PROTEINS: Structure, Function, and Genetics 17:232–251 (1993)

Conformational Analysis of Protein Structures Derived From NMR Data

Malcolm W. MacArthur[1,2] and Janet M. Thornton[1]

[1]Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, London WC1E 6BT, England; and [2]Crystallography Department, Birkbeck College, London WC1E 7HX, England

ABSTRACT **A study is presented of the conformational characteristics of NMR-derived protein structures in the Protein Data Bank compared to X-ray structures. Both ensemble and energy-minimized average structures are analyzed. We have addressed the problem using the methods developed for crystal structures by examining the distribution of ϕ, ψ, and χ angles as indicators of global conformational irregularity. All these features in NMR structures occur to varying degrees in multiple conformational states. Some measures of local geometry are very tightly constrained by the methods used to generate the structure, e.g., proline ϕ angles, α-helix ϕ,ψ angles, ω angles, and C_α chirality. The more lightly restrained torsion angles do show increased clustering as the number of overall experimental observations increases. ϕ, ψ, and χ_1 angle conformational heterogeneity is strongly correlated with accessibility but shows additional differences which reflect the differing number of observations possible in NMR for the various side chains (e.g., many for Trp, few for Ser). In general, we find that the core is defined to a notional resolution of 2.0 to 2.3 Å. Of real interest is the behavior of surface residues and in particular the side chains where multiple rotameric states in different structures can vary from 10% to 88%. Later generation structures show a much tighter definition which correlates with increasing use of J-coupling information, stereospecific assignments, and heteronuclear techniques. A suite of programs is being developed to address the special needs of NMR-derived structures which will take into account the existence of increased mobility in solution.** © 1993 Wiley-Liss, Inc.

Key words: protein structure, NMR, conformation, ϕ,ψ distribution, χ_1 distribution, heterogeneity

INTRODUCTION

There has recently been much discussion about protein structure coordinates derived from solution nuclear magnetic resonance spectroscopy and how they compare with those determined by X-ray crystallography (see reviews by Wagner et al.[1] and by Billeter[2]). The main discussion revolves around the fundamental question of whether the structures obtained by the two methods can be expected to be the same or indeed whether in the case of NMR it is possible to represent the structure by a unique set of coordinates as is usually (but not always) the practice with X-ray structures. The question is often asked whether the crystal structure is a true picture of the molecule as it exists in solution under physiological conditions. One of the advantages of NMR structure determination is that these conditions are more closely matched by those existing during the experiment, thereby eliminating the possibility of perturbations of the molecular geometry by the crystalline environment. Accordingly, much effort has recently gone into studying the NMR structures and how they compare with those solved by X-ray crystallography. Two main approaches have been employed in these evaluations.

1. Simulated NMR determinations. The NMR structures are determined from a set of distance and angle constraints by generating computationally models which fit these constraints. To test the power of this approach, datasets of constraints extracted from well-resolved crystal structures have been used as input data.[3,4] The results suggest that crystal structures can be faithfully reproduced in this way, and are largely independent of the mathematical techniques used.

2. Direct comparison of structures independently determined by both NMR and X-ray crystallography, e.g., tendamistat,[5] interleukin-1β,[6] bovine pancreatic trypsin inhibitor,[7] metallothionein,[8] HPr,[9,10] and interleukin-8.[11] Providing the structures have been correctly determined using both methods the comparisons show that the fold and conformations of interior residues and hydrogen bonded secondary structures are very similar. However, it is generally concluded that protein surfaces have different struc-

Received May 17, 1993; revision accepted July 23, 1993.

Address reprint requests to Dr. Janet M. Thornton, Biomolecular Structure and Modeling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, England.

© 1993 WILEY-LISS, INC.

ture and dynamic properties in crystals and in solution.[12] There are several examples where the differences appear to be genuine and significant (e.g., interleukin-8,[11] inflammatory protein C3α[13,14]) and can be ascribed to the effects of crystal packing.

Over 100 protein structures have been solved by NMR including more than 30 whose coordinates are now available in the Brookhaven Protein Databank,[15] and an increasingly pressing need exists for some quantitative means of assessing their relative accuracy and precision and how they compare with X-ray structures. In the case of the latter, the resolution and *R*-factor have commonly been used as a measure of quality. In the case of NMR the root mean square deviation (rmsd) is sometimes used in this way. However, it could also be taken as a measure of the differential mobility of some parts of the structure, or a reflection of the thoroughness of sampling of conformational space producing structures all of which are consistent with the experimental constraints. In addition, there appears to be no final agreement on the best method of calculating this figure and its presentation.

In this paper we describe a conformational analysis of the NMR-derived structures. We have shown for X-ray-derived structures that an association exists between the standard deviations about their idealized means of many lightly restrained parameters such as dihedral angles and hydrogen bonds, and resolution and *R*-factor.[16] For example the degree of clustering within the most energetically favored regions of the Ramachandran plot increases as the resolution of the structure increases. In the absence of the usual X-ray crystal structure quality indicators, such measures can be used for NMR structures to evaluate the quality in terms of the global and local geometry. As with protein X-ray crystallography much of the *covalent* structure is determined by input parameters giving tight restraints on bond lengths and angles. In NMR structure simulations, even the dihedral angles are often restrained by the force field, which will be reflected in the final structures.

DATA AND METHODS

All coordinate data were taken from the July 1992 release (including prerelease) of the Brookhaven Protein Structure Databank.[15] These coordinates were for 22 ensembles of which 12 had separate entries for averaged energy-minimized structures. A further 9 structures have entries for averaged energy-minimized coordinates for which the corresponding ensemble data are unavailable. Secondary structure assignments were made using the Kabsch and Sander algorithm.[17] Stereochemical parameters and derived structural data were calculated directly by the use of in-house programs. The program PROCHECK[18] and modifications of it developed for NMR ensembles were used to analyze the coordinate sets of the ensembles and energy minimized average structures.

Solvent accessibilities were calculated using the method of Lee and Richards[19] as implemented in the program ACCESS.[20] The percentage relative accessibility is defined[21] as the ratio × 100 of the water accessible surface area of the side chain *R* to that of its accessible surface area in the tripeptide –Gly–R–Gly– where $\phi = -140°$, $\psi = 135°$, $\chi_1 = -120°$, and $\chi_{1+n} = 180°$. As a consequence of the variability in conformation within some ensembles the relative accessibility for a given side chain can also vary considerably and so the mean value across the ensemble was used where approprate. It is to be noted that the mean of the accessibilities across the ensemble may have quite large standard deviations for some of the residues which are less well defined. For example it can be as high as a maximum of 28.4% as is found for Asp-11 in a loop region of 2SH1. They are therefore not identical to the accessibility values calculated from the averaged energy minimized structure.

The NMR structures studied in this analysis are listed in Table I[22–48] and include both ensembles and averaged energy-minimized coordinate sets. It gives basic information as well as the year of publication and the methods employed in the determination of each structure.[49–59] The data were extracted from the coordinate file where given, or from the original reference. Also shown are the RMSD values and the number of residues used in their calculation. These were taken directly from the original publications or derived from the data therein. For some proteins the data were not available. Table II gives further experimental data for the set of 22 ensembles together with derived parameters.

