Protein Engineering vol.10 no.11 pp.1241–1248, 1997

Protein distance constraints predicted by neural networks and probability density functions

O.Lund[1,3], K.Frimand[1], J.Gorodkin[1], H.Bohr[1], J.Bohr[2], J.Hansen[1] and S.Brunak[1]

[1]Center for Biological Sequence Analysis and [2]Department of Physics, The Technical University of Denmark, DK-2800 Lyngby, Denmark

[3]To whom correspondence should be addressed

We predict interatomic C^{α} distances by two independent data driven methods. The first method uses statistically derived probability distributions of the pairwise distance between two amino acids, whilst the latter method consists of a neural network prediction approach equipped with windows taking the context of the two residues into account. These two methods are used to predict whether distances in independent test sets were above or below given thresholds. We investigate which distance thresholds produce the most information-rich constraints and, in turn, the optimal performance of the two methods. The predictions are based on a data set derived using a new threshold which defines when sequence similarity implies structural similarity. We show that distances in proteins are predicted more accurately by neural networks than by probability density functions. We show that the accuracy of the predictions can be further increased by using sequence profiles. A threading method based on the predicted distances is presented. A homepage with software, predictions and data related to this paper is available at http://www.cbs.dtu.dk/services/CPHmodels/.

Keywords: distance prediction/neural network/pair density function/protein structure/threading

Introduction

It is widely believed that the tertiary structure of proteins is determined by the primary structure (Anfinsen, 1973). Determination of tertiary protein structure from the sequence can be seen as consisting of two tasks: (i) the definition of an energy or cost function that gives the native conformation a lower energy or cost than all other conformations; (ii) the development of algorithms that, given such a cost or energy function, can find the correct conformation. Both the precision of the currently used potentials and the computer time needed to simulate protein folding are at present bottlenecks for *ab initio* calculation of protein structure (Karplus and Petsko, 1990; Elofsson *et al*., 1995).

Two main types of potentials have been applied to evaluate the 'nativeness' of protein conformations: classical empirical potentials such as CHARMM (Brooks *et al*., 1983) and pair potentials based on the distribution of distances in proteins (Tanaka and Scheraga, 1976; Sippl, 1990). For some proteins the CHARMM potential failed to distinguish between correctly and incorrectly folded protein models (Novotny *et al*., 1984). This prompted a pursuit for alternative energy measures (Kocher *et al*., 1994). However, a genetic algorithm could find conformations with lower cost than the native structure, when using pair potentials as the cost function (Elofsson *et al*., 1995). These observations indicate that the quality of the potentials is a highly problematic part of the protein structure prediction problem today.

If a sequence similar protein with known structure exists, homology modeling is probably still the most powerful method for determining the approximate structure of a protein from its sequence (Blundell *et al*., 1987; Moismann *et al*., 1995). A similar sequence with a known structure can be found for approximately one out of seven of the newly determined sequences (Bork *et al*., 1992). Loops and insertions are still difficult to model and often no improvement is made in relation to the initial model when the sequence identity is in the order of 30% or less (Moismann *et al*., 1995).

Many methods have been proposed for predicting the structure from sequences for which no significantly similar sequence with known structure exists (Eisenhaber *et al*., 1995). One popular technique is that of threading a sequence through a structure (Novotný *et al*., 1984; Hendlich *et al*., 1990; Bowie *et al*., 1991; Jones *et al*., 1992; Miyazawa and Jernigan, 1996). A public 'blind' test has shown that the threading methods, in some cases, can lead to the correct conformation (Lemer *et al*., 1995). However, these methods can only be applied if a similar protein structure is known.

A general method would be to generate distance constraints and subsequently use these in an algorithm that computes the folded structure. Interatomic distances in proteins can be predicted by methods using the distribution of distances in proteins with known structures (Tanaka and Scheraga, 1976; Wako and Scheraga, 1982a; Miyazawa and Jernigan, 1985; Sippl, 1990; Maiorov and Crippen, 1992; Grossman *et al*., 1995; Huang *et al*., 1995; Mirny and Shaknovich, 1996), or using correlated mutations (Göbel *et al*., 1994; Shindyalov *et al*., 1994; Taylor and Hatrick, 1994). Recently, a superior performance was reported from using a combination of the two (Thomas *et al*., 1996), and a combination of correlated mutations with other sources of sequence information (Olmea and Valencia, 1997). Another approach has been to predict β-sheet tertiary structure (Lifson and Sander, 1980; Kikuchi *et al*., 1988; Hubbard, 1994). Estimated distances based on statistical studies of protein structures have been used to determine the approximate structure of proteins (Wako and Scheraga, 1982b; Yčas, 1990; Wako and Kubota, 1991; Seitoh *et al*., 1993; Monge *et al*., 1994; Mumenthaler and Braun, 1995; Aszódi *et al*., 1995; Skolnick *et al*., 1997). These methods, based on distance distributions, have been the most successful means of obtaining protein structures from sequences with little similarity to sequences for which the structure is known. These methods, however, do not take into account the sequence context around the amino acid pairs. Neural networks, using a string of amino acids as input, have proven highly successful in the prediction of secondary structure in proteins (Rost and Sander, 1995). Neural networks trained on homologous sequences have previously been used

© Oxford University Press

to predict distances between amino acids and these predictions have in turn been used to determine the structure of proteins (Bohr *et al*., 1990, 1993; Reczko and Bohr, 1994). Neural networks have also been used to predict β-strand contact patterns (Krogh and Riis, 1996) and to represent empirical protein potentials (Grossman *et al*., 1995).

We have described how distance intervals suitable for structure prediction may be defined (Reese *et al*., 1996), and also how the structure of proteins may be derived from a limited set of constraints (Lund *et al*., 1996). Here we define a threshold above which sequence similarity implies structural similarity and analyze the effect of using different alignment methods, matrices and gap penalties. This threshold is then used to extract a set of non-sequence similar protein chains from the Brookhaven Protein Data Bank (PDB) (Bernstein *et al*., 1977). We divide this set into a training set, which we use to develop methods for predicting distances in proteins, and a test set with which we evaluate the performance of the methods.

We compare the predictions of distance inequalities made from two data driven methods: artificial neural networks and pair density functions. The predictions of distance inequalities in proteins presented here gave a leading edge performance. The predicted distance inequalities might enhance the performance of methods like threading, *ab initio* folding and homology modeling.

