Protein Science (1995), *4*:863–871. Cambridge University Press. Printed in the USA.

Predicting the helix packing of globular proteins by self-correcting distance geometry

CH. MUMENTHALER AND W. BRAUN

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

(RECEIVED January 4, 1995; ACCEPTED February 15, 1995)

Abstract

A new self-correcting distance geometry method for predicting the three-dimensional structure of small globular proteins was assessed with a test set of 8 helical proteins. With the knowledge of the amino acid sequence and the helical segments, our completely automated method calculated the correct backbone topology of six proteins. The accuracy of the predicted structures ranged from 2.3 Å to 3.1 Å for the helical segments compared to the experimentally determined structures. For two proteins, the predicted constraints were not restrictive enough to yield a conclusive prediction. The method can be applied to all small globular proteins, provided the secondary structure is known from NMR analysis or can be predicted with high reliability.

Keywords: DIAMOD; distance geometry in torsion angles; helix bundles; multiple sequence alignment; tertiary structure prediction; variable target function method

The prediction of the three-dimensional structure of a protein from its amino acid sequence is still one of the great unsolved problems in macromolecular structural biology (Cohen & Kunz, 1989). Recently, a new automated approach to this problem was developed and applied to myohemerythrin (Hänggi & Braun, 1994). First, a protein sequence data bank is searched for sequences that are similar to the given one. The multiple aligned sequence family is then screened by the program MULTAN to detect sequence positions that are likely to be buried or solvent exposed. The procedure is based on the observations that the three-dimensional structures of homologous proteins are similar (Chothia & Lesk, 1986) and that polar amino acid residues tend to be solvent exposed more frequently than nonpolar residues (Hubbard & Blundell, 1987). Because the polarity of structurally important residues is highly conserved in the aligned sequences, the preference to be buried or solvent exposed can be detected more easily in a family of aligned sequences (Holbrook et al., 1990; Benner et al., 1994; Donnelly et al., 1994). MULTAN can also predict the secondary structure (Hänggi & Braun, 1994).

The segments of secondary structures and the inside/outside preferences of individual residues can be translated into dihedral angle constraints and distance constraints, respectively, for distance geometry calculations (Havel et al., 1983; Braun, 1987). Distance geometry in torsion angle space (Braun & Gō, 1985; Güntert et al., 1991) is particularly suitable for modeling protein structures because the three-dimensional structures are calculated with standard bond length and bond angle geometry, and the steric hindrance of all atoms is explicitly included. The resulting structures are therefore stereochemically correct. The high efficiency of this computational method has been shown in the calculations of protein structures from NMR data (Braun, 1991).

Distance geometry methods have been used in the past in the modeling of homologous proteins (Havel & Snow, 1991; Šali & Blundell, 1993). In these studies, spatial constraints were derived from a known three-dimensional structure within the family of homologous proteins. In contrast, we deduce our spatial constraints directly from the multiple aligned sequences, and therefore we do not need a known three-dimensional structure. Because our predicted distance constraints contain more errors, we have to detect and eliminate wrong distance constraints during the distance geometry calculations.

We focus our study on the packing of α-helical segments because we do not want to combine it with the question of the accuracy and reliability of secondary structure prediction. Therefore, we assume that the helical segments are known and the backbone dihedral angles of the residues in these segments are constrained to values as found in ideal α-helices. Backbone dihedral angles of the residues in the loops were constrained according to the known distribution maps of the ϕ and ψ angles

Reprint requests to: W. Braun, Institut für Molekularbiologie und Biophysik, ETHZ, CH-8093 Zürich, Switzerland; e-mail: braun@mol.biol.ethz.ch.

of the amino acid residues (Kamimura & Takahashi, 1994). The χ^1 dihedral angles of the residues in the helical segments were constrained to the preferred rotamers of the side chains in the helices (Dunbrack & Karplus, 1994).

Distance constraints were obtained by a statistical study on a set of high-resolution X-ray structures, where we excluded our test proteins. Inconsistent distance constraints were automatically detected by an analysis of residual constraint violations and eliminated during the calculations in an iterative way. This self-correcting algorithm, implemented in the distance geometry program DIAMOD (Hänggi & Braun, 1994), proves to be a powerful tool to accurately predict three-dimensional protein structures.

We assess the strengths and limitations of our method in a test set of eight small α-helical proteins with different topologies. All test proteins consist of a single domain with a well-defined hydrophobic core. Even in the case of only a few helices, very different global folds with good hydrophobic packing and low conformational energies can be generated (Chou et al., 1988; Tufféry & Lavery, 1993; Mumenthaler & Braun, 1995). The correct handedness of helix bundle structures represents a further difficulty in predicting the global fold of these small helical proteins. Our completely automated method calculated the backbone fold of six proteins with the highest accuracy and reliability reached so far without using three-dimensional structures of homologous proteins.

Results

Application to eight α-helical proteins

The three-dimensional structures of our test proteins (Table 1) were previously determined by X-ray diffraction method or NMR spectroscopy. The pheromone Er-10 (Brown et al., 1993), the DNA-binding domain of the c-myb proto-oncogene product (Ogata et al., 1992), and the *Antennapedia* homeodomain (Qian et al., 1989) contain a three-helix packing motif. Myohemerythrin (Sheriff et al., 1987), the J-domain of DnaJ from *Escherichia coli* (Szyperski et al., 1994), the de novo designed protein FELIX (Hecht et al., 1990), and calbindin (Svensson et al., 1992) are all four-helix bundles. The DNA binding domain of the repressor protein from the P434 phage (Mondragon et al., 1989) consists of five helices. The helix-packing topologies of all of these proteins are different, with the exception of c-myb and *Antp* homeodomain. These proteins have, however, only a low (28%) amino acid sequence identity.

