Estimating polypeptide α-carbon distances from multiple sequence alignments

András Aszódi and William R. Taylor

Laboratory of Mathematical Biology, National Institute for Medical Research,
The Ridgeway, Mill Hill, London NW7 1AA, UK

Received 14 December 1994

The hydrophobic amino acids that make up the core of a protein can be expected to be closer together than the rest of the residues in the molecule and are likely to remain conserved during evolution due to their important role. In the present study, a general theoretical framework is provided for estimating interresidue distances from residue hydrophobicity and conservation deduced from multiple alignments. While the accurate prediction of individual distances by statistical procedures is theoretically impossible, the method is able to match the distribution of predicted distances to a prescribed distribution with good accuracy.

1. Introduction

The three-dimensional conformation of a protein can be deduced from a set of interresidue distances by various methods collectively known as Distance Geometry techniques [1]. These procedures rely on distance constraints which are obtained from experimental data such as NOE distance estimates or from general stereochemical considerations. In the majority of cases, however, the scarcity of reliable distance data constitutes a serious obstacle and may render the whole approach unfeasible. Therefore, the prediction of interresidue distances is of considerable theoretical as well as practical importance.

In the present study we describe a prediction method that is based on the hydrophobic effect, one of the key factors of protein structure formation. In qualitative geometric terms, the hydrophobic effect can be regarded as the tendency of hydrophobic amino acid side chains to cluster together within the core of a folded protein molecule. Therefore, hydrophobic amino acids in the core can be expected to be closer together than the rest of the residues in the protein molecule. Also, the hydrophobic amino acids which make up the hydrophobic core are likely to remain conserved during evolution due to their important role, and could be detected by sensitive multiple alignment techniques [2]. By combining these two factors, our approach attempts the prediction of interresidue distances as a function of average residue hydrophobicity and conservation. The transform function that maps hydrophobicity and conservation values to distance estimates is determined in such a way that the distribution of the resulting estimates matches a predetermined distance distribution.

2. Methods

2.1. DISTANCE ESTIMATION FROM MULTIPLE ALIGNMENTS

The $C_\alpha - C_\alpha$ distances between hydrophobic residues in proteins tend to be comparatively small, due to the hydrophobic effect that leads to the formation of a buried hydrophobic core inside the molecule. When the sequences of several related proteins with presumably similar three-dimensional structure are available, the alignment of these sequences can highlight the structurally important amino acids as these usually occupy conserved positions. To quantify this effect, a score was introduced that gave high values for conserved and hydrophobic amino acid pairs. A transform function was then used to convert these score values into pairwise distance estimates so that their distribution matched a theoretical or observed interresidue distance distribution.

2.1.1. The hydrophobic packing score

The conservation g_i at the ith position of an alignment of N sequences was measured by an unweighed average of the pairwise similarity scores of all amino acids in the position:

$$g_i = \frac{2}{N(N-1)} \sum_{j=1}^{N-1} \sum_{k=j+1}^{N} M(R_{ij}, R_{ik}), \qquad (1)$$

where R_{ij} is the type of the amino acid in the jth sequence in alignment position i and $M(\cdot, \cdot)$ is an entry in an amino acid similarity matrix [3]. The similarity matrix was scaled so that all entries were $\geqslant 0$ by subtracting the smallest entry from all the others, and gaps were skipped in the summation. For a single sequence ($N = 1$) all g_i values were set to 1. The conservation value was normalised by dividing it by the maximal score encountered among the amino acid pairs:

$$c_i = \frac{g_i}{max_{j<k} M(R_{ij}, R_{ik})}, \qquad (2)$$

ensuring that $0 \leqslant c_i \leqslant 1$. This measure gave 0 for positions which contained only one amino acid and $N - 1$ gaps, while totally conserved positions made up exclusively of one kind of amino acid and no gaps gave $c_i = 1$. Following [4], the *pairwise hydrophobic packing score* was then defined as

$$h_{ij} = c_i h_i + c_j h_j, \qquad (3)$$

where c_i is the conservation and h_i is the average Levitt hydrophobicity [5] of the amino acids in the ith sequential position.

2.1.2. *Converting hydrophobic packing scores into expected distances*

The hydrophobic packing score gave a high value for conserved hydrophobic pairs while the expected distances for these should be low. Consequently, a monotonically decreasing transform function, depending on a parameter vector $\boldsymbol{p} = (p_1, p_2, \ldots)$ was needed that converted the hydrophobic scores (h_{ij}) into expected distances (d_{ij}):

$$d_{ij} = T(h_{ij}, \boldsymbol{p}) . \tag{4}$$

[6] used the transform function

$$T_{old}(h, \boldsymbol{p}) = p_1 h_{ij}^{-p_2} - p_3 , \tag{5}$$

in which the negative exponent $-p_2$ ensured the inversion of the measure. However, this function had the undesirable property of having a discontinuity at $h = 0$. Consequently, an alternative formula

$$T_{new}(h, \boldsymbol{p}) = -p_1 h_{ij}^{p_2} + p_3 \tag{6}$$

was also tested which is continuous over $h \geqslant 0$. The parameter estimation was constrained by the requirement that all three parameters should be positive for both functions.

2.2. ESTIMATION OF THE PARAMETERS OF THE TRANSFORM FUNCTION

The parameters in the transform function had to be adjusted so that the distribution of the expected distances should match a prescribed distribution of $C_\alpha - C_\alpha$ distances. The most straightforward approach would have been to apply the transform function to all hydrophobic scores, calculate the distribution of the expected distances and evaluate the fit. However, this procedure should have been repeated several times in each cycle of the nonlinear parameter estimation algorithm, which, given the large number (usually at least 10^4) of the packing scores, would have resulted in unacceptably long execution times.