Materials and methods

Protein structure data

Two data sets were extracted from the Brookhaven Protein Data Bank, release 76 containing 4432 entries. Set I was extracted in order to establish a threshold above which sequence similarity implies structural similarity. This threshold was used to generate Set II: a low similarity data set used to develop and validate methods for predicting distances in proteins. Entries were excluded from Set I if (i) they were not determined by X-ray diffraction (796 entries), since no commonly accepted measure of quality is available for NMR or theoretical model structures. (ii) The secondary structure of the proteins could not be assigned by the program DSSP (Kabsch and Sander, 1983) (765 entries), since we wanted to use the DSSP assignment to quantify the secondary structure identity in the pairwise alignments. (iii) The proteins had any physical chain breaks (defined as neighboring amino acids in the sequence having C^{α}-distances exceeding 4.0 Å (732 entries). (iv) They had a resolution greater than 1.8 Å (3466 entries), since resolutions better than this enable the crystallographer to remove most errors from the model. Exclusion of the above data gave 795 entries with 1035 chains of high quality. Of these, chains with a length of less than 30 amino acids were also discarded (93 chains). The final Set I consisted of 942 chains.

To generate a set of non-sequence similar protein chains (Set II) we extracted a new basic set of data from PDB. In this selection we did not apply the same strict criteria for inclusion as in Set I, for reasons of statistics. We accepted resolutions up to 2.5 Å (658 entries discarded) and structures determined by NMR, but excluded entries if the DSSP program detected chain breaks or incomplete backbones (658 entries) leaving us with 4319 chains. A representative set with low pairwise sequence similarity was selected by running algorithm #1 of Hobohm *et al*. (1992). The sequences were aligned using the local alignment program, *ssearch* (Myers and Miller, 1988;

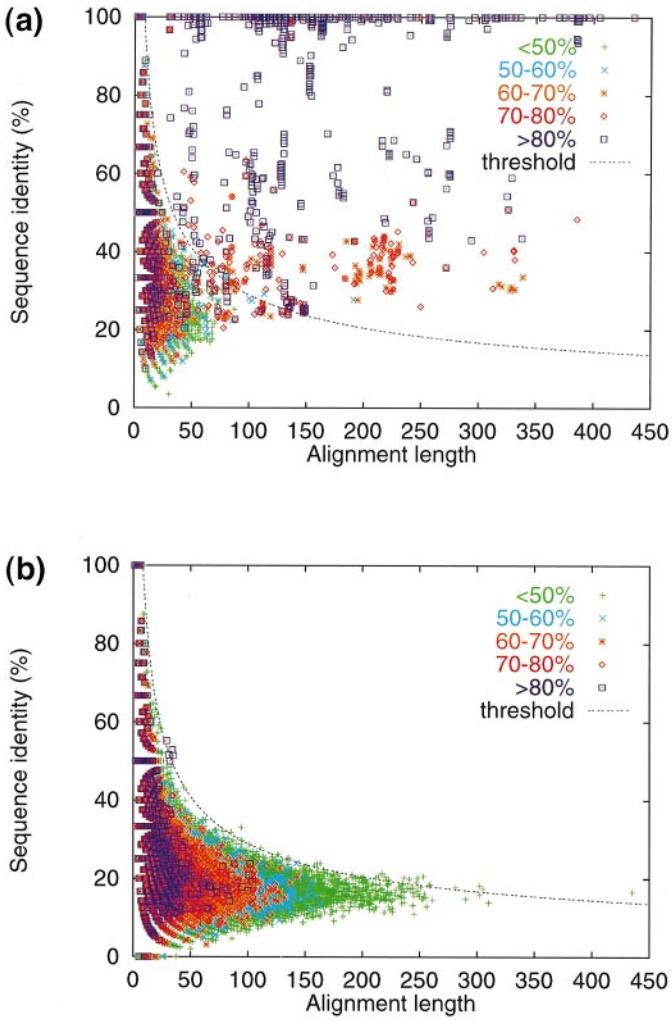

Fig. 1. DSSP secondary structure identity in alignments as a function of the alignment length and the percent sequence identity. In these calculations we used a pam120 matrix with opening gap penalty $f = -12$ and gap elongation penalty $g = -4$. (**a**) The 942 sequences of Set I. (**b**) The order of the amino acids in the sequences in Set I has been shuffled.

Pearson, 1990) using the pam120 amino acid substitution matrix (Dayhoff *et al*., 1978), with gap penalties –12, –4. As a cutoff for sequence similarity we applied a threshold for when sequence similarity implies structure similarity (see below). Finally, we obtained Set II, consisting of 525 distinct protein chains containing 105 773 amino acids to be employed for the statistical examination and prediction algorithm development. The chains in Set II were divided randomly into a training set of 420 chains and a test set of 105 chains. This was done in such a way that the distribution of sequence lengths were approximately equal in the two sets.

Derivation of a sequence similarity threshold

We aligned the 942 sequences in Set I, all against all, and evaluated the percentage of DSSP (Kabsch and Sander, 1983) secondary structure identity as a function of the alignment length L and the sequence identity I_{seq} (Figure 1). Most of the alignments had either a short length or a low percentage of sequence identity (Figure 1a). In only a small fraction of the alignments a combination of a high percentage of sequence identity and a long alignment length was seen. In almost all of these alignments there was a high percentage of secondary structure identity. When the order of the amino acids in each

of the sequences were shuffled no alignments could be found with a combination of a high percentage identity and a long alignment length (Figure 1b). This shows that alignments with a combination of long length and high percentage of sequence identity implies structural similarity, whilst alignments with either of the two may be found in alignments of shuffled sequences. Like Sander and Schneider (1991), we divided the alignments into two groups depending on whether their secondary structure identity in the alignments was above or below 70%. We chose threshold curves of the form $I_{seq} = K/\sqrt{L}$, where K is a variable to be optimized. This functional form fitted well to the boundary of the area of alignments of shuffled sequences (Figure 1b). To obtain the optimal threshold curve we determined, as a function of K, the number of alignments above a given threshold curve which had a secondary structure identity above 70% (true positives), and the number of alignments above a given threshold curve which had secondary structure identity below 70% (false positives).

In this study we used substitution matrices from the pam series (Dayhoff *et al*., 1978). The pam20, pam120 and pam250 matrices were taken from the fasta package, and the pam350 was taken from the clustalW package (Thompson *et al*., 1994) and changed into the fasta matrix format. We also used the blosum50 matrix (Henikoff and Henikoff, 1992) from the fasta package and identity matrices, either taken directly from the fasta package, or modified from the fasta package to obtain identity matrices with different diagonal and off-diagonal substitution scores. The alignments were performed using both the rigorous Smith–Waterman algorithm (Smith and Waterman, 1981) implemented in the *ssearch* program as well as the *fasta* program (Pearson and Lipman, 1988; Pearson, 1990). The *ssearch* program found more true positives as a function of the number of false positives than the *fasta* program and was thus found to perform significantly better.