In all eight proteins, the majority of residues were correctly predicted to be buried or solvent exposed (Table 1). After the forth DIAMOD cycle we selected the 25 structures with lowest target function values from the final 50 structures. The target function values used in DIAMOD indicate how well the calculated structures fulfill the input constraints (Hänggi & Braun, 1994). For four proteins, Er-10, c-myb, *Antp* homeodomain,

Table 1. *Predicted inside/outside residues, distance and dihedral angle constraints, and RMSDs of the predicted structures from the NMR/X-ray structures*

			Prediction[c]						Constraints						
			Inside residues			Outside residues			Number[d]		% Correct[e]			RMSD (Å)	
Protein name[a]	Length	No. of seq.[b]	T	C	W	T	C	W	Dist.	Angle	Start	End	Clust[f]	p.a.[g]	⟨m⟩[h]
Pheromone Er-10	38	5	10	5	4	9	6	1	114	61	53	62	1	1.5	3.0
C-myb	51	31	11	5	2	17	14	2	173	93	29	52	1	1.3	3.1
Antp homeodomain	68	90	9	7	1	24	14	3	350	133	52	69	1	1.8	2.4
Myohemerythrin	118	7	16	15	0	16	6	3	564	211	56	74	1	1.9	2.3
DnaJ	75	29	8	6	0	21	15	1	265	133	52	59	2	2.6	2.7
FELIX	79	1	15	10	2	21	14	1	505	145	53	65	2	1.7	2.9
Repressor (434)	69	5	10	8	1	19	7	2	301	113	73	77	–	7.3	–
Calbindin	75	13	11	10	0	24	12	3	488	141	68	67	–	7.2	–

[a] PDB codes of the protein structures are: 1ERP (Er-10), 1POM (C-myb), 1HOM (*Antp*), 2MHR (myohemerythrin), 1FLX (FELIX), 1R69 (434 repressor), and 4ICB (calbindin). With the exception of C-myb and FELIX, these are NMR/X-ray structures used as reference structures in this work. 1POM (C-myb) is a model structure known to have small deviations from the NMR structure (Murthy, 1993). FELIX is a protein that was artificially designed to have a left-handed 4 α-helical topology. Even though no NMR or X-ray structure is available yet, chemical evidence suggests that the protein does indeed have the predicted fold (Hecht et al., 1990).

[b] Number of homologous sequences used for the prediction.

[c] Total number (T) of residues predicted to be inside or outside; number of residues correctly (C) and wrongly (W) predicted.

[d] Number of distance and angle constraints for use in distance geometry calculations.

[e] Percentage of correct distance constraints before and after the DIAMOD calculations using the self-correcting distance constraint algorithm. Angle constraints usually were correct within a tolerance range of ±30°.

[f] Number of clusters among the 25 final DIAMOD structures with lowest target function values.

[g] Average pairwise RMSD value among the 10 structures with lowest target function values.

[h] RMSD of the predicted structure to the reference NMR/X-ray structure obtained by superimposing all backbone atoms of the helical regions. At the end of cycle IV, the structures were sorted with respect to their target function values. The 10 with lowest rank were then selected. The predicted structures are the mean structures of these 10. The helical regions are: L2–L8, E12–C19, and K24–W32 for Er-10; E8–L21, W25–L31, and D37–S46 for C-myb; R11–H22, R29–L39, and E43–E60 for *Antp* homeodomain; E19–R37, A41–A64, V71–I84, and A93–K108 for myohemerythrin; Y6–V12, E18–K31, and K41–L57 for DnaJ; E3–L18, E23–I35, A42–T57, and Q63–H70 for FELIX; S1–Q12, N16–Q22, Q28–N36, L45–A51, and V56–N61 for the 434 repressor protein; S2–K16, S24–F36, T45–D54, and S62–Q75 for calbindin.

and myohemerythrin, the self-correcting distance geometry calculations converged to a unique and well-defined cluster of correct structures. For two proteins, DnaJ and FELIX, two clusters could be found within the 25 selected structures. In both cases, the structures in the correct clusters had significantly lower final target function values. All 25 structures of the six proteins were sorted according to their final target function values. As a representative predicted structure, we chose the mean structure of the 10 best ranking structures. The RMS deviation (RMSD) values for all backbone atoms in the helices of these predicted structures are about 3 Å compared to experimentally determined structures (Table 1).

In Figure 1 we show the distribution of the RMSD values of all 25 individual structures for the six proteins Er-10, C-myb, *Antp* homeodomain, myohemerythrin, DnaJ, and FELIX. Almost all structures have RMSD values below 4 Å. In the case of Er-10, there are no outliers (Fig. 1A). Only for the protein FELIX did a substantial fraction of structures have a clearly wrong fold, a right-handed helix bundle, with RMSD values above 10 Å. For all six proteins the 10 structures with lowest target function values had the correct fold, as illustrated by the hatched area in Figure 1.