Instead of this "brute force" approach, a simpler procedure was chosen. Since the hydrophobic packing scores h_{ij} were mapped to the expected distances d_{ij} by the transform function $T(h, \boldsymbol{p})$ (eq. (4)), the relationship between the distribution of a random variable and the distribution of its function (which is another random variable) could be exploited (see, e.g., [7]). Keeping in mind that $T(h)$ is monotonic decreasing, the probability density function (p.d.f.) of the packing scores $f(h)$ can be expressed in terms of the p.d.f. of the expected distances $g(d)$ as

$$f(h) = g(T(h, \boldsymbol{p})) \left| \frac{\partial T}{\partial h} \right| . \tag{7}$$

Similarly, the cumulative probability function (c.d.f.) of the packing scores $F(h)$ is related to the c.d.f. of the expected distances $G(d)$ as

$$F(h) = 1 - G(T(h,\boldsymbol{p})) . \tag{8}$$

Both eq. (7) and eq. (8) specify an "ideal" packing score distribution determined by the target distribution of the expected distances (fig. 1). The estimation of the parameter vector $\boldsymbol{p}$ in $T(h,\boldsymbol{p})$ could then be carried out by fitting the actual distribution of the h_{ij} scores to this "ideal" distribution using either the p.d.f.'s or the c.d.f.'s. The nonlinear regression algorithm operated on the mappings eq. (7) or eq. (8) instead of the transfer function $T(h,\boldsymbol{p})$. The p.d.f. or c.d.f. of the expected distances was approximated by cubic splines to make the mappings continuous. The parameter estimation was carried out by the standard Gauss–Newton–Mar-

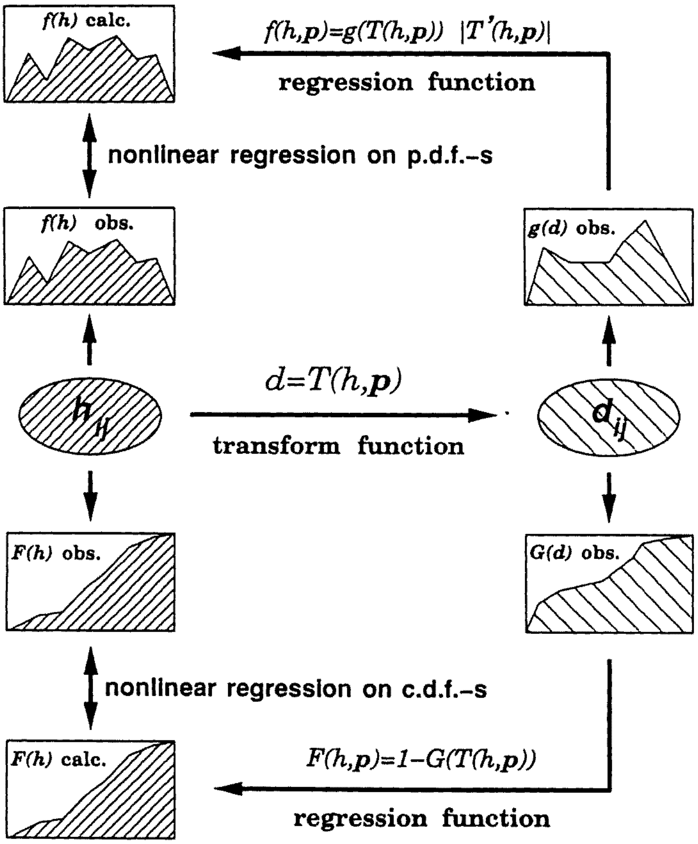


Fig. 1. Scheme of the transform function parameter estimation. The set of hydrophobic packing scores h_{ij} is mapped to the set of expected distances d_{ij} by the transform function $d = T(h,\boldsymbol{p})$. The estimation of the parameters $\boldsymbol{p}$ is carried out either by fitting the observed p.d.f. $f(h)$ to the ideal p.d.f. calculated from the p.d.f. of the expected distances $g(d)$ (upper part of the diagram), or by fitting the corresponding c.d.f.'s (lower part of the diagram). The two procedures are mathematically equivalent. See text for details.

quardt nonlinear regression method [8,9]. The algorithm was started with several different sets of initial parameters to avoid local minima. The quality of the estimation was measured by the residual deviation, the estimated standard deviations of the parameters and their estimated correlation coefficients.

2.3. THEORETICAL DISTANCE DISTRIBUTIONS

2.3.1. Spherical approximation

Protein molecules were approximated by spheres so that the volume of the sphere was the same as the expected volume of the molecule. This volume depends on the chain length only, since the residue density (the number of C_α atoms per unit volume) ρ is approximately constant [10]. The expected radius of the sphere corresponding to the protein is

$$R_{exp} = \left(\frac{3N}{4\pi\rho}\right)^{1/3}, \tag{9}$$

where N is the number of residues and $\rho = 6.3 \times 10^{-3}$ Å^{-3} [11].

2.3.2. The distance distribution within a sphere

The probability of finding two points, **A** and **B** within a sphere of radius R which lie less than D distance apart can be broken down into two separate probabilities. Firstly, the probability of finding a point **A** which is closer to the centre O of the sphere than a is the ratio of the volume of the sphere with radius a to the volume of the encompassing sphere:

$$Pr(r_A \leqslant a) = F_A(a) = \frac{a^3}{R^3}, \tag{10}$$

and the density function of this distribution is

$$f_A(a) = \frac{dF_A}{da} = \frac{3a^2}{R^3}. \tag{11}$$

Secondly, the probability of finding another point **B** which is less than D distance apart from **A** is the ratio of the volume of the sphere with radius D (centered on **A**) to the volume of the large sphere. This reasoning, however, is valid only if $a \leqslant R - D$ and the small sphere around **A** is completely contained by large sphere (Case I in fig. 2). In this case, we obtain (analogously to eqs. (10) and (11)):

$$Pr(D_{AB} \leqslant D, a \leqslant R - D) = F_{D,I}(D) = \frac{D^3}{R^3}, \tag{12}$$

$$f_{D,I}(D) = \frac{dF_{D,I}}{dD} = \frac{3D^2}{R^3}. \tag{13}$$

For the partially overlapping case (Case II in fig. 2), when $a > R - D$, the volume of the small sphere *within* the large sphere is

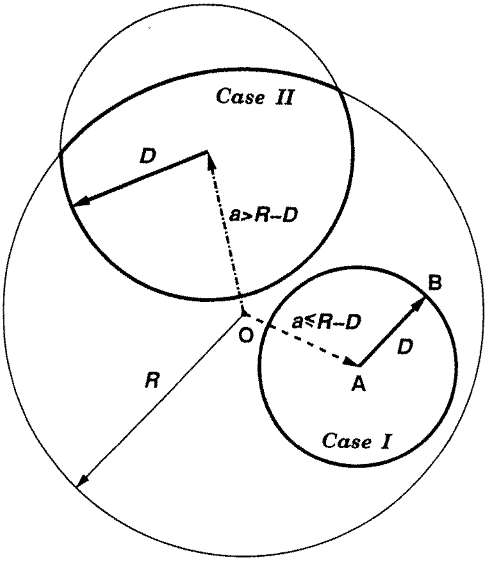


Fig. 2. The calculation of the distribution of distances within a uniform sphere of radius R. For a given distance D, first a point **A** is chosen whose distance from the origin is a. Then a second point **B** is picked, from a distance D away from **A**. The probability of this second event is proportional to the volume of the thick sphere on the right side if $a \leqslant R - D$ (Case I). If $a > R - D$, then the probability is proportional to the volume of the intersection of the small sphere with the large sphere (Case II, thick outline). See text for details.

