On the Problem of Comparing Protein Structures
Development and Applications of a New Method for the Assessment of Structural Similarities of Polypeptide Conformations

MANFRED J. SIPPL

Universität Salzburg, Institut für Allgemeine Biologie
Biochemie und Biophysik Abteilung Biochemie
Erzabt-Klotz Strasse 11, A-5020 Salzburg, Austria

(Received 9 June 1981, and in revised form 13 November 1981)

The development of prediction schemes and the search for evolutionary relationships amongst proteins require reliable methods for the comparison of polypeptide structures. It is shown that methods which attempt to describe structural similarities by a single value generally do not yield reliable estimates of the relatedness of two conformations. A new method is reported, called the $D_k$ procedure, which yields a spectrum of deviations between two structures. Each particular $D_k$ value is a measure of the similarity of the diagonals of the distance matrices of the compared conformations, $k$ being the distance of the diagonals relative to the main diagonal.

The method has the following features. (1) $D_k$ is independent of chain length; (2) the method yields the relatedness of two conformations in terms of different structural levels; (3) $D_k$ is a high-speed algorithm; (4) the $D_k$ deviations of random structures from any particular conformation are predictable.

The following applications are reported. (1) The success of both secondary and tertiary structure predictions are measured in terms of a reliable quality index. (2) The route of a conformation during simulation studies is followed on different structural levels, which exhibit the characteristics of the simulation. (3) The significance of hypotheses on protein folding subject to prediction schemes can be established. (4) A priori information (fixing pieces of secondary structure derived from X-ray investigations during prediction) is extractable from the predicted structures. (5) The evolutionary relatedness of two nucleotide binding proteins is established.

The simplicity and speed of the $D_k$ procedure allow its implementation even on minicomputers.

1. Introduction

The development of algorithms for simulating the folding process of proteins and the prediction of protein tertiary structure as a function of amino acid sequence require methods for the comparison of computed structures with the crystallographic co-ordinates. The same problem arises for the question of structural evolution and structural similarity of different proteins with known X-
ray structure. The methods available to tackle these problems may be arranged in two groups.

(1) Methods based on interatomic distances. (1.1) The method reported by Levitt (1976) is a measure for the similarity of the interatomic distances of two structures, and is usually applied to prediction studies.

(2) Methods based on the rotation of a reference frame. (2.1) The conventional root-mean-square deviation of two optimally oriented sets of co-ordinates (Levitt, 1976) and (2.2) the method reported by Rossmann & Argos (1976), which requires as an essential component the search through the space of the three Eulerian angles. Methods (2.1) (in modified versions) and (2.2) have been applied to a number of proteins with known X-ray structure and a number of relationships have been established (Rossmann & Argos. 1976,1977: Remington & Matthews. 1978,1980: Matthews et al., 1981: McLachlan. 1979; McLachlan et al., 1980).

Cohen & Sternberg (1980a) have investigated methods (1.1) and (2.1). They concluded that the rotation method is a more efficient procedure for the evaluation of structural similarity than the interatomic distance method. A main point in their argument is that a method based on interatomic distances does not distinguish between a structure and its mirror image, in contrast to the rotation method, which generally indicates a certain deviation between image and mirror image. They suggested applying the rotation method to evaluate the success of prediction studies.

Nevertheless, there remains uncertainty in interpreting prediction success, as reviewed by Banaszak et al. (1981). It seems that most prediction schemes so far reported produce structures whose deviation from the native state does not improve beyond a certain limit. This limit is of the order of 5 Å for the interatomic distance method in the case of the pancreatic trypsin inhibitor (Levitt, 1976; Kuntz et al., 1976; Cohen & Sternberg, 1980a). The appearance of a limit is remarkable because there are very different approaches to the construction of prediction algorithms. However, the actual extent of the agreement of the predicted structures with the native state is not evaluated satisfactorily by the interatomic distance method or the rotation method. As will be shown in this paper, the lack of the detectable differences in the qualities of the various prediction schemes is due to the nature of the applied distance measures. Both the interatomic distance method and the rotation method are oversensitive to the overall agreement of two structures, which causes secondary structure agreement to be undetectable when the overall structures do not match. These findings are in agreement with the striking observation that no improvement of prediction success is detectable when one incorporates information like pieces of secondary structure, taken from X-ray analysis, into the algorithm (Havel et al., 1979; Cohen & Sternberg, 1980a).

Methods used to compare structures representing protein chains should possess the following features.

(1) Compatibility: it should be possible to compare simulations and predictions applied to proteins of different chain length.

(2) Separation of structural levels: it should be possible to compare structures on different levels of organization.

(3) The ability to test hypotheses: the method should provide a clear answer to
the question of whether a particular hypothesis on chain folding improves prediction success when it is translated to a prediction scheme. This is an essential requirement for the development of powerful algorithms.

(4) The possibility of assessing the influence of a priori information on prediction success. When one puts information like correct pieces of secondary structure, derived from X-ray investigations, into the procedure it should be possible to extract this information from the final prediction success, because this information makes no contribution to the quality of the algorithm.

(5) Speed: the method has to save computing time.

A new method is reported. the diagonal deviation of kth order (called here the $D_k$ procedure), which possesses the features stated above. The $D_k$ procedure uses subsets of the distance matrices of the compared structures. The distance matrices are formed by the intramolecular distances $d_{i,j}$ between the Cα co-ordinates. The subsets used by the $D_k$ procedure are the diagonals of the distance matrices formed by the elements $d_{i,i+k}$, where $k$ represents the order of the diagonals relative to the main diagonal. This decomposition yields a spectrum of $D_k$ values, which give an accurate description of the structural similarities of two conformations. The $D_k$ method is tested on an example where the native PTI molecule is distorted by a rotation around an internal bond, which clearly demonstrates the capability of the $D_k$ procedure. A method is reported. and a derivation given in the Appendix, which allows computation of the expected values of the $D_k$ deviations of any conformation from a set of random conformations. The values obtained are in excellent agreement with the $D_k$ values obtained by direct calculation from a random population. Thus, the random values of the $D_k$ deviations of any conformation are predictable. The same procedure shows why α-helix and β-sheet-rich proteins differ considerably from random analogues in their $D_k$ deviations. The usefulness of the $D_k$ procedure as a tool for the development of prediction schemes is illustrated with a "primitive" folding algorithm. A random conformation is forced to fold to a compact globular shape. and the route of the chain during the simulation is traced by means of the $D_k$ procedure. The $D_k$ trajectories thus obtained clearly exhibit the construction and features of the algorithm.

The fact that methods based on interatomic distances do not distinguish between a conformation and its mirror image is not a problem of practical interest to protein structure comparisons. This is shown by energetic considerations based on the observation that the φ, ψ angles defining a conformation are transformed by a reflection to locations in the φ–ψ map which, in almost all cases, correspond to extremely high backbone energies.

Although this paper is concerned mainly with applications of the $D_k$ procedure to prediction studies, an application to the search of evolutionary relationships among proteins is presented for completeness. In a preliminary application to the GPDHase and LDHase molecules, the $D_k$ method, in conjunction with the segment

† Abbreviations used: PTI, bovine pancreatic trypsin inhibitor; MBN, myoglobin (sperm whale. met); FDX, ferredoxin (Peptococcus aerogenes); REI, immunoglobulin (B-J fragment REI); LYZ, lysozyme (hen egg-white); HMR, hemerythrin; GLC, glucagon; GPDHase, glyceraldehyde-3-phosphate dehydrogenase; LDHase, lactate dehydrogenase. The crystallographic co-ordinates were supplied by the Protein Data Bank (Bernstein et al., 1977).
approach reported by Remington & Matthews (1978). is shown to be capable of establishing evolutionary relationships. In contrast to the time-consuming procedures available in this field (Rossmann & Argos, 1976; Remington & Matthews, 1978), the \( D_k \) procedure is a high-speed algorithm and, due to its simplicity, may be implemented even on minicomputers.

2. Definitions

The definitions of the two procedures commonly applied to protein structure comparisons, the root-mean-square deviation based on interatomic distances and the root-mean-square deviation based on the rotation method, are recalled. Subsequently, the diagonal deviation of \( k \)th order is defined: a new procedure, whose properties are discussed extensively in this paper.

Let the structure \( S \) be the set of \( N \) points \( p_i \) given in Euclidian space \( R^3 \) with co-ordinates:

\[
p_i = (x_i, y_i, z_i).
\]

where \( p_i \) represents the co-ordinates of the C\(^\alpha\) carbon of residue \( i \) from the amino acid sequence of a particular protein.

The distance \( d_{i,j} \) between two points \( p_i \) and \( p_j \) of \( S \) is given by:

\[
d_{i,j} = [(x_i-x_j)^2 + (y_i-y_j)^2 + (z_i-z_j)^2]^{1/2}. \tag{1}
\]

The elements \( d_{i,j} \) form the distance matrix of \( S \), which is symmetric, of size \( N \times N \) and has zeros in the main diagonal.