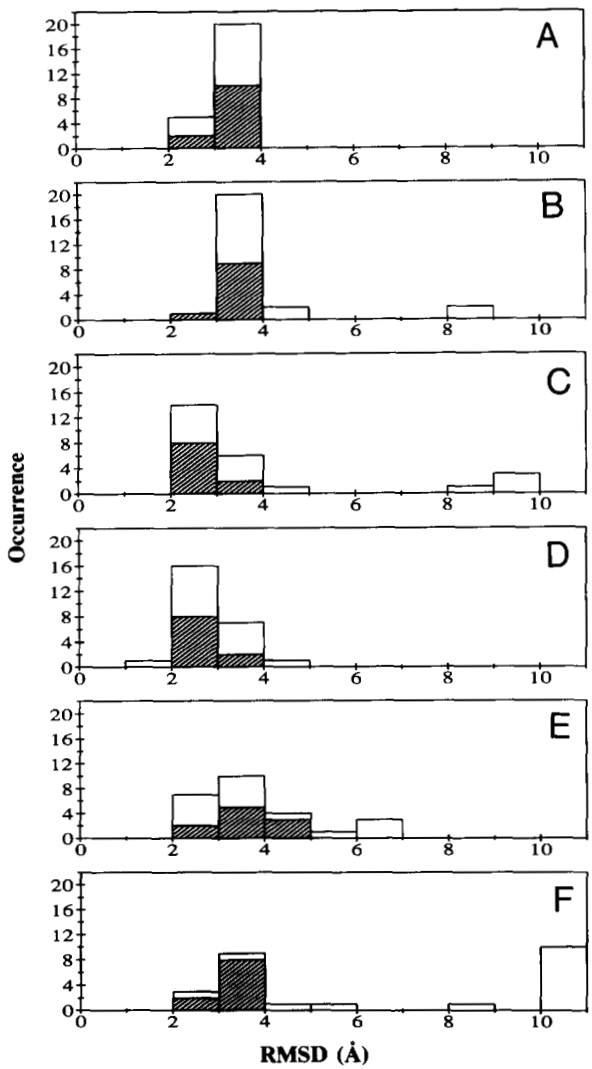

Fig. 1. Distribution of RMSDs of the 25 structures with lowest target function values for **(A)** Er-10, **(B)** C-myb, **(C)** *Antp* homeodomain, **(D)** myohemerythrin, **(E)** DnaJ, and **(F)** FELIX after the DIAMOD cycle IV. Each of the structures were superimposed with the NMR/X-ray reference structures on all backbone atoms of the α-helical regions, and the RMSD values of these atoms were calculated. Hatched areas show the distribution of the 10 structures with lowest target functions.

Convergence of the DIAMOD cycles

During the self-correcting DIAMOD cycles, the number of correct constraints increased significantly for these six proteins. Constraints with large errors were particularly well detected: the number of violations above 9 Å in all six proteins decreased from 73 for the initial constraints to 22 after the fourth DIAMOD cycle. The residue-based correction in the first DIAMOD cycle only changed constraints in the calculations where inside residues were wrongly predicted, i.e., in Er-10, C-myb, *Antp* homeodomain, and 434 repressor. For all eight proteins, 57 constraints changed from the upper limit constraint list to the lower limit constraint list during this first cycle. This led to 53 correct constraints and 4 wrong constraints with relatively small violations of less than 3 Å compared to the NMR/X-ray structure.

The automatic elimination of errors in the constraint list dramatically improved the quality of the calculated structures as illustrated in Figure 2 for the DIAMOD cycles I, III, and IV in the calculation of myohemerythrin. A well-defined four-helix bundle clearly emerges after the third and fourth cycle.

The procedure did not converge for calbindin and the 434 repressor. No clusters could be found in the final structures for these proteins and the pairwise RMSD values among the 10 structures with lowest target function values were greater than 7 Å. The relatively high number of correct initial constraints for these two proteins indicates that the applied distance constraints were not sufficiently restrictive. Due to the larger number of degrees of freedom that both of these proteins with short helices and long loops possess, DIAMOD found many different folds that fulfill the input constraints with low violations. The self-correcting algorithm based on the detection of large violations could therefore not operate properly. A few specific distance constraints might improve the convergence of the calculations, as has been previously shown (Hänggi & Braun, 1994).

Robustness and accuracy of the prediction

We analyzed the distribution of the final target function values to study the robustness of our method. The target function values achieved by the best structures after the fourth DIAMOD cycle are listed in Table 2. The best ranking structure always had the correct global fold with an RMSD around 3 Å compared to the reference structure. We searched for the best ranking structure, which is not included in the correct cluster, and listed its target function value, rank, and RMSD value. We found that the target function values of these wrong folds are typically a factor of 2 higher than the lowest target function values. These values can therefore be used in a relative scale as a quality criterion to select correct structures.

Figures 3 and 4 show the high accuracy of the predicted structures. All the helical segments of the predicted structures have the same orientation as in the reference structures. The figures also demonstrate the diversity of the backbone folds for the six

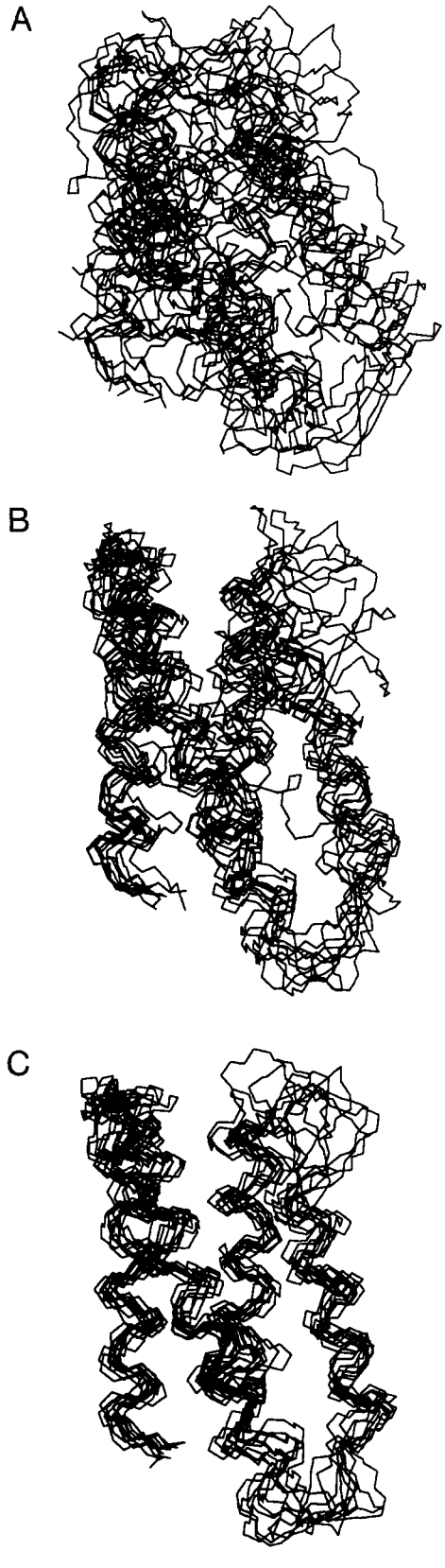

Fig. 2. Effects of the self-correcting algorithm in the calculation of myohemerythrin. The 10 structures with lowest target function after the DIAMOD cycles (**A**) I, (**B**) III, and (**C**) IV are superimposed with their backbone atoms in the α-helical regions.

proteins where the DIAMOD cycles converged. Our method correctly predicted the right-handed four-helix bundle of myohemerythrin, as well as the left-handed bundle of the protein FELIX (Fig. 4).