NMR ENSEMBLES AND MULTIPLE CONFORMATIONS

By contrast with the single structure usually generated from an X-ray crystallographic analysis, for each NMR structure an ensemble of coordinates is generated which fit the data. Some or all of these coordinates are then deposited in the Brookhaven Protein Databank File. In both X-ray and NMR-derived structures the root mean square deviation (rmsd) is commonly reported as a measure of the precision. In our analysis of X-ray crystal structures[16] we used the means and linear standard deviations of the parameters within the structure as measures to assess the global conformational state of the molecule. NMR ensembles, however, frequently show much conformational heterogeneity across the ensemble and parameters such as the dihedral angles ϕ, ψ, and χ for a given residue can vary over the whole range 0° to 360°. Clearly in this case the use of linear statistics would be inappropriate, and as a measure of the variability in torsion

TABLE I. NMR Structures in Dataset

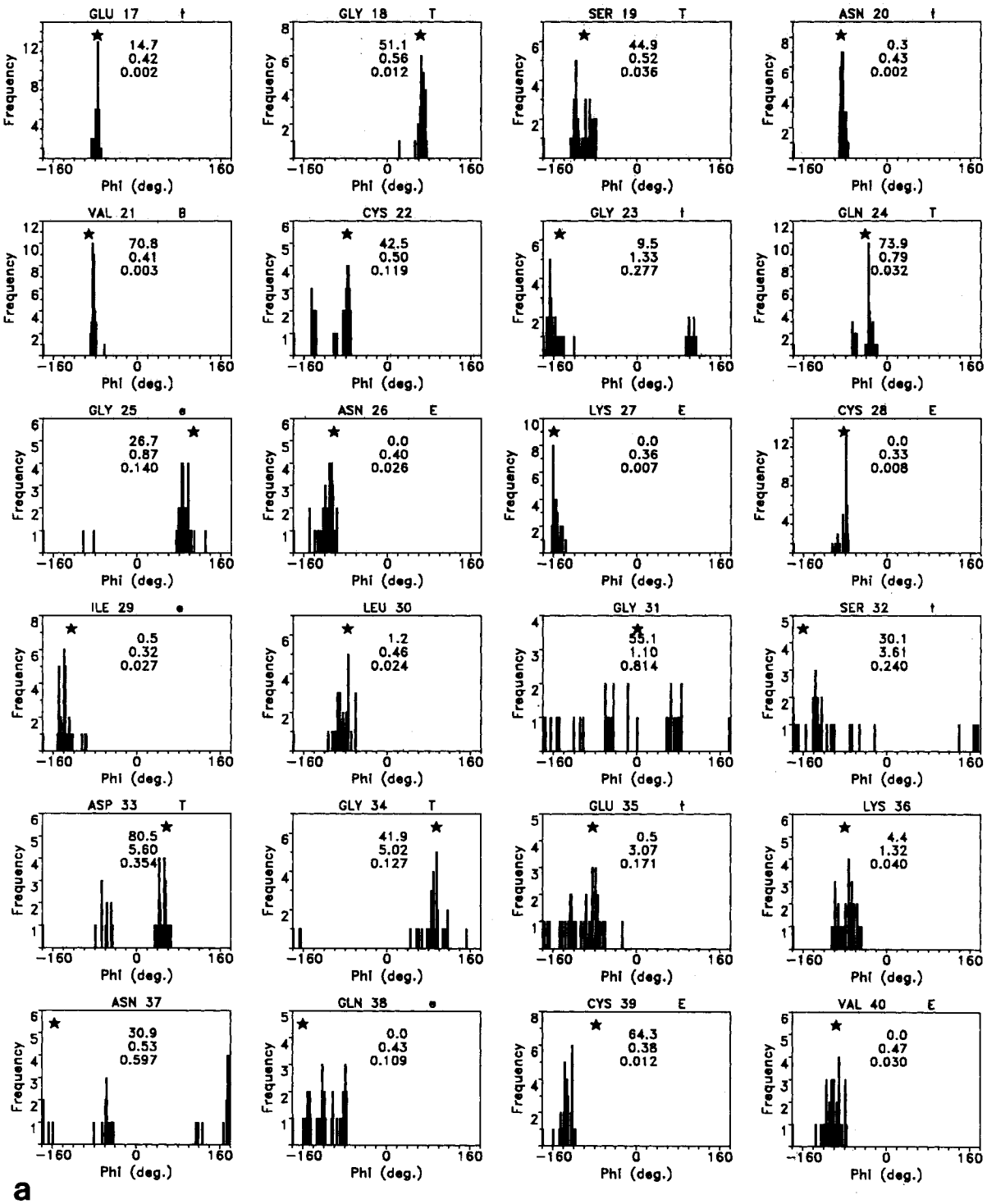
Brookhaven code							
Ensembles	av/EM	Year	Protein name	No. of models	No. of residues	NMR method	Backbone RMSDs*
1aps[22]		1992	Acylphosphatase	5	98	DISMAN + GROMOS	2.5 (98)
1atx[23]		1989	Sea anemone toxin ATX 1A	8	46	DISMAN[49]/HABAS[30] + AMBER[50]	1.4 (35)
1c5a[24]		1990	Complement C5A (des-Arg)	41	73	Modeled on C3A† + CHARMM[51]	1.41 (65)
1ego[25]		1991	Glutaredoxin (oxidized)	20	85	DIANA[30] + X-PLOR[52]	1.56 (85)
1egr[26]		1991	Glutaredoxin (reduced)	20	85	DIANA + X-PLOR	1.42 (85)
1hom[27]		1990	*Antennapedia* homoedomain	19	68	DISMAN/HABAS + AMBER	0.9 (53)
1trx[28]		1990	Thioredoxin (reduced)	10	108	DISGEO[53] + AMBER	0.8 (106)
1znf[29]		1989	Zinc finger (xfin)	37	26	DISGEO + AMBER	1.17 (26)
2hoa[30]		1991	*Antennapedia* homeodomain (C39S)	19	68	DIANA + AMBER	0.76 (53)
2znf[31]		1990	Zinc finger HIV *gag* protein	16	18	DSPACE[54]/BKCALC[55] + MD	0.2–0.8 (18)
4tgf[32]		1990	Transforming growth factor α	4	50	DSPACE/BKCALC + X-PLOR	—
1bus[33]	2bus[33]	1985	Proteinase inhibitor 2A	5	57	DISGEO/CONFOR[56] + EM	3.1‡ (57)
1gb1[34]	2gb1[34]	1991	Protein G (B1 domain)	60	56	DISGEO/STEREOSEARCH[57] + X-PLOR	0.41 (56)
2ait[35]	3ait,4ait[35]	1989	Tendamistat	9	74	DISMAN + AMBER,FANTOM[58]	0.85 (68)
2bds[36]	1bds[36]	1989	BDS-I (sea anemone)	42	43	DISGEO + X-PLOR	0.96 (43)
2cbh[37]	1cbh[37]	1989	Cellobiohydrolase	41	36	DISGEO/STEREOSEARCH + X-PLOR	0.80 (36)
2hir[38]	5hir[38]	1989	Hirudin	32	65	DISGEO + X-PLOR	1.37 (49)
2il8[39]	1il8[39]	1990	Interleuken-8	30	144	DISGEO/STEREOSEARCH + X-PLOR	0.58 (134)
2sh1[40]	1sh1[40]	1990	Neurotoxin I (sea anemone)	8	48	DISMAN/DISGEO + GROMOS[59]	1.30 (39)
4hir[38]	6hir[38]	1989	Hirudin mutant (K47E)	32	65	DISGEO + X-PLOR	1.27 (49)
4trx[41]	3trx[41]	1991	Thioredoxin (reduced)	33	105	DISGEO/STEREOSEARCH + X-PLOR	0.57 (104)
4znf[42]	3znf[42]	1990	Zinc finger (enhancer binding)	41	30	DISGEO/STEREOSEARCH + X-PLOR	0.58 (28)
	1cti[43]	1989	Trypsin inhibitor (cucurbit)	5§	29	DISGEO + X-PLOR	0.45 (28)
	1mca[44]	1991	*Monocyte* chemoattractant	—	146	Modeled on interleukin-8 + CHARMM	—
	1mhu[45]	1990	Human metallothionein α-domain	10§	31	DISMAN	1.5 (31)
	1mrb[46]	1990	Rabbit metallothionein α-domain	20§	31	DISMAN	1.4 (31)
	1mrt[47]	1990	Rat metallothionein α-domain	10§	31	DISMAN	1.7 (31)
	2eti[48]	1991	Trypsin inhibitor (*Ecballium elaterium*)	—	28	DISGEO + AMBER	—
	2mhu[45]	1990	Human metallothionein β-domain	10§	30	DISMAN	2.6 (30)
	2mrb[46]	1990	Rabbit metallothionein β-domain	20§	31	DISMAN	3.0 (31)
	2mrt[47]	1990	Rat metallothionein β-domain	10§	30	DISMAN	2.0 (30)

*These are the averages of the pairwise values for backbone atoms as reported in the original publications. When the only figure reported was computed by pairwise comparison with an average structure the conversion factor $(2n/n-1)^{1/2}$ was used to derive the values shown. Given in parentheses are the number of residues used in the RMSD calculation as reported by the authors of the structure.