The side diagonal \( B_k \) of the \( k \)th order is defined as the set of elements \( d_{i,j} \) with the property \( j - i = k \), which are arranged parallel to the main diagonal. There are \( N-k \) elements \( d_{i,i+k} \) in \( B_k \). The root-mean-square deviation based on the optimal superposition of two structures is defined as (McLachlan, 1979; Remington & Matthews, 1978; Cohen & Sternberg, 1980a):

\[
R(S, S) = \min_M \left\{ \frac{1}{N} \sum_{i=1}^{N} (\tilde{p}_i - p_i)^2 \right\}^{1/2}, \tag{2}
\]

where \( S \) and \( S \) have the same size \( N \). In equation (2), the co-ordinates of both structures are defined in reference systems with centres of gravity \( \bar{G} \) and \( G \):

\[
G = \frac{1}{N} \sum_{i=1}^{N} p_i, \tag{3}
\]

shifted to the origin. The problem is to find a rotation \( M \) in terms of an orthonormalized \( 3 \times 3 \) matrix, so that the points \( p_i \) of \( S \) multiplied by \( M \) minimize the expression in equation (2). This yields the best superposition of \( S \) on \( \bar{S} \).

The root-mean-square deviation based on the interatomic distances is defined by (Levitt, 1976):

\[
D(S, S) = (1/N) \left[ \sum_{i,j} (\bar{d}_{i,j} - d_{i,j})^2 \right]^{1/2}. \tag{4}
\]
The diagonal deviation of order \( k \) is defined by:

\[
D_k(\mathcal{S}, S) = \left[ \frac{1}{N-k} \sum_{i=1}^{N-k} (d_{i,i+k} - d_{i,i+k})^2 \right]^{1/2}.
\] (5)

The elements \( d_{i,i+k} \) of diagonal \( B_k \) are restricted to a well-defined domain in the case of protein structures. The lower bound \( b_l(k) \) is equivalent to the smallest possible contact distance of two amino acid residues, measured from the \( C^\alpha \) centres and independent of \( k \): i.e. \( b_l(k) = \text{constant} \). The upper bound \( b_u(k) \) represents the distance of two terminal \( C^\alpha \) carbons in a fully extended chain of length \( k+1 \), and is approximately a linear function of \( k \):

\[
b_u(k) = ck.
\] (6)

In equation (6), the constant \( c \) is set to 3.8 Å, the distance of two adjacent \( C^\alpha \) carbons in the amino acid sequence. For an exact determination of \( b_u(k) \), one simply has to construct an extended chain and to compute the corresponding distances. The elements \( d_{i,i+k} \) assume values in the range \([b_l(k), b_u(k)]\) and the size of this domain increases with increasing values of \( k \). The same applies to the differences \( d_{i,j} - d_{i,j} \) in equations (4) and (5). The contributions of differences belonging to diagonals of high order \( k \) are generally overestimated with respect to those of low order when inserted in equation (4). The elements \( d_{i,i+k} \) in the neighbourhood of the main diagonal carry information on secondary structure, whereas the outer diagonals carry mainly information about the overall shape of the molecule. The conclusion must be drawn that \( D(\mathcal{S}, S) \) reflects the deviation of the overall shapes of two molecules. The properties of \( R(\mathcal{S}, S) \) cannot be discussed in the same way. However, Cohen & Sternberg (1980a) found a linear relationship:

\[
D(\mathcal{S}, S) = 0.75 \cdot R(\mathcal{S}, S) + 0.19,
\]

which is a strong indication that the qualitative behaviour of both methods is similar. The characteristic feature of the rotation method is the superposition of the two compared structures. This is best carried out by matching the overall shapes of two molecules.

The \( D(\mathcal{S}, S) \) measure may be corrected in two ways. One way is to normalize the elements \( d_{i,i+k} \) of different diagonals \( B_k \) by the size of their domain of variation:

\[
d'_{i,i+k} = d_{i,i+k} / (b_u(k) - b_l(k)).
\] (7)

in order to assign equal weights to all elements in equation (4). Another method, realized by equation (5), is to restrict the discussion of structural similarity to the diagonals \( B_k \). The set of \( D_k \) values for \( k = 1, \ldots, N-1 \) thus obtained is defined as the \( D_k \) spectrum of the deviations of two structures.

Note that the value of \( D_{k+1} \) is bounded by \( D_k \), due to the triangle inequalities:

\[
|d_{i,i+k} - d_{i,i+k+1}| \leq d_{i+k,i+k+1} = d_{i,i+1} \sim 3.8 \text{ Å}
\] (8)

(for \text{trans} configurations of the peptide bond), so that the \( D_k \) spectrum is a function of bounded variation, which means that neighbouring diagonals have comparable \( D_k \) values, and that the \( D_k \) spectrum changes slowly as a function of \( k \).
3. Generation of Random Analogues

Various attempts have been made to develop computer algorithms for the prediction of secondary and tertiary structures of proteins as a function of the amino acid sequence (Levitt, 1976; Kuntz et al., 1976; Cohen & Sternberg, 1980a; Tanaka & Scheraga, 1975; Burgess & Scheraga, 1975; Chou & Fasman, 1974; Robson & Suzuki, 1976; for reviews, see Chou & Fasman, 1978; Nemethy & Scheraga, 1977; Banaszak et al., 1981).

Investigation of the range of random predictions is a prelude to any structure prediction of a particular protein. This is usually achieved (for tertiary structure predictions) by the generation of a set of random analogues and subsequent calculation of the deviations of the random set from the native target conformation in terms of the D or R procedure (Cohen & Sternberg, 1980a).

A Monte Carlo procedure is used here, related to the procedure reported by Jordan et al. (1978) for the simulation of amylose chains, with the following properties.

(1) The geometry of the polypeptide backbone is fully conserved. The variables are the dihedral angles $\varphi$ and $\Psi$ ($\omega$ was held constant at 180°). The conformations are therefore suitable representations of real protein chains, and there exists a unique and invertible relation between the simplified model (no side-chains considered) and the backbone structure of polypeptides.

(2) The distributions of the dihedral angles must reflect the distributions found in protein structures determined by X-ray crystallography. The PCIL0-approximations reported by Pullman & Pullman (1974) are in good agreement with these distributions, and may be used instead.

(3) Collisions in the simulated structures are prohibited. The smallest approach of two $C^\alpha$ carbons is set to 3.5 Å.

Property (2) is achieved in the following way. The energy of a particular conformation $C$ is represented by:

$$E(C) = \sum_{i=1}^{N} E^a(\varphi_i, \Psi_i)$$

(N is chain length, and $a = 1, \ldots, 20$, is the type of amino acid residue). The energy of the backbone conformation is separated and the residues behave in this approximation as independent systems, i.e. no non-bonded (besides steric collisions) or solvent interactions are included.

Chain conformations represented in the Monte Carlo sample must be distributed in a manner consistent with the Boltzmann factors of their respective conformational energies. The density function $F^a$ of the dihedral angles of a particular residue $a$ is defined by:

$$F^a(\varphi, \Psi) = 1/Z^a \exp \left( -E^a(\varphi, \Psi)/kT \right).$$

where $Z^a$:

$$Z^a = \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \exp \left( -E^a(\varphi, \Psi)/kT \right) d\varphi d\Psi$$

is a constant; i.e. the state integral.
Equations (10) and (11) may be approximated by:

\[ F^a(\varphi_m, \Psi_n) = \frac{1}{Z^a} \exp \left( -E^a(\varphi_m, \Psi_n)/kT \right) \]  

and

\[ Z^a = \sum_{m=1}^g \sum_{n=1}^g \exp \left( -E^a(\varphi_m, \Psi_n)/kT \right). \]  

In this way, a discrete state space is produced with \( G^2 \) conformations \( C_{m,n}^a \) allowed for each amino acid. This is equivalent to a square lattice with grid interval \( g \). Particular values for \( G \) and \( g \) used in this paper are \( G = 12 \) and \( g = 30^\circ \). The grid points of this lattice are represented by:

\[ \varphi_m = g(m - 1), \quad \Psi_n = g(n - 1), \]

corresponding to the conformations \( C_{m,n}^a \).

The frequency \( R_{m,n}^a \) of the conformation \( C_{m,n}^a \) is determined by the number of residues with \( \varphi \), \( \Psi \) values in the range:

\[ \varphi_m - g/2 \leq \varphi < \varphi_m + g/2, \quad \Psi_n - g/2 \leq \Psi < \Psi_n + g/2, \]

and is normalized by \( H_{m,n}^a = R_{m,n}^a / H^a \), with \( H^a \) being the total frequency of amino acid \( a \) in the analysed proteins. Setting \( H^a \sim Z^a \), and \( H_{m,n}^a \sim \exp \left( -E(\varphi_m, \Psi_n)/kT \right) \), the normalized frequencies \( H_{m,n}^a \) are equivalent to the Boltzmann factors \( F^a(\varphi_m, \Psi_n) \) of the conformations \( C_{m,n}^a \). The numerical values of the distributions \( H_{m,n}^a \) may be obtained from the Protein Data Bank (Bernstein et al., 1977).