The decision on which alignment matrix to choose depends on the acceptable fraction of errors. When accepting less than 3% false positives we found that the number of true positives was maximized when using the pam120 matrix with gap penalties –12, –4 together with the threshold curve $I_{seq} = 290/\sqrt{L}$ as shown in Figure 1a. An almost identical performance was obtained using the alignment score $A_{sco} = 60$ as threshold.

The secondary structure identity was also calculated using a three state secondary structure assignment rather than the eight state assignment assigned by the DSSP program. This was done by maintaining the helix (H) and extended sheet (E) assignments and converting all other assignments to coil (C). Such assignment is identical to the one used by Sander and Schneider (1991), (Schneider, personal communication). Using these re-assignments, only five alignments above the threshold curve had a secondary structure identity of less than 70%. Note that Figure 1 was made using the full eight state DSSP secondary assignments, and that more than five alignments according to this assignment scheme have a secondary structure identity of less than 70%.

In order to establish the correspondence between sequence identity and structural identity, we also calculated the root mean square (r.m.s.) of distances of C^{α} atoms of equivalent amino acids in the alignments. The r.m.s. of distances for M pairs of amino acids is defined as (Wako and Kubota, 1991)

$$\left[\sum_{i=1}^{M} \sum_{j=1}^{i-1} \frac{(d_{1,ij} - d_{2,ij})^2}{((M^2 - M)/2)} \right]^{1/2} \qquad (1)$$

where $d_{1,ij}$ and $d_{2,ij}$ are the distances between the C^{α} atoms of the corresponding amino acids i and j in the first and the second sequence, respectively. The relation between $\sqrt{L}I_{seq}$ and the r.m.s. of distances is shown in Figure 2. The points to the right of the vertical dotted line represent alignments with a sequence similarity above our threshold ($I_{seq} = 290/\sqrt{L}$), and the points above the horizontal dotted line represent alignments in which the r.m.s. of distances is more than 2.5 Å. Less than 2% of the alignments which had a similarity above the threshold had an r.m.s. of distances of more than 2.5 Å. This confirms the above results using secondary structure assignments that most of the alignments with a similarity above our threshold are structurally similar.

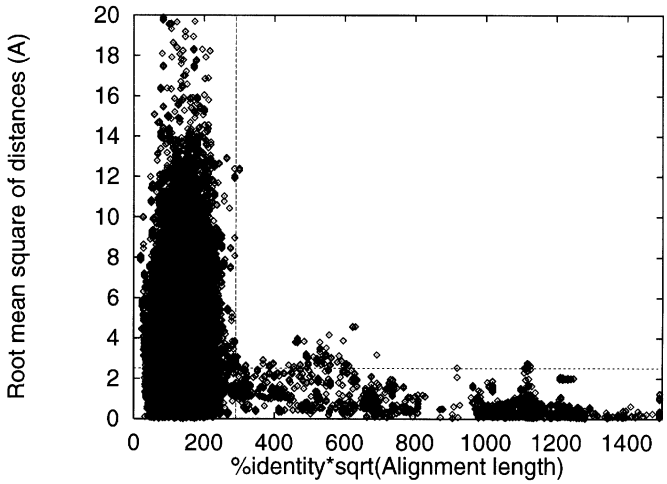

Fig. 2. Root mean square of distances of equivalent C^{α} atoms in the alignments of 942 sequences as a function of $\sqrt{L}I_{seq}$. The vertical line corresponds to the sequence-similarity-implies-structural-similarity threshold $\sqrt{L}I_{seq} = 290$ and the horizontal line is at 2.5 Å.

Mean distances between amino acids in proteins

If proteins are assumed to be spherical and the amino acids are randomly distributed in the sphere the mean distance between two amino acids can be calculated as

$$2 \int_0^R \int_0^q \frac{3r^2}{R^3} \frac{3q^2}{R^3} \left[\int_0^{\pi} \frac{\sin\phi}{2} \sqrt{(q - r\cos\phi)^2 + (r\sin\phi)^2}\, d\phi \right] dr dq = \tfrac{36}{35} R, \qquad (2)$$

where R is the radius of the sphere. The factors $3r^2/R^3$ and $3q^2/R^3$ are the derivatives of the probabilities for finding a point within spheres of radius r and q, respectively. The two outer integrals sum over all pairs of points within a sphere of radius R. The expression inside the square brackets is the mean distance between a point on the spheric shell with radius r and a point on the spheric shell with radius $q > r$. If it is assumed that an amino acid on average occupies a volume of $V_a = 161$ Å^3 (Creighton, 1984), then the mean distance d_m between two amino acids in an L amino acid long protein chain is

$$\tfrac{36}{35} R = \tfrac{36}{35} \sqrt[3]{\frac{3V_a L}{4\pi}} \approx 3.47 \sqrt[3]{L} \text{Å} \qquad (3)$$

If proteins are assumed to be shaped like rods with a length $3.8L$Å, the mean distance between amino acids should be

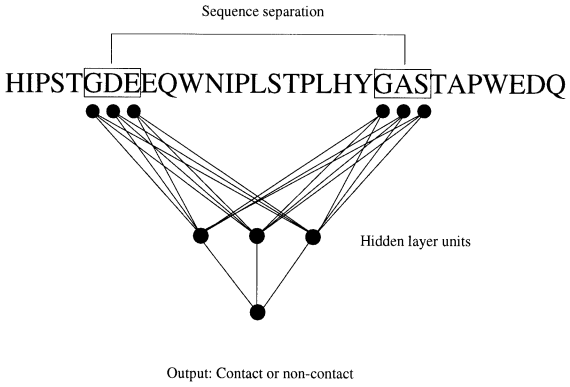

Fig. 3. Schematic drawing of the neural network architecture used.

$$\frac{3.8}{3} L\text{Å} \approx 1.27\, L\text{Å}. \quad (4)$$

We also estimated the mean distance d_m^N between amino acids separated by a given number of amino acids in the sequence (in the following referred to as the sequence separation). If we assume that the N amino acids, after a given amino acid, form a sphere with volume $V_a N$ Å^3, and the Nth amino acid is placed at maximal separation within this sphere, then the mean distance between two amino acids separated by N amino acids is $2\sqrt[3]{(3V_aN)/(4\pi)} = 6.75\sqrt[3]{N}$Å.

