ELSEVIER

FEBS Letters 344 (1994) 147–153

FEBS 13981

FEBS LETTERS

Pattern recognition and self-correcting distance geometry calculations applied to myohemerythrin

G. Hänggi, W. Braun*

Institut für Molekularbiologie und Biophysik, Eidgenössische Technische Hochschule – Hönggerberg, CH-8093 Zürich, Switzerland

Received 24 February 1994; revised version received 25 March 1994

Abstract

A topological list, consisting of segments of regular secondary structures and a list of buried and solvent accessible residues, is automatically predicted from multiple aligned sequences in a protein family. This topological list is translated into geometric constraints for distance geometry calculation in torsion angle space. A new self-correcting distance geometry method detects and eliminates false distance constraints. In an application to the four-helix bundle protein, myohem-erythrin, the right-handed global fold was correctly reproduced with a root-mean-square deviation of 2.6 Å, when the topological list was derived from the X-ray structure. A predicted topological list, coupled with constraints from the residues in the active site of myohemerythrin, predicted the correct fold with a root-mean-square deviation of 4 Å for backbone atoms.

Key words: Distance geometry; Helix bundle; Multiple sequence alignment; Pattern recognition; Tertiary structure prediction

1. Introduction

Multiple aligned sequences in a family of homologous proteins have been used in the past to improve the prediction of secondary structures in proteins [1–5]. The predictions achieved with these methods are significantly better than predictions based on single sequences [6,7], but still contain errors. We want to combine those predictions with three-dimensional structure calculations where methods and experiences are quite limited [8–10]. In this paper we describe a new approach and report on the first experience with the four-helix bundle protein, myohemerythrin. A new software tool MULTAN detects significant amino acid sequence motifs in a set of aligned sequences. This information is then automatically translated into distance and dihedral angle constraints to calculate an ensemble of three-dimensional structures by a distance geometry method in torsion angle space. The constraints set is refined by using geometry calculations in a self-correcting way.

The variable target function method in torsion angle space [11,12] is a frequently used tool to calculate three-dimensional structures of proteins from nuclear magnetic resonance data [13–16]. Protein structures are calculated with standard bond length and bond angle geometry [17] in a hard sphere model for all atoms from given distance and dihedral angle constraints. The variable target function method has also been used for predicting the tertiary structure of the DNA binding domain of the transcriptional transactivator c-myb [9]. The input for this calculation was prepared by manual inspection of a series of homologous proteins, whereas in our approach the input data are generated automatically.

We show the convergence property of the self-correcting distance geometry method in an application to the four-helix bundle protein, myohemerythrin. Starting from 50 random structures of this protein, the self-correcting distance geometry calculations converged to the correct right-handed topology of the four helices, when the constraints were derived from the secondary structure prediction, the inside-outside prediction and active site constraints.

2. Materials and methods

2.1. Generation of the topological list

The basic steps of our procedure are outlined in Fig. 1. We applied our approach to the four α-helix bundle protein myohemerythrin from *Themiste zostericola*, where a high resolution X-ray structure [18] is available to test the accuracy of the prediction. Homologous protein sequences were identified in the Swissprot and PIR sequence data banks and aligned by the PILEUP tool (GCG software package [19]) using the similarity matrix of Risler et al., [20]. The set consisted of 12 amino acid sequences with 38% to 69% amino acid identity to the sequence of *Themiste zostericola*.

Simplified Chou-Fasman rules [21] were implemented in MULTAN to predict helical segments from the aligned sequences. Subgroups of amino acids were defined as α-helix forming residues H = {A, L, I, E, Q, K, F, W, M} and turn forming residues as T = {G, P, S, D, N, K, Y, C}, as the conformational parameters for these residues exceed 1.0 in these secondary structures [22]. At each column in the multiple aligned sequences the number of α-helix forming residues H and of turn forming residues T is then counted. To be of significance, these numbers must be greater than the theoretical expectation values for the subgroups given the sequence similarities q_k.

* Corresponding author. Fax: (41) (1) 371 4873.

SSDI 0014-5793(94)00366-4

The theoretical expectation values for the individual amino acid residues can be calculated from the sequence similarities q_k, in percent of identical amino acids, of the homologous sequences k = 2,..,13 relative to the first sequence of *Themiste zostericola* (q_1 = 1) as follows:

$$\langle N_A \rangle = \sum_{k=1}^{M} q_k \quad (1)$$

$$\langle N_A \rangle = m_{BA} \sum_{k=1}^{M} (1 - q_k) \quad (2)$$

A and B represent two amino acids. Eq. **(1)** applies if the amino acid A occurs at the current position in the first sequence. Eq. **(2)** applies, if a different amino acid B is at the current position in the first sequence. The probability m_{BA} for substituting the amino acid B with the amino acid A is taken from Risler et al. [20].