Particular values of \( \varphi_i \) and \( \Psi_i \) (the index \( i \) refers to eqn (9)) are assigned by a random number generator in the following way. The conformations \( C_{m,n}^a \) are renumbered:

\[ C_1 = C_{1,1}, \quad C_{G+1} = C_{2,1}, \]
\[ C_2 = C_{1,2}, \quad C_{G+2} = C_{2,2}, \]
\[ \ldots, \quad \ldots \]
\[ C_g = C_{1,g}, \quad C_{G^2} = C_{G,g}, \]

and the associated frequencies are mapped onto the interval \([0, 1]\) by

\[ I_0 = 0, \]
\[ I_1 = I_0 + H_{1,1}, \]
\[ I_2 = I_1 + H_{1,2}, \]
\[ \ldots \]
\[ I_{G^2} = I_{G^2-1} + H_{G,g}. \]

The probability of finding an equally distributed random number (on the interval \([0, 1]\)) in the range \( I_{k-1}, I_k \) is equal to \( I_k - I_{k-1} \), and therefore:

\[ I_k - I_{k-1} = I_{k-1} + H_{m,n} - I_{k-1} = H_{m,n}, \]

where \( m \) and \( n \) are determined by:

\[ m = \text{int} \left( \frac{(k-1)}{G} \right) + 1 \quad \text{and} \quad n = k - (m-1)G. \]

The polypeptide chain is generated by inserting the \( \varphi_i \) and \( \Psi_i \) values obtained by the above process into the structure-building algorithm (e.g. see Hopfinger, 1973).
Amino acids are tested for collisions according to the distance \( d_{ij} \) of the C° carbons. If a contact less than 3.5 Å occurs between residues \( i \) and \( j \), some residue between \( i \) and \( j \) in the amino acid sequence is picked by a random process, and new values for \( \varphi \) and \( \psi \) of this residue are chosen by the algorithm.

Through a simple modification, it is possible to extend the discrete state space to the whole conformation space. When \( \varphi, \psi \) pair is determined by the process described above, the final choice is located by:

\[
\varphi = (\varphi - y/2) + yg \\
\psi = (\psi - y/2) + yg,
\]

where \( x \) and \( y \) are again random variables uniformly distributed on \([0, 1]\). All random conformations discussed in this paper are generated by this relaxed procedure.

4. Behaviour of \( R \) and \( D_k \) under Reflection

By the definition of \( D_k \) given in equation (5), it is clear that the native structure \( B \) and the associated mirror image \( I \) have \( D_k \) deviations \( D_k(B, I) = 0 \), whereas \( R(B, I) \) will generally indicate a certain deviation between both conformations.

Cohen & Sternberg (1980a) produced two sets of random conformations: one set by a (limited) perturbation of the native backbone conformation of the PTI molecule, and a second set by perturbing the mirror image of the native conformation. In the first case, they obtained a good linear correlation between the \( R \) and \( D \) deviations of the conformations of the perturbed set from the native conformation (in this case the \( D \) and \( R \) values go to zero simultaneously with decreasing perturbation). In the second case, this relation was lost for \( D \) values less than 6 Å (here \( D \) again goes to zero, whereas \( R \) goes to 7.76 Å, due to \( R(B, I) = 7.76 \), with decreasing perturbation). The authors concluded that the \( R \) method, which is sensitive to the handedness of two conformations, is a better method of assessing structural relatedness than methods based on interatomic distances.

A more detailed picture of the image–mirror image relations of protein conformations can be obtained when the effect of the reflection on the C° backbone is discussed in terms of the dihedral angles \( \varphi, \psi \) and \( \omega \). The specification of valence bonds, valence angles and torsion angles is equivalent to the co-ordinate representation of protein conformations. By energetic considerations, valence bonds and valence angles are fixed so that the only variables defining a particular conformation are the torsion angles. Let the vector \( T \) be the set of all backbone angles of a particular polypeptide or protein with components:

\[
T = (\psi_1, \omega_2, \varphi_2, \psi_2, \omega_3, \ldots, \psi_{N-1}, \omega_N, \varphi_N),
\]

where the index specifies the position of the aligned residue in the amino acid sequence. Then every vector \( Z \) with components:

\[
Z = (n_1 \pi, n_2 \pi, n_3 \pi, \ldots),
\]

where

\[
n_i = \ldots, -2, -1, 0, 1, 2, \ldots
\]
with the same dimension as $T$ represents a centre of reflection in the space of torsion angles. Note that valence bonds and valence angles are not affected by a reflection. Thus, the torsion angle representation of the mirror image of a particular conformation is found by adding twice the vector $(Z - T)$ to $T$:

$$T' = T + 2(Z - T) = 2Z - T.$$ 

which is equivalent mod $2\pi$ to the reflection through the special centre $Z = (0, 0, \ldots)$ and $T' = -T$ (i.e. only the paths of reflections are different, not the end points).

It is noted that $L$-amino acids are changed by a reflection to $D$-amino acids. The $C^\alpha$ backbone of the mirror image in the sense of Cohen & Sternberg (1980a), however, is actually the $L$-analogon of the $D$-protein produced by the reflection.

All residues other than glycine have a highly asymmetric energy surface with respect to the centre $(0, 0)$ of the $\varphi$-$\Psi$ map. which is reflected by the asymmetric $\varphi$, $\Psi$ distributions of the amino acids of all proteins so far investigated by X-ray analysis. Thus, in general, a reflection will result in an extremely high conformational energy of the $L$-analogon. because most $\varphi$, $\Psi$ pairs of the mirror image will assume values in an energetically unfavourable or forbidden region. Only in special cases will the original conformation and the $L$-analogon of the mirror image have comparable energies. (1) A glycine homopolyamide, where the original and the mirror image are identical. (2) Conformations where almost all $\varphi$, $\Psi$ pairs lie close to or at any centre of reflection $Z$. In such cases, the $\varphi$, $\Psi$ values of the image and mirror image lie close together in the $\varphi$-$\Psi$ map and, by the smoothness of the energy surfaces, they may have comparable energy (consider the extreme case of a fully extended chain with $T' = T$). (3) A mixture of (1) and (2).

Now consider a variation $T + \varepsilon$ of the vector $T$ defining the $L$-analogon of the mirror image of the PTI molecule (which has unfavourable backbone energy). Again, by the smoothness of the energy function, there is a certain region around $T'$ corresponding to high conformational energies. Thus, the members of the set derived from the mirror image of the PTI conformation by Cohen & Sternberg (1980a) are likely to be high energy species (especially those with $D$ values beyond $6$ Å), and are far from any structures that will be produced by serious prediction schemes or Monte Carlo studies. Thus, situations where ambiguities may arise when a procedure based on interatomic distances is applied to structure comparisons of proteins will, in practice not occur, because the $L$-analogon of the mirror image is in most cases not contained in the set of energetically admissible conformations. A lot of information concerning the characteristics of the $R$ and $D_k$ methods can be obtained when the reflection is considered as a continuous process. and the $R$ and $D_k$ deviations (from the native conformation) corresponding to structures on the trajectory:

$$T(t) = T + t(2Z - 2T)$$

$$(T(0) = T, \quad T(1) = T' = -T)$$

are plotted as a function of the parameter $t$. as shown in Figure 1 for the PTI molecule. The centre of reflection $Z$ is chosen to be $\varphi_i = \psi_i = \pi$ and $\omega_i = \pi(\omega_i > 0)$ or $\omega_i = -\pi(\omega_i < 0)$, which represents an extended structure. The assignment of the $\omega$
values ensures that the trajectory runs through the trans conformation of the peptide bond. The three vertical lines in Figure 1 correspond (from left to right) to the native, extended and L-analagon, respectively. In Figure 1(b) to (d), the R curve is superimposed on the D_k curves, which is achieved by normalizing the absolute maxima of all curves equal to one. The mean values of random analogues (see section 6, below) are included as horizontal lines. Surprisingly, not only are the D_k curves symmetric to the centre of reflection, but the R curve also has a distinct symmetry. Only at the point corresponding to T a narrow region exists where the symmetry is broken. This region is not accessible to L-conformations, due to high backbone energies. Figure 1(d) shows that the D_{30} and R curves are very similar, indicating a similar information content (i.e. comparison of the overall agreement) of both measures. At the points t = a and t = b, D_{30} and R indicate good agreement between the native structure and the conformations T(a) and T(b) but D_5 and D_{10} clearly demonstrate bad agreement in the short (secondary structure) and medium range, respectively. Studies using other trajectories through different centres of reflection yield the same results.

5. Behaviour of D_k and R under Internal Rotations

Figure 2 demonstrates the behaviour of R and D_k as functions of an internal rotation θ in the PTI molecule. The rotation axis was chosen as the virtual bond between the adjacent Cα carbons of Ala25 and Lys26. The value of R as a function
Fig. 2. Behaviour of $R$ and $D_k$ as a function of an internal rotation $\theta$ in the PTI molecule. The $R$ curve is drawn on the back of the box. The continuous and broken lines on the back correspond to the means of PTI random analogues with $R = 15.71$ Å (without compactness constraint) and $R = 11.19$ Å (with compactness constraint), as reported by Cohen & Sternberg (1980a).