Prediction of distances

We have used pair density functions and artificial neural networks to predict whether distances in the test set were above or below a given distance threshold. The pair density functions were used to predict distances in the test set in the following way: for each of the 400 types of amino acid pairs *ab* at a given sequence separation N, we counted the number of distances F_N^{ab} in the training set above the threshold and the distances C_N^{ab} below the threshold. If F_N^{ab} was larger than C_N^{ab}, then distances in the test set between the amino acid pair *ab*, at sequence separation N, were predicted to be larger than threshold (non-contact), otherwise the distance was predicted to be lower than the threshold (contact). The distance between a particular pair of amino acids, at a given sequence separation, is thus always predicted with the same outcome.

We used standard neural networks without hidden units or with one layer of hidden units and adjusted the weights by conventional back propagation (Rummelhart, 1986). For details of the implementation of neural networks to analyze biological sequences see for example Brunak *et al.* (1991). The main novel feature of the neural network architecture was the two window input layer. In the schematic illustration shown in Figure 3 the distance is predicted between D (in GDE) and A (in GAS), which have the sequence separation 16. The two symmetric windows both have a size of three amino acids.

Evaluation of results

The Mathews correlation coefficient C (Mathews, 1975) was used to evaluate the performance of the networks and the pair density functions

$$C = \frac{P_x N_x - N_{fx} P_{fx}}{\sqrt{(N_x + N_{fx})(N_x + P_{fx})(P_x + N_{fx})(P_x + P_{fx})}} \quad (5)$$

Here, we use the following notation: P_x: true positive (experimentally contact, predicted contact); N_x: true negative (experimentally non-contact, predicted non-contact); P_{fx}: false positive (experimentally non-contact, predicted contact) and N_{fx}: false negative (experimentally contact, predicted non-contact). $C = 1$ and $C = -1$ correspond to a perfect and a completely wrong prediction, respectively.

The percentage of correct predictions $(P_x + N_x)/(P_x + N_x + P_{fx} + N_{fx})$ is also used.

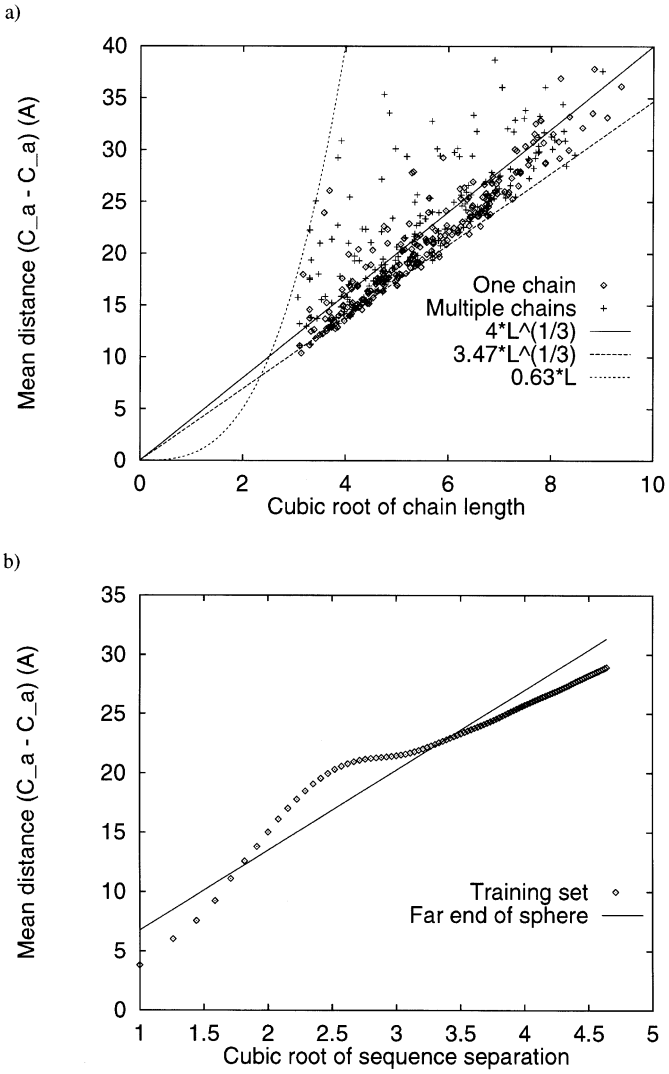

Fig. 4. Distribution of distances in the proteins in our training set. (**a**) Mean distance as a function of the cubic root of the sequence length for chains from PDB entries with one chain; and from PDB entries with multiple chains. (**b**) Mean distance as a function of the cubic root of the sequence separation.

Results

Derivation of distance thresholds

We first studied the distribution of distances between amino acids in the proteins in the training set in order to derive distance thresholds for the predictions. The theoretically derived expression for spherical proteins $3.47\sqrt[3]{L}$Å corresponds well to the lower limit of the mean distances d_m in the proteins of the training set (Figure 4a). This is expected since the spheric form minimizes the average distances between points enclosed in a given volume. The points below the line may correspond to proteins with a closer packing, or with a high fraction of small amino acids.

A least squares fit of the data from the training set to the line $d_m = \alpha\sqrt[3]{L}$ gave the relation $d_m = 3.96\ \sqrt[3]{L}$Å $\approx 4\ \sqrt[3]{L}$Å (see Figure 4a). We will use $4\ \sqrt[3]{L}$Å as one of the thresholds for predicting distances, since this choice for large sequence separations should ensure that approximately as many distances are above as below the threshold. The points far above this line represent the 'rod' shaped proteins, and almost all of these correspond to chains from PDB entries with more than one chain. There are only 12 single chain proteins above a line defined by $d_m = 4.5\ \sqrt[3]{L}$Å. Six of these are protein fragments, three are multimeric in their natural environment, two are metal-binding proteins and one is a calcium binding protein. Thus, it is likely that all these chains are stabilized by other molecules, in their natural environment.

If proteins are assumed to be rod-shaped the mean distance between amino acids should be $1.27L$Å. Figure 4a shows that $0.5\times1.27L = 0.63L$Å is an upper limit to the mean distances between amino acids in the proteins of the training set. All proteins in the training set are thus less than half as long as they could be if they were rod shaped. Note that this line is curved in Figure 4a because the *x*-axis is the cubic root of the chain length.