†The homologous crystal structure of complement factor C3A was used as a model.

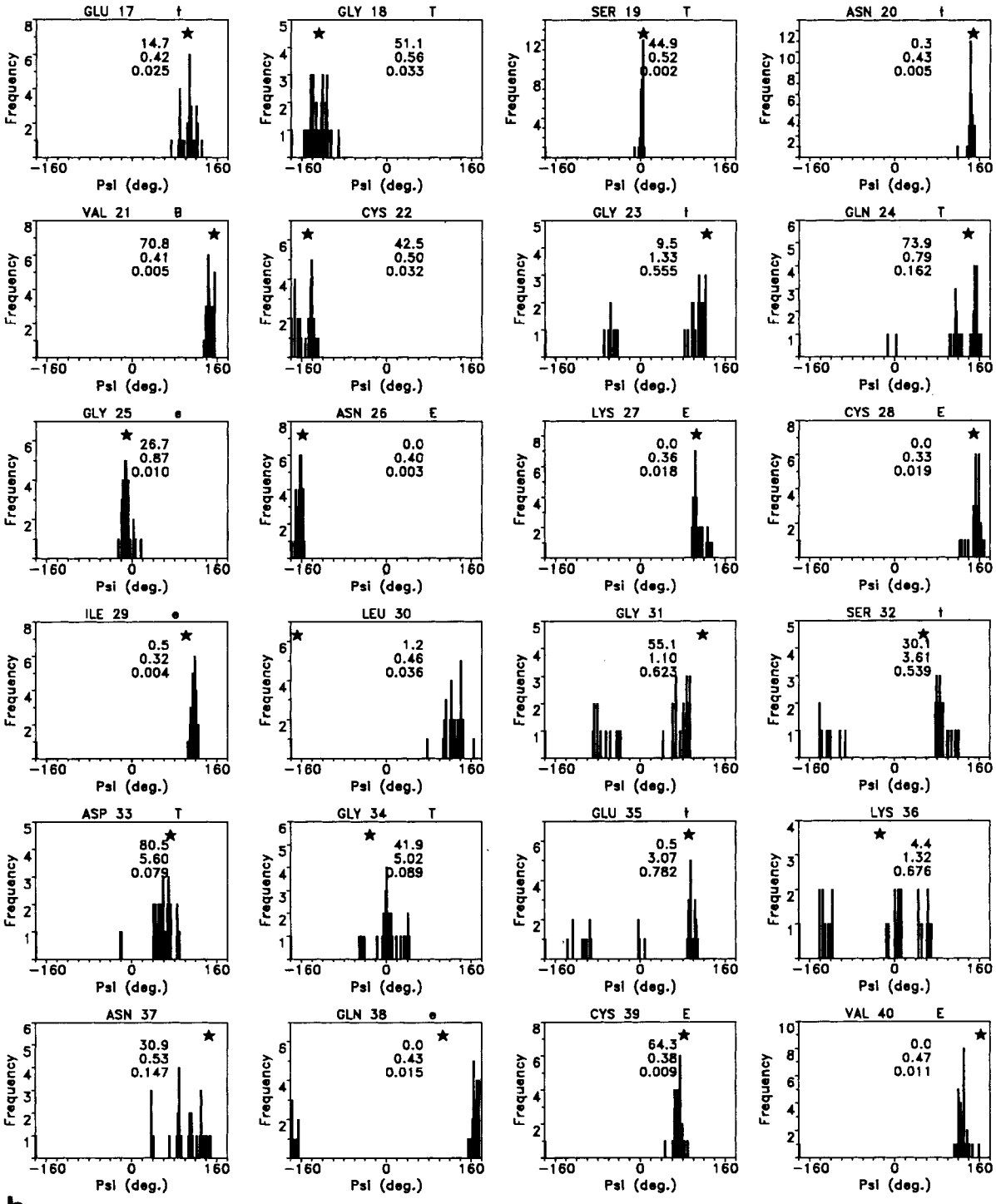
‡The RMSD value shown here is for all heavy atoms.

§For these proteins only 1 set of coordinates was available.



a

Fig. 1. (a) Frequency distribution of ϕ angles for a set of selected residues from the 32 NMR models of hirudin (2hir). The captions above the boxes indicate the residue, the PDB sequence number, and the secondary structure. Notation for secondary structure is that of Kabsch and Sander with lower case letter specifying the initial and final residues of the secondary structure element. Shown within the graph area in descending order are the main chain relative percentage accessibilities as defined in the text, the C_α rmsd values from the energy minimized model 5hir of the ensemble average, and the circular variance across the ensemble. An asterisk (*) marks the ϕ value in 5 hir. (b) Frequency distribution of the ψ angles for the same set of residues shown in (a). Note the large spread of values in the loops at Gly-31, Ser-32 and Glu-35, Lys-36 indicating mobility, or uncertainty due to lack of data. Figure 1b appears on page 236.



b

Fig. 1b. Legend appears on page 235.

TABLE II. Experimental Data and Derived Parameters for the Dataset of 22 NMR Ensembles

PDB code	φ CV*	ψ CV*	% Ram†	χ_1 CV*	% mro‡	χ SD§ (deg)	Constraints				
							Dist/res**	φ	ψ	χ_1	ssa††
1aps	0.179	0.185	65.9	0.375	81.0	15.9	12.0	59	—	—	—
1atx	0.147	0.165	67.1	0.186	48.5	21.0	6.5	32	—	25	12
1bus	0.187	0.233	50.4	0.409	79.5	22.8	4.8	34	—	5	—
1c5a	0.047	0.110	84.7	0.318	88.7	18.0	4.2	25	—	2	—
1ego	0.102	0.122	74.5	0.240	62.3	16.0	10.3	56	—	18	—
1egr	0.077	0.102	78.6	0.193	53.6	14.4	10.4	68	68	55	35
1gb1	0.009	0.015	81.0	0.054	10.9	19.0	16.3	54	51	39	33
1hom	0.103	0.112	76.0	0.317	61.3	21.6	10.3	—	—	—	37
1trx	0.043	0.032	84.8	0.205	56.1	14.1	11.8	41	—	2	2
1znf	0.160	0.136	76.5	0.293	73.9	11.2	6.1	12	—	—	—
2ait	0.089	0.092	67.4	0.209	57.9	20.7	12.5	50	—	36	—
2bds	0.082	0.058	67.1	0.083	44.8	28.9	11.9	23	—	21	—
2cbh	0.015	0.013	67.8	0.048	18.5	23.0	16.4	33	24	25	15
2hir	0.092	0.142	38.8	0.205	61.5	26.1	10.9	26	—	18	—
2hoa	0.167	0.130	81.3	0.232	56.5	21.6	14.1	61	61	49	35
2il8	0.018	0.038	81.9	0.073	19.4	17.0	13.2	136	122	104	38
2sh1	0.110	0.131	62.5	0.299	77.8	17.0	5.2	15	—	—	—
2znf	0.080	0.091	51.8	0.213	42.9	18.1	12.3	—	—	—	—
4hir	0.074	0.095	53.0	0.216	61.5	30.5	10.6	26	—	18	—
4tgf	0.170	0.167	48.3	0.202	51.2	24.6	—	—	—	—	—
4trx	0.011	0.015	76.0	0.108	25.8	15.7	19.4	98	71	72	25
4znf	0.017	0.037	51.2	0.099	28.6	16.1	16.5	28	14	21	17
Overall	0.090	0.101	65.6	0.208	53.8	19.7	11.2	—	—	—	—

*Circular variance.
†Percentage of residues in the most favored regions of the Ramachandran plot.
‡Percentage of residues which have χ_1 multiple rotameric occupancy.
§Pooled standard deviations of the uniquely defined χ_1 rotamers.
**Number of distance constraints per residue (including hydrogen and disulfide bonds).
††Number of stereospecific asignments.

angle space we have instead used the *circular variance* defined by Mardia[60] as

$$\mathrm{Var}(\theta) = 1 - R_{av}$$

where $R_{av} = R/n$, n being the number of members in the ensemble and R is given by the expression

$$R^2 = (\sum_{i=1}^{n} \cos\theta_i)^2 + (\sum_{i=1}^{n} \sin\theta_i)^2.$$

The circular variance can be regarded as a vector sum and can range between 0 and 1. A low value indicates a tight clustering of the values about the mean.