$D_k$ curves are plotted inside the box for $k = 5, 10, 15, \ldots, 50$. The curves contained in the left face correspond to the $180^\circ$ derivative. Continuous line: $D_k$ spectrum of the $180^\circ$ analogon deviations from the native state, defined by eqn (5). Broken line: random spectrum of the native conformation defined by eqn (16). The distance of both curves (for a particular $k$ value) indicates the significance of the relatedness of the $180^\circ$ analogon to the native conformation with respect to random deviations.

of $\theta$ is plotted on the back of the box shown in Figure 2. $R$ increases very rapidly to a maximum of 15 Å at about $180^\circ$ distortion. This value is in the range of the mean value of 15.71 Å of PTI random analogues without the compactness constraint reported by Cohen & Sternberg (1980a). This means, in terms of the $R$ method, that the structure generated by the $180^\circ$ rotation is not similar to the native PTI conformation. This is clearly in contradiction to the construction: most of the backbone structure of the distorted PTI molecule is in the native conformation. The N and C-terminal segments are identical to the native state. What has been measured by the $R$ method is the deviation of the overall shapes of both structures. The distance matrix shows the origin of this misleading interpretation: the elements comprising the hatched area in Figure 3 are influenced by the rotation. The greater the order of the side diagonal $B_k$, the more elements of $B_k$ are changed. Due to the expanding domain of the $d_{i,i+k}$ elements, as discussed in section 2. above, the degree of the alteration of these elements affected by the rotation rises. (Note that the $180^\circ$ derivative is far from the mirror image of the PTI conformation.)

Figure 2 shows the behaviour of the $D_k$ method for $k = 5, 10, \ldots, 50$ as a function of $\theta$. The spectrum of the $D_k$ values of the $180^\circ$ derivative (continuous line) and the associated random spectrum (broken line), which is defined in the next section, are plotted on the left-hand side of the box. At low values of $k$ ($k \leq 8$), the $D_k$ deviation is significantly beyond the corresponding random values (for all $\theta$ terms). The $D_k$ spectrum then rapidly approaches the random spectrum and falls off again at
higher levels. The $D_k$ curves in the range of $D_{30}$ are again similar to the $R$ curve, which underlines the equivalence of the two measures.

By inspection of the stereo drawings of the native conformation and the $180^\circ$ derivative shown in Figure 4, it is seen that the latter has maintained a relatively compact conformation. In particular, the N and C termini of both conformations are in similar spatial proximity (though not orientation). The two $\beta$-strands

Fig. 3. Schematic representation of the effect of an internal rotation on the distance matrix. The hatched area represents the elements of the matrix that are changed by the rotation. $B_0$, main diagonal; $B_k$, a particular side diagonal.

Fig. 4. Stereo drawings of (a) the native PTI backbone conformation and (b) the $180^\circ$ analogon. The C$^\alpha$ carbons of Ala25 and Lys26 defining the rotation axis are marked by circles. The N termini are labelled N.
adjacent to the rotation axis lie parallel in the native conformation but are distant in the 180° derivative. The $D_k$ spectrum responds to these changes with random deviations in the medium range ($10 < k \leq 25$) and low deviations in the (far) long range ($k > 30$).

If significant similarities of two structures have been established by the $D_k$ procedure, meaningful information can be obtained when the elements of the low-deviation diagonals are compared in the form of a difference plot, as a function of the amino acid sequence. Such a plot is shown in Figure 5 for $k = 5$ in the above example. The location of the rotation axis is easily found, and the plot shows that the conformations of the two halves of the 180° analogon are identical to the native conformation or vice versa. More generally, the difference plot will indicate whether the similarities are distributed uniformly along the sequence or whether they are centred at particular positions.

The conclusion must be drawn from the above discussion that a measure describing structural deviations by a single value will not suffice to yield an accurate description of structural similarity. The $R$ and $D$ methods, for instance, will indicate only the overall agreement of two structures.

Note that for the assessment of similarities of two structures, a few $D_k$ values will suffice. (1) The $D_{k+1}$ deviation is bounded by $D_k$, i.e. the $D_k$ spectrum varies slowly, and (2) by increasing $k$ the structural levels change slowly. The range for secondary structure comparisons (short range) extends up to $k = 10$, whereas the domains $10 < k \leq 25$ and $k > 25$ may be assigned to the medium and long range, respectively (these boundaries are fluid and somewhat arbitrarily derived from the spectrum shown in Fig. 2). Thus, only a few $D_k$ values (at least one out of each domain) will suffice to provide a good survey of the structural similarities of two conformations.

6. Results on Random Structures

Random analogues of PTI and MBN were generated by the relaxed procedure described in section 3. above. The mean $M$ and standard deviation $\sigma$ of the $R$ values
<table>
<thead>
<tr>
<th>$k$</th>
<th>PTI†</th>
<th>MBN</th>
<th>FDX</th>
<th>REI</th>
<th>LYZ</th>
<th>HMR</th>
<th>GLC</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.91 ± 0.06</td>
<td>0.80 ± 0.04</td>
<td>0.92 ± 0.07</td>
<td>0.95 ± 0.05</td>
<td>0.87 ± 0.05</td>
<td>0.90 ± 0.05</td>
<td>0.93 ± 0.12</td>
</tr>
<tr>
<td>3</td>
<td>2.38 ± 0.22</td>
<td>2.79 ± 0.14</td>
<td>2.29 ± 0.19</td>
<td>2.51 ± 0.16</td>
<td>2.51 ± 0.15</td>
<td>2.80 ± 0.17</td>
<td>2.77 ± 0.30</td>
</tr>
<tr>
<td>4</td>
<td>3.37 ± 0.35</td>
<td>3.49 ± 0.22</td>
<td>3.99 ± 0.30</td>
<td>3.70 ± 0.27</td>
<td>3.33 ± 0.24</td>
<td>3.56 ± 0.26</td>
<td>3.25 ± 0.42</td>
</tr>
<tr>
<td>5</td>
<td>4.07 ± 0.39</td>
<td>3.36 ± 0.26</td>
<td>3.67 ± 0.39</td>
<td>4.75 ± 0.35</td>
<td>3.80 ± 0.30</td>
<td>3.54 ± 0.29</td>
<td>3.00 ± 0.52</td>
</tr>
<tr>
<td>6</td>
<td>4.84 ± 0.46</td>
<td>3.08 ± 0.30</td>
<td>4.29 ± 0.48</td>
<td>5.84 ± 0.43</td>
<td>4.25 ± 0.34</td>
<td>3.84 ± 0.33</td>
<td>3.43 ± 0.69</td>
</tr>
<tr>
<td>8</td>
<td>6.29 ± 0.67</td>
<td>4.54 ± 0.41</td>
<td>5.27 ± 0.77</td>
<td>7.06 ± 0.57</td>
<td>5.27 ± 0.46</td>
<td>4.71 ± 0.46</td>
<td>4.09 ± 0.95</td>
</tr>
<tr>
<td>10</td>
<td>7.41 ± 0.89</td>
<td>5.26 ± 0.53</td>
<td>6.19 ± 1.04</td>
<td>9.11 ± 0.71</td>
<td>6.33 ± 0.63</td>
<td>5.59 ± 0.58</td>
<td>4.93 ± 1.28</td>
</tr>
<tr>
<td>15</td>
<td>9.40 ± 2.05</td>
<td>7.63 ± 0.96</td>
<td>9.53 ± 2.31</td>
<td>10.46 ± 0.90</td>
<td>9.21 ± 1.17</td>
<td>8.32 ± 1.09</td>
<td>8.00 ± 2.09</td>
</tr>
<tr>
<td>20</td>
<td>13.91 ± 3.64</td>
<td>9.01 ± 1.46</td>
<td>13.68 ± 3.00</td>
<td>11.90 ± 1.79</td>
<td>11.43 ± 1.94</td>
<td>11.19 ± 1.71</td>
<td>11.03 ± 5.10</td>
</tr>
<tr>
<td>40</td>
<td>19.51 ± 9.43</td>
<td>19.49 ± 3.70</td>
<td>23.81 ± 11.11</td>
<td>22.15 ± 5.99</td>
<td>19.27 ± 4.03</td>
<td>21.55 ± 5.90</td>
<td>— —</td>
</tr>
</tbody>
</table>

$L^\dagger$ 58 153 54 107 129 113 29

† The PTI deviations are measured from a set of 100 PTI random analogues, the MBN, FDX, REI, LYZ, HMR and GLC deviations from a set of 100 MBN analogues.

‡ $L$, chain lengths of the protein.
of 100 PTI analogues are 16.29 and 3.11 Å, respectively, and are in good agreement with the unrestricted random analogues reported by Cohen & Sternberg (1980a) with parameters 15.71 ± 3.63.