To elucidate the influence of our χ^1 and loop angle constraints on the handedness, we have repeated the calculations for the two proteins excluding these angle constraints (see Table 2). The results show that the differences in the target function values between correct and wrong folds become much smaller, but the correct fold still scores best. Because we have not included specific rules for α-helical packing, the correct handedness for both proteins is a consequence of quite subtle differences in the distribution of hydrophobic and hydrophilic residues in the α-helical segments.

The Ramachandran plots of the calculated structures are similar to those obtained from high-resolution X-ray and NMR structures. The maximal van der Waals violation in the 25 best structures varied from 0.05 Å to 0.3 Å for the six proteins where our procedure converged. The use of all atoms in the calculation did not cause a prohibitive computational burden. The four DIAMOD cycles of all 50 structures required 0.5–3 h of cpu time on a Cray Y-MP, depending on the size of the protein.

Discussion

The automatic generation of distance and dihedral angle constraints with MULTAN in combination with DIAMOD correctly predicted a variety of different folds of small helical proteins. Particularly intriguing is the correct prediction of a left-handed fold for the protein FELIX and a right-handed fold for the protein myohemerythrin. The calculations performed for these two proteins without the dihedral angle constraints in the loop regions and the χ^1 angle constraints for the side chains in the helical regions show that these constraints have a positive influence

Table 2. *Target function values of best ranking structure and best ranking wrong structure*

	Best ranking structure[a]		Best ranking wrong structure[c]		
Protein	TF[b]	RMSD (Å)	TF[b]	Rank	RMSD (Å)
Er-10	0.11	2.9	–[d]	–	–
C-myb	1.22	3.2	4.39	15	8.0
Antp homeodomain	3.18	2.5	5.37	11	4.3
Myohemerythrin	3.94	2.4	7.55	15	4.1
DnaJ	1.36	3.6	3.47	6	4.8
FELIX	4.96	3.0	10.03	13	10.9
Myohemerythrin[e]	5.68	3.0	7.87	5	6.9
FELIX[e]	8.89	2.9	9.36	4	10.6

[a] Structure with lowest target function.
[b] Target function values. Only constraint violations were considered. Van der Waals violations were not taken into account.
[c] Best ranking structure that was not identified as part of the correct structure cluster (see text for definition of cluster).
[d] All 25 structures with lowest target function values were part of the correct cluster.
[e] Calculations excluding χ^1 and loop angle constraints.