For short sequence separations we will apply thresholds specific for the sequence separations. The mean distance d_m^N between amino acids with a sequence separation of N scales approximately with the cubic root of the sequence separation (Figure 4b). The points on this curve represent averages for all 420 proteins in the training set. Although the relation $d_m^N = 6.75\ \sqrt[3]{N}$Å derived earlier fits these data reasonably well, we will use the mean distances derived directly from the training set as thresholds.

Predictions of distances

We first evaluated the ability of the neural networks and the pair density functions to predict whether distances in the independent test set were larger or smaller than the mean distance d_m^N in the training set for a given sequence separation N. Pair potentials could correctly predict whether distances were larger or smaller than d_m^N in at least 54.9% of the test examples, depending on the separation between the amino acids in the sequence (Figure 5). The correlation coefficients reached a maximum of only 0.21 for these predictions. We evaluated the performance of neural networks with input window sizes from 2 up to 46. For short sequence separations, we found that the optimal window size was 18 (i.e. two input windows each of width nine amino acids). The windows were centered around each of the two amino acids between which the distance was to be predicted. This is not surprising, since window sizes of 9–13 are good for secondary structure predictions. Using neural networks with a window size of 18, and five hidden units, more than 57.4% of the distances were predicted correctly and correlation coefficients of up to 0.42 were obtained. For sequence separations 2–100 the neural networks had on average correlation coefficients which were more than twice as large as those from pair density functions.

A random prediction of the constraints will be 50% correct on average. The number of distances N_d in the test sets varied between 11 437 and 20 684 depending on the sequence separation. For each of the sequence separations tested, the number of correct predictions N_c was more than eight standard deviations above 50% (assuming that the number of correct predictions follows a Poisson distribution: $N_c - N_d/2 > 8\sqrt{(N_c + N_d/2)/2}$). Calculation of the correlation between the predicted constraints and the correct constraints using Chi-square statistics (Press *et al.*, 1992) yielded p values less than 10^{-15} for all sequence separations. Prediction of distance inequalities in proteins is a very difficult task, and the obtainable correlation coefficients may seem small. However, the predictions presented here are significantly better than random.

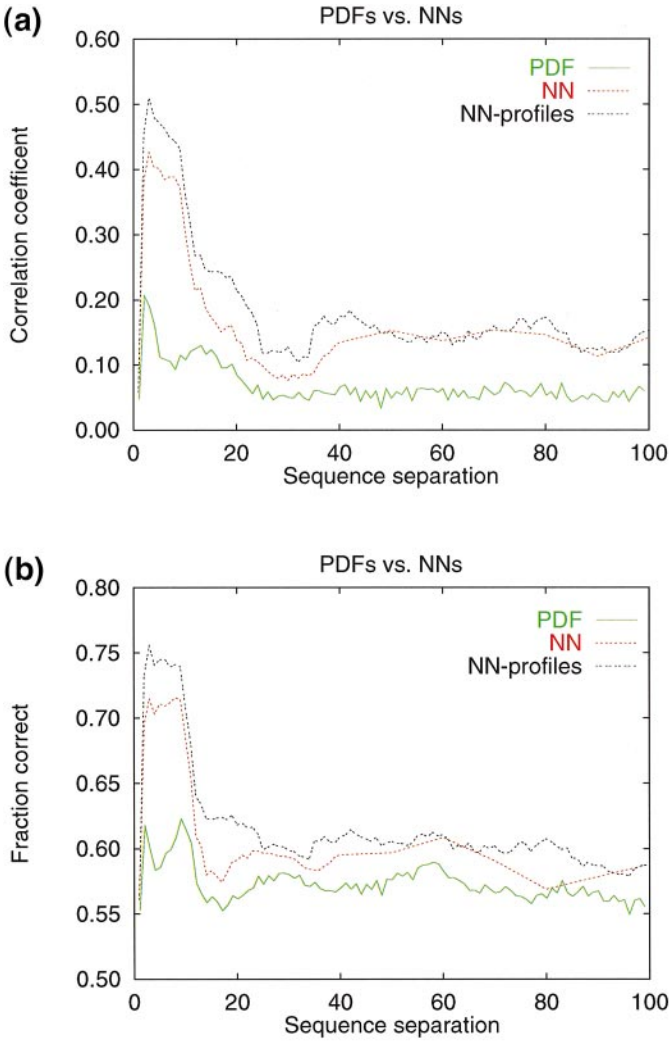

Fig. 5. Performance of neural networks with 18 amino acids in the input windows and five hidden units, (NNs) and the pair density functions (PDFs) as a function of the sequence separation. The mean distance at the given sequence separation in the training set was used as thresholds. Performance when using NN predictions on sequence profiles is also shown. (**a**) Correlation coefficients. (**b**) Fraction of distances predicted correctly.

For relatively large sequence separations we found that neural networks with 30 amino acid input windows were optimal. When using 30 amino acids in the input window, the correlation coefficients and the fraction of correctly predicted distances of the networks declined for small sequence separations, relative to neural networks with a window size of 18, but were markedly better for sequence separations 10–50. For larger sequence separations there was no significant difference in the performance of networks with 18 and 30 amino acids in the input windows. For sequence separations 2–100 the neural networks with 30 amino acids in the input window on average predicted 3.9% more of the distances correctly than did pair density functions. Using neural networks without

hidden layers led to a decrease in correlation coefficients and ten hidden units did not lead to improvement.

To evaluate the effect of using different distance thresholds, we compared the performance of pair density functions and neural networks using the fixed thresholds of 5.8, 7.1, 9.5 and 11.0 Å (Reese *et al.*, 1996). For sequence separations larger than 7, the use of any of these thresholds led to a much poorer performance compared with the use of the sequence separation specific thresholds d_m^N. For these sequence separations most of the decrease in the performance could be avoided by using two additional fixed thresholds of 14.5 and 21.9 Å. The latter threshold equals the average of the mean distances d_m for all the proteins in the training set.

For sequence separations of 5 and 6, the mean distances in the training set were 11.1 and 12.6 Å, respectively. For these sequence separations the threshold of 11.0 Å (Reese *et al.*, 1996) gave approximately the same performance.

For sequence separations of 2, 3 and 4 the thresholds corresponding to the mean distances in the training set (6.03, 7.56 and 9.25 Å, respectively) performed significantly better than the thresholds of 5.8, 7.1 and 9.5 Å. The similarity of these two sets of thresholds shows that the predictability for small sequence separations is very sensitive to the choice of thresholds. This is most likely because the distance distributions for small sequence separations are very narrow.