The validity of equations **(1)** and **(2)** can be seen by a recursive procedure. If we have only two homologous sequences (k = 2) with sequence similarity q_2 of the second sequence relative to our first reference sequence (q_1 = 1), and we assume that the first sequence has a certain amino acid A at a given position, then the probabilities $p^{(2)}(N_A = 0)$, $p^{(2)}(N_A = 1)$ and $p^{(2)}(N_A = 2)$ to find zero, one or two occurrences of amino acid A in the two sequences at this position are:

$$p^{(2)}(N_A = 0) = 0 \quad (3)$$

$$p^{(2)}(N_A = 1) = 1 - q_2 \quad (4)$$

$$p^{(2)}(N_A = 2) = q_2 \quad (5)$$

The theoretical expectation value for the occurrence of amino acid A in the two sequences is therefore

$$\langle N_A \rangle^{(2)} = \sum_{m=0}^{2} p^{(2)}(N_A = m) \times m = 1 + q_2 \quad (6)$$

The probabilities $p^{(k+1)}(N_A = m)$ for k+1 sequences can be derived from the probabilities $p^{(k)}(N_A = m)$

$$p^{(k+1)}(N_A = m) = p^{(k)}(N_A = m)(1 - q_{k+1}) + p^{(k)}(N_A = m - 1)q_{k+1} \quad (7)$$

From this equation follows:

$$\langle N_A \rangle^{(k+1)} = \langle N_A \rangle^{(k)} + q_{k+1} \quad (8)$$

Equations **(6)** and **(8)** prove equation **(1)**. If a different amino acid B is in the first sequence by counting the occurrence of A, we have to use $m_{BA}(1-q_k)$ as the probability to find A in the sequences k = 2,3,.... Therefore we have to apply equation **(2)** in that case.

The expectation value for a subgroup of amino acids is calculated by the sum of the expectation values of individual amino acids. If the helix forming subgroup occurs more often in the aligned sequences than the theoretical expectation value indicates, the helix group occurs significantly at this position. To predict a helix, the helix forming group must significantly occur in three out of five contiguous positions and for the turns in four contiguous positions.

The prediction of residues to be buried or solvent exposed follows the same strategy. From the statistical study of Hubbard and Blundell [23] we define two subgroups of residues, potentially being buried, i = {C, M, I, V, L, W, F}, or potentially solvent exposed o = {K, R, E, N, T, S, Q, P, D}. At each position in the sequence the number of i residues and o residues are then counted in the aligned sequences and compared to the theoretical expectation values for the two subgroups i and o, given the sequence similarities q_k of the sequences k relative to the first sequence. If amino acids of the groups i or o occur more often than the theoretical expectation values of their groups at a certain position, this position is predicted as being buried or solvent exposed.

2.2. Translation to distance geometry input constraints

The backbone dihedral angles in a α-helical segment are constrained by $-58 < \varphi < -56$ and $-48 < \psi < -46$. An upper bound of 15 Å for the distance between reference atoms for side chains of all residues which have been included in the 'inside' list, are used. These reference atoms are Q^{α} for Gly, Q^{β} for Ala, C^{β} for Ser, Asn, Asp, Thr and Cys, C^{γ} for Pro, Gln, Glu, Met, Trp and His, $C^{\gamma 1}$ for Ile, QQG for Val, C^{δ} for Lys and Arg, QQD for Leu, and QR for Phe and Tyr. The reference atoms Q are pseudo atoms as used by distance geometry calculations from NMR data [24]. Between all residues of the 'outside' list, separated in the sequence by more than 10 residues, a lower bound of 15 Å for the distance between the reference atoms are used.

2.3. Distance geometry calculations

Distance geometry calculations in torsion angles [11] are done with the program DIAMOD, a modified version of DIANA [12]. A new target function and an automatic detection of inconsistent upper and lower bounds have been included. The target function used in DIAMOD has been constructed such that large violations of upper and lower bounds do not dominate the non-linear fit procedure. We have chosen for the contribution TF_i of a violated constraint b_i to the target function the form

$$TF_i = \sqrt{1 + v_i^2} - 1 \quad (9)$$

$$v_i = \frac{d_i^2 - b_i^2}{b_i^2} \quad (10)$$

Equations **(9)** and **(10)** apply for a violated upper bound, i.e. $d_i > b_i$. Otherwise the contribution is zero. Analogous expressions are used for the lower bound. The contributions from the dihedral angles and the van der Waals repulsion are treated as usual with distance geometry calculations in torsion angles [24]. Calculations are done including all heavy atoms of myohemerythrin.

