen for the premature terminations discussed here.

Triose phosphate isomerase is shown in Figure 5 in two separate panels for clarity. Both the regularities and irregularities of this eight-stranded β barrel structure can be seen. There are seven parallel, H-bonded, strand pairs just off the diagonal. The eighth pair (8,1) which closes the barrel, is shown in the upper right corner of the figure. The helices, which represents the return from each strand to the next, show similar interaxial angles between adjacent pairs all the way around the barrel. The four small "extra" helices make very different angles and are part of the irregular aspects of this structure. The strand–helix and helix–strand relations are shown in Figure 5b. The barrel strands can be identified by the H-bonded strand pair symbols, and the angles that they make with the helices that they are connecting can be seen in the appropriate boxes. The strands and helices are all roughly antiparallel but show considerable angular variation.

Specified Substructures

As an example of the homologous substructure search, we examined the Protein Data Bank[16] for occurrences of the helix-turn-helix motif observed in several DNA binding proteins.[17–20] The composite submatrix was composed of both pairs of helices in trp repressor (residues 68–75, 79–86) and CAP (residues 169–176, 180–187). The angle made between the axes of the two helices of the helix-turn-helix motif as determined by the helix and strand relationship

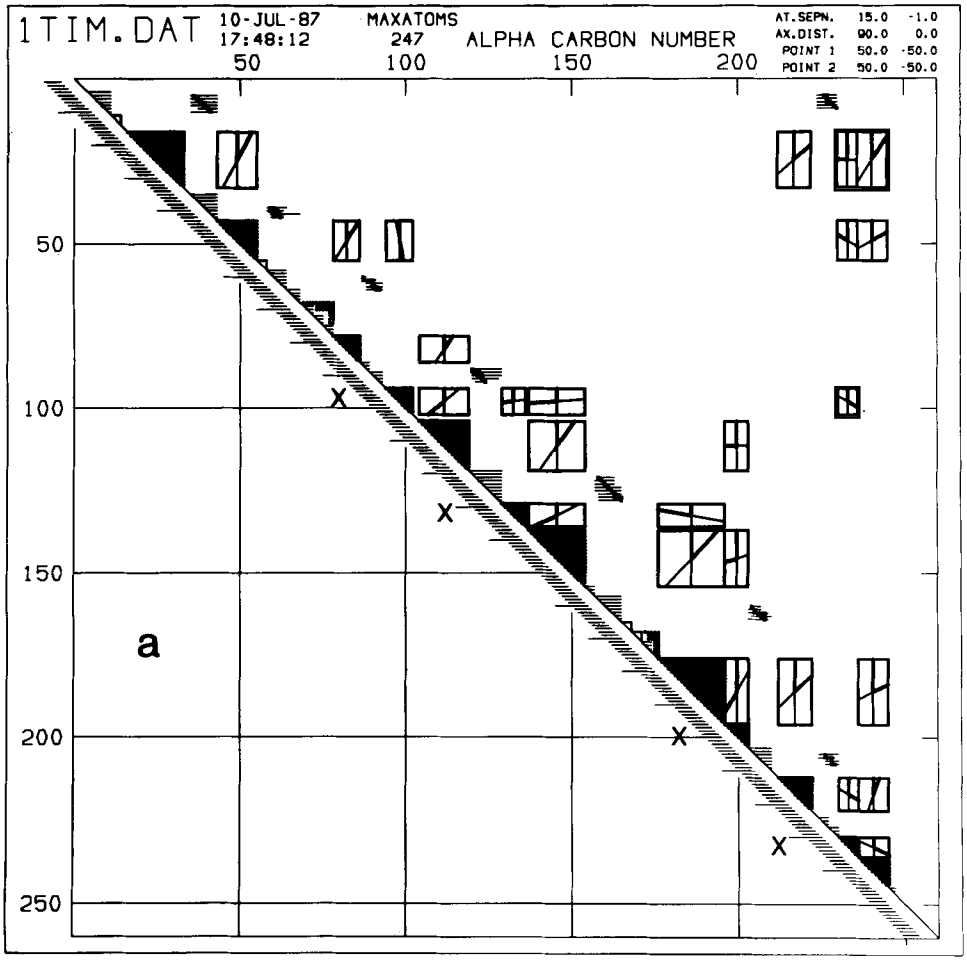

Fig. 5. Character matrix representation of triose phosphate isomerase, PDB file–ITIM.[36] **a:** Secondary structure, H bonded strand pairs and helix–helix interaxial angles. **b:** Same as a, but with helix–strand and strand-helix angles shown.