Variability of φ,ψ Angles Across the Ensemble

The presence of multiple conformations is most strikingly seen when presented graphically as in Figure 1a and b. These show the distribution of the φ and ψ values, respectively, across the ensemble for selected residues in the structure. The example chosen here represents a fragment of the hirudin (2HIR) sequence, a typical structure of medium precision for which 32 models are given in the Brookhaven databank. The values shown for accessibility and C_α rmsd for each residue in Figure 1 are those for the energy-minimized structure of the averaged ensemble coordinates. It clearly illustrates the correlation of surface exposure and regular secondary structure with the incidence of multiple conformers across the ensemble.

It will be seen that, except for Gly-23 the two turns between Glu-17 and Gln-24 are well defined about the same position in all 32 models. There then follows a β-strand where the φ,ψ values for a given residue across the ensemble are even more tightly clustered. This is reflected in the low rmsd values, the small circular variance, and the low exposure values for the completely buried section of backbone from Asn-26 to Leu-30. The next β-strand antiparallel to the above from Gln-38 to Gly-42 also exhibits a small spread of values. Linking the two strands the loop from Leu-30 to Asn-37, however, shows a considerably greater degree of conformational variability. Again this is reflected in the higher rmsd values and larger circular variance.

A comparison of the distributions for the φ and the ψ angles for all the proteins in the dataset (Table II) shows that overall the variability is usually slightly smaller for φ than ψ, the mean circular variances being 0.090±0.059 and 0.101±0.060, respectively. The proline φ angle distribution for the NMR dataset (−65.8°±12.7°) is remarkably similar to that ob-

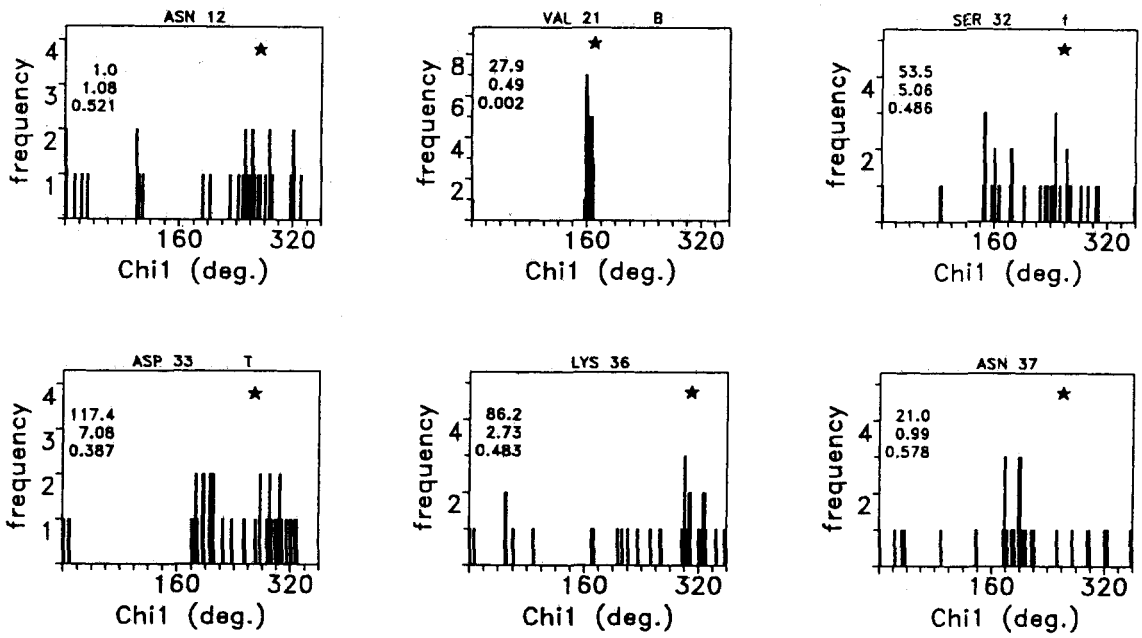


Fig. 2. Frequency distribution of χ_1 angles for a set of side chains selected to highlight the range in variabilities. Val-21 is very well defined, and the circular variance correlates with a small C_γ rmsd and low exposure. The short polar side chains exhibit disorder at all levels of exposure, and the loop-exposed lysine 36 shows a large spread in its χ_1 values. See legend to Figure 1.

served in crystal structures ($-65.4° \pm 11.2°$),[16] but this may reflect the model generation process where the proline ϕ angle is often fixed, such as in procedures which operate in torsional space, e.g., for tendamistat.

Variability of χ_1 Angles Across the Ensemble

The conformational heterogeneity is even more evident in the side chains where the χ_1 angles are observed in many cases to adopt all possible rotameric positions for a given residue across the ensemble. For example, Figure 2 shows the distribution of χ_1 values across the ensemble for selected residues in the structure of hirudin (2HIR). The values shown for accessibilities and C_γ rmsd values are those of the averaged energy minimized structure (5HIR). In carefully refined crystal structures, at least 10% of all residues were reported to exist in more than one distinct conformation in torsion angle space and most of these were surface residues.[61] In crystal structures it is often observed that a few side chains on the surface have little electron density, presumably because they do not adopt a single conformation. However the majority of side chains in crystal structures are assigned a single χ_1 value and in an analysis of high resolution structures χ_1 values were found to cluster strongly in one of the three favored energy wells g^-, t, and g^+. In these structures the pooled standard deviation from the optimal torsion angles ($\sim +60°$, $\sim +180°$, and $\sim -60°$) was found to be only $\pm 15°$, implying that the need to minimize the local conformational energy is a powerful factor even in the complex environment of a protein's core. For an ensemble of structures derived by NMR in which many side chains in the ensemble of structures adopt conformers over the whole of χ_1 space, such a standard deviation would have little meaning. At a grosser level therefore we have calculated the percentage of side chains in each protein for which χ_1 adopts more than one state (g^-, t, or g^+) across the ensemble (see Table II, %mro column). For the 22 proteins in our NMR dataset, this figure ranges from 10.9 to 88.7%.

CIRCULAR VARIANCE AS A CONTINUOUS MEASURE OF VARIABILITY

Since the values of a χ_1 angle for a particular residue can range over 360° within a given ensemble the continuous measure used here is the circular variance as defined above. A low value for the circular variance indicates a tight clustering of the χ_1 values about the mean of one of the three rotameric states (g^-, t, or g^+), and a value greater than 0.1 would generally suggest occupancy of more than one conformation in the χ_1 distribution. The value of the circular variance will depend on the shape of the distribution and its modality and there is no simple relationship with the linear standard deviation. However, a typical distribution within one energy well of 100 residues with a linear standard deviation of 25° as might be found for all the χ_1 angles in a medium resolution crystal structure would be equiv-

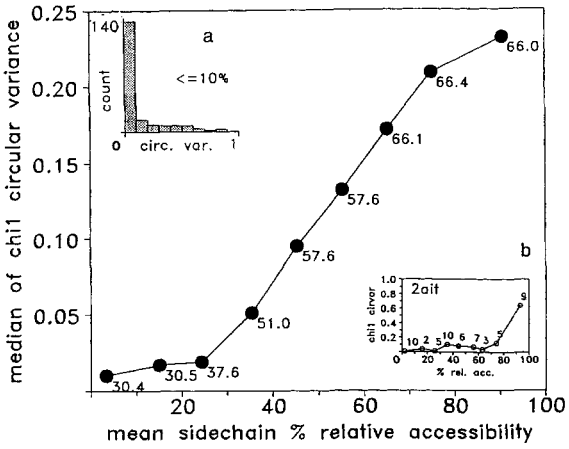


Fig. 3. Plot of χ_1 circular variance against mean percentage relative accessibility for side chains across the ensemble (excluding Gly, Ala, and Pro), to illustrate the correlation between disorder and extent of surface exposure. The plot is for 1,148 residues from the 22 ensembles. The numbers below the curve are the percentages of multiple rotameric occupancy within the successive 10% accessibility intervals. The median of the χ_1 circular variance within each accessibility range is used rather than the mean because of the skewed nature of its distribution. This is shown in the inset in (**a**) for the range 0≤10.0% accessibility. Shown in the inset in (**b**) is the analogous plot of variability versus exposure for an individual ensemble, tendamistat. For most individual ensembles, however, the association tends to be very irregular due largely to the small molecular size.

alent to a circular variance of 0.09, while an evenly spread distribution of 60 values over the 120° range of a χ_1 rotamer would give a circular variance of 0.17. The average circular variance for all χ_1 angles in each protein is shown in Table II. As expected, it is highly correlated with the percentage multiple rotamer occupancy (%mro) and ranges from 0.054 to 0.409 with a mean value of 0.208.