A second new feature in DIAMOD is an automatic identification method for inconsistent constraints. An ensemble of 50 structures is calculated, each starting from a random structure. Within good final structures, i.e. final structures with a target function less than a cut-off, probable inconsistent constraints are automatically identified if they are violated in more than 50% of the structures by more than 1 Å. We assume that if upper bounds are violated in the majority of final structures with low target function values, these constraints cannot be fulfilled in combination with all the other constraints. So the actual distances in the correct final structures must be larger. Inconsistent upper bounds are used in a next cycle as lower bounds and vice versa. After each cycle the original constraint data set as generated by TRANSLATE is automatically checked for inconsistency by the calculated structures. When we discard instead of exchange the inconsistent constraints in test calculations, we observed inferior convergence behaviour. In practice we found that four to six cycles are sufficient to considerably improve the constraints as well as the structures by the exchange procedure.

2.4. Generating data sets X1, X2 and X3 with topological information from the X-ray structure

First, we wanted to know to what extent the X-ray structure can be reproduced with a correct topological list. Therefore, we assigned the four α-helical segments, Glu-19 to Arg-37, Ala-41 to Ala-64, Val-71 to Ile-81 and Ala-93 to Lys-108 as in the X-ray structure. Residues were considered to be solvent exposed, if the accessible surface area of the side chain atoms in the X-ray structure exceeds 50% of the average value in the tripeptide Gly-X-Gly in an ensemble of 30 random conformations for residue type X. Residues with an accessible surface area of less than 20% of the average value were considered to be buried. Accessible surface areas were calculated with the program ANAREA [25].

Although both the helical segments and the inside/outside list, were extracted from the X-ray structure, the generated constraints were not completely consistent with the X-ray structure due to the uniform constraints of 15 Å. Therefore a corrected constraint set X1 was generated by exchanging those upper and lower bounds which are not compatible with the X-ray structure. The data set X2 contains no such corrections. Both data sets have the same dihedral angle constraints (144) but different upper and lower bounds. In X1, there are 295 upper and 377 lower bounds and in X2, there are 406 upper and 266 lower bounds.

In the third data set, X3, we added to X2 further constraints from

the active site residues of myohemerythrin. The active centre consists of two Fe atoms octahedral co-ordinated to the residues His-25, -54, -73, -77, and -106 and Glu-58 and Asp-111. These residues are conserved in the 13 sequences, except for Glu-58, where in 3 sequences the sequence Gln-58–Glu-59 prevails. The detailed atomic three-dimensional structure of the Fe cluster was not used. We included distance constraints between all potential Fe-binding atoms of these 7 Fe-binding residues as a uniform upper limit of 7.2 Å and a uniform lower limit of 2.8 Å.

2.5. *Data sets P1,P2,P3 and P4 derived from the predicted topological list*

In the data set P1 the inside and outside residues were predicted by MULTAN (Table 1) and the helical segments were assigned as in the X-ray structure. Data set P1 contains 231 upper and 351 lower limit distance constraints. The same constraints and the addition of the constraints for the Fe cluster leads to data set P2. In the data set P3 constraints were generated from the predicted helical segments and the predicted inside/outside residues. In data set P4 the constraints for the Fe cluster were added to the constraints from P3. In the distance geometry calculations with P4, the active site constraints were weighted with a factor of 50 relative to the other constraints. The number of upper constraints are 253 for P3 and 287 for P4, the number of lower constraints are 272 and 306, respectively.

3. Results and discussion

3.1. *Prediction of the topology list by MULTAN*

The predictions of the helical segments, the buried and solvent accessible residues for myohemerythrin of *Themiste zostericola* are listed in Table 1. Six helical segments E24-I36, N43-V48, T52-S68, H73-S88, V96-V104 and H106-F122 are predicted. These segments compare fairly well with the four helical segments of the X-ray structure [18]. The secondary structure prediction of Chou-Fasman [22] and Garnier-Osguthorpe-Robson [26] as implemented in the GCG software package [19] are included for comparison. The results given by MULTAN are quite encouraging, especially if one considers that we did not optimize the parameters for the prediction rules.

The buried residues are also quite successfully predicted. If an inside residue is predicted, the residue at that position or the neighbouring residue is indeed buried in