method described above is 105°. This angle is characteristic of class 1–4 α helix packing,[21] and thus a search based solely on the helix–helix orientation will yield many structures that are not connected by a turn. In order to restrict our search to helices connected by no more than six residues, the distance matrix of candidate proteins was searched near the diagonal only. Seven proteins in the January 15, 1987 version of the Protein Data Bank have at least one substructure with a score (equation 1) less than 2.5. These substructures were then examined by performing a least-squares superposition on the trp repressor helices. Four of these substructures had an rmsd of less than 1.8 Å upon superposition with the trp repressor helices: L7/L12 50S ribosomal protein (69–76, 80–87),[22] cytochrome c peroxidase (92–99, 104–111) and (164–171, 155–172),[23] and citrate synthase (103–100, 91–98).[24] The "turn" in the above examples ranges from one to four residues in length. In the case of citrate synthase and the second occurrence in cytochrome c peroxidase, the helices superimpose very well but occur in the opposite order as in the protein sequence. In other words, the polypeptide link between the helices occurs at the end opposite to the turn in the standard helix-turn-helix motif. This example illustrates the ability of the method to identify homologous substructures even in the presence of transpositions in the protein sequence.

DISCUSSION

The specification of the starting and ending residues in each secondary element will obviously depend on both the quality of the coordinate set and on the mismatch level selected for rejection of the mask fit. Since the secondary segment is required to be continuous, the first match failure terminates the segment. For poorly refined structures at an early stage in a structure determination, the mismatch limits in this program may have to be set quite generously to get a reasonable coverage of the secondary structure; by the same token, inadvertent terminations can be expected. Inspection of both the secondary structure list