Correlation of Side Chain Disorder With Accessibility

The correlation between the variation in the degree of multimodality with surface exposure is most strikingly apparent when the circular variance is plotted against accessibility. This is shown in Figure 3 for 1,148 residues (Gly, Ala, and Pro excluded) from the 22 ensembles. The median of the circular variance of the set of χ_1 values within successive 10% accessibility intervals is plotted against the percentage relative accessibility. It clearly shows the trend toward greater rotameric heterogeneity with increasing side chain exposure. This tendency is also evident in some of the individual structures (Fig. 3b). Many, however, show a condiderable degree of disorder throughout.

Residue Type and χ_1 Conformational Variability

Table III gives the mean circular variance, the percentage multiple rotameric occupancy, mean percentage relative side chain accessibilities, and number of nonlabile hydrogen atoms per residue (i.e., those which are potentially visible using NMR) for the different amino acid types for the subset of 19 nonhomologous structures from the 22 ensembles. As expected, the variability as estimated by the circular variance and percent multiple rotamer occupancy is greatest for the residue types which tend to be most frequently exposed on the surface, but the overall correlation with accessibility is rather weak ($r = 0.65$ for 16 degrees of freedom). If however, the variability is plotted against accessibility (Fig. 4a), a division into three distinct categories becomes evident: (1) Cys and Ser, (2) the strictly hydrophobic Phe, Ile, Leu, and Val, and (3) the polar and charged residues Trp, Tyr, His, Thr, Gln, Asp, Glu, Arg, Asn, Lys, and Met. Tryptophan, with a mean circular variance across all ensembles of 0.009 and a unique occupancy rate of 93.8%, is especially well defined, as are the other aromatics and large hydrophobics, which is consistent with observations from crystal structures (Fig. 4b). All these have in addition more nonlabile hydrogen atoms which have the potential to provide additional ^1H–^1H distance information from the NOE spectrum. Conversely, cysteine and serine with only two nonlabile hydrogens exhibit a relatively much greater degree of variability. This is in marked contrast to the situation in crystal structures where cysteine (average accessibility 8.0%) has the lowest apparent disorder as measured by its mean side chain B-value (11.0 Å^2), and serine though generally more exposed (54.8%) is also relatively well defined (17.6 Å^2). Aspartate also has only two nonlabile hydrogens (the amino hydrogens of asparagine though labile are often NMR-observable) and is more exposed on average in NMR structures than serine, but nevertheless has a lower χ_1 variance (0.257 as against 0.392), which is the opposite to that seen in crystal structures (corresponding average B-values are 22.1 and 17.6 Å^2). The crystallographic B-value versus accessibility shows some correlation ($r = 0.65$), and also an effect reflecting the length of the side chain, i.e., the number of rotatable χ angles.

Of the remainder, methionine shows the most anomalous behavior in that, while possessing seven nonlabile hydrogens and being generally classified as hydrophobic, it appears in these NMR ensembles as having the most exposed side chain and the least well defined χ_1 angle. In crystal structures methionine is much more buried, but does show a high average B-value. The variability in NMR-defined methionines may simply reflect the greater mobility and shortage of experimental data.

Disulfide Bonds

Of particular interest in a study of the side chains in NMR ensembles are the disulfide bonds. It is well established from crystal structure stereochemical

TABLE III. Relationship of Residue Type With Accessibility and Variability Across the Ensemble

Residue type	Nonlabile hydrogens	Num. res.	%*	mn χ_1 circvar.	SD	% mro†	Mean % rel. acc.‡	SD
Arg	6	52	5.7	0.289	0.265	71.2	60.4	21.9
Asn	2	48	5.2	0.268	0.275	62.5	62.4	26.3
Asp	2	60	6.6	0.257	0.237	65.0	57.6	16.5
Cys	2	71	7.8	0.150	0.184	52.1	14.0	26.9
Gln	4	45	4.9	0.254	0.243	71.1	56.0	22.7
Glu	4	75	8.2	0.269	0.236	72.0	59.1	22.7
His	4	24	2.6	0.117	0.237	29.2	48.0	21.1
Ile	9	49	5.4	0.091	0.166	28.6	30.2	28.6
Leu	9	63	6.9	0.100	0.170	33.3	27.3	27.5
Lys	8	103	11.2	0.280	0.245	69.9	63.1	19.7
Met	7	14	1.5	0.371	0.226	85.7	65.1	27.6
Phe	7	42	4.6	0.044	0.143	9.5	24.2	29.4
Ser	2	63	6.9	0.392	0.301	76.2	39.9	33.8
Thr	4	71	7.7	0.151	0.250	38.0	52.6	27.0
Trp	7	16	1.7	0.009	0.022	6.2	35.3	24.0
Tyr	6	47	5.1	0.105	0.192	29.8	43.7	22.3
Val	7	73	8.0	0.188	0.284	38.4	31.9	32.8

*Number of residues as a percentage of total.
†Percentage multiple rotameric occupancy of the χ_1 angle observed for the residue across all ensembles.
‡Relative accessibility.

analysis that the χ_3 angle C_β–S–S–C_β in a disulfide bridge can adopt either a right or left handed conformation.[62,63] In high resolution crystal structures these show strong clustering around the right-handed (χ_3 = 96.8° ± 10.1°) and left-handed (χ_3 = −85.8° ± 8.6°) conformers.[16] In the absence of β-methylene stereospecific assignments, definition of the disulfide bond geometry must present considerable difficulty as in addition there will be no direct information from NOE (nuclear Overhauser enhancement) measurements on the two central sulfur atoms. Eleven of the 22 structures studied have a total of 30 disulfide bonds. Figure 5 shows the frequency distribution of the χ_3 angles for these disulfides. They have a rather poor clustering about the preferred values and in particular the atypical concentration around −150° must represent a very distorted geometry. Of the 30 disulfides, only 7 are uniquely assigned to either the left-handed or right-handed conformer across the ensemble. It is to be noted that disulfide bonds are in general the most buried features observed within protein structures and frequently the entire bridge substructure has zero percentage relative accessibility. The χ_1 and χ_2 angles are also found in a great variety of different conformations in the NMR structures and show little or no clustering about preferred values as observed in crystal structures.[62,63] 2D heteronuclear ^{14}C–H COSY (correlated spectroscopy) experiments have provided some evidence for limited conformational switching. For example, in bovine pancreatic trypsin inhibitor (BPTI) it is thought that the 14–38 disulfide bond flips at elevated temperatures.[64] However, since it is known that disulfide bonds usually confer stability on proteins, probably stabilizing their correct topology by counteracting conformational changes, the extensive variability observed in the NMR structures may be due to a lack of data, although genuine structural disorder has been suggested.[7]

CONFORMATIONAL ANALYSIS OF THE MODELS IN AN ENSEMBLE

Criteria Used to Assess Coordinate Geometry

Based on our analysis of crystal structures in the Protein Databank, we developed several criteria which can be used to assess the stereochemistry and internal coordinates of any given protein molecule.[16] These include consideration of the following:

1. Main chain bond lengths and bond angles.
2. Peptide bond ω angle distribution.
3. C_α tetrahedral distortion as measured by the virtual dihedral angle C_α–N–C′–C_β.
4. φ,ψ Ramachandran plot distribution.
5. χ_1 standard deviations.
6. χ_2 standard deviations.
7. Main chain hydrogen bond energies.
8. Nonbonded close contacts.