Table 1
Prediction of the topology of myohemerythrin with the program MULTAN

```
      sequence GWEIPEPYVWDESFRVFYEQLDEEHKKIFKGIFDCIRD.NSAPNLATLVKVTTNHFTHEE
Secondary Structure
      alph      H H     H   H  H   H  H H HH  HH  H H   H  H HH     H HH HH
                                          HHHHHHHHHHHHH       HHHHHHH   HHHHHHHH
      turn     T T TTTT  T T    TT  TT  T   TT        TT T  T    T      T
               TTTT
      ---------------------------------------------------------------------------
      Pred.        TTTT              HHHHHHHHHHHHH       HHHHHH   HHHHHHHH
      CF[a]         hhhhhhh BBBBBHHHHHHHHHHHHHHBBBBBBT Tt TTBBBBBBBBttHHHHHH
      GOR[b]          TTT   TT HHHHHHHHHHHHHHHHBBBBBTTT T    BBBBBBB     HHHH
      x-ray[c]      TTTTT TTTTT   HHHHHHHHHHHHHHHHHHHHT TTHHHHHHHHHHHHHHHHHH
Inside
      Pred.     i i     i   i  i   i      ii  ii ii        i  i  i    ii
      x-ray[d]    i  i i   i i    ii  i  ii ii ii         i  i   i  ii  i
Outside
      Pred.      o ooo   ooo oo  o  ooo oo  o         oo oo oo oo o o   ooo
      x-ray[d]  ooo  oo o  oo   o oo  o   o  o      oo   oo  oo  o  oo  o
      sequence GWEIPEPYVWDESFRVFYEQLDEEHKKIFKGIFCDIRD.NSAPNLATLVKVTTNHFTHEE
```

```
      sequence AMM.DAAKYSEVVP.HKKMHKDFLEKIGGLSAPVDAKNVDYCKEWLVNHIKGTDFKYKGKL
Secondary Structure
      alph      HH HH H H      HHH H HH HH  HH        H HH HH   HH H HH H HH
               HHHHHHHHHH    HHHHHHHHHHHHHHHHH       HHHHHHHHH HHHHHHHHHHHH
      turn           T T TT    TT   T     TT TT     T  T TT   T  T  T TTTTT
                                                                      TTTTT
      ---------------------------------------------------------------------------
      Pred.    HHHHHHHHHH    HHHHHHHHHHHHHHHHH       HHHHHHHHH HHHHHHTTTTT
      CF[a]    HHH HHHHHHHHTT HHHHHHHHHHHH   hhhhhhh TTtt  BBBBBB  tt   TT
      GOR[a]   HHH HHHHHHHHHH HHHHHHHHHHTT       TTTTTTHHHHHHH TTTTTTTTTTT
      x-ray[c] HHH TTTTTTTTHH HHHHHHHHHHHTTTTTTTT HHHHHHHHHHHHHHHH  HHHTTTT
Inside
      Pred.     ii                  ii  i  i       i     iii  i  i i     i
      x-ray[d]   i   i i  i   i  i  i   i  i i i   ii  i  iii iii ii  i   i
Outside
      Pred.       o oo    o   oo  o      o  o    o o o  oo   o  oo o o o o
      x-ray[d]  o    o o oo oo   o  oo  o  oo o o ooo  o   o          o  oo
      sequence AMM.DAAKYSEVVP.HKKMHKDFLEKIGGLSAPVDAKNVDYCKEWLVNHIKGTDFKYKGKL
```

[a]Predicted from the single sequence by Chou and Fasman [22] as part of the GCG package [19].
[b]Predicted from the single sequence by the GOR method [26] as part of the GCG package [19].
[c]According the X-ray structure [18].
[d]Determined from solvent accessible areas of the residues in the X-ray structure.

the X-ray structure. Only three residues (W2, F17 and T56) were predicted to be buried which were classified as outside in the X-ray structure. Eight residues were predicted to be outside that are classified as inside in the X-ray structure. The number of errors is still small as compared to the total number of residues. If we include further 34 correct constraints from residues known to be in the active site (data set P4), the self-correcting distance geometry calculations largely eliminated these errors (Fig. 3D and Fig. 4F).

3.2. *Convergence of the self-correcting distance geometry calculations in data set X2*

DIAMOD calculations of 50 structures with the manually corrected data set X1 lead to a narrow bundle of the best 25 structures with a root-mean-square deviation of 1.2 Å of the mean structure to the X-ray structure. A correct data set with uniform constraints of 15Å is therefore sufficient to reproduce the backbone fold quite precisely.

Data set X2, which is not completely consistent with the X-ray structure, was used to check the convergence property of the self-correcting DIAMOD cycles. The convergence of the calculated structures towards the X-ray structure is shown in Fig. 2. After each cycle the 25 best structures of 50 calculated structures are selected and the root-mean-square deviations (rmsd) of all back-