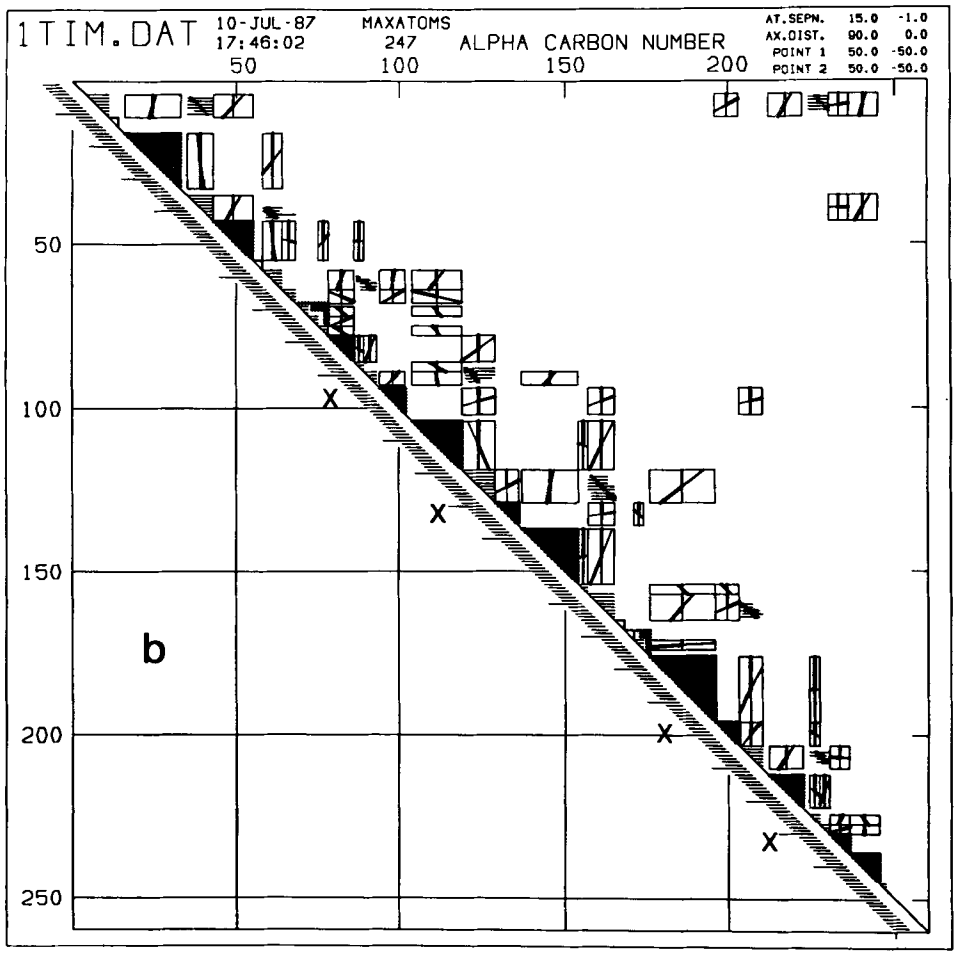

Fig. 5. Continued.

and the supersecondary parameters should usually permit a self-consistent description to be derived. Fortunately in most instances the change in the rms fit is quite abrupt. The segment lengths are thus clearly defined for a reasonable range of cutoff values. The mask-fitting procedure is more objective and appears at least as good as visual inspection. It is, of course, very much more rapid.

The I/O events in the program are by far the most time consuming. On a VAX 8800 the CPU time for determining the secondary structure is about 20 ms/100 Cα atoms, and is roughly linear with chain length since only elements within 50 positions of the diagonal of the distance matrix need to be checked. At this speed it would be practical, for example, to use this program as a subroutine in a molecular dynamics package to monitor the secondary structure at frequent intervals. The CPU time for the supersecondary structure listing is somewhat more variable, but for a full listing about 100 ms would be required for 100 atoms. The time will be roughly proportional to the square of the number of individual secondary elements in the chain.

Analysis and Comparison of Structures

The analysis of a structure begins with the identification of the secondary elements, through use of distance masks of ideal structures. In the program outlined in this paper the decisions are made solely on the basis of Cα interatomic distances; however, other procedures could be used, and a cross-check between several might be best. Methods based on hydrogen-bonds, torsional angles, and Cβ positions are among the possible extensions. These procedures would enhance the delineation of α and 3_{10} helices and allow sharp turns to be sorted into different types. Additionally, more specialized masks could be employed, e.g., separate distances for parallel and antiparallel sheets. However, in the early stages of structure analysis a Cα chain may be the only reliable model available.

With the secondary structure established, the next level of analysis defines the relations between pairs of secondary elements. As shown in Figure 1, a total of six parameters is required to specify such a relation with the elements considered as straight segments. In principal, except for the residues not assigned to a

secondary structure type, the complete tertiary structure of a protein can be specified by the full set of pairwise relations between secondary segments.

A supersecondary structure consists of two or more secondary elements with defined geometrical relations between them. The distance relations between a pair of elements appears in a rectangular submatrix, or "box," in the off-diagonal region of the full distance matrix. A box of distances derived from a model supersecondary element can be tested against the distance matrix by looking for an acceptable rms difference as the mask is translated over the DM. The efficiency of the search can be improved by only searching those regions that contain the distances between the appropriate secondary structures.

If the model supersecondary element contains three secondary segments, then it will be represented by three rectangular boxes. The positions of the boxes are controlled by the sequence position of the secondary elements. With an unknown structure there may be insertions or deletions in the sequence between the secondary segments. Such supersecondary structures may be represented in an abbreviated form as discussed below.