Main Chain

In a first attempt to consider NMR structures we decided to apply the criteria defined for X-ray crystal structures directly to each member of an ensemble. Figure 6a–c compares the results for 2BDS, represented by an ensemble of 42 models and the more recently determined 60 model ensemble of 1GB1. In Figure 6a each point represents the percentage of residues in an individual model in the ensemble which fall within the most energetically favored re-

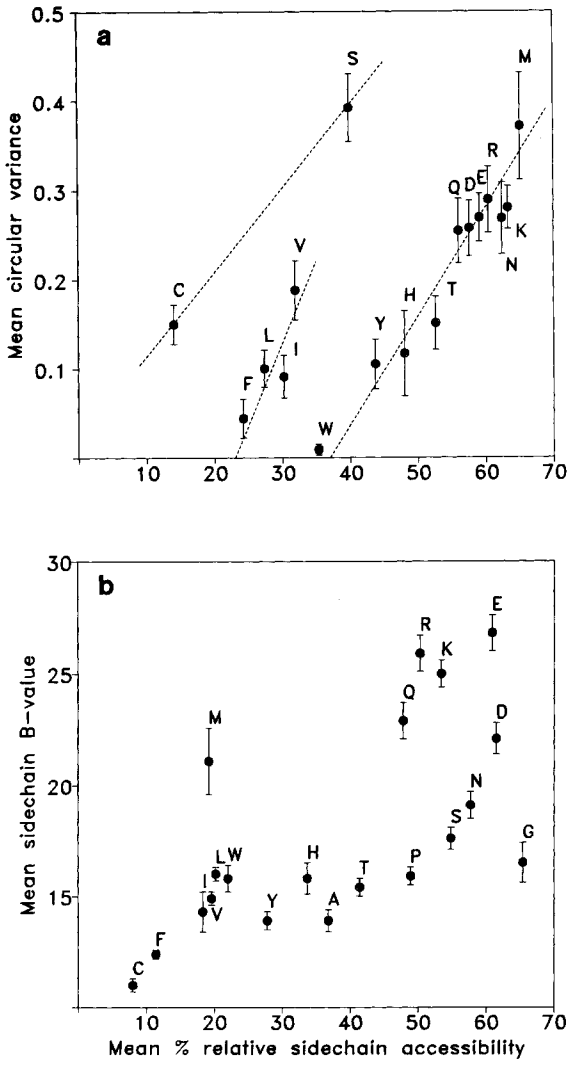


Fig. 4. (**a**) Relationship between residue type, χ_1 variability and percentage relative side chain exposure in NMR ensembles. The plot is for the subset of 19 nonhomologous structures where 1ego, 1hom, 2hir, and subunit B of 1il8 were excluded. Residue types are indicated by their one letter code. Gly, Ala, and Pro were not included. The subdivision into three clusters is emphasized by the dashed lines. Of interest is the generally poor definition of the cysteine and serine. This may be due to their small number of nonlabile hydrogen atoms. (**b**) Relationship between residue type, mean side chain *B*-value, and percentage relative side chain exposure for a set of 17 high-resolution, well-refined, small crystal structures. These are 1bp2, 1crn, 1ctf, 1ecd, 1fd2, 1fx1, 1hoe, 1ilb, 1lz1, 1r69, 1sn3, 1ubq, 2lh3, 2ovo, 5pti, 5rxn, and 7rsa (as designated by their PDB codes). It is to be noted that in all the calculations of percentage relative accessibility, the C_α is included as part of the side chain.

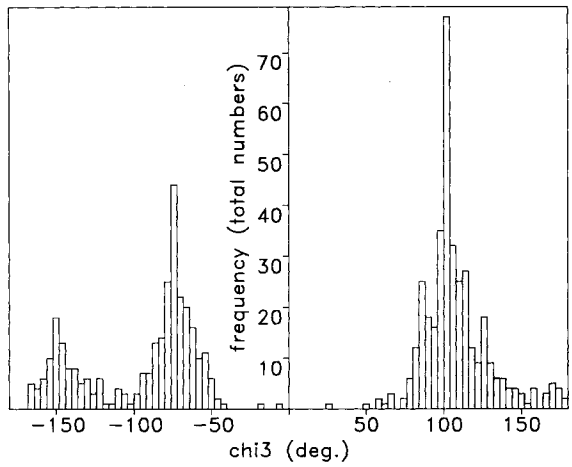


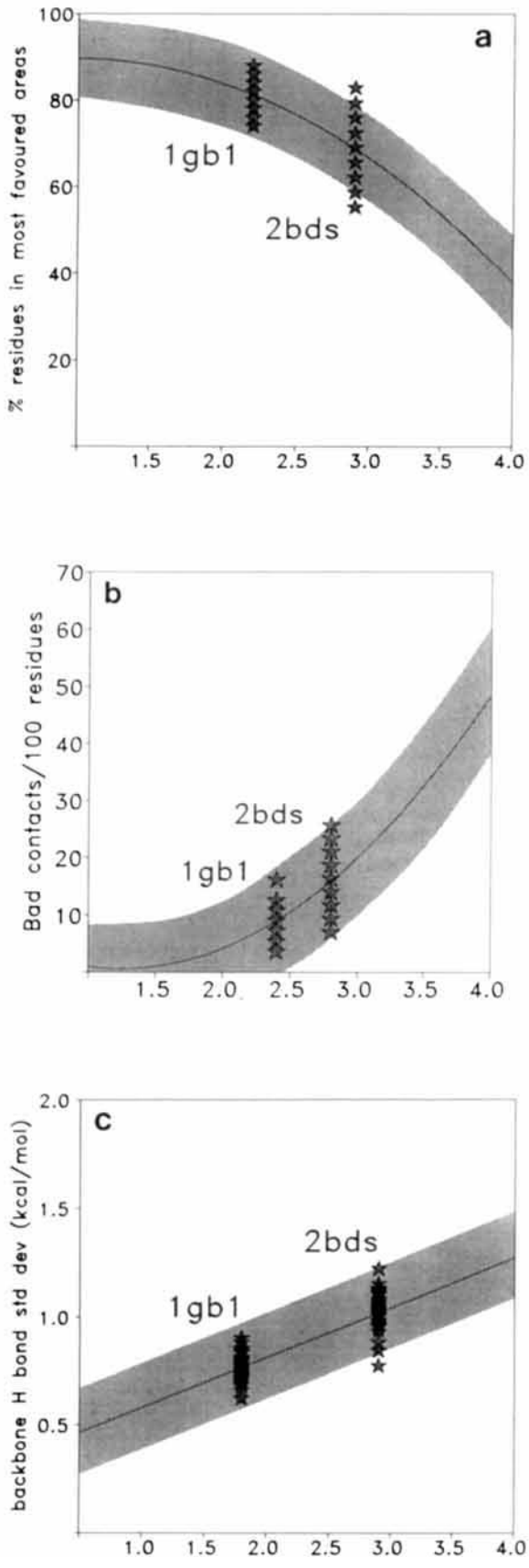
Fig. 5. Distribution of χ_3 disulfide angles. This indicates a large number of disulfides which have a very distorted geometry, in particular the substantial cluster near $\chi_3 = -150°$ and a smaller number near $+150°$. The plot is for a total of 696 angles representing 30 disulfide bonds from 11 proteins out of the starting dataset of 22 ensembles.

gions of ϕ,ψ space. These regions are defined on the Ramachandran map in Figure 7 which was derived from crystallographic data extracted from the Brookhaven databank. The continuous curved line in Figure 6a represents the correlation which is observed in crystal structures between percentage of residues in the most highly favored regions and resolution. The points are plotted so that the regression line intersects the set at about its midpoint. Thus the ensemble as a whole and the individual members by extrapolation can be related to a notional resolution. For 2BDS the notional resolution derived from the ϕ,ψ angle distribution is around 3.0 Å though the individual models range from 2.4 to 3.6 Å. In 1GB1 the ϕ,ψ angles cluster more tightly in the most favored regions and correspond to a notional resolution of 2.2 Å with a range of 1.4 to 2.7 Å. In an analogous manner Figure 6b and c gives an evaluation of the nonbonded bad contacts and main chain hydrogen bond energies, respectively. The assessments clearly establish 1GB1 as the better defined model according to these measures.