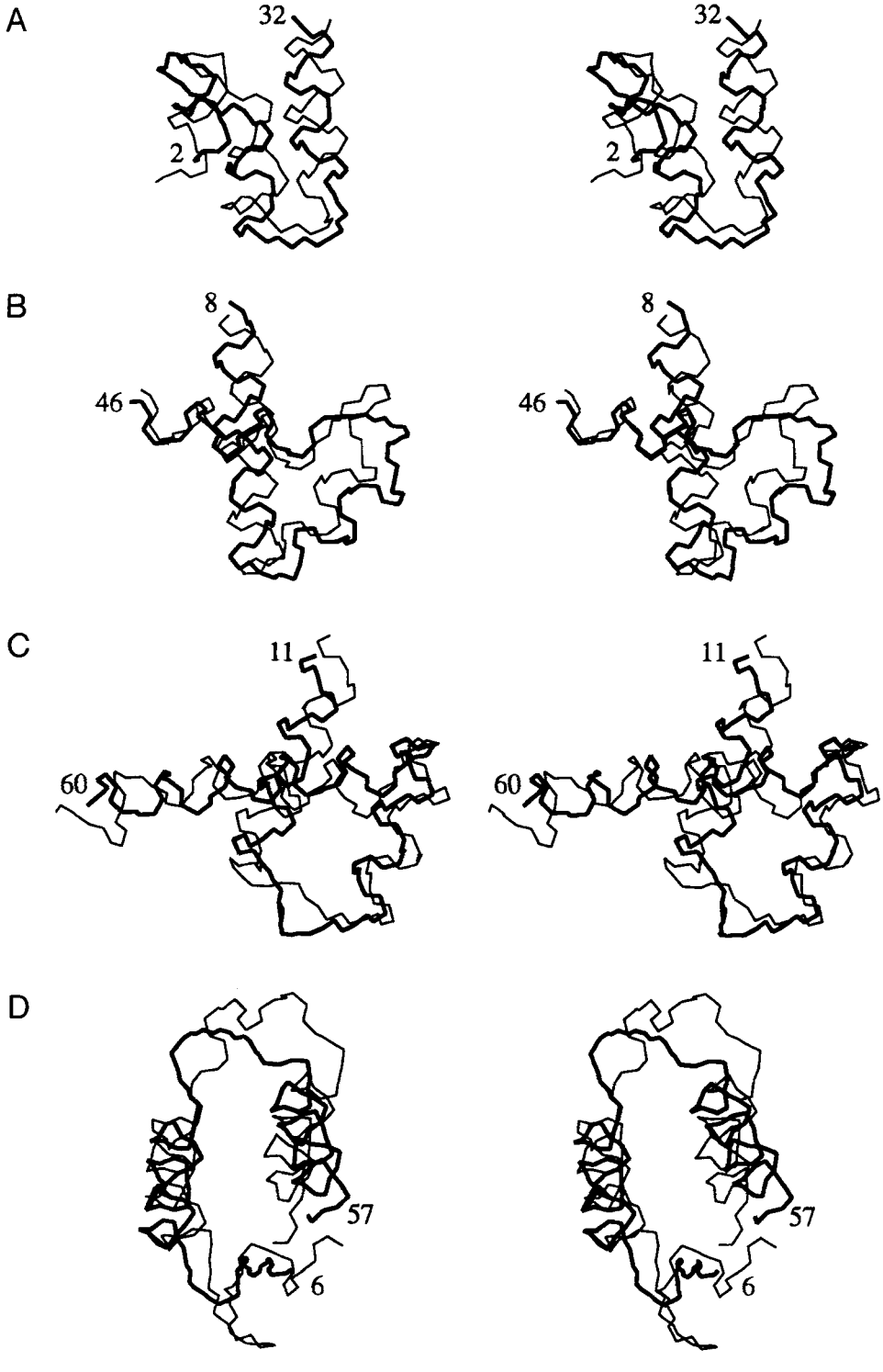

Fig. 3. Superposition of the predicted structures (bold lines) with the NMR/X-ray reference structures (thin lines) for the proteins (**A**) Er-10, (**B**) C-myb, (**C**) *Antp* homeodomain, and (**D**) DnaJ. Predicted structures are the mean structures of the 10 DIAMOD structures with lowest target function. The reference structure of C-myb is the model structure deposited in the Brookhaven Protein Data Bank, which is known to have relatively small deviations to the NMR solution structure (Murthy, 1993). N- and C-terminal ends outside the helical region are not shown.

on the prediction, but the best scoring structures calculated without these constraints still have the correct folds. This indicates that the distribution of hydrophilic and hydrophobic residues along the sequence determines the helix packing in these two proteins. There is a quantitative difference in the robustness of the prediction for the two proteins. For myohemerythrin, no left-handed structures were found among the best 25 final structures in the calculations with and without the dihedral angle constraints, whereas for FELIX a cluster with a substantial number of right-handed structures was observed (Fig. 1D,F).

Further improvements of our method can be expected at different levels. The accuracy of the method should improve with the increasing number of sequences in the sequence data banks. Intrinsic limitations arise if conserved charged residues, classi-

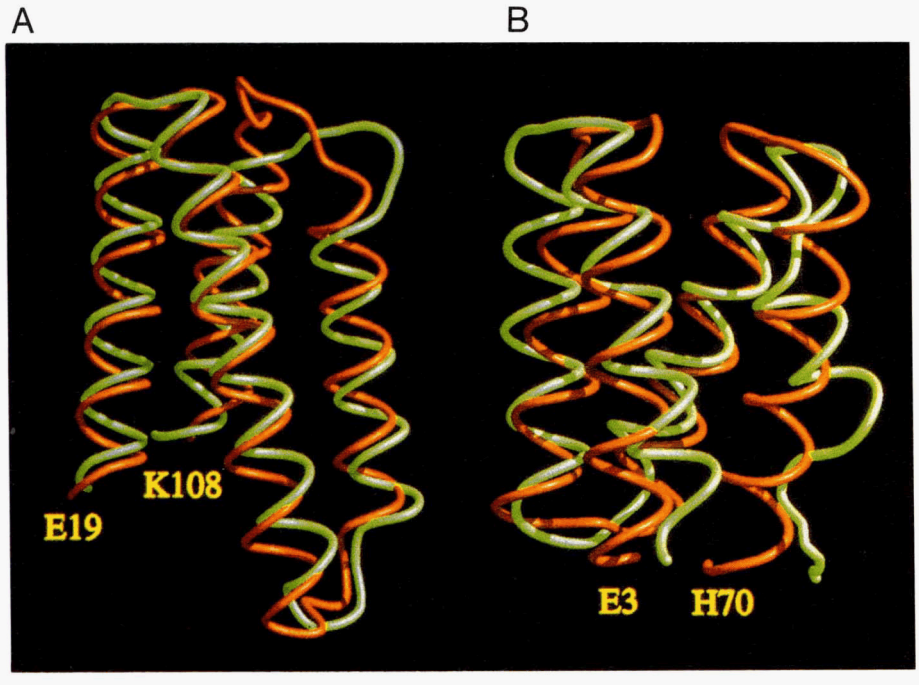

Fig. 4. Correct prediction of right- and left-handed helix bundle topology by the DIAMOD calculations. **A:** Predicted structure of myohemerythrin (green) was superimposed on the X-ray structure (orange). **B:** Predicted structure of FELIX (green) is superimposed on the model (1FLX) as deposited in the PDB (orange). Figure prepared with the program Midas-Plus (Ferrin et al., 1988).

fied as potentially solvent exposed, form salt bridges in the interior of the protein. Analysis of correlated mutation might be able to detect such specific pairs (Shindyalov et al., 1994; Taylor & Hatrick, 1994). In addition, residue-type specific distance constraints, derived from simplified residue-pair potentials as used in the inverse folding problem, might lead to convergence for more complex protein folds (Bowie et al., 1991; Casari & Sippl, 1992; Jones et al., 1992; Maiorov & Crippen, 1992; Godzig et al., 1993).

In our earlier work (Hänggi & Braun, 1994), we generated upper limit distance constraints between inside residues and lower limit distance constraints between outside residues. Experience, however, has shown that incorrect upper limit constraints could be detected much more easily than incorrect lower limit constraints during the DIAMOD calculations. This might be due to the van der Waals repulsion of the individual atoms, which defines an absolute limit to the protein packing, whereas lower limit distance constraints are more easily fulfilled, yielding less densely packed structures.

Related methods have been published. The first predictions for α-helical proteins used helix-packing rules (Richmond & Richards, 1978; Cohen et al., 1979). The procedure used in a prediction of the three-dimensional structure of the transcriptional transactivator c-myb (Frampton et al., 1991) was not automated. Another approach, based on the embedding algorithm for C^{α} atoms (Taylor, 1993), requires starting models with approximately correct fold.