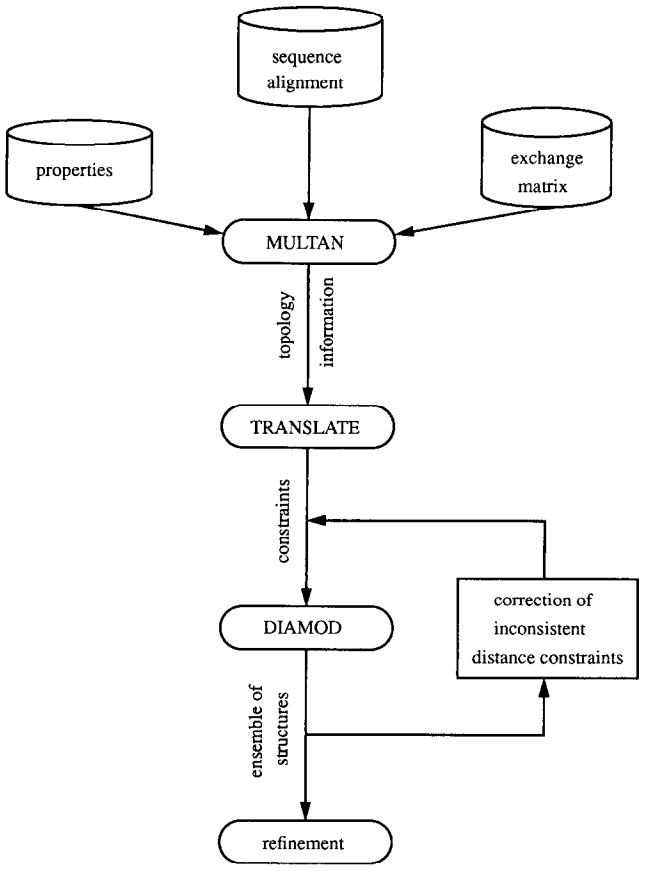

Fig. 1. Flow chart for automatic prediction of tertiary structure of proteins. Boxes and arrows indcate programs and data files, respectively. The multiple sequence alignment was done with the GCG package [19]. The program MULTAN predicts secondary structures and a list of buried and solvent-accessible residues. The program TRANSLATE generated dihedral angle constraints and upper and lower bounds for distances from these predictions. DIAMOD automatically improves the distance constraints set for 3-dimensional consistency. In the last step, the structures can be improved by energy refinement [17,37] or molecular dynamics calculations [38–40].

Table 2
Overview of the distance geometry calculations for the different data sets

Data set	X1	X2	X3
<*m*> to X-ray	1.2	2.6	2.3
av/dev to X-ray	1.8 ± 0.3	3.7 ± 2.0	2.8 ± 0.5
av/dev to <*m*>	1.4 ± 0.3	2.6 ± 2.0	1.7 ± 0.4
pairwise average	2.0 ± 0.4	3.8 ± 2.7	2.4 ± 0.7

Data set	P1	P2	P3	P4
<*m*> to X-ray	3.0	2.7	9.0	4.0
	7.3	7.4		
	11.5	10.6		
av/dev to X-ray	3.5 ± 0.4	3.0 ± 0.3	10.1 ± 1.5	4.6 ± 0.3
	7.3 ± 0.0	7.5 ± 0.2		
	11.5 ± 0.1	10.8 ± 0.6		
av/dev to <*m*>	1.8 ± 0.5	1.4 ± 0.4	4.8 ± 2.3	2.4 ± 0.5
	0.5 ± 0.0	1.4 ± 0.3		
	1.5 ± 0.5	2.1 ± 1.2		
pairwise average	2.6 ± 0.8	2.0 ± 0.5	6.8 ± 3.2	3.4 ± 0.7
	1.0 ± 0.0	2.3 ± 0.5		
	2.2 ± 0.7	3.1 ± 1.8		

Backbone root-mean-square deviations in Å of the 25 best DIAMOD structures. Deviations are calculated including all backbone atoms in the helical segments Glu^{19} to Arg^{37}, Ala^{41} to Ala^{64}, Val^{71} to Ile^{81} and Ala^{93} to Lys^{108}. <*m*> denotes the mean structure of the 25 best structures. The average values and standard deviations of the rmsd values of the individual structures are denoted by av/dev. They are calculated with respect to the mean structure <*m*> or to the X-ray structure, X-ray [18].

bone atoms of the helical segments E19-R37, A41-A64, V71-I81 and A93-K108 are calculated with respect to the X-ray structure. The distribution of these rmsd values are shown in white boxes and for the 10 best structures in hatched boxes. Whereas after the first cycle, i.e. after the calculation with the uncorrected data set, a quite broad distribution is observed, the overwhelming majority of the structures converge towards the X-ray structure within less than 4 Å after the sixth cycle. Thus four cycles might be sufficient in practice to obtain a reasonably good set of structures. In the calculation including additional active site constraints (data set X3) all 25 best structures have an rmsd value less than 5 Å (Fig. 3A).

3.3. *Distance geometry calculations with data sets P1 to P4*

The results of the distance geometry calculations with all data sets are summarized in Table 2. For the data sets