Figure 8 summarizes the Ramachandran plot distribution assessment for all the ensembles, where the percentages for each ensemble as a whole are placed on the regression line derived from crystal structure data.

Influence of exposure

As noted earlier for side chains, conformational disorder is greatest toward the exposed surface and such a simple comparison with crystal structures may not be a fair one. In order to investigate the influence of exposure on the backbone, the quality of the Ramachandran plot distribution was examined at different levels of accessibility within the structure. This is summarized for the 22 ensembles in Figure 9 where the percentage of residues falling in the most favored regions is plotted against absolute C-α accessibility. A plot of 17 well-refined, high-resolution small crystal structures is shown for comparison. It is seen that whereas the curve for the crystal dataset is more or less flat, that for the NMR en-



sembles begins to fall off significantly at medium C-α exposure from an initial value, which would typically be expected from a crystal structure of 2.0–2.5 Å resolution.

Side Chains

In Figure 10a the plotted values represent the pooled χ_1 standard deviations for the individual members of each of the two ensembles 1GB1 and 2BDS. It is known that χ_1 angles adopt one of three preferred conformers, $g^-(\sim+60°)$, $t(\sim+180°)$, or $g^+(\sim-60°)$. For each conformer the standard deviation of the values from their average was calculated for a given model. To calculate a measure for the model as a whole the data for the three conformers were pooled, and a weighted average was calculated.

$$\chi_1(\text{pooled}) = \frac{SD_{gm} \times n_{gm} + SD_t \times n_t + SD_{gp} \times n_{gp}}{n_{gm} + n_t + n_{gp}}$$

where SD_{gm}, SD_t, and SD_{gp} are the standard deviations for each of the energy minima about the preferred values derived from high-resolution crystal data, and n_{gm}, n_t, and n_{gp} are the corresponding totals. The continuous straight line is the regression line which relates χ_1 (pooled) to resolution in crystallographic structures. As before, the ensemble of structures is plotted so that the regression line intersects the clusters as near as possible to their mean values. For 2BDS this corresponds to a notional resolution of approximately 4.2 Å with a spread of 3.6 to 4.8 Å, and for 1GB1 the values range from 1.8 to 2.7 Å about a mean of approximately 2.3 Å. It is significant that the ensemble notional "resolution" predicted from the χ_1 data is somewhat

Fig. 6. Application of crystal structure assessment criteria to the NMR ensemble's main chains. In (**a**) the curve shows the correlation derived from crystal structures, between resolution and degree of clustering in the most favored regions of ϕ,ψ conformational space. (CORE regions as defined in Fig. 7.) The shaded area in the above represents the standard deviation about the nominally quadratic regression line. The graph illustrates the percentage of residues in the CORE regions for each member of the 42 model 2bds ensemble and the 60 models of 1gb1 superimposed on the regression line and arranged so that the clusters are approximately bisected. The number of residues in an individual member falling within the CORE regions varies within a limited range among the different models in the ensemble. In the case of 2bds they range between 16 and 24. Thus, many values are identical and the 42 models of 2bds are represented by just 9 points. The values are calculated as percentages of the total number of residues excluding Gly, Pro, and the N and C termini. (**b**) The number of interatomic bad contacts (defined here as a distance ≤2.6 Å) between nonbonded atoms separated by more than four covalent bonds is reduced in crystal structures with improved resolution and increased refinement. In general, NMR solution structures have few bad contacts possibly due to the use of high repulsive force constants in the simulation, or an inherently greater intramolecular looseness. (**c**) In X-ray crystal structures the standard deviation of backbone hydrogen bond energies demonstrate a linear correlation with resolution within the range of interest. In this example the midpoints of the clusters suggest a moderate degree of uniformity in the energies.

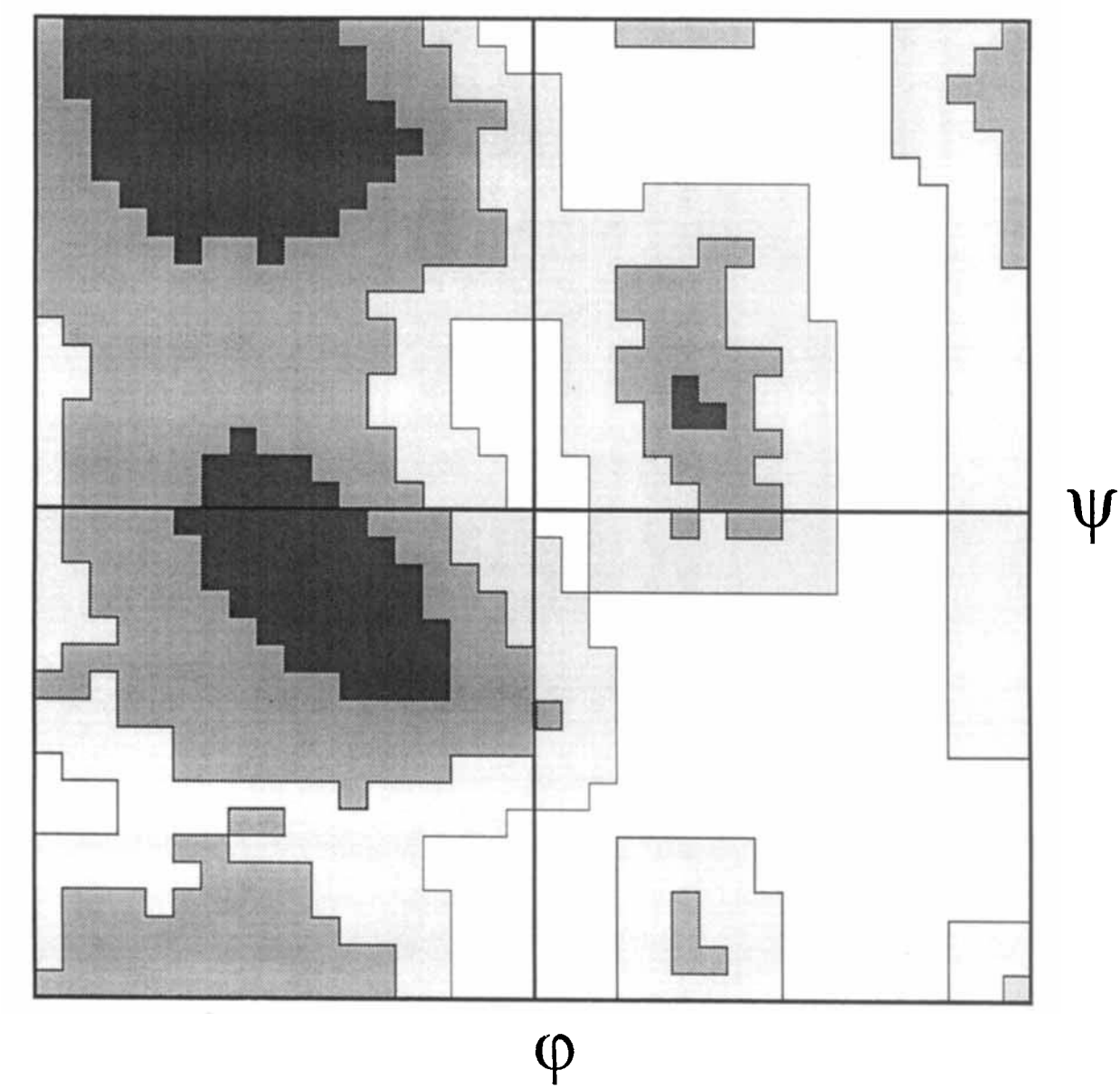


Fig. 7. Ramachandran type map derived from the probability density distribution of ϕ,ψ angles from 462 protein crystal structures. The boundaries are arbitrarily defined according to the density within the $10° \times 10°$ pixels. The most highly favored regions thus defined are shaded dark gray and referred to in the text as Ramachandran CORE regions.

lower than that indicated by the Ramachandran plot assessment. This is in agreement with the circular variance figures which show that the side chain χ_1 angles are less well defined than the backbone ϕ,ψ angles about their respective energy minima. Perhaps of greater significance is the more pronounced improvement in the χ_1 angle distribution relative to that in the ϕ,ψ distribution (Fig. 6a) on going from 2BDS to 1GB1. This probably reflects the greater influence on χ_1 angle definition of the increased number of stereospecific assignments in 1GB1 which was made possible by the use of conformational database grid search and matching. In contrast, the χ_2 angles (Fig. 10b) show little improvement in the spread of their values. It is of interest to note that for the averaged energy minimized structures (1BDS, 2GB1) the values fall very close to the middle of the ensemble distributions. The reservations expressed about calculating standard deviations for ϕ,ψ values apply even more strongly to these χ_1 values.