The genetic algorithm (Dandekar & Argos, 1994) has also been used to fold four-helix bundle proteins. For idealized bundles, where the amino acid sequence was replaced by a sequence matching perfect amphipathic wheels for the helices, the structures obtained with the highest score were quite close to correct right-handed bundles. With the amino acid sequences of the cytochrome b_{562}, cytochrome c', and myohemerythrin, the correct folds were found only in about half of the simulations. The final scores obtained by the genetic algorithm were higher by only about 1.5% on average for the correct folds compared to the failures. These small differences would not justify rejection of the structures with the incorrect fold.

In an impressive work based on lattice Monte Carlo simulations (Kolinski & Skolnick, 1994), three small protein structures could be successfully predicted with an accuracy of 3–4 Å for the C^{α} atoms. The protein model consisted of C^{α} atoms on a lattice and spheres representing residue side chains. The method does not require information about the secondary structure and is therefore comparable to our calculations with DnaJ and FELIX. Correct folds could be distinguished by lower energy values. Because the potential employed by this Monte Carlo method has not yet been tested on a large variety of different folds, it remains to be seen if the heuristically chosen weight parameters have general validity.

Overall, the results of the self-correcting distance geometry calculations are very promising for further applications in the prediction of three-dimensional protein structures. Other methods have not yet been applied to a comparable variety of different helical folds. A reliable prediction of all α-helical folds based on the amino acid sequence and the knowledge of helical segments is a realistic challenge for automated prediction methods in the near future.

Methods

Prediction of buried and solvent-exposed residues

Homologous amino acid sequences to each of the eight protein sequences were identified in the PIR, MIPSX, and Swiss-Prot sequence databases and aligned with the PILEUP tool of the GCG (Genetics Computer, Inc.) software package (Devereux et al., 1984). Residues were then predicted to be buried or solvent exposed with the program MULTAN (Hänggi & Braun, 1994). The similarity matrix of Risler et al. (1988) was used in the multiple alignment. From a statistical study (Hubbard &

Blundell, 1987), we defined two subgroups of residues, one containing potentially buried residues, i = {C, M, I, V, L, W, F}, and one containing potentially solvent-exposed residues, o = {K, R, E, N, T, S, Q, P, D}. At each sequence position, the number of i and o residues was counted and compared to the theoretical expectation values for the two subgroups i and o, given the sequence similarities of the homologous sequences relative to the first sequence. Every sequence position where one of these numbers exceeded the theoretical expectation value of the corresponding group was then predicted to be inside or outside.

Dihedral angle constraints

Helical segments were assumed as experimentally determined except for the J-domain of DnaJ and the designed protein FELIX, where the secondary structure was also predicted by MULTAN. The information on the helical segments gave us constraints on the backbone and the χ^1 side-chain dihedral angles. Dihedral angle constraints for the backbone dihedral angles ϕ and ψ were applied for all residues in α-helical regions ($-58° < \phi < -56°$ and $-48° < \psi < -46°$). In addition, ϕ angles of all other residues were restricted to $-180° < \phi < 0°$ with the exception of Gly, Asn, Asp, Ser, Cys, Ala, His, and Lys, following a recent statistical study (Kamimura & Takahashi, 1994). Side-chain rotamer preferences for residues in helical segments were included as χ^1 dihedral angle constraints of $-240° < \chi^1 < 0°$ for Phe, Tyr, His, Trp, Lys, Arg, Met, Glu, and Gln. These constraints are fulfilled in 95% of known three-dimensional globular protein structures (Dunbrack & Karplus, 1994). The same statistical study suggested a dominant single rotamer preference for Val ($-240° < \chi^1 < -120°$) and Ile ($-120° < \chi^1 < 0°$). Dihedral angle constraints were weighted with higher priority than distance constraints.

Distance constraints

For all predicted inside-inside, inside-outside, and outside-outside residue pairs located in helices, we applied upper limit distance constraints between reference points representing the residue side chains. These reference points 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 points Q are pseudoatoms as used by distance geometry calculations from NMR data. The average distances between the reference points of inside-inside pairs, inside-outside pairs, and outside-outside pairs were calculated by a statistical survey of 24 proteins from the Brookhaven Protein Data Bank (Bernstein et al., 1977). Pairs of residues that are separated by less than 10 sequence positions and where one residue is outside were excluded from the statistical study. Residues were classified as "inside" if their solvent-accessible surface area in the tertiary structure was less than 20% of a "random coil" value, and as "outside" if their solvent-accessible surface area was more than 50% of this reference value. The "random coil" value of a residue X is the average solvent-accessible surface area of X in the tripeptide Gly-X-Gly in an ensemble of 30 random conformations.

The average distances were fitted by second-order polynomials as a function of the number of residues N. The resulting calibration curves for $d_{ii}(N)$, $d_{io}(N)$, and $d_{oo}(N)$ were used as upper limit constraints for our eight test proteins. The three polynomials are

$$d_{ii}(N) = 3.2 + 0.116 \cdot N - 2.34 \cdot 10^{-4} \cdot N^2$$

$$d_{io}(N) = 8.7 + 0.119 \cdot N - 2.09 \cdot 10^{-4} \cdot N^2$$

$$d_{oo}(N) = 13.7 + 0.109 \cdot N - 1.52 \cdot 10^{-4} \cdot N^2$$

for inside-inside, inside-outside, and outside-outside residue pairs (Fig. 5). Distances are measured in units of 1 Å.