Metamatrix Abstraction of a Structure

The direct matching of the DMs described above requires that the lengths of the secondary segments in the test and target structures be identical or very similar. No match within acceptable mismatch limits will be found if that is not true. Distances within a DM can also be sensitive to small changes in orientation or the average separation of two segments. However, many structural motifs involve general relations between the elements rather than their precise size. To search for such patterns, one can use a more abstract version of the DM, which we call a metamatrix or MM.

Once the secondary and pairwise supersecondary relations are determined, the MM can be constructed. Each element of this matrix now represents not an atom but an entire secondary segment. The diagonal elements contain the segment type identifier (helix, strand, turn, or loop) and may include the actual length of the segment. The off-diagonal elements, which describe the relation between segment pairs, will contain one to six parameters (Fig. 1) depending on the type of metamatrix. An example of the simplest type of MM is the characterization of β-sheet topologies: each diagonal element is a β-strand (of unspecified length) and each off-diagonal element contains one parameter and describes the interaction between two strands as parallel h-bonding, anti-parallel h-bonding, or non-h-bonding. More complex metamatrices are not easy to show in a useful two-dimensional format, but should be easy to use computationally for pattern searches. An example of a MM with two parameters per off-diagonal element would be the characterization of helix–helix interactions by 1) the distance between helices and 2) the angle between the helix axes. The single or multiple pair searches described above for the DM submatrices can now be carried out on the MM. These comparisons are based not on atomic distance or positions but on the derived parameters, and should speed the search for any particular specified pattern.

ACKNOWLEDGMENTS

We thank Dr. J.W. Ponder for the coordinates of idealized secondary structures, Prof. T.A. Steitz and P.B. Sigler for Cα coordinates of CAP and trp repressor, respectively, and J. Mouning and M. Bannon for help in preparing the manuscript. This work was supported by a grant from the Institute of General Medical Sciences to F.M.R. (GM22778).

REFERENCES

1. Ponder, J.W., Richards, F.M. Tertiary template for proteins—use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. 193:775–791, 1987.
2. Blumenthal, L.M. "Theory and Application of Distance Geometry." New York: Chelsea, 1970.
3. Phillips, D.C. Development of crystallographic enzymology. Biochem. Soc. Symp. 31:11–28, 1970.
4. Nishikawa, K., Ooi, T., Ysogai, Y., Saito, N. Tertiary Structure of Proteins. I. Representation and Computation of the Conformations. J. Phys. Soc. Jpn. 32:1331–1337, 1972.
5. Padlan, E.A., Davies, D.R. Variability of three-dimensional structure in immunoglobulins. Proc. Natl. Acad. Sci. U.S.A. 72:819–823, 1975.
6. Kundrot, C.E., Richards, F.M. Crystal structure of hen egg-white lysozyme at a hydrostatic pressure of 1000 atmospheres. J. Mol. Biol. 193:157–170, 1987.
7. Frauenfelder, H., Hartmann, H., Karplus, M., Kuntz, I.D., Kuriyan, J., Parak, F., Petsko, G.A., Ringe, D., Tilton, R.F., Connolly, M.L., Max, N. Thermal expansion of a protein. Biochemistry 26:254–261, 1987.
8. Elber, R., Karplus, M. Multiple conformational states of proteins: A molecular dynamics analysis of myoglobin. Science 235:318–321, 1987.
9. Jones, T.A., Thirup, S. Using known substructures in protein model building and crystallography. EMBO J 4:819–822, 1986.
10. Liebman, M.N. Quantitative analysis of structural domains in proteins. Biophys J. 32:213–215, 1980.
11. Liebman, M.N. Correlation of structure and function in biologically active small molecules and macromolecules by distance matrix partitioning. In: "Molecular Structure and Biological Activity." Griffin, J.F., Duax, W.L., eds. New York: Elsevier. 1982:193–211.
12. Ramachandran, G.N., Sasisakharan, V. Conformation of polypeptides and proteins. Adv. Protein Chem. 23:283–437, 1968.
13. Sibanda, B.L., Thrornton, J.M. β-hairpin families in globular proteins. Nature 316:170–174, 1985.
14. Leszczynski, J.F., Rose, G.D. Loops in globular proteins: A novel category of secondary structure. Science 23:849–855, 1986.
15. Kelley, R.F., Richards, F.M. Replacement of proline 76 with alanine eliminates the slow kinetic phase in thioredoxin folding. Biochemistry 26:6765–6774, 1985.
16. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, 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. Shevitz, R.W., Otwinoski, Z., Joachmaik, A., Lawson, C.L., Sigler, P.B. The three dimensional structure of trp repressor. Nature 317:782–786, 1985.
18. Anderson, W.F., Ohlendorf, D.H., Takeda, Y., Matthews, B.W. Structure of the cro repressor from bacteriophage and its interaction with DNA. Nature 290:754–758, 1981.