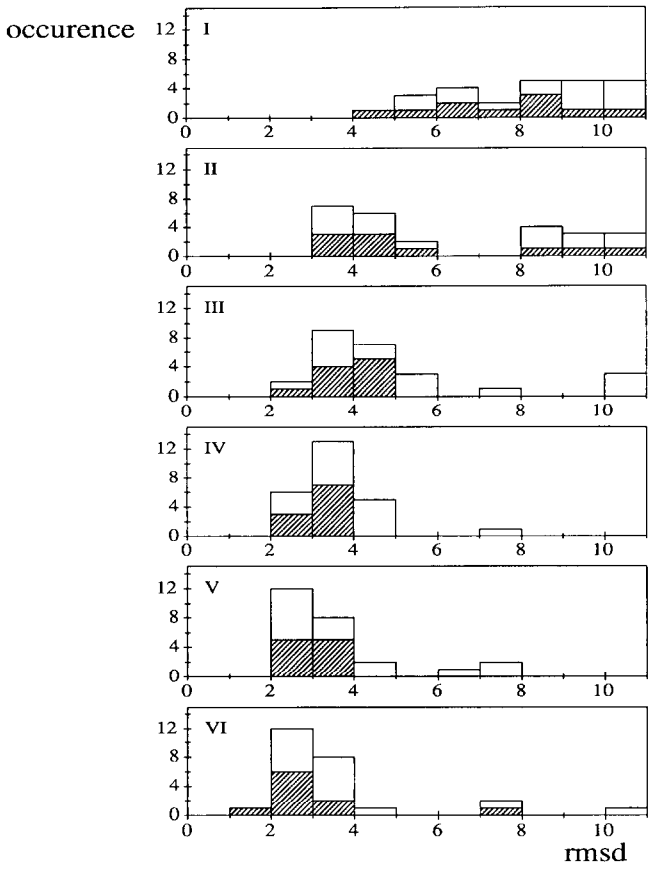

Fig. 2. Convergence of the self-correcting distance geometry calculations for data set X2. The distribution of the backbone root-mean-square deviations (rmsd) for the four helical segments of the 25 best structures (white boxes) and of the 10 best structures (grey boxes) to the X-ray structure of myohemerythrin [18] is given after cycles I to VI. The last column includes all structures with a rmsd greater than 10 Å.

P1 and P2, three groups of structures could be clearly identified in the final structures. One group always contained the correct right-handed topology with rmsd values of the mean structure to the X-ray structure of 3.0 Å and 2.7 Å, respectively. There was no convergence with the completely predicted data set P3. This could also be seen from high final target function values. However, when 34 constraints from residues known to be in the active site, were added (data set P4), all 25 best structures form one group. The mean structure deviates from the X-ray structure by only 4 Å. The distribution of the rmsd values for the data sets P1, P2 and P4 are shown in Fig. 3. Fig. 3A shows as a reference the result from the data set X3. For the data sets P1 and P2 (Fig. 3B,C) three distinct groups of structures were observed. For the data set P4 (Fig. 3D) a unique group was observed.

The backbone fold of some predicted structures from the various data sets are compared to the X-ray structure in Fig. 4. Fig. 4A shows the right-handed up-and-down helix bundle of the X-ray structure with labels for the first helix (E19 to R37). The global fold of the mean structure of all 25 calculated structures of the data set X1 (Fig. 4B) is identical to the X-ray structure. Figs. 4C, 4D and 4E show the differences of the three groups of structures (Table 2 and Fig. 3B) when the secondary structural elements are taken from X-ray structure and the inside/outside residues are predicted by MULTAN (data set P1). The mean structure is calculated for each of the three groups. The three groups correspond to the right-handed (Fig. 4C), right-handed zig-zag (Fig. 4D) and the left-handed helix bundle (Fig. 4E). Fig. 4F shows the mean structure of the 25 best structures of data set P4. This structure has the same right-handed topology as the X-ray structure.

We have shown as a result of the calculation with the data set X2 that the packing of the four helices of myohemerythrin into a right-handed bundle is uniquely determined by the list of correct inside and outside residues. Self correcting distance geometry calculations could reconstruct the correct backbone topology of the four helices with a rmsd value of 2.6 Å. With the predicted list of inside and outside residues, the best 25 structures could be classified in three well defined groups representing the right-handed, the left-handed and the right-handed zig-zag bundle. It is known and confirmed by this calculation that the differences in hydrophobic packing or conformational energies between these different