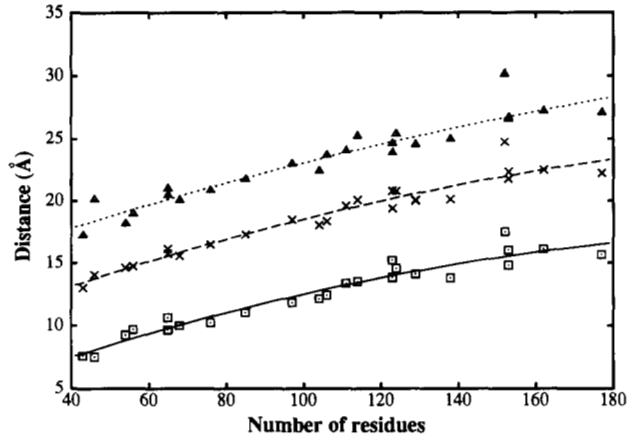

Fig. 5. Average distances between the reference points of side chains (see text) of inside-inside pairs (squares), inside-outside pairs (crosses), and outside-outside pairs (triangles) in 24 high-resolution three-dimensional protein structures. Second-order polynomials, least-squares fitted to the data points, define the calibration curve for upper limit distance constraints of inside-inside pairs (continuous line), inside-outside pairs (dashed line), and outside-outside pairs (dotted line).

The following proteins (PDB code), which represent a set of α-, β-, and mixed α/β proteins and cover the residue range from 40 to 180, were used in the statistical study: sea anemone antiviral protein (1BDS), crambin (1CRN), rubredoxin (5RXN), ovomucoid third domain (2OVO), scorpion neurotoxin (2SN3), barley chymotrypsin inhibitor (2CI2), c-terminal domain of ribosomal protein L7LI2 (1CTF), ubiquitin (1UBQ), high potential iron protein (1HIP), HIV protease (3HVP), ribonuclease T1 (2RNT), ferredoxin (5FD1), rice cytochrome *c* (1CCR), prealbumin (2PAB), pseudoazurin (2PAZ), phospholipase A2 (1BP2), ribonuclease A (3RN3), hen egg white lysozyme (1LYZ), azurin (2AZA), flavodoxin (3FXN), tumor necrosis factor (1TNF), interleukin-1β (2I1B), lupin leghemoglobin (2LH4), dihydrofolate reductase (3DFR), elongation factor TU (1ETU).

Self-correcting distance geometry calculation

Distance geometry calculations in torsion angle space were performed with the program DIAMOD (Hänggi & Braun, 1994). DIAMOD, based on DIANA (Güntert et al., 1991), contains an iterative algorithm to detect and correct inconsistent distance constraints. Starting from random structures, every DIAMOD cycle calculates an ensemble of 50 structures and counts distance constraint violations greater than 1 Å. With the number of these violations, new constraints are produced for the next DIAMOD

cycle. In the first cycle, a residue-based correction is performed where DIAMOD calculates the average number of violations of all constraints belonging to a specific residue. Residues with violations in more than 45% of the structures are assumed to be predicted incorrectly. Their upper limit distance constraints are then treated in the following way: if a constraint is individually violated in more than 50% of the structures, it is assumed that the most likely distance is indeed greater than this upper bound. Therefore, this constraint is used as a lower limit constraint in the next cycle. All the other distance constraints of the incorrectly predicted residue are discarded. In the following DIAMOD cycles II, III, and IV, all constraints that are violated in more than 55%, 50%, and 45% of the structures, respectively, are shifted from the upper limit constraint list to the lower limit constraint list and vice versa.

Cluster analysis

A three-dimensional cluster analysis is performed on the structures resulting from the last DIAMOD cycle to find out whether the structures have converged to one or two distinct and well-defined folds or if they are different from each other. The two structures that have the smallest RMSD value among all pairs of structures form a cluster core if the RMSD value is less than 3 Å. New structures are added, as long as every new structure has an average RMSD of less than 3 Å, to all other structures already included in the cluster. If no further structures can be added, the procedure is repeated with the remaining structures. Clusters were only counted as such if they contained more than two structures.

Acknowledgments

We acknowledge financial support to Ch.M. by the ETHZ. We thank Dr. C.H. Schein for critical reading of the manuscript, Dr. P. Güntert for help in the analysis of structures, and M. Pellecchia and Dr. Th. Szyperski for making the atomic coordinates of DnaJ available to us. The use of the computing facilities of the IPS of the ETHZ is gratefully acknowledged.

References

Benner SA, Badcoe I, Cohen M, Gerloff DL. 1994. Bona fide prediction of aspects of protein conformation. *J Mol Biol 235*:926–958.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Bowie JU, Lüthy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science 253*:164–170.

Braun W. 1987. Distance geometry and related methods for protein structure determination from NMR data. *Q Rev Biophys 19*:115–157.

Braun W. 1991. Distance geometry in torsion angle space: New developments and applications. In: Hoch JC, Poulsen FM, Redfield C, eds. *Computational aspects of the study of biological macromolecules by NMR*. New York: Plenum Press. pp 199–208.

Braun W, Gō N. 1985. Calculation of protein conformations by proton–proton distance constraints. *J Mol Biol 186*:611–626.

Brown LR, Mronga S, Bradshaw RA, Ortenzi C, Luporini P, Wüthrich K. 1993. Nuclear magnetic resonance solution structure of the pheromone Er-10 from the ciliated protozoan *Euplotes raikovi*. *J Mol Biol 231*: 800–816.

Casari G, Sippl MJ. 1992. Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. *J Mol Biol 224*:725–732.

Chothia C, Lesk AM. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J 5*:823–826.

Chou KC, Maggiora GM, Némethy G, Scheraga HA. 1988. Energetics of the structure of the four-α-helix bundle in proteins. *Proc Natl Acad Sci USA 85*:4295–4299.

Cohen FE, Kuntz ID. 1989. Tertiary structure prediction. In: Fasman GD, ed. *Prediction of protein structure and the principles of protein conformation*. New York: Plenum Press. pp 647–705.

Cohen FE, Richmond TJ, Richards FM. 1979. Protein folding: Evaluation of some simple rules for the assembly of helices into tertiary structures with myoglobin as an example. *J Mol Biol 132*:275–288.