$$V_{II} = \frac{\pi}{12a}[a^4 - 6a^2(D^2 + R^2) + 8a(D^3 + R^3) - 3(R^2 - D^2)^2] \tag{14}$$

[12], the corresponding probability is (cf. eq. (12)):

$$Pr(D_{AB} \leqslant D, a > R - D) = F_{D,II}(D) = \frac{3V_{II}}{4\pi R^3}, \tag{15}$$

which yields the density function by differentiating with respect to D:

$$f_{D,II}(D) = \frac{dF_{D,II}}{dD} = -\frac{3D}{4aR^3}[(a - D)^2 - R^2]. \tag{16}$$

Since picking the points **A** and **B** are independent events, the joint probability density function will be the product of the two separate density functions $f_A(a)$ and $f_D(D)$:

$$f(a, D) = \begin{cases} f_A(a)f_{D,I}(D) & \text{if } a \leqslant R - D, \\ f_A(a)f_{D,II}(D) & \text{if } a > R - D, \end{cases} \tag{17}$$

from which the density function for D can be obtained by integration:

$$\begin{aligned} f(D) &= \int_0^{R-D} f_A(a) f_{D,I}(D)\, da + \int_{R-D}^{R} f_A(a) f_{D,II}(D)\, da \\ &= \frac{9}{R^6} \int_0^{R-D} a^2 D^2\, da - \frac{9}{4R^6} \int_{R-D}^{R} D[(a-D)^2 - R^2]\, da \\ &= \frac{3D^2}{16R^6} (D^3 - 12R^2 D + 16R^3)\,. \end{aligned} \tag{18}$$

The density function is 0 outside the interval $[0 \ldots 2R]$, it is unimodal but it is not symmetric around $D = R$ (fig. 3). The first and second moments (cf. [6]) are

$$\bar{D} = \int_0^{2R} Df(D)\, dD = \frac{36}{35} R\,, \tag{19}$$

$$\overline{D^2} = \int_0^{2R} D^2 f(D)\, dD = \frac{6}{5} R^2\,. \tag{20}$$

The cumulative distribution function can be obtained from eq. (18) by integration:

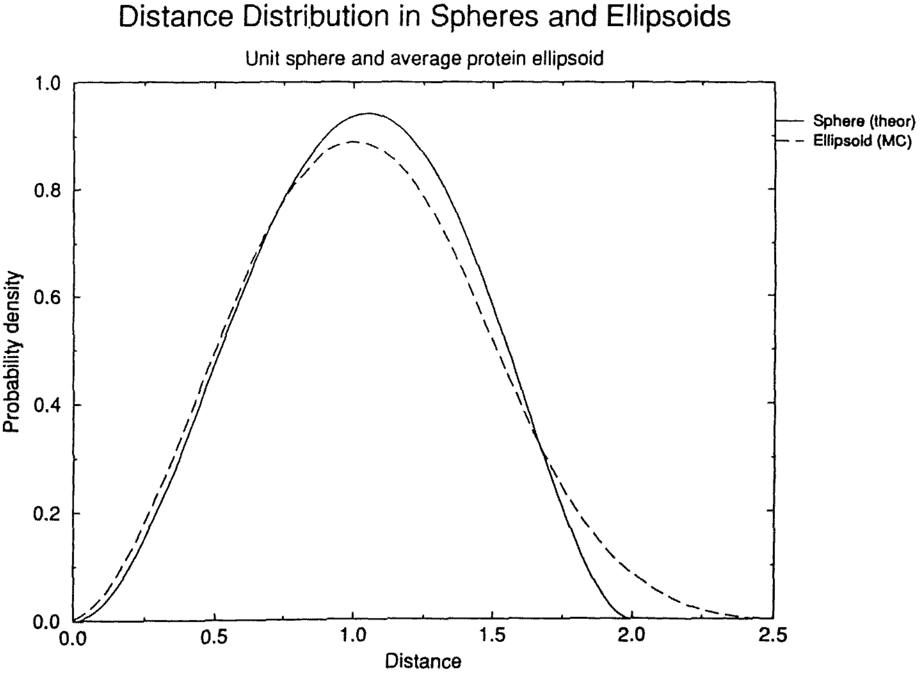


Fig. 3. Distance distributions in spheres and ellipsoids. The probability density functions of the distance distribution within a unit sphere (solid line) and within an ellipsoid of the same volume, axis ratios 1.00 : 0.75 : 0.60 (dashed line) are shown. The ellipsoid distribution has a distinct "tail" in the high distance region, indicating that an ellipsoid can accommodate longer distances than a sphere of equal volume.

$$F(d) = d^3\left(1 - \frac{9}{16}d + \frac{d^3}{32}\right), \tag{21}$$

where d is the normalised distance D/R.

2.3.3. *Ellipsoidal approximation*

A more accurate approximation of the molecular shape of proteins was achieved by fitting ellipsoids. Following [13], the ellipsoid was centred on the centroid of the molecule, its semiaxes collinear with the principal axes of inertia. The length of the semiaxes A, B and C were proportional to the three moments of inertia (the square root of the eigenvalues of the moment matrix). The semiaxes were then scaled by the same amount so that the resulting ellipsoid contained 90% of the point set. This scaling was performed to compensate for the concave crevices present on most protein surfaces. The semiaxis length ratios of the average ellipsoid calculated from an 84-protein dataset [14] are $A : B : C = 1.00 : 0.75 : 0.60$.

The derivation of the density function f_D of the distribution of distances within an ellipsoid could be performed, at least in theory, similarly to that of eq. (18), but the calculations are much more tedious. Instead, the distribution was approximated by a simple Monte Carlo simulation. The ellipsoid in question was scaled so that its volume was set equal to that of a unit sphere ($R = 1$) and then random points were chosen within it, using a pseudo-random number generator [15]. The main qualitative difference between the density distributions within spheres and ellipsoids was that the ellipsoid distributions had distinct "tails" in the region of large distances (fig. 3) due to the fact that an ellipsoid can accommodate longer distances than a sphere with the same volume.