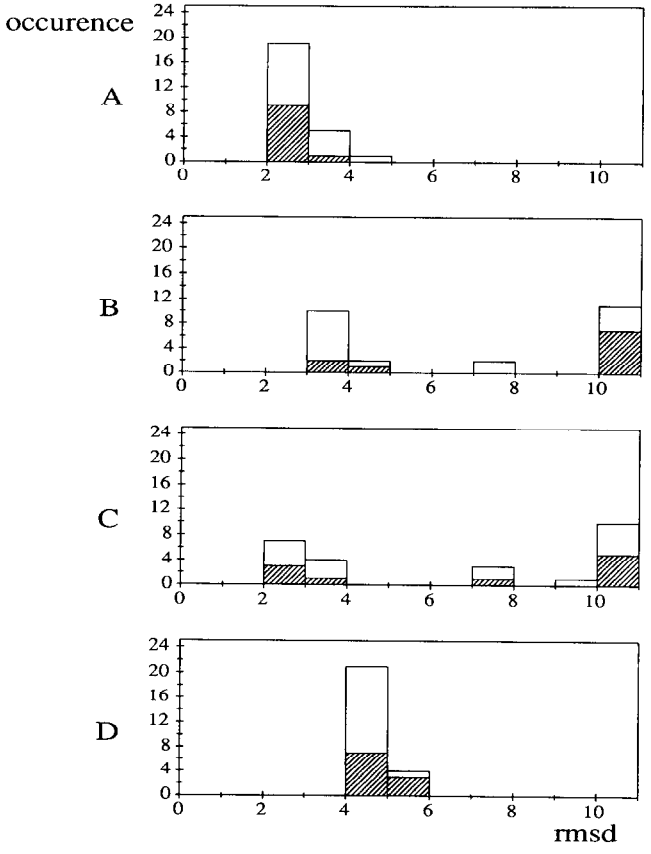

Fig. 3. The distribution of the rmsd deviations of the 25 best structures (white boxes) and 10 best structures (grey boxes) for four different data sets after the final self-correcting cycle. (A) Data set X3. (B) Data set P1.(C) Data set P2. (D) Data set P4.

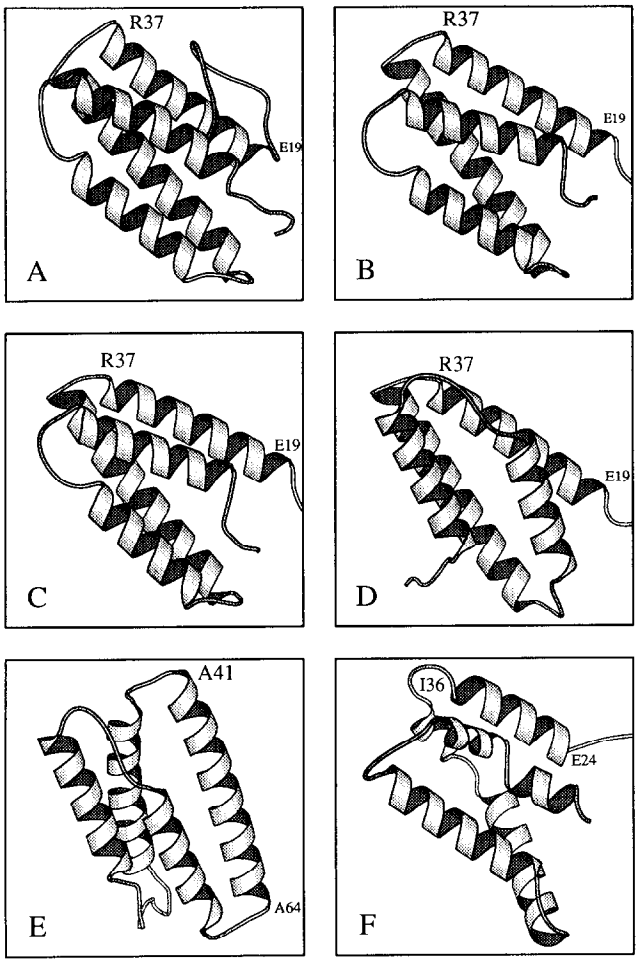

Fig. 4. MOLSCRIPT representation [41] of the backbone structures. (A) The X-ray structure of myohemerythrin [18]. (B) The mean structure of the 25 best structures as calculated from data set X1. (C,D,E) The mean structures of the three groups (see text) from data set P1. (F) The mean structure of the 25 best structures calculated with data set P4.

topologies are quite subtle [8,27,28]. Predicted helical segments, a list of predicted inside and outside residues, and the list of active site residues of the protein myohemerythrin produced a single group of calculated structures with the correct right-handed topology.

Protein modelling by homology has been extensively explored in the past [29]. Basically, it needs the three-dimensional structure of at least one member of the protein family. Recently automated methods have been described for recognizing structural conserved regions and then transferring these conserved structural segments from the known structure to the unknown protein structure by spatial constraints [30–36]. These methods usually achieve quite high accuracy of about 1 Å for the core region of the protein. The spatial constraints applied in these methods are derived from a given tertiary structure and are to a large extent consistent constraints. In our case these constraints are predicted and are thus prone to errors. Therefore one needs a reliable method to detect inconsistent constraints.

We propose a simple self-correcting distance geometry approach to tackle this problem. The self-correcting method needs a minimum amount of correct constraints in the initial data set for convergence. In practice, this information can often be obtained from biochemical studies or preliminary nuclear magnetic resonance studies. We have shown that our approach successfully predicts the correct global fold of the four helix bundle protein, myohemerythrin, with such additional information. Future studies with other protein folds should elucidate the general strengths and limitations of this approach.

Acknowledgements: We thank Rudi Baumann for his technical assistance with the calculations, and Dr. C. Schein for fruitful discussions and carefully reading the manuscript. The computing facilities of the IPS of the ETH Zürich are gratefully acknowledged.