We proceeded to use the pair density functions and the neural networks to predict whether distances in the test set were above or below $4\sqrt[3]{L}$Å. This threshold was chosen because it corresponds to the estimated mean distance in protein chains with a chain length of L. Using such a threshold thus ensures that there are approximately as many distances above as below the threshold for all large sequence separations N. The pair density functions could predict whether distances were above or below this threshold with correlation coefficients up to 0.074, and predicted the distances correctly in more than 52% of the examples in all test sets only. Using neural networks the distances could be predicted with correlation coefficients of up to 0.2. For sequence separations between 50 and 100, the protein length dependent threshold $4\sqrt[3]{L}$Å led on average to better correlation coefficients than the one based on the average distance for a given sequence separation d_m^N.

The results above show that the optimal threshold is a function of the sequence separation. For all the described thresholds the neural networks performed better than the pair potentials, both in terms of the correlation coefficients and in terms of the fraction of predictions which were correct.

We also investigated if the performance of the algorithm could be enhanced by training and testing one neural network using data from several sequence separations rather than using a fixed separation. Training one network on sequence separations of 48 to 52 and testing on the corresponding test sets gave a 5% higher correlation coefficient than using only sets corresponding to a sequence separation of 50.

Predictions for whole proteins

We subsequently set up a program to predict distance inequalities for whole proteins. Individual neural networks trained on sequence separations 2–20 were used to predict inequalities for these separations. For sequence separation intervals 21–24, 25–34, 35–44, . . . and 95+ the inequalities were predicted by neural networks trained on sequence separations 20, 30, 40 . . . and 100, respectively. All networks had 18 amino acids in the input windows and five hidden units. This program predicted 55.6% of the inequalities in the test set correctly, and the average correlation coefficient for the 105 proteins in the test set was 0.174.

In order to assess quantitatively the significance of the neural network output, we used the training set to establish a relation between the neural network output o and the probability p that two amino acids are closer than a given distance threshold. For each sequence separation, the network predictions were used to construct a table relating o and p, using bin sizes of 0.01 for o. Using this table to convert the neural network output to probabilities, 61.7% of the inequalities in the test set were predicted correctly, and the average correlation coefficient for the 105 proteins in the test set was 0.204.

To test if the overall performance was indeed reliable for sequences not used in the training or testing of the networks, we extracted 131 new sequences from the latest release 79 of PDB. All 208 sequences had sub-threshold sequence similarity to the 525 sequences in the training and test sets. The simple non-profile based scheme predicted 61.7% of the inequalities in this set correctly, and the average correlation coefficient for the 208 proteins was 0.189. This confirmed the reliability of the evaluation of the performance. Thirty-nine of the 208 new sequences belong, according to the SCOP (Murzin *et al.*, 1995) classification of protein structures to fold classes not present in our set of 525 proteins used for training and testing of the proteins. The percentage of correct predictions in this subset was 59.8% and the average correlation coefficient was 0.200. This shows that the performance of our algorithm is also sustained on proteins with no homology with the proteins used to develop the algorithm.

It has been reported that the accuracy of prediction schemes can be dramatically increased when using profiles of aligned sequences (Gribskov *et al.*, 1987; Rost and Sander, 1993) as input rather than single sequences. To test this we aligned each of the test sequences using *fasta* with default parameters against Swiss-Prot. All sequences reported by *fasta* with expectation values less than 0.01 were included in a profile. Sequences fulfilling this criterion were subsequently aligned to the query sequence using the program *align* from the *fasta* package. Regions corresponding to gaps in the query sequence were removed and regions with gaps in the database sequence were replaced with the corresponding amino acids from the query sequence. On the average there were 17 sequences in each profile. We calculated the probability of contact as an average of the predictions on the sequences in the profile. In this way 63.2% of the inequalities in test set were predicted correctly, and the average correlation coefficient for the 105 proteins was 0.233. The largest increase in performance was observed for sequence separation 17, where the percentage point of correctly predicted inequalities increased by 5 and the correlation coefficient increased by 0.09 (see Figure 5). These improvements are comparable to those found when using profiles for secondary structure prediction (Rost and Sander, 1993).

Threading

One application of the predicted distances is to align a query sequence against all entries in a database of structures with the aim of finding which structure the sequence is most likely to adopt (threading). For this task, we define a score which is large if the predicted distances between a residue in the query sequence and other residues in this sequence are similar to the

distances between a residue in a database structure and other residues in the structure. More specifically, we define a score $S_{m,n}$ for aligning residue m in the query sequence with a residue n in the database sequence. The score is defined as a sum over log odds ratios for sequence separations up to l,

$$S_{m,n} = \sum_{k \in (-l,-l,+1\ldots,-2,2,3\ldots l)} \begin{cases} \log(\frac{q_{m,m+k}}{p_s}) & \text{for } d_{n,n+k} = 1 \\ \log(\frac{1-q_{m,m+k}}{1-p_s}) & \text{for } d_{n,n+k} = 0 \end{cases} \quad (6)$$

$q_{i,j}$ is the predicted probability of a contact between residue i and j. $d_{i,j}$ is unity if residues i and j in the database sequence are in contact and zero otherwise. p_s is the probability for contact between any pair of amino acids at that sequence separation. In the following we have set $p_s = 0.5$.

Each of the sequences in the test set were used as a query sequence and threaded against a database consisting of all the sequences in the test set, using the score defined above with $l = 20$. In these alignments, gaps were only allowed at the ends of the database sequences, and these gaps were unpenalized. Sixty-seven of the 105 sequences (63.8%) could find their own structure using the non-profile based score. This compares favorably with the 58% previously reported for a C^{α} atom score based on pair density functions (Kocher *et al*., 1994).

We also evaluated the ability of our method to find the approximate structure from the sequences. For this task we used the SCOP database (Murzin *et al*., 1995) which classifies proteins according to their structure. Proteins in the same SCOP family have a clear evolutionary relationship. A superfamily is defined as a set of families between which there is low sequence similarity, but where structural and functional features suggest a common evolutionary origin. The 35 test set sequences, which had another sequence belonging to the same SCOP superfamily in the test set, were extracted for this task. We subsequently threaded each of these sequences against the other 34 and found the top scoring one among these. The success of the threading was defined as the number of top scoring sequences which is in the same SCOP superfamily as the query sequence. Using a global alignment algorithm (Needleman and Wunsch, 1970) with opening gap penalty –12 and elongation penalty –4, we could find the correct superfamily for 13 of the 35 sequences using our profile based score. Using a Blosum50 substitution matrix together with the same alignment procedure, 14 of the sequences were identified correctly. The performance of our threading method is thus roughly equal to that found by a conventional alignment method. For this task, it has been reported that a potential based on predicted secondary structure has a performance that is approximately 35% worse than a normal amino acid substitution matrix (Fischer and Eisenberg, 1996). Using a 50/50 combination of our potential with a Blosum50 matrix, we found the correct superfamily for 15 of the sequences.