2.4. DATA COLLECTION

2.4.1. *α-Carbon distance statistics*

The distribution of $C_\alpha - C_\alpha$ distances in folded polypeptides was approximated by processing a subset of monomeric protein structures which were well-resolved (better than 2.6 Å), non-homologous (less than 25%) and the chain lengths fell between 100 and 200 residues. 35 proteins (Brookhaven codes 155C, 1ACX, 1CD4, 1ECA, 1FD2, 1FKF, 1GCR, 1I1B, 1IFB, 1L58, 1LH1, 1LZ1, 1MBA, 1MBD, 1PAZ, 1RBP, 1RNH, 1YCC, 2CDV, 2FCR, 2RHE, 2RNT, 2SGA, 2SNS, 2SSI, 2STV, 3DFR, 3FGF, 4BP2, 4CPV, 4FXN, 4TNC, 5P21, 7RSA, 8DFR) satisfied these criteria, giving a total of 349268 distances. Distances between C_α atoms of sequential neighbours were not included in the analysis (this distance is 3.8 Å in all polypeptides). The c.d.f. of the distance distribution $G(D)$ was approximated by a cubic spline fitted to the 100-bin cumulative histogram of the distance data.

2.4.2. *Multiple alignments*

Multiple alignments were constructed by the MULTAL program [2]. The sequences of a set of proteins with known structures were compared to the rest of the sequence database and then aligned to the most similar entries. The hydrophobic packing scores were then calculated from each alignment (see section 2.1.), see table 1.

Table 1

Protein	Sequences	Total length	Average length	Conservation
1BP2	10	126	119.40	0.746
1CCR	10	118	107.10	0.707
1CTF	10	129	116.90	0.731
1ECA	9	163	148.22	0.737
1GCR	10	183	166.00	0.713
1LZ1	10	148	139.70	0.768
1MBA	5	147	145.20	0.937
2AZA	10	185	141.20	0.598
2MHR	5	118	115.00	0.813
2PAB	6	147	138.17	0.857
2RHE	10	142	112.10	0.569
2RNT	10	113	104.30	0.761
4CPV	10	111	109.00	0.815
4FD1	6	113	102.00	0.713
4FXN	8	146	140.62	0.656
7RSA	10	158	129.60	0.676

2.5. IMPLEMENTATION

The multiple alignment program MULTAL [2] was written by W.R.T. in traditional C and run on a Silicon Graphics Challenge L server. The distance distribution approximation and nonlinear parameter estimation programs were written by A.A. in ANSI C and run on a Silicon Graphics Indigo workstation. The graphs were produced by P.J. Turner's ACE/gr freeware visualisation package (Version 2.10).

3. Results

3.1. TRANSFORM FUNCTION PARAMETER ESTIMATION

3.1.1. Numerical properties

While in strictly theoretical terms the use of p.d.f.'s or c.d.f.'s for the estimation of transform function parameters are equivalent, in practice the method based on the fit of cumulative distribution functions proved superior to its counterpart that used probability density functions. The data used in the estimation were inherently noisy and this noise was amplified by the p.d.f.'s and "smoothed out" by the c.d.f.'s which was not totally unexpected, given that the p.d.f. of a distribution is the derivative of the corresponding c.d.f., and that differentiation enhances noise. While some carefully chosen weighting schemes can sometimes improve the quality of the estimation, in our case it would have been difficult to compensate for the amplification of noise inherent in the p.d.f. transformation. Consequently, the results presented in this study were generated by fitting c.d.f.'s with uniform weights.

The choice of the transform function T also influenced the estimation. The function T_{old} (eq. (5)) was used in previous studies [6], but some of its numerical properties were undesirable. Not only had it a discontinuity at $h = 0$ which was inconvenient since the hydrophobic packing scores could very well be zero, but also its rate of change in the range of small $h - s$ was very high, which led to a noise amplification phenomenon similar to the effect of derivation outlined above. The alternative transform function T_{new} defined by eq. (6), on the other hand, was defined for all non-negative h values and changed more smoothly in the range of interest except for an initial sharp drop when $h \ll 1$. In general, T_{old} did not give satisfactory results in preliminary trials, therefore T_{new} was used instead.

3.1.2. Numerical results

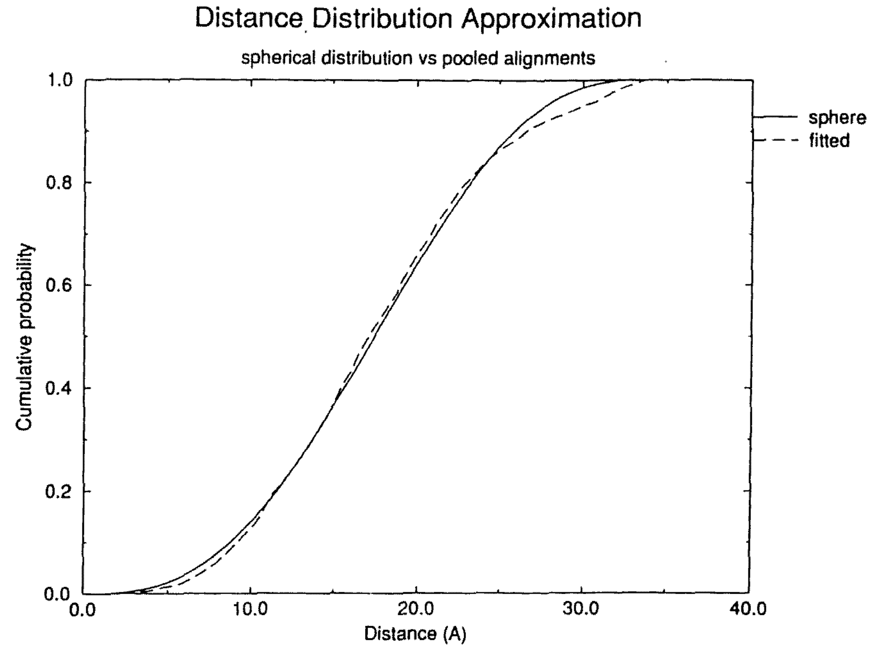
The c.d.f. of the distribution of hydrophobic packing scores pooled from 16 multiple alignments was fitted to the c.d.f. of three different distributions:

1. the distribution of distances within a sphere (eq. (21)) that had a radius $R_{exp} = 16.89$ Å calculated from the average sequence length (eq. (9));
2. the distribution of distances within an ellipsoid that had the same volume as the sphere above and the ratio of its semiaxes were 1.00 : 0.75 : 0.60 (the average protein ellipsoid);
3. the observed distribution of α-carbon distances in 35 non-homologous proteins with chain lengths between 100 and 200 residues.