Table I contains the mean values of the $D_k$ deviations of different proteins from random analogues for various values of $k$. It is recognized that the means for particular $D_k$ deviations are similar in all cases, although the chain length ranges from 29 residues in GLC to 153 residues in MBN. This reflects the fact that $D_k$ is independent of the chain length. This measure therefore provides a basis for a unified description of structure comparisons of proteins of different size. Nevertheless, there are differences between the means of $D_k$ for particular diagonals: for example, in PTI and MBN. It is apparent, however, that proteins with similar secondary structure, like MBN and HMR, behave almost identically. and it will be shown below that these differences indeed depend on secondary structure content.

The distributions of the $d_{i,i+k}$ elements of the PTI random analogues for $k \geq 4$ are nearly exactly Gaussian. The parameters $M_k$ and $\sigma_k$ of these distributions are listed in Table 2 for $2 < k \leq 50$ and the distributions for $k = 5$ and 10 are shown in Figure 6. Some values of the distributions of the MBN analogs are listed in Table 3 for comparison. The respective parameters do not differ significantly (which would have been expected at least for low diagonals), indicating that the $\phi, \Psi$ distributions alone, used to construct the Monte Carlo samples, do not reproduce the $\alpha$-helix forming potential of the MBN amino acid sequence. In the following

Table 2

Means and standard deviation of the elements $d_{i,i+k}$ of diagonal $B_k$ found in PTI random analogues

<table>
<thead>
<tr>
<th>$k$</th>
<th>$M$</th>
<th>$\sigma$</th>
<th>$k$</th>
<th>$M$</th>
<th>$\sigma$</th>
<th>$k$</th>
<th>$M$</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>---</td>
<td>---</td>
<td>21</td>
<td>25.241 ± 7.800</td>
<td>41</td>
<td>36.265 ± 12.677</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6.007 ± 0.680</td>
<td>22</td>
<td>25.927 ± 8.051</td>
<td>42</td>
<td>36.758 ± 12.871</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>7.735 ± 1.628</td>
<td>23</td>
<td>26.592 ± 8.292</td>
<td>43</td>
<td>37.258 ± 13.092</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>9.290 ± 2.182</td>
<td>24</td>
<td>27.231 ± 8.537</td>
<td>44</td>
<td>37.766 ± 13.325</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>10.708 ± 2.954</td>
<td>25</td>
<td>27.860 ± 8.774</td>
<td>45</td>
<td>38.268 ± 13.574</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>15.392 ± 4.177</td>
<td>29</td>
<td>30.295 ± 9.794</td>
<td>49</td>
<td>40.032 ± 14.258</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>17.365 ± 4.844</td>
<td>31</td>
<td>31.386 ± 10.355</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>18.290 ± 5.177</td>
<td>32</td>
<td>31.899 ± 10.636</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>19.176 ± 5.475</td>
<td>33</td>
<td>32.398 ± 10.911</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>20.030 ± 5.771</td>
<td>34</td>
<td>32.893 ± 11.188</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>20.389 ± 6.070</td>
<td>35</td>
<td>33.383 ± 11.448</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>21.622 ± 6.374</td>
<td>36</td>
<td>33.880 ± 11.681</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>22.377 ± 6.685</td>
<td>37</td>
<td>34.351 ± 11.894</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>23.104 ± 6.984</td>
<td>38</td>
<td>34.797 ± 12.090</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>23.825 ± 7.275</td>
<td>39</td>
<td>35.259 ± 12.288</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>24.339 ± 7.542</td>
<td>40</td>
<td>35.752 ± 12.477</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
discussion, the parameters derived from the PTI sample are used. The parameters of the Gaussian distributions of the $d_{i,i+k}$ elements can be used to compute the expected value of the $D_k$ deviation of any protein (note that these distributions are independent of the amino acid sequence, at least in the MBN and PTI random populations). First, consider the case of a polypeptide chain that consists entirely of one secondary structure type. The $\varphi, \Psi$ (and $\omega$) values are constant, and the resulting conformation is a helix: ($-57, -47$) and ($180, 180$), for instance. denote

<table>
<thead>
<tr>
<th>$k$</th>
<th>PTI</th>
<th>MBN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M$</td>
<td>$\sigma$</td>
</tr>
<tr>
<td>2</td>
<td>6.01</td>
<td>± 0.68</td>
</tr>
<tr>
<td>7</td>
<td>13.21</td>
<td>± 3.37</td>
</tr>
<tr>
<td>12</td>
<td>18.29</td>
<td>± 5.18</td>
</tr>
<tr>
<td>17</td>
<td>22.38</td>
<td>± 6.60</td>
</tr>
<tr>
<td>22</td>
<td>25.93</td>
<td>± 8.05</td>
</tr>
<tr>
<td>27</td>
<td>29.07</td>
<td>± 9.27</td>
</tr>
<tr>
<td>32</td>
<td>31.90</td>
<td>± 10.84</td>
</tr>
<tr>
<td>37</td>
<td>34.35</td>
<td>± 11.80</td>
</tr>
<tr>
<td>42</td>
<td>36.76</td>
<td>± 12.87</td>
</tr>
<tr>
<td>47</td>
<td>39.17</td>
<td>± 13.95</td>
</tr>
</tbody>
</table>

†See the footnote to Table 1.
an $\alpha$-helix and extended chain, respectively, with undefined length. The distance matrices of helices are doubly symmetric; i.e. all elements of the diagonal $B_k$ are identical, say $d'_k$. Now consider the definition in equation (5) of the $D_k$ deviation in the case of a helix and a random structure:

$$D_k(H, R) = \left[ \frac{1}{(N-k)} \sum_{i=1}^{N-k} (d'_k - \bar{d}_{i,i+k})^2 \right]^{1/2}. \tag{14}$$

The elements $d'_k$ are constant, and the $\bar{d}_{i,i+k}$ terms have Gaussian distributions. The expected value of $v_k^2 = (d'_k - \bar{d}_{i,i+k})^2$ is the second moment of the function $F_k(v)$, which is derived in the Appendix, and computed as:

$$\langle v_k^2 \rangle = \sigma_k^2 + (d' - M_k)^2.$$

Inserting this in equation (14) yields:

$$\langle D_k(H, R) \rangle = \left[ \frac{1}{(N-k)} \sum_{i=1}^{N-k} (\sigma_k^2 + (d'_k - M_k)^2) \right]^{1/2},$$

$$= \left[ \sigma_k^2 \right]^{1/2}. \tag{15}$$

which is independent of chain length. Equation (15) shows that different secondary structures, which have different $d'_k$ values, have different $D_k$ deviations from random structures.

By considerations similar to those leading to equation (15), we arrive at the expected value of the $D_k$ deviations of two random structures (see the Appendix):

$$\langle D_k(R_i, R_j) \rangle = \sqrt{2} \sigma_k.$$

For the general case of any conformation, we obtain:

$$\langle D_k(S, R) \rangle = \left[ \frac{1}{(N-k)} \sum_{i=1}^{N-k} (\sigma_k^2 + (d_{i,i+k} - M_k)^2) \right]^{1/2}. \tag{16}$$

which is easy to evaluate. Table 4 shows the excellent agreement of the $D_k$ values predicted by equation (16) with the deviations found by explicit computation of the mean value from the random populations. The predictions are less reliable in the far long range (maximum error, 12°). The number of $d_{i,i+k}$ elements contained in $B_k$ decreases with increasing values of $k$ (the last diagonal consists of one element), and hence the significance of the $D_k$ deviations is diminishing with increasing $k$. This effect also introduces uncertainty in the predictions due to the small sample space of both the random population (the direct calculation) and the Gaussian distributions (predictions).

Equation (16), which is characteristic for a particular conformation, is defined as the random spectrum of the structure $S$. The numerical values of the random spectrum depend on the parameters $M_k$ and $\sigma_k$, derived from a Monte Carlo sample. These parameters are substantially stable under variations of the amino acid sequence (compare Tables 2 and 3), but they depend on the features of the Monte Carlo procedure used to construct the random sample. The procedure presented in section 3. above, has all the characteristics required to generate genuine C$\alpha$
Table 4

Predicted† $D_k$ deviations of various proteins from random analogues

<table>
<thead>
<tr>
<th>$k$</th>
<th>PTI</th>
<th>MBN</th>
<th>FDX</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R_{\xi}$</td>
<td>$P_{\xi}$</td>
<td>$E_{|}$</td>
</tr>
<tr>
<td>2</td>
<td>0:91</td>
<td>0:91</td>
<td>0:0</td>
</tr>
<tr>
<td>3</td>
<td>2:38</td>
<td>2:38</td>
<td>0:0</td>
</tr>
<tr>
<td>5</td>
<td>4:07</td>
<td>4:04</td>
<td>0:7</td>
</tr>
<tr>
<td>6</td>
<td>4:84</td>
<td>4:79</td>
<td>1:0</td>
</tr>
<tr>
<td>10</td>
<td>7:41</td>
<td>7:25</td>
<td>2:1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$k$</th>
<th>REI</th>
<th>HMR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R$</td>
<td>$P$</td>
</tr>
<tr>
<td>2</td>
<td>0:95</td>
<td>0:95</td>
</tr>
<tr>
<td>3</td>
<td>2:51</td>
<td>2:46</td>
</tr>
<tr>
<td>4</td>
<td>3:70</td>
<td>3:67</td>
</tr>
<tr>
<td>5</td>
<td>4:75</td>
<td>4:77</td>
</tr>
<tr>
<td>6</td>
<td>5:84</td>
<td>5:86</td>
</tr>
<tr>
<td>8</td>
<td>7:66</td>
<td>7:77</td>
</tr>
<tr>
<td>15</td>
<td>10:46</td>
<td>10:61</td>
</tr>
<tr>
<td>20</td>
<td>11:99</td>
<td>12:15</td>
</tr>
</tbody>
</table>

†Predicted via eqn (16) using the PTI distributions of Table 2.
‡$D_k$ values obtained directly from a random sample (same as Table 1).
§Predicted $D_k$ values.
∥Error of the prediction in percentage terms, $E = 100(R - P)/R$.

backbone conformations of polypeptide chains. (1) The exact backbone configuration is preserved. No simplification is introduced that may violate the constraints imposed on the peptide bond by valence bonds and valence angles, and may thus confine or falsify the accessible state space of polypeptides. (2) The random conformations have favourable backbone energies. (3) The conformations have no close contact less than 3.5 Å.