Dandekar T, Argos P. 1994. Folding the main chain of small proteins with the genetic algorithm. *J Mol Biol 236*:844–861.

Devereux J, Häberli P, Smithies O. 1984. A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res 12*:387–395.

Donnelly D, Overington JP, Blundell TL. 1994. The prediction and orientation of α-helices from sequence alignments: The combined use of environment-dependent substitution tables, Fourier transform methods and capping rules. *Protein Eng 7*:645–653.

Dunbrack RL Jr, Karplus M. 1994. Conformational analysis of the backbone-dependent rotamer preferences of protein side chains. *Nature Struct Biol 1*:334–339.

Ferrin TE, Conrad CH, Laurie EJ, Langridge R. 1988. The MIDAS display system. *J Mol Graphics 6*:13–27.

Frampton F, Gibson TJ, Ness SA, Doderlein G, Graf T. 1991. Proposed structure for the DNA-binding domain of the Myb oncoprotein based on model building and mutation analysis. *Protein Eng 4*:891–901.

Godzig A, Kolinski A, Skolnick J. 1993. De novo and inverse folding predictions of protein structure and dynamics. *J Comput Aided Mol Des 7*:397–438.

Güntert P, Braun W, Wüthrich K. 1991. Efficient computation of three-dimensional protein structures in solution from NMR data using the program DIANA and the supporting programs CALIBA, HABAS, and GLOMSA. *J Mol Biol 217*:517–530.

Hänggi G, Braun W. 1994. Pattern recognition and self-correcting distance geometry calculations applied to myohemerythrin. *FEBS Lett 344*: 147–153.

Havel TF, Kuntz ID, Crippen GM. 1983. The theory and practice of distance geometry. *Bull Math Biol 45*:665–720.

Havel TF, Snow ME. 1991. A new method for building protein conformations from sequence alignments with homologues of known structure. *J Mol Biol 217*:1–7.

Hecht MH, Richardson JS, Richardson DC, Ogden RC. 1990. De novo design, expression and characterization of Felix: A four-helix bundle protein of native-like sequence. *Science 249*:884–891.

Holbrook SR, Muskal SM, Kim SH. 1990. Predicting surface exposure of amino acids from protein sequence. *Protein Eng 3*:659–665.

Hubbard TJP, Blundell TL. 1987. Comparison of solvent-inaccessible cores of homologous proteins: Definitions useful for protein modelling. *Protein Eng 1*:159–171.

Jones DT, Taylor WR, Thornton JM. 1992. A new approach to protein fold recognition. *Nature 358*:86–89.

Kamimura M, Takahashi Y. 1994. Phi-psi conformational pattern clustering of protein amino acid residues using the potential function method. *CABIOS 10*:163–169.

Kolinski A, Skolnick J. 1994. Monte Carlo simulations of protein folding. II. Application to protein A, ROP, and crambin. *Proteins Struct Funct Genet 18*:353–366.

Maiorov VN, Crippen GM. 1992. Contact potential that recognizes the correct folding of globular proteins. *J Mol Biol 227*:876–888.

Mondragon A, Subbiah S, Almo SC, Drottar M, Harrison SC. 1989. Structure of the amino-terminal domain of the phage 434 repressor at 2.0 Å resolution. *J Mol Biol 205*:189–200.

Mumenthaler Ch, Braun W. 1995. Folding of globular proteins by energy minimization and Monte Carlo simulations with hydrophobic surface area potentials. *J Mol Modeling 1*:1–10.

Murthy K. 1993. Molecular astrology: The case of the Myb DNA binding domain. *Protein Eng 6*:129–131.

Ogata K, Hojo H, Aimoto S, Nakai T, Nakamura H, Sarai A, Ishii S, Nishimura Y. 1992. Solution structure of a DNA-binding unit of Myb: A helix-turn-helix-related motif with conserved tryptophans forming a hydrophobic core. *Proc Natl Acad Sci USA 89*:6428–6432.

Qian YQ, Billeter M, Otting G, Müller M, Gehring W, Wüthrich K. 1989. The structure of the *Antennapedia* homeodomain determined by NMR spectroscopy in solution: Comparison with prokaryotic repressors. *Cell 59*:573–580.

Richmond TJ, Richards FM. 1978. Packing of α-helices: Geometrical constraints and contact areas. *J Mol Biol 119*:537–555.

Risler JL, Delorme MO, Delacroix H, Henaut A. 1988. Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix. *J Mol Biol 204*: 1019–1029.

Šali A, Blundell TL. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol 234*:779–815.

Sheriff S, Hendrickson WA, Smith JL. 1987. Structure of myohemerythrin in the azidomet state at 1.7/1.3 Å resolution. *J Mol Biol 197*:273–296.

Shindyalov IN, Kolchanov NA, Sander C. 1994. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng 7*:349–358.

Svensson LA, Thulin E, Forsen S. 1992. Proline *cis–trans* isomers in calbindin D_{9k} observed by X-ray crystallography. *J Mol Biol 223*:601–606.

Szyperski T, Pellecchia M, Wall D, Georgopoulos C, Wüthrich K. 1994. NMR structure determination of the *Escherichia coli* DnaJ molecular chaperone: Secondary structure and backbone fold of the N-terminal region (residues 2–108) containing the highly conserved J domain. *Proc Natl Acad Sci USA 91*:11343–11347.

Taylor WR. 1993. Protein fold refinement: Building models from idealized folds using motif constraints and multiple sequence data. *Protein Eng 6*:593–604.

Taylor WR, Hatrick K. 1994. Compensating changes in protein multiple sequence alignments. *Protein Eng 7*:341–348.

Tufféry P, Lavery R. 1993. Packing and recognition of protein structural elements: A new approach applied to the 4-helix bundle of myohemerythrin. *Proteins Struct Funct Genet 15*:413–425.