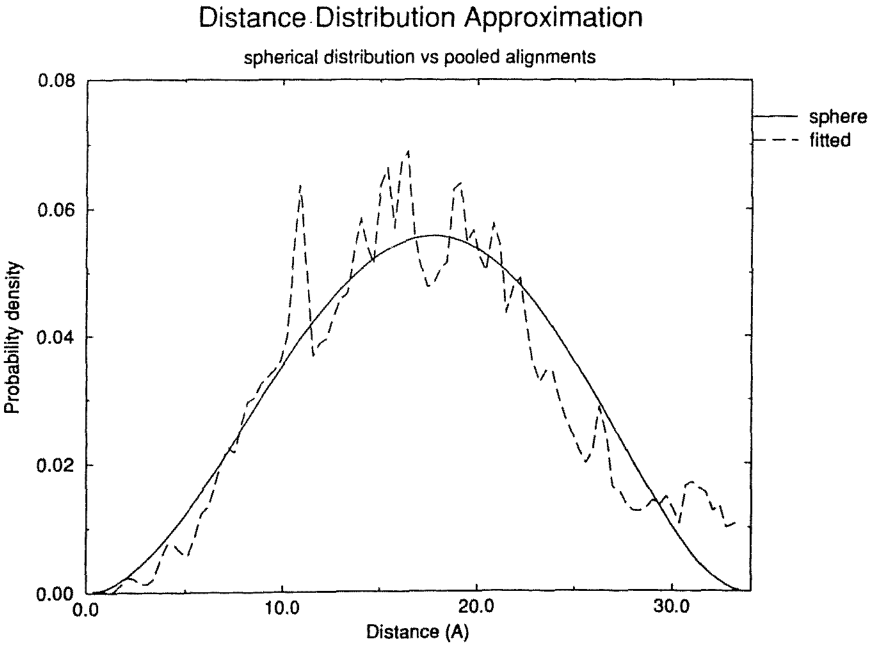
For all three cases, the observed and calculated p.d.f.'s were obtained from the corresponding c.d.f.'s by numerical derivation to facilitate comparison (fig. 4). The estimated values of the parameters and their standard deviations as well as the residual deviations are given in table 2.

Table 2

Parameter	Sphere	Ellipsoid	Observed α-carbon
p_1	19.3 ± 0.6	31.9 ± 0.9	30.3 ± 0.9
p_2	0.339 ± 0.008	0.220 ± 0.006	0.264 ± 0.006
p_3	37.4 ± 0.6	50.2 ± 0.9	50.5 ± 0.8
Residual	0.011	0.015	0.004



a



b

Fig. 4. Transform function parameter estimation by fitting distance distributions. The expected distances were characterised by spherical (a, b), ellipsoidal (c, d) and experimental α-carbon distance (e, f) distributions, respectively (solid lines). The fitted distributions (dashed lines) were calculated using the transform function with the estimated parameters. The p.d.f.'s (b, d, f) are much "noisier" than the corresponding c.d.f.'s (a, c, e).

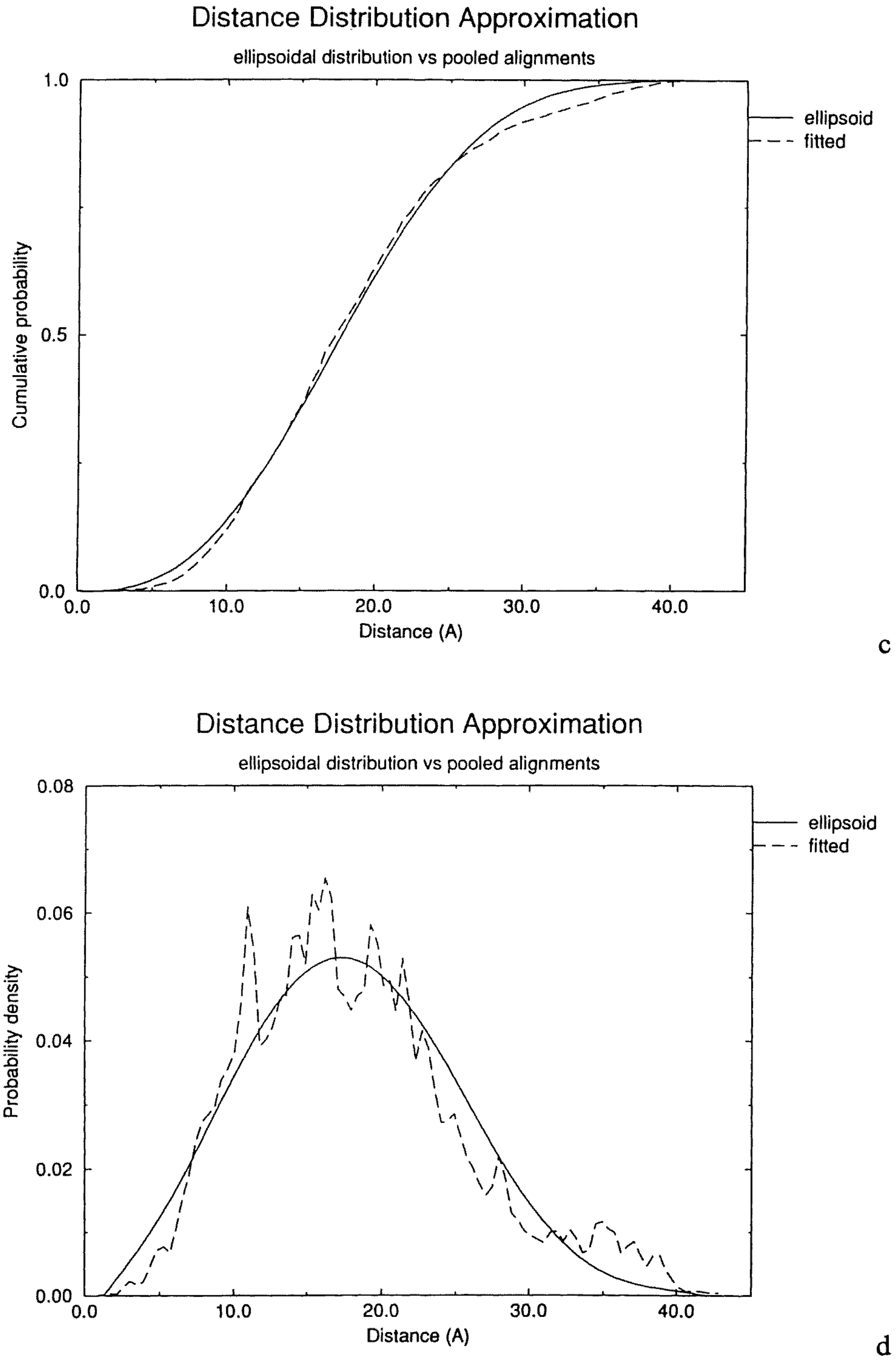
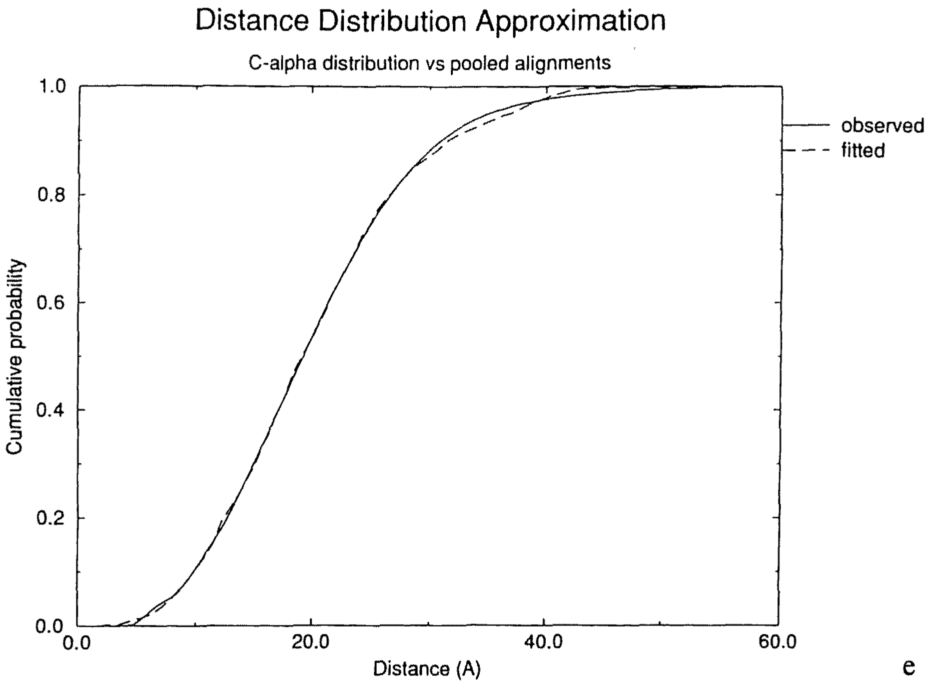