All further constraints are subject to prediction schemes, and if they correspond to useful hypotheses they will lower any of the $D_k$ deviations of a predicted fold or the mean value of a predicted sample compared to the random spectrum.

It may be desirable to establish, in a statistical sense, more reliable parameters $M_k$ and $\sigma_k$ via the construction of an enlarged random sample that contains analoga
CORRESPONDING TO DIFFERENT PROTEINS AND HENCE AMINO ACID SEQUENCES. IT IS EXPECTED, HOWEVER, THAT THE PARAMETERS Μ K AND Α K DERIVED FROM THE PTI SAMPLE WILL NOT CHANGE SUBSTANTIALLY.

7. RESPONSE OF D K AND R TO PRESET SECONDARY STRUCTURE AND THE EVALUATION OF PREDICTION SUCCESS

In attempting to fold proteins from extended chain conformations to the native state, researchers frequently fix secondary structure elements using the crystallographic assignments. Cohen & Sternberg (1980a) and Havel et al. (1979) reported that there is hardly any improvement of prediction success, even if large regions of the backbone structure are preset. It is striking that the D and R procedures do not respond when more than 70% of the correct backbone angles (as in the case of MBN) are supplied to the algorithm, compared to the values obtained from pure random analogues.

To investigate the response of the D k method to preset secondary structure, a set of MBN analogues with preset helices was constructed. The φ, Ψ values corresponding to α-helical residues (residues 3 to 18, 20 to 35, 36 to 42, 51 to 57, 58 to 77, 86 to 94, 100 to 118 and 125 to 148) were assigned to −77 and −47, respectively (at the helix termini, only one dihedral angle is fixed). The assigned φ, Ψ values are not identical to the crystallographic results but this choice corresponds to the situation when one has a reliable secondary structure prediction.

Table 5 contains the results of this study in terms of the D k and R deviations of

**Table 5**

Means of the D k deviation of MBN random analogues and MBN random analogues with preset helices from the native conformation

<table>
<thead>
<tr>
<th>k</th>
<th>MBN, †</th>
<th>MBN, ‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.89</td>
<td>0.55</td>
</tr>
<tr>
<td>3</td>
<td>2.79</td>
<td>1.43</td>
</tr>
<tr>
<td>4</td>
<td>3.49</td>
<td>1.86</td>
</tr>
<tr>
<td>5</td>
<td>3.36</td>
<td>1.95</td>
</tr>
<tr>
<td>6</td>
<td>3.80</td>
<td>2.27</td>
</tr>
<tr>
<td>8</td>
<td>4.54</td>
<td>2.84</td>
</tr>
<tr>
<td>10</td>
<td>5.26</td>
<td>3.45</td>
</tr>
<tr>
<td>15</td>
<td>7.62</td>
<td>5.41</td>
</tr>
<tr>
<td>20</td>
<td>9.90</td>
<td>7.32</td>
</tr>
<tr>
<td>30</td>
<td>14.41</td>
<td>12.40</td>
</tr>
<tr>
<td>40</td>
<td>19.49</td>
<td>19.20</td>
</tr>
<tr>
<td>50</td>
<td>25.56</td>
<td>26.96</td>
</tr>
<tr>
<td>75</td>
<td>30.58</td>
<td>33.79</td>
</tr>
<tr>
<td>100</td>
<td>38.50</td>
<td>41.85</td>
</tr>
<tr>
<td>R</td>
<td>27.17</td>
<td>28.20</td>
</tr>
</tbody>
</table>

† D k means of a pure random sample (same as Table 1).
‡ D k means of a random sample with preset helices.
§ R deviation.
the random population with preset helices from the native MBN conformation. The
$R$ method indicates a slight deterioration in the agreement of the secondary
structure analogues with the native state compared to the pure random set: i.e. $R$
fails to exhibit the excellent agreement of two structures in the short range. Again,
The $D_k$ procedure yields an accurate description: good agreement in the short range
and bad agreement in the long range. The long rigid helices of MBN force the
random analogues with preset secondary structure to more open conformations
compared with pure random analogues.

A simple quality index for prediction studies may be defined in terms of the $D_k$
procedure:

$$Q_k' = [(D_k(N, R) - D_k(N, P))/D_k(N, R)] \times 100,$$

where $N, R$ and $P$ correspond to the native, random (eqn (16)) and predicted
conformation, respectively.

Assume that the helices of the random sample of MBN are predicted by a
prediction algorithm. Then the success of the prediction on $k = 3$ will be:

$$Q_3' = [(2.79 - 1.43)/2.79] \times 100 = 48.7\%,$$

which, however, does not agree very well with about 70\% correct secondary
structure. A better estimate is obtained by:

$$Q_k = \{[(D_k(N, R) - D_k(N, P))/D_k(N, R)]^{1/2}\} \times 100,$$

(17)

which yields 69.8\% for the above example. The exact estimate of the prediction
success by equation (17) may be due to the quadratic nature of the $D_k$ method.

8. Monitoring Folding Studies

To demonstrate the usefulness of the $D_k$ procedure as a powerful tool for the
development of prediction algorithms and folding simulations, a simple though
instructive folding study is performed. The vector $T_0$ of dihedral angles defining a
start conformation (taken from the random set of PTI analogues) is perturbed by a
vector $P$ with components:

$$p_i = 5^{\circ}(x - 1/2),$$

when $i$ matches $\varphi$ or $\Psi$ and $p_i = 0$ if $i$ corresponds to $\omega$. $x$ is a random number
uniformly distributed on $[0, 1]$.

The core of the algorithm (i.e. the functional) is the search for the $C^\alpha$ atom that
has the greatest distance $r_m$ from the centre of gravity $G$ defined by equation (3).
Beginning with the first perturbation $P_1$, the $r_m$ value of the conformation defined
by $T_0 + P_1$ is compared to $r_m(T_0)$. If $r_m(T_0) > r_m(T_0 + P_0)$, the new conformation is
adopted. Generally:

$$T_{n+1} = T_n, \quad r_m(T_n + P_{n+1}) \geq r_m(T_n)$$

$$T_{n+1} = T_n + P_{n+1}, \quad r_m(T_n + P_{n+1}) < r_m(T_n),$$

with the restriction that no $C^\alpha$ contact less than 3.5 Å occurs in the conformation of
Fig. 7. The development of $D_{30}$, $D_{15}$, $D_5$, $R$ and $r_m$ during the folding of a random conformation to a compact form (see the text for a detailed description of the simulation). Only the 66 successful trials (out of 300) are shown.

$T_n + P_{n+1}$. After $n = 300$ perturbations, the algorithm is stopped. This procedure applied to the PTI random set produced a number of structures with $R$ deviations less than 7.5 Å from the native state.

The $D_k$ method yields a reliable description of the conformational changes occurring during the simulation process. Figure 7 and Table 6 give a report of the evolution of the $D_5$, $D_{15}$, $D_{30}$, $R$ and $r_m$ values during the minimization of $r_m$. Whereas $R$ and $D_{30}$ decrease monotonically, there is no significant effect on $D_5$ and $D_{15}$, thus unmasking the construction of the algorithm. This behaviour is characteristic for the whole set of runs (from which the run reported in Fig. 7 and Table 6 is chosen randomly). No improvements in the medium and short range and fast decrease of the long range deviations are produced.

Serious algorithms should therefore exhibit improvements in the short and medium range, and it will obviously be interesting to report a $D_k$ trace of the various prediction algorithms so far reported.