19. McKay, D.B., Weber, I.T., Steitz, T.A. Structure of catabolite gene activator protein at 2.9 Å resolution incorporation of amino acid sequence and interactions with cyclic amp. Nature290:744–749, 1981.
20. Pabo, C.O., Lewis, M. The operator-binding domain of repressor: Structure and DNA recognition. Nature 298:443–447, 1982.
21. Chothia, C. Principles that Determine the Structure of Proteins. Annu. Rev. Biochem. 53:537–572, 1984.
22. Leijonmarck, M., Liljas, A. Structure of the C-terminal domain of the ribosomal protein L7/L12 from *Escherichia coli* at 1.7 Å. J. Mol. Biol. (in press), 1987.
23. Finzel, B.C., Poulos, T.L., Kraut, J. Crystal structure of yeast cytochrome c peroxidase refined at 1.8 Å resolution. J. Biol. Chem. 259:13027–13036, 1984.
24. Remington, S., Wiegand, G., Huber, R. Crystallographic refinement and atomic models of two different forms of citrate synthase at 2.7 and 1.7 Å resolution. J. Mol. Biol. 158:111–152, 1982.
25. Lewis, P.N., Momany, F.A., Scheraga, H.A. Folding of polypeptide chains in proteins: A proposed mechanism for folding. Proc. Natl. Acad. Sci. U.S.A. 68:2293–2297, 1971.
26. Kuntz, I.D. Protein folding. J. Am. Chem. Soc. 94:4009–4012, 1972.
27. Crawford, J.L., Lipscomb, W.N., Schellman, C.G. The reverse turn as a polypeptide conformation in globular proteins. Proc. Natl. Acad. Sci. U.S.A. 70:538–542, 1973.
28. Levitt, M., Greer, J. Automatic identification of secondary structure in globular proteins. J. Mol. Biol. 114:181–239, 1977.
29. Rose, G.D., Seltzer, J.P. A new algorithm for finding the peptide chain turns in a globular protein. J. Mol. Biol. 113:153–164, 1977.
30. Chou, P.Y., Fasman, G.D. β-turns in proteins. J. Mol. Biol. 115:135–175, 1977.
31. Kolaska, A.S., Ramabraham, V., Soman, K.W., Reversals of polypeptide chain in globular proteins. Int. J. Pept. Protein Res. 16:1–11, 1980.
32. Ramakrishman, C., Soman, K.V. Identification of secondary structures in globular proteins—a new algorithm. Int. J. Pep. Protein Res. 20:218–237, 1982.
33. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
34. Hohne, E., Kretschmer, R.G. Description of secondary structures in proteins. Stud. Biophys 108:165–186, 1985.
35. Stenkamp, R.E., Siekes, L.C., Jensen, L.H. Adjustment of restraints in the refinement of methemerythrin and azidomethemerythrin at 2.0 Ångstroms resolution. Acta Cryst., Sect. B 39:697–703, 1983.
36. Banner, D.W., Bloomer, A.C., Petsko, G.A., Phillips, D.C., Wilson, I.A. Atomic coordinates for triose phosphate isomerase from chicken muscle. Biochem. Biophys. Res. Commun. 72:146–155, 1976.