Fig. 4. (Continued.)



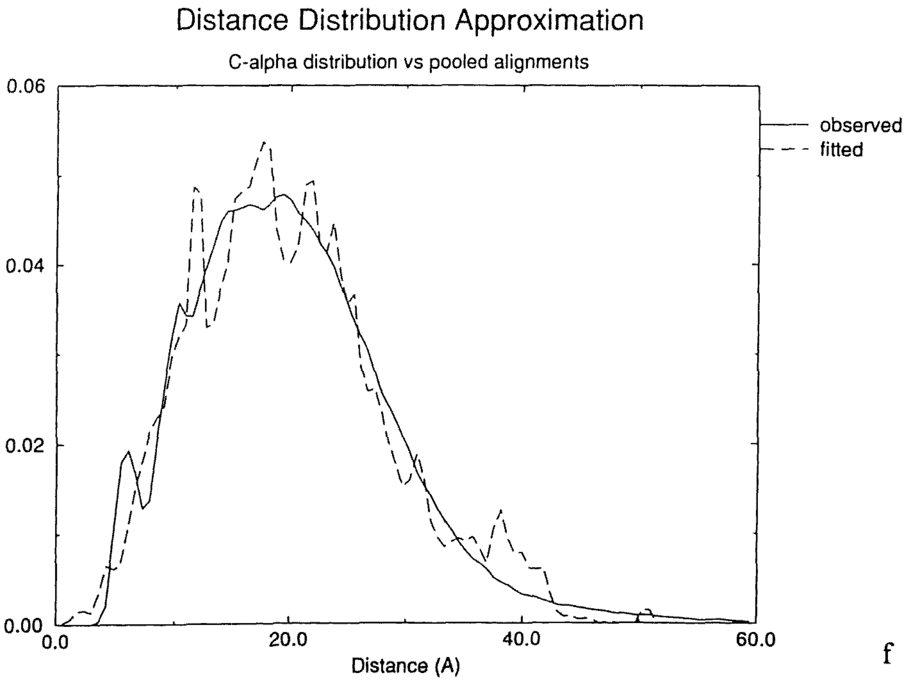


Fig. 4. (Continued.)

The best fit was obtained with the distribution of the "real" α-carbon distances. The spherical and ellipsoidal distributions gave less satisfactory results as judged from the graphs and the residual deviation values. The parameters estimated from the ellipsoidal distribution were much closer to the values obtained from the fit to the observed α-carbon distance distribution than to the spherical approximation data. The graphs of the transform functions (fig. 5) were similar in all cases, suggesting that the transformation was not particularly sensitive to variation in the parameters. The length of the alignment was more important for the quality of the estimation than the number of sequences in the alignment.

3.2. APPLICATION TO MODELLING

We have developed a distance geometry-based algorithm, DRAGON that folds up model polypeptide chains into compact globules with distinct hydrophobic cores [11]. The interresidue distance estimation technique described above was built into this program and 25 model structures for the protein 4CPV were generated, using a multiple alignment of 10 homologous sequences. The pooled interresidue distance distribution for the models matched that of the native 4CPV structure (fig. 6(a)), indicating that the estimation provided a reliable guidance to the model-