9. Application of the $D_k$ Procedure to the Nucleotide Binding Proteins
   GPDHase and LDHase

As stated in the Introduction, this paper is concerned mainly with the applications of the $D_k$ method to simulation studies. Nevertheless, it is considered
Table 6

Start and end parameters of the minimization of the maximal \( C^\alpha \) vector of a random structure

<table>
<thead>
<tr>
<th>( P )</th>
<th>Start</th>
<th>End</th>
<th>Max( \dagger )</th>
<th>Min( \ddagger )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r_m )</td>
<td>37.19</td>
<td>14.27</td>
<td>37.19</td>
<td>14.27</td>
</tr>
<tr>
<td>( R )</td>
<td>15.51</td>
<td>9.26</td>
<td>15.51</td>
<td>9.12</td>
</tr>
<tr>
<td>( D_4 )</td>
<td>2.47</td>
<td>2.47</td>
<td>2.50</td>
<td>2.42</td>
</tr>
<tr>
<td>( D_3 )</td>
<td>4.61</td>
<td>4.59</td>
<td>4.78</td>
<td>4.48</td>
</tr>
<tr>
<td>( D_2 )</td>
<td>7.19</td>
<td>7.69</td>
<td>7.69</td>
<td>7.18</td>
</tr>
<tr>
<td>( D_{10} )</td>
<td>8.28</td>
<td>8.39</td>
<td>8.41</td>
<td>7.96</td>
</tr>
<tr>
<td>( D_{15} )</td>
<td>10.46</td>
<td>7.64</td>
<td>10.66</td>
<td>7.51</td>
</tr>
<tr>
<td>( D_{20} )</td>
<td>15.38</td>
<td>5.05</td>
<td>15.55</td>
<td>5.05</td>
</tr>
<tr>
<td>( D_{25} )</td>
<td>19.83</td>
<td>3.72</td>
<td>20.24</td>
<td>3.55</td>
</tr>
<tr>
<td>( D_{40} )</td>
<td>13.39</td>
<td>9.88</td>
<td>13.39</td>
<td>6.67</td>
</tr>
</tbody>
</table>

\( \dagger \) See the text for the definitions of the parameters \( P \).
\( \ddagger \) Maximum and minimum values of the parameters during the simulation.

It is desirable to report an explicit example of the applicability of the method to evolution research. Without discussing the special problems that arise in this field, the \( D_k \) method is applied to the comparison of GPDHase and LDHase. Buehner et al. (1973) have shown that there is a strong homology of the nucleotide binding domains of both molecules, whereas the lengths and positions of the secondary structure segments of both domains differ considerably.

By analogy with the procedure reported by Remington & Matthews (1978), the amino acid sequences of two proteins with lengths \( N_1 \) and \( N_2 \) are dissected to \((N_1 - L)\) and \((N_2 - L)\) segments of length \( L \), but the resulting segments are compared by the \( D_k \) procedure. The symbol \( D_k(i, j) \) denotes the \( D_k \) deviation of two segments \( i \) and \( j \), where \( i \) and \( j \) label the first (N-terminal) residues of the compared segments with respect to the amino acid sequence of the two proteins. \( D_k(i, j) \) thus compares the segment \((i, i+1, \ldots, i+L-1)\) of the first protein with the segment \((j, j+1, \ldots, j+L-1)\) of the second. Note that \( k \leq L-1 \) is a necessary condition. The \( D_k(i, j) \) values obtained constitute a comparison map (or matrix) with dimension \((N_1 - L)(N_2 - L)\). Remington & Matthews (1978) reported the comparison map obtained by the \( R \) method as a contour map. An alternative representation can be given by specifying some cut-off value \( s \) and inserting some marker (say a cross) at positions \((i, j)\) of the map with \( D_k(i, j) \leq s \). At regions with structural homology, islands will occur in the comparison map. Islands that extend from bottom left to top right (see Fig. 8(a)) indicate homology of segments longer than specified by \( L \). Generally, if \( D_k(i+m, j+m) \leq s \) holds for \( m = 0 \) to \( M \), the homology extends over \( L + M \) residues.

Figure 8(a) shows the comparison map of the GPDHase and LDHase molecules with parameters \( N_1 = 333 \) (GPDHase), \( N_2 = 329 \) (LDHase), i.e. the whole molecules are compared. \( L = 60 \), \( k = 20 \) and \( s = 5 \) Å. The expected island marked A corresponds to the superposition of the nucleotide binding domains of both
Fig. 8. $D_k$ maps resulting from the comparison of the GPDHase and LDHase conformations. (a) $N_1 = 333$ (GPDHase), $N_2 = 329$ (LDHase), $L = 60$, $k = 20$ and $s = 5$ Å. The window corresponds to the region that is further analysed in (b), (c) and (d). The island marked A corresponds to the known relatedness of the nucleotide binding domains of both molecules. (b) $N_1 = 220$ (GPDHase), $N_2 = 210$ (LDHase), $L = 60$, $k = 20$, $s = 5.5$ Å. The higher cut-off value $s$ causes an enlargement of the islands contained in (a), and the appearance of new ones. (c) $N_1 = 220$, $N_2 = 210$, $L = 60$, $k = 5$, $s = 3.2$ Å. (d) Intersection of (b) and (c). Note that for the intersection of 2 maps, they must have the same $L$ value.

molecules and is in agreement with the results obtained by Rossmann & Argos (1976) and Remington & Matthews (1980). This island with $M = 24$ equivalences residues 1 to $1 + L + M = 85$ of GPDHase with 22 to $22 + L + M = 106$ of LDHase.

However, the significance of the islands in Figure 8(a) is not clear, because one $D_k$ value generally will not suffice to yield reliable structural equivalences. The
window shown in Figure 8(a) corresponding to residues 1 to 220 of GPDHase and 1 to 210 of LDHase, which contains most islands, has therefore been analysed further. Figure 8(b) and (c) show the maps of $D_{20}^{60}(s = 5.5 \text{ Å})$ and $D_{5}^{60}(s = 3.2 \text{ Å})$, and the intersection $D_{20}^{60} \cap D_{5}^{60}$ is shown in Figure 8(d). Only two islands remain in Figure 8(d): A, the known relationship and B, a small island relating the sequence 73 to 137 (GPDHase) to 115 to 179 (LDHase). $R^{60}(73, 115) = 6.45 \text{ Å}$ is relatively small too, yielding a good superposition of the segments, as shown in Figure 9(a). An excellent superposition is obtained by $R^{40}(80, 122)$, and Figure 9(b) clearly exhibits the same folding pattern of both segments. Note that the assignment corresponding to island B does not agree with the alignments suggested by Rossmann et al. (1974), where a segment beginning with residue 87 of GPDHase is aligned with residue 89 of LDHase, whereas B equivalences residue 80 of GPDHase with residue 122 of LDHase. Thus the alignments A and B may indicate a major deletion of the GPDHase molecule relative to LDHase or an insertion in LDHase relative to GPDHase.

Compared to the time-consuming procedures reported by Rossmann & Argos (1976) and Remington & Matthews (1978), computation of the $D_k$ map is very fast. The elements $d_{i,i+k}$ of both structures (i.e. $N_1 + N_2 - 2k$ distances) have to be
Fig. 9. (a) Best superposition \( R = 6.45 \text{ Å} \) of the segment 115 to 174 of LDHase (bottom) on 73 to 132 for GPDHase (top) corresponding to island B (Fig. 8(d)). (b) Best superposition \( R = 3.74 \text{ Å} \) of the segment 122 to 161 for LDHase (bottom) on 80 to 119 for GPDHase (top). The N termini of both segments are labelled by circles.

computed only at the beginning of the computation and can be stored easily in computer memory. The remaining algorithm requires \((N_1 - L)(N_2 - L)\) times the evaluation of:

\[
D^k(i, j) = \frac{1}{c} \left[ \sum_{n=0}^{L-k-1} (d_{i+n,i+k+n} - d_{j+n,j+k+n}) \right]^{1/2},
\]

where \(c = (L - k)^{1/2} \). A considerable saving in computing time is achieved when \(D^k(i, j)\) is computed along the diagonals:

\[
(m, j + m - 1) \quad m = 1, 2, \ldots, N_2 - j + 1
\]

and

\[
(i + m - 1, m) \quad m = 1, 2, \ldots, N_1 - i + 1.
\]

running from bottom left to top right with respect to the comparison matrix.
Consider the square of $D_k^2(i, j)$:

$$c^2(D_k^2(i, j))^2 = \sum_{n=0}^{L-k-1} (d_{i+n, i+k+n} - d_{j+n, j+k+n})^2.$$  \hspace{1cm} (18)

Then $(D_k^2(i+1, j+1))^2$ is obtained from equation (18) by:

$$c^2(D_k^2(i+1, j+1))^2 = c^2(D_k^2(i, j))^2 - (d_{i, i+k} - d_{j, j+k})^2 + (d_{i+L-k, i+L} - d_{j+L-k, j+L})^2.$$  \hspace{1cm} (19)

The speed and simplicity of the $D_k^2(i, j)$ algorithm allows its implementation even on minicomputers.