Discussion

Using carefully prepared data, we defined a sequence-similarity-implies-structural-similarity threshold and used this threshold to generate a non-sequence similar set of proteins. The main result in this paper could be summarized by stating that neural networks trained and tested on this data were better at predicting distances than a method based on pair density functions, and that the predicted distances can be used in a threading algorithm.

It is well established that a large sequence similarity implies structural similarity (Chothia and Lesk, 1986; Sander and Schneider, 1991), but the quantitative aspects are still not fully understood. In contrast to the study of Sander and Schneider (1991) we found that for long alignment overlaps, the size of the safe region, where sequence similarity implies structural similarity, still depends on the alignment length. In their study, the similarity threshold was constant at 25% identical amino acids when the alignment length L was larger than 80 amino acids. We demonstrate that the safe region is confined by the non-constant identity equivalent to $290/\sqrt{L}$. With this criterion, a sequence similarity below 25% will for sufficiently long alignments still imply structural similarity. The boundary of the safe region was shown to correspond well to the maximal similarity found by alignments of shuffled sequences.

We found that the best alignment matrix for the task of finding the maximal number of alignments with a secondary structure identity of more than 70% was the pam120 matrix. It has been argued on theoretical grounds that a matrix with an entropy similar to that of pam120 should be optimal for database searches (Altshul, 1991). This corresponds well to our finding that identity matrices with an entropy close to that of pam120 also performed well. Johnson and Overington (1993) compared a number of matrices in order to find which one was most suitable for sequence comparisons. They found the pam120 matrix to be among the ones exhibiting better performances in their study. In an assessment of alignment matrices the blosum50 matrix performed better than the pam120 matrix (Vogt *et al*., 1995). The reason behind this discrepancy may be that the entropy of the blosum50 matrix was more suited for their particular test scheme. For the task of finding signal peptide cleavage sites, for example, matrices with an entropy of approximately 3.0 [a (6,–6) identity matrix and pam20 matrix] were found to be the optimal choice (Nielsen *et al*., 1996).

When using neural networks and pair density functions to predict distances in proteins we found that, for sequence separations up to 50 amino acids, the best results were obtained using thresholds equal to the mean of all distances at that sequence separation in the training set. A good performance could also be obtained with our previously defined thresholds (Reese *et al*., 1996) supplemented by two additional thresholds of 14.5 and 21.9 Å. For larger sequence separations the best results were obtained with the protein chain length dependent threshold $4\sqrt[3]{L}$Å. This indicates that, for large sequence separations, distances are governed more by the size of the protein than by the sequence separation. The performance of the neural networks could be further increased by training and testing simultaneously on data from several sequence separations.

The best distance threshold is the one which results in most information. The information is maximized if there are an equal number of distances above and below the threshold and the optimum threshold is therefore given by the median of the distance distribution. Due to the approximate symmetry of the distributions of distances between amino acids in proteins the mean and the median of the distributions are close to each other and for practical reasons the mean may be used instead of the median. It is a common belief that, for protein structure determination, knowing that two amino acids are close to each other is more useful than knowing that amino acids are far apart. However, if constraints between all pairs of amino acids are considered, the optimal distance threshold for determination of protein structure is close to the mean (Bohr *et al*., 1993).

Predicting constraints with 100% accuracy may not be necessary in order to determine the structure of proteins. The success of alignment and threading algorithms, for example, is not based on the unambiguous recognition of a small number of matches, but on the significance of the summed score over all aligned positions.

It has recently been shown that a combination of pair potentials and correlated mutations is better at predicting contacts in proteins than pair potentials alone (Thomas *et al*., 1996). We also found that the performance of our method could be further enhanced by predicting on sequence profiles (Gribskov *et al*., 1987; Rost and Sander, 1993) rather than on single sequences.

The distances predicted by neural networks can potentially be used to construct improved potentials that can enhance the performance of threading and protein folding algorithms. They may also find use in homology modeling of loops and insertions/deletions.

Acknowledgements

The authors are indebted to David Ussery for his careful and critical review of the work. This work was supported by The Danish 1991 Pharmacy Foundation and The Danish National Research Foundation.