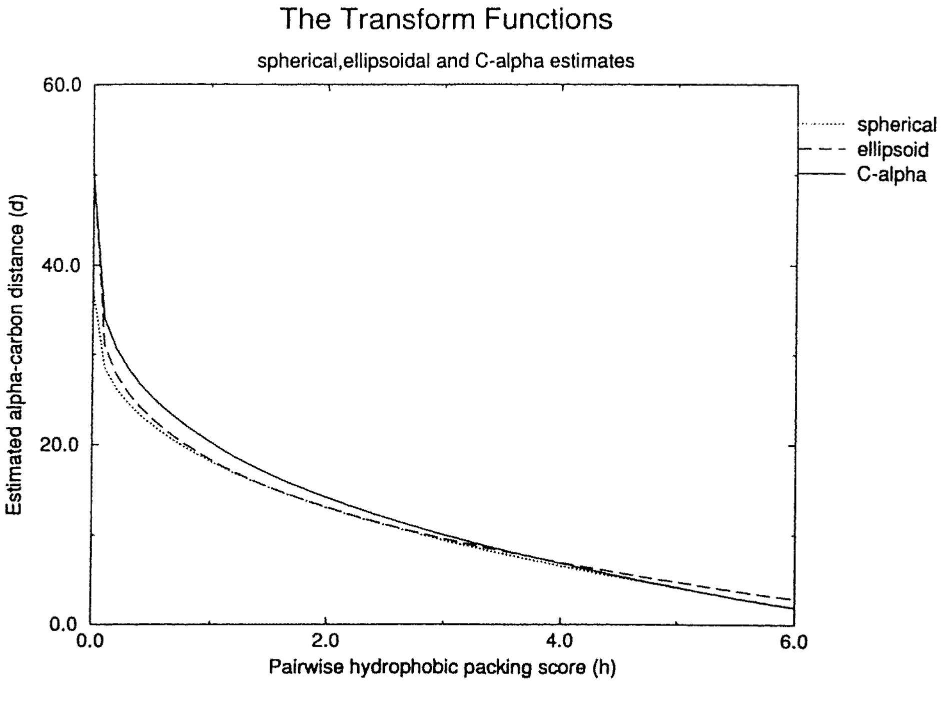


Fig. 5. The transform functions estimated from the spherical (dotted line), ellipsoidal (dashed line) and experimental α-carbon (solid line) distance distributions. The graphs are very similar, indicating that the method was not particularly sensitive to the choice of the ideal distribution.

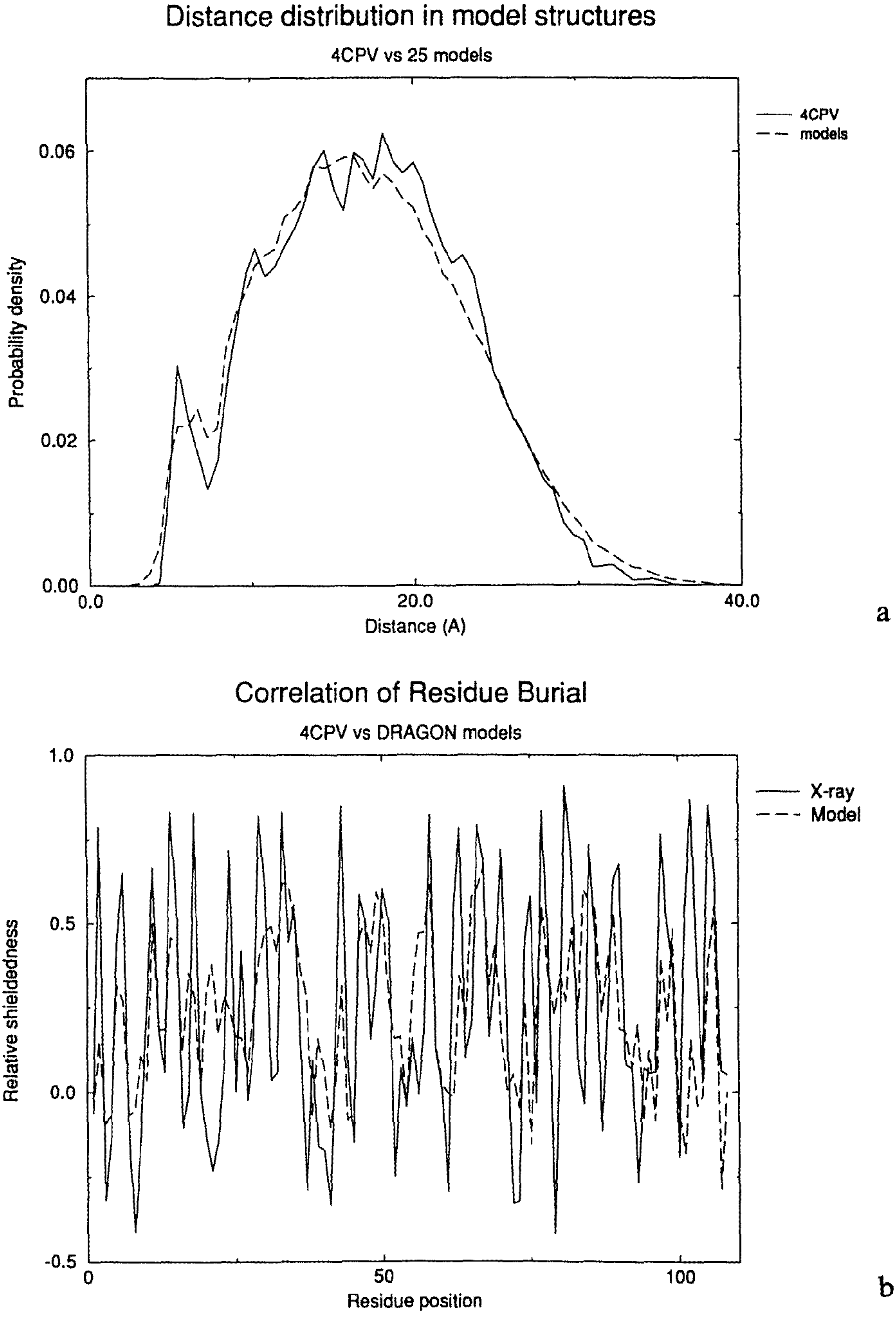


Fig. 6. Application of the distance prediction method to structural modelling. 25 models of carp parvalbumin (4CPV) were generated by the program DRAGON. (a) Distance distribution. The p.d.f. of interresidue distances in carp parvalbumin (solid line) is matched by the pooled p.d.f. of distances from the model structures (dashed line). (b) Correlation of residue burial. The relative shieldedness of the residues in carp parvalbumin (solid line) is approximately matched by the average shieldedness values from the 25 model structures (dashed line).

ling program. The solvent accessibility of each residue (approximated by the "cone" algorithm, [11]) in the models was also averaged and compared to that of the corresponding residue in the native structure (fig. 6(b)). The average model accessibilities reproduced the native accessibilities fairly well (correlation coefficient $R = 0.51$), indicating that the method was capable of predicting the "core membership" of the individual amino acids.