References

[1] Crawford, I.P., Niermann, T. and Kirschner, K. (1987) Proteins 2, 118–129.
[2] Zvelebil, M.J., Barton, G.J., Taylor, W.R. and Sternberg, M.J.E. (1987) J. Mol. Biol. 195, 957–961.
[3] Niermann, T. and Kirschner, K. (1991) Protein Eng. 4, 359–370.
[4] Benner, S.A., Cohen, M.A. and Gerloff, D. (1993) J. Mol. Biol. 229, 295–305.
[5] Gerloff, D.L., Jenny, T.F., Knecht, L.J., Gonnet, G.H. and Benner, S.A. (1993) FEBS Lett. 318, 118–124.
[6] Boscott, P.E., Barton, G.J. and Richards, W.G. (1993) Protein Eng. 6, 261–266.
[7] Rost, B. and Sander, C. (1993) J. Mol. Biol. 232, 584–599.
[8] Cohen, F.E. and Kuntz, I.D. (1989) in: Prediction of Protein Structure and the Principles of Protein Conformation (Fasman, G.D. Ed.) Plenum Press, New York. pp. 647–705.
[9] Frampton, J., Gibson, T.J., Ness, S.A., Doderlein, G., and Graf, T. (1991) Protein Eng. 4, 891–901.
[10] Taylor, W.R. (1993) Protein Eng. 6, 593–604.
[11] Braun, W. and Gō, N. (1985) J. Mol. Biol. 186, 611–626.
[12] Güntert, P., Braun, W. and Wüthrich, K. (1991) J. Mol. Biol. 217, 517–530.
[13] Braun, W., Wagner, G., Wörgötter, E., Vasak, M., Kägi, J.H.R. and Wüthrich, K. (1986) J. Mol. Biol. 187, 125–129.
[14] Kline, A.D., Braun, W. and Wüthrich, K. (1988) J. Mol. Biol. 204, 675–724.
[15] Braun, W. (1991) in: Proc. NATO Workshop: 'Computational Aspects of the Study of Biological Macromolecules by NMR' (Hoch, J. Ed.) Plenum Press, New York.
[16] Widmer, H., Widmer, A. and Braun, W. (1993) J. Biomol. NMR 3, 307–324.
[17] Némethy, G., Pottle, M.S. and Scheraga, H.A. (1983) J. Phys. Chem. 87, 1883–1887.
[18] Sheriff, S., Hendrickson, W.A. and Smith, J.L. (1987) J. Mol. Biol. 197, 273–296.
[19] Devereux, J., Haeberli, P. and Smithies, O. (1984) Nucleic Acid Res. 12, 387–395.
[20] Risler, J.L., Delorme, M.O., Delacroix, H. and Henaut, A. (1988) J. Mol. Biol. 204, 1019–1029.
[21] Fasman, G.D. (1989) Prediction of Protein Structure and the Principles of Protein Conformation, Plenum Press, New York.

[22] Chou, P.Y. and Fasman, G.D. (1978) Adv. Enzymol. 47, 45–148.
[23] Hubbard, T.J.P. and Blundell, T.L. (1987) Protein Eng. 1, 159–171.
[24] Braun, W. (1987) Quart. Rev. Biophys. 19, 115–157.
[25] Richmond, T.J. (1984) J. Mol. Biol. 178, 63–89
[26] Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) J. Mol. Biol. 120, 97–120.
[27] Efimov, A.V. (1982) Mol. Biol. 16, 271–281.
[28] Chou, K.-C., Maggiora, G.M., Némethy, G. and Scheraga, H.A. (1988) Proc. Natl. Acad. Sci. USA 85, 4295–4299.
[29] Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E. and Thornton, J.M. (1987) Nature 326, 347–352.
[30] Greer, J. (1990) Proteins 7, 317–334
[31] Schiffer, C.A., Caldwell, J.W., Kollman, P.A. and Stroud, R.M. (1990) Proteins 8, 30–43.
[32] Havel, T.F. and Snow, M. (1991) J. Mol. Biol. 217, 1–7.
[33] Overington, J., Donnelly, D., Johnson, M.S., Sali, A. and Blundell, T.L. (1992) Protein Sci. 1, 216–226.
[34] Sali, A. and Blundell, T.L. (1993) J. Mol. Biol. 234, 779–815.
[35] Havel, T.F. (1993) Mol. Simul. 10, 175–210.
[36] Srinivasan, S., March, C.J. and Sudarsanam, S. (1993) Protein Sci. 2, 227–289.
[37] von Freyberg, B. and Braun, W. (1993) J. Comp. Chem. 14, 510–521.
[38] Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M. (1983) J. Comp. Chem. 4, 187–217.
[39] van Gunsteren, W.F. and Berendsen, H.J.C., GROMOS Library Manual, Biomos, Groningen, 1987.
[40] Weiner, P.K. and Kollman, P.A. (1981) J. Comp. Chem. 2, 287–303.
[41] Kraulis, P. (1991) J. Appl. Crystallogr. 24, 946–950.