References

Altshul,S.F. (1991) *J. Mol. Biol*., **219**, 555–565.
Anfinsen,C. (1973) *Science*, **181**, 223–230.
Aszódi,A., Gradwell,M.J. and Taylor,W.R. (1995) *J. Mol. Biol*., **251**, 308–326.
Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F.,Jr, Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol*., **112**, 535–542.
Blundell,T.L., Sibanda,B.L., Sternberg,M.J. and Thornton,J.M. (1987) *Nature*, **326**, 347–352.
Bohr,H., Bohr,J., Brunak,S., Cotterill,R.M.C., Fredholm,H., Lautrup,B. and Petersen,S.B. (1990) *FEBS Lett*., **261**, 43–46.
Bohr,J., Bohr,H., Brunak,S., Cotterill,R., Fredholm,H., Lautrup,B. and Petersen,S.B. (1993) *J. Mol. Biol*., **231**, 861–869.
Bork,P., Ouzounis,C., Sander,C., Scharf,M., Schneider,R. and Sonnhammer,E. (1992) *Nature*, **358**, 287.
Bowie,J., Luthy,R. and Eisenberg,D. (1991) *Science*, **253**, 164–170.
Brooks,B.R., Burccoleri,R.E., Olafson,B.D., States,D.J., Swaminathan,S. and Karplus,M. (1983) *J. Comput Chem*., **4**, 187–217.
Brunak,S., Engelbrecht,J. and Knudsen,S. (1991) *J. Mol. Biol*., **220**, 49–65.
Chothia,C. and Lesk,A.M. (1986) *EMBO J*., **5**, 823–826.
Creighton,T.E. (1984) *Proteins*. W.H.Freeman and Company, New York.
Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) *Atlas of Protein Sequence and Structure*, 5, Suppl. 3, pp. 345–352.
Eisenhaber,F., Persson,B. and Argos,P. (1995) *Crit. Rev. Biochem. Mol. Biol*., **30**, 1–94.
Elofsson,A., Le Grand,S.M. and Eisenberg,D. (1995), *Proteins*, **23**, 73–82.
Fischer,D. and Eisenberg,D. (1996) *Protein Sci*., **5**, 947–955.
Göbel,U., Sander,C., Schneider,R. and Valencia,A. (1994) *Proteins*, **18**, 309–317.
Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
Grossman,T., Farber,R. and Lapedes,A. (1995), *Ismb*, **3**, 154–161.
Hendlich,M., Lackner,P., Weitckus,S., Flöckner,H., Froschauer,R., Gottsbacher,K., Casari,G. and Sippl,M.J. (1990) *J. Mol. Biol*., **216**, 167–180.
Henikoff,S. and Henikoff,J.G. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
Hobohm,U., Scharf,M., Schneider,R. and Sander,C. (1992) *Protein Sci*., **1**, 409–417.
Huang,E.S., Subbiah,S. and Levitt,M. (1995) *J. Mol. Biol*., **252**, 709–720.
Hubbard,T. (1994) In Lathrop,R.H. (ed.) *Proceedings of the Biotechnology Computing Track, Protein Structure Prediction MiniTrack of the 27th HICSS*. IEEE Computer Society Press, pp. 336–354.
Johnson,M.S. and Overington,J.P. (1993) *J. Mol. Biol*., **233**, 716–738.
Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992), *Nature*, **358**, 86–89.
Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.
Karplus,M. and Petsko,G.A. (1990) *Nature*, **347**, 631–634.
Kikuchi,T., Nemethy,G. and Scheraga,H.A. (1988) *J. Protein Chem*., **7**, 473–490.
Kocher,J.-P., Rooman,M. and Wodak,S. (1994) *J. Mol. Biol*., **235**, 1598–1613.
Krogh,A. and Riis,S.K. (1996) In Tourelzky,D.S., Mozer,M.C. and Hasselmo,M.E. (eds) *Advances in Neural Information Processing Systems 8*, MIT Press, in press.
Lemer,C.M.-R., Rooman,M.J. and Wodak,S.J. (1995) *Proteins*, **23**, 337–355.
Lifson,S. and Sanders,S. (1980) *J. Mol. Biol*., **139**, 627–639.
Lund,O., Hansen,J.E., Brunak,S. and Bohr,J. (1996) *Protein Sci*., **5**, 2217–2225.
Maiorov,V.N. and Crippen,G.M. (1992) *J. Mol. Biol*., **227**, 876–888.
Mathews,B.W. (1975) *Biochim. Biophys. Acta*, **405**, 442–451.
Mirny,L.A. and Shaknovich,E.I. (1996) *J. Mol. Biol*., **264**, 1164–1179.
Miyazawa,S. and Jernigan,R.L. (1985) *Macromolecules*, **18**, 534–552.
Miyazawa,S. and Jernigan,R.L. (1996) *J. Mol. Biol*., **256**, 623–644.
Moismann,S., Meleshko,R. and James,M.N.G. (1995) *Proteins*, **23**, 301–217.
Monge,A., Friesner,R.A. and Honig,B. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 5027–5029.
Mumenthaler,C. and Braun,W. (1995) *Protein Sci*., **4**, 863–871.
Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol*., **247**, 536–540.
Myers,E.W. and Miller,W. (1988) *Comput. Applic. Biosci*., **4**, 11–17.
Needleman,S.B. and Wunsch,C.D. (1970) *J. Mol. Biol*., **48**, 443–453.
Nielsen,H., Engelbrecht,J., von Heijne,G. and Brunak,S. (1996) *Proteins*, **24**, 165–177.
Novotný,J., Bruccoleri,R. and Karplus,M. (1984) *J. Mol. Biol*., **177**, 787–818.
Olmea,O. and Valencia,A. (1997) *Folding Design*, **2**, S25–S32.
Pearson,W.R. (1990) *Methods Enzymol*., **183**, 63–98.
Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
Press,W., Teukolsky,S., Vetterling,W. and Flannery,B. (1992) *Numerical Recipes*. University Press, Cambridge.
Reczko,M. and Bohr,H. (1994) In Bohr,H. and Brunak,S. (eds) *Protein Structure by Distance Analysis*. IOS Press, Amsterdam, pp. 87–97.
Reese,M.G., Lund,O., Bohr,J., Bohr,H., Hansen,J.E. and Brunak,S. (1996) *Protein Engng*, **9**, 733–740.
Rost,B. and Sander,C. (1993) *J. Mol. Biol*., **232**, 584–599.
Rost,B. and Sander,C. (1995) *Proteins*, **23**, 295–300.
Rummelhart,D.E. and McClelland,J.L. (1986) *Parallel Distributed Processing*. MIT Press, Boston.
Sander,C. and Schneider,R. (1991) *Proteins*, **9**, 56–68.
Seitoh,S., Nakai,T. and Nishikawa,K. (1993) *Proteins*, **15**, 191–204.
Shindyalov,I.N., Kolchanov,N.A. and Sander,C. (1994) *Protein Engng*, **7**, 349–358.
Sippl,M.J. (1990) *J. Mol. Biol*., **213**, 859–883.
Skolnick,J., Kolinski,A. and Ortiz,A.R. (1997) *J. Mol. Biol*., **265**, 217–241.
Smith,T.F. and Waterman,M.S. (1981) *J. Mol. Biol*., **147**, 195–197.
Tanaka,S. and Scheraga,H.A. (1976) *Macromolecules*, **9**, 945–950.
Taylor,W.R. and Hatrick,K. (1994) *Protein Engng*, **7**, 341–348.
Thomas,D.J., Casari,C. and Sander,C. (1996) *Protein Engng*, **9**, 941–948.
Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) *Nucleic Acids Res*., **22**, 4673–4680.
Vogt,G., Etzold,T. and Argos,P. (1995) *J. Mol. Biol*., **249**, 816–831.
Wako,H. and Kubota,Y. (1991) *J. Protein Chem*., **10**, 233–243.
Wako,H. and Scheraga,H.A. (1982a) *J. Protein Chem*., **1**, 5–45.
Wako,H. and Scheraga,H.A. (1982b) *J. Protein Chem*., **1**, 85–117.
Yčas,M. (1990) *J. Protein Chem*., **9**, 177–200.

Received January 27, 1997; revised July 22, 1997; accepted July 28, 1997