4. Discussion

4.1. DISTANCE DISTRIBUTIONS

The interresidue distance distribution in protein molecules is determined by the packing of main and side chains within the protein interior. A detailed model of packing of amino acids should take the following factors into account:

1. The residues in the polypeptide chain have finite volumes and are practically impenetrable. This *excluded volume effect* means that two atoms cannot approach each other any closer than a well-defined lower limit (the sum of their van der Waals radii).
2. The volumes of the various amino acid side chains are different.
3. The molecular shapes of the amino acid residues are rather complex and a spherical approximation is not always justifiable.
4. The shape of globular proteins could better be approximated by an ellipsoid than by a sphere.

The theoretical distribution of distances within a uniform sphere given by eq. (18) was very simple to evaluate and could serve as a quick approximation. For the sake of simplicity, however, the details of packing outlined above were ignored in the derivation. More accurate descriptions of the packing of identical impenetrable spheres, both theoretical [16,17] as well as experimental [18] are able to reproduce the peaks of the distribution at short distances corresponding to local order in the arrangement. While these investigations represent an important contribution towards the understanding of the structure of dense molecular aggregates, the application of results to protein molecules is by no means straightforward.

The p.d.f. of the protein α-carbon distance distribution showed several interesting features. The overall shape of the curve was similar to the p.d.f. of the ellipsoid distance distribution, indicating that the ellipsoidal approximation was more accurate than the spherical method. The most conspicuous features were the distinct peaks at 5.5 and 11.0 Å which corresponded to the first and second "coordination spheres" around the amino acids. These peaks were not present in the theoretical distance distributions where uniform density was assumed, suggesting that these approximations are valid only if global molecular properties are analysed. The most reliable approach, therefore, seems to be to use the α-carbon distance distribution obtained from a nonhomologous subset of PDB entries whose chain lengths roughly correspond to the length of the chains in the prediction.

4.2. INTERRESIDUE DISTANCE PREDICTION

The method of estimating α-carbon distances from conserved hydrophobicity scores was originally applied by Taylor [4,6] to build structural models of proteins from multiple sequence alignments. While the underlying idea remains the same, in the present study we attempted to give the method a solid theoretical basis. Rather than scaling the hydrophobic scores so that the first two moments of the distribution of the predicted distances matched those of the observed distribution, the fit between two distributions was carried out. The main advantage of this new approach is that it retained the information which would otherwise have been discarded by the calculation of moments only. The procedure enabled the systematic assessment of transform functions on a quantitative basis.

4.2.1. Numerical properties

The method described in the present study proved to be reasonably robust. Firstly, the c.d.f.'s used in the transform acted as smoothing filters which reduced noise to a considerable extent as opposed to the p.d.f.-based approach which was much more sensitive to noise. Secondly, the distributions were calculated from pairwise distance data and packing scores obtained from alignments. From an alignment that contained N positions, $N(N-1)/2$ pairwise packing scores could be calculated, providing a large amount of raw data for the estimation of distributions. The quality of the parameter estimation did not depend heavily on the number of sequences in the alignment for the same reason: when the packing scores were calculated from subalignments of $M/2$ sequences taken from an alignment of M sequences, the residual deviation did not change dramatically.

4.2.2. Limitations

Perhaps the most important limitation of the present scheme is that in general it is impossible to predict *all* individual interresidue distances. Although the distribution of expected distances could accurately be reproduced, the actual distance between an arbitrary pair of residues in a particular protein depends on many other stereochemical factors. Even if we know that residue X always favours the company of residue Y, and therefore the expected X–Y distance is small, there might be other, unrelated Y's far away in another region of the protein whose distances from X are essentially random and cannot be reliably predicted by any scheme.

4.2.3. Applications

The method for fitting distributions of random variables which are functions of each other is fairly general and therefore can easily be applied to a wide variety of prediction schemes. In particular, the approach described above can readily be applied in distance-based protein structure prediction methods. By reproducing the observed α-carbon distance distribution, an appropriate scaling of estimated distances can be achieved. In our example calculations, even the "core membership" of the amino acids in 4CPV was predicted at a reasonable accuracy. Of course the native structure of 4CPV could not be reproduced from the predicted distances alone, but the method helped the distance geometry algorithm to make educated guesses about unknown distances and build a realistic hydrophobic core. These results indicate that our prediction scheme can play a useful role in protein modelling applications.

Acknowledgements

The authors are grateful to Dr. N. Goldman for his valuable comments and suggestions.

References

[1] G.M. Crippen and T.F. Havel, *Distance Geometry and Molecular Conformation* (Chemometrics Research Studies Press, Wiley, New York, USA, 1988).

[2] W.R. Taylor, J. Mol. Evol. 28 (1988) 161.

[3] M.O. Dayhoff, R.M. Schwartz and B.C. Orcutt, in: *Atlas of Protein Sequence and Structure, Vol. 5, Suppl. 3*, ed. M.O. Dayhoff (Nat. Biomed. Res. Foundation, Washington DC, USA, 1978) p. 345.

[4] W.R. Taylor, Protein Eng. 4 (1991) 853.

[5] M. Levitt, Biochemistry 17 (1978) 4277.

[6] W.R. Taylor, Protein Eng. 6 (1993) 593.

[7] B.W. Lindgren, *Statistical Theory* (Macmillan, New York, USA, 1976).

[8] D.W. Marquardt, SIAM J. Appl. Math. 11 (1963) 431.

[9] P. Valkó and S. Vajda, *Müszaki-tudományos feladatok megoldása személyi számítógéppel* (in Hungarian) (Müszaki Könyvkiadó, Budapest, Hungary, 1987).

[10] L.M. Gregoret and F.E. Cohen, J. Mol. Biol 219 (1991) 109.

[11] A. Aszódi and W.R. Taylor, Biopolymers 34 (1994) 489.

[12] I.N. Bronshtein and K.A. Semendyayev, *Handbook of Mathematics* (Verlag Harri Deutsch, Frankfurt am Main, Germany, 1985).

[13] W.R. Taylor, J.M. Thornton and W.G. Turnell, J. Mol. Graphics 1 (1983) 30.

[14] A. Aszódi and W.R. Taylor, Protein Eng. 7 (1994) 633.

[15] W.H. Press, B.P. Flannery, S.A. Teukolsky and W.T. Vetterling, *Numerical Recipes: The Art of Scientific Computing* (Cambridge University Press, Cambridge, UK, 1992).

[16] G. Mason, Nature 217 (1968) 733.

[17] J. Edelman, Biopolymers 32 (1992) 3.

[18] G.D. Scott, Nature 194 (1992) 956.