10. Conclusion

Robson & Osguthorpe (1979) remarked that describing the overall agreement between simulated folds and the observed native structures is complex and has not been resolved. It is claimed here that this problem is satisfactorily solved by the $D_k$ procedure. The features of the $D_k$ method are briefly recalled with respect to the requirements (1) to (5) listed in the Introduction.

(1) Compatibility: $D_k$ is independent of chain length.

(2) Separation of structural levels: this is an immediate consequence of the construction of $D_k$.

(3) The ability to test hypotheses is demonstrated via the folding simulation presented in section 8, above, and through the response of $D_k$ on formation of correct secondary structure.

(4) Extraction of a priori information: if one uses a priori information, like pieces of secondary structure, which can be determined only by X-ray analysis, then the extent of this information should be extracted from the final prediction success. This can be done by deleting from $D_k$ those elements $d_{i, i+k}$ of the distance matrices that are fixed by this information and are unrelated to the quality of the prediction method. For these elements $(d_{i, i+k} - d_{i, i+k})^2 = 0$, and the denominator in equation (5) has to be replaced by $(N - k - t)$, with $t$ being the number of involved elements.

(5) $D_k$ is a high-speed algorithm: a few $D_k$ values, corresponding to different structural levels, suffice to specify structure similarities.

The success of secondary structure predictions is difficult to assess (for a review see Schulz & Schirmer. 1979). As shown in section 7, above, a quality index $Q_3$ can be calculated by fixing the predicted secondary structure segments during the generation of a Monte Carlo sample, computing the $D_3$ deviation from the resulting set and inserting in equation (17). This yields correctly the secondary structure content of the predicted conformations. However, this is a long-winded and time-consuming procedure. A simple and fast estimate is represented by equation (16). When a secondary structure prediction assigns secondary structure elements, then it also assigns the $d_{i, i+k}$ elements contained in the predicted segments. The $D_k(N, P_s)$ deviation ($P_s$ stands for a conformation partially defined by a secondary structure prediction) can be evaluated in the following way. The $d_{i, i+k}$ elements of
the native state and the $d_{i,i+k}$ elements contained in the predicted segments are
known and $(d_{i,i+k} - d_{i,i+k})^2 = v_i^2$ is definite in this case. For the undefined elements
$d_{i,i+k}$ (which can be assumed to be random). $v_i^2 = \sigma^2 + (d_{i,i+k} - M)^2$, which is the
expected value of $v_i^2$ as a function of $d_{i,i+k}$. Combined with equation (16), this
yields:

$$D_k(N, P_s) = \left[ \frac{1}{(N-k)} \sum_{i=1}^{N-k} v_i^2 \right]^{1/2}, \quad (20)$$

where

$$v_i^2 = (d_{i,i+k} - d_{i,i+k})^2, \quad \text{when } d_{i,i+k} \text{ is predicted}$$

$$v_i^2 = \sigma^2 + (d_{i,i+k} - M), \quad \text{when } d_{i,i+k} \text{ is undefined.}$$

Using the helix assignments of MBN presented in section 7. above, equation (20)
yields $D_3(N, P_s) = 1.46$, which is almost identical to $D_3 = 1.43$ (Table 5) derived
from the Monte Carlo set with the MBN helices preset. Equation (17) in
conjunction with equation (20) yields the quality index $Q_3 = 69\%$ in a direct way.
The success of secondary and tertiary structure predictions can thus be evaluated
by the same measure. $D_k$.

As mentioned previously, the development of prediction schemes is particularly dependent on reliable similarity measures for polypeptide
conformations. A reliable and simple method, the $D_k$ procedure, is now available.

Appendix

In the following, the expected values of $v_k^2 = (d_{i,i+k} - d_{i,i+k})^2$ required in section 6
of the main text are derived.

Case 1. $d_{i,i+k} = d'$ = constant. and the $d_{i,i+k}$ terms have Gaussian distributions.
i.e.:

$$f_k(d_{i,i+k}) = f(d) = (\sigma \sqrt{2\pi})^{-1} \exp \left[ -d^2/2\sigma^2 \right].$$

where $k$ is dropped for simplicity, and $M_k = 0$ without loss of generality. The distribution function $F(v)$ of $v = |d' - d|$ (for definition see also Figure 10) is
introduced by:

$$F(v) = f(d' - v) + f(d' + v)$$

$$= (\sigma \sqrt{2\pi})^{-1} \left[ \exp \left[ -(d' - v)^2/2\sigma^2 \right] + \exp \left[ -(d' + v)/2\sigma^2 \right] \right].$$

$F(v)$ is a function of $v$ and $d'$, but the latter is assumed constant as mentioned above.
above.

The expected value of $v^2$ is the second moment of $F(v)$:

$$\langle v^2 \rangle = \int_0^\infty v^2 F(v) \, dv / \int_0^\infty F(v) \, dv$$

$$= (\sigma \sqrt{2\pi})^{-1} \int_0^\infty v^2 F(v) \, dv. \quad (21)$$
i.e. the denominator is equal to one. The nominator (the leading constant is dropped temporarily) splits up to:

\[
\int_0^\infty v^2 F(v) \, dv = \int_0^\infty v^2 \{ \exp \left\{ -(v+d')^2/2\sigma^2 \right\} + \exp \left\{ -(v-d')^2/2\sigma^2 \right\} \} \, dv = \\
= \int_0^\infty (v-d')^2 \exp \left\{ -v^2/2\sigma^2 \right\} \, dv + \int_0^\infty (v+d')^2 \exp \left\{ -v^2/2\sigma^2 \right\}.
\]

Let \( h(v) = \exp \left\{ -v^2/2\sigma^2 \right\} \). then:

\[
\int_0^\infty v^2 F(v) \, dv = \int_{-d'}^\infty v^2 h(v) \, dv + \int_{d'}^\infty v^2 h(v) \, dv
\]

\[
= -2d' \int_{-d'}^\infty vh(v) \, dv + 2d' \int_{d'}^\infty vh(v) \, dv + \int_{d'}^\infty h(v) \, dv + d'^2 \int_{-d'}^\infty h(v) \, dv
\]

(22)

The single terms are integrable in closed form:

\[
\int_{-d'}^\infty v^2 h(v) \, dv + \int_{d'}^\infty v^2 h(v) \, dv = \sigma^3 \sqrt{2\pi}
\]

(23)

\[
2d' \left( \int_{-d'}^\infty vh(v) \, dv - \int_{-d'}^\infty vh(v) \, dv \right) = 0
\]

(24)

\[
d'^2 \left( \int_{-d'}^\infty h(v) \, dv + \int_{d'}^\infty h(v) \, dv \right) = d'^2 \sigma \sqrt{2\pi}
\]

(25)

Collecting equations (23), (24), and (25), and reintroducing the constant \((\sigma \sqrt{2\pi})^{-1}\), yields:

\[
\int_0^\infty v^2 F(v) \, dv = \langle v^2 \rangle = \sigma^2 + d'^2.
\]
Reintroducing \( k \) and \( M_k \neq 0 \) yields, finally:

\[
\langle v_k^2 \rangle = \sigma_k^2 + (d' - M_k)^2. \tag{26}
\]

**Case II.** \( d_{i,i+k} \) and \( d_{i,i+k} \) both have Gaussian distributions. The equivalent to \( F(v) \) is, in this case, the autocorrelation function of the Gaussian distribution:

\[
A(v) = \int_{-\infty}^{\infty} f(x)f(x+v) \, dx = (\sigma \sqrt{2\pi})^{-2} \exp \left[ -v^2/4\sigma^2 \right].
\]

Again:

\[
\langle v^2 \rangle = \frac{\int_{0}^{\infty} v^2 A(v) \, dv}{\int_{0}^{\infty} A(v) \, dv}. \tag{27}
\]

The nominator is thus computed as:

\[
\int_{0}^{\infty} v^2 \exp \left[ -v^2/4\sigma^2 \right] \, dv = 2\sigma \int_{0}^{\infty} (2\sigma v)^2 \exp \left[ -v^2 \right] \, dv
\]

\[
= 2\sigma^3 \sqrt{\pi}. \tag{28}
\]

and the denominator as:

\[
\int_{0}^{\infty} \exp \left[ -v^2/4\sigma^2 \right] \, dv = 2\sigma \int_{0}^{\infty} \exp \left[ -v^2 \right] \, dv = \sigma \sqrt{\pi}. \tag{29}
\]

Inserting equations (28) and (29) into equation (27) yields:

\[
\langle v_k^2 \rangle = 2\sigma_k^2. \tag{30}
\]

The square-roots of equations (26) and (30) represent the expected values of the \( D_k \) deviations of a helix from a population of random structures, and the \( D_k \) deviations of a random structure from a random population, respectively. In the case of a general conformation with \( d_{i,i+k} \neq d_{j,j+k} (i \neq j) \), the random spectrum is found by evaluating equation (26) for each \( d_{i,i+k} \) term and inserting in equation (5), as shown by equation (16).

I thank Professor H. B. Strack and Professor P. Zinterhof for their support, encouragement and discussions; and Dr J. Stegbuchner for discussions of the mathematical aspects of this work.

**REFERENCES**


Edited by R. Huber