Figure 11 shows the plot of the χ_1 angle mean

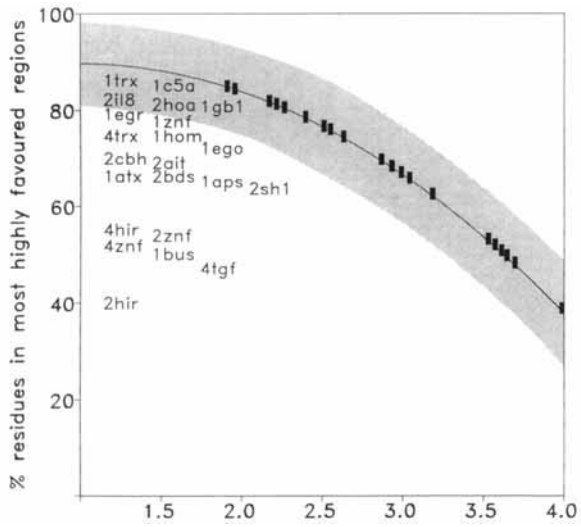


Fig. 8. Ramachandran plot distribution assessment for the 22 ensembles listed in Table I. As described in Figure 6a the percentage of residues found in the most favored Ramachandran regions for each structure is matched with the regression line from crystallographic data which relates the percentage to resolution.

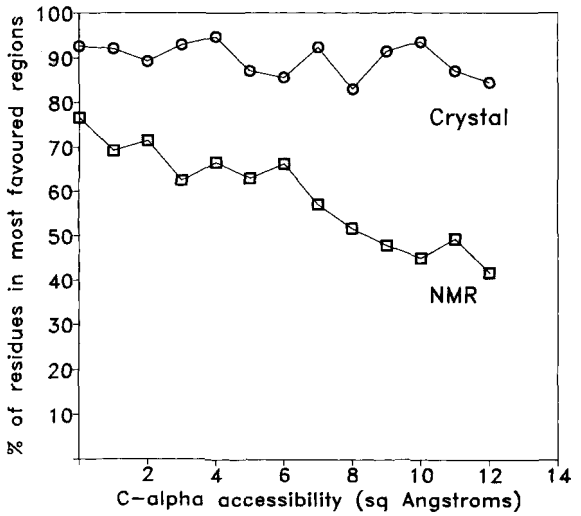
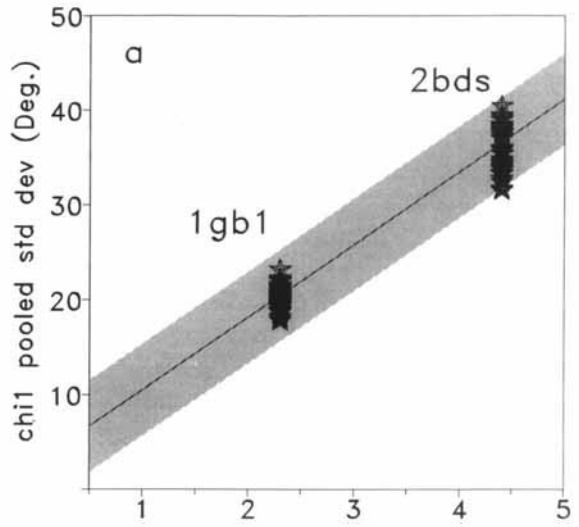


Fig. 9. Correlation of Ramachandran plot CORE region distribution with degree of exposure of the C_α atoms in the set of 22 ensembles. Shown also is the comparable plot for the set of 17 high-resolution, well-refined, small crystal structures listed in Figure 4b.



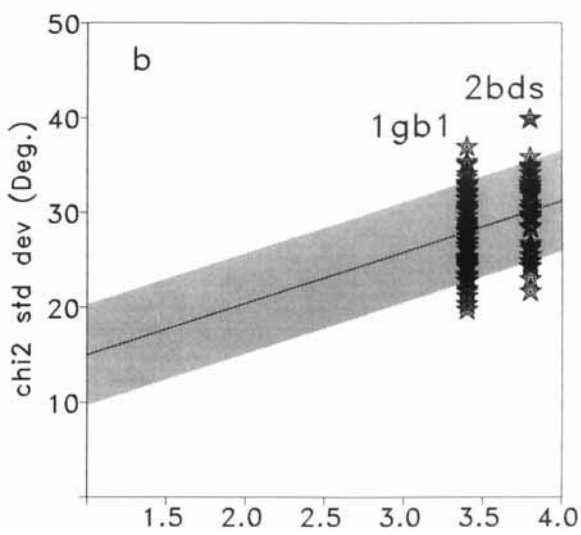


Fig. 10. Application of crystal structure assessment criteria to the NMR ensemble side chains. (**a**) The curve shows the linear regression for the pooled standard deviations of χ_1 for the g^+, trans, and g^- rotamers versus crystal structure resolution, with the shaded area indicating the standard deviation about the line. Shown superimposed on the line are the clusters of values representing the pooled χ_1 standard deviations for individual members of the 1gb1 and 2bds ensembles. (**b**) The analogous clusters for the χ_2 trans angles of the ensembles superimposed on the corresponding crystal derived regression line.

value for the pooled standard deviations of each ensemble in the dataset overlaid on the crystal structure regression line which relates the χ_1 standard deviation to resolution. It should be noted that in the calculation of the standard deviations for this plot, only the rotamers uniquely defined across the ensembles were considered. The values are given in Table II.

These data can also be used for choosing the model which is the "best" or the most representative of the ensemble. One could, for example, choose the one which comes closest to the consensus for the ϕ,ψ and χ_1 criteria. A review of the different approaches for making such a choice has been carried out by Sutcliffe.[65]

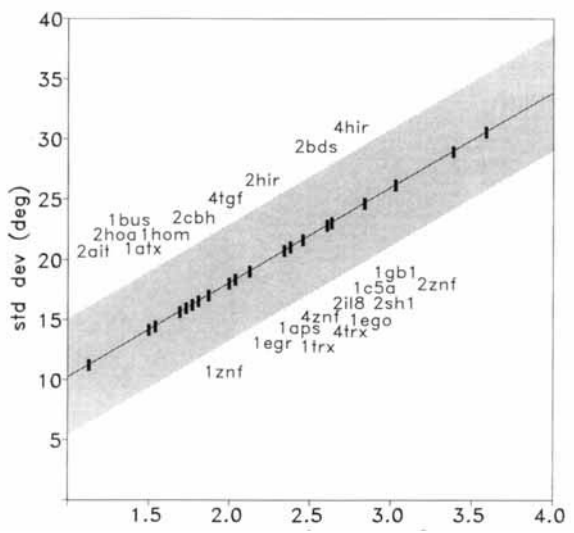


Fig. 11. Plot relating the χ_1 pooled standard deviations (Table II) of the 22 ensembles to the crystal structure regression line. The captions are drawn on approximately the same level as the points to which they correspond.

CORRELATION OF CONFORMATIONAL STATISTICS OF ENSEMBLE WITH NMR DATA

Backbone

The number of NOE and other distance contraints employed in the structure determination is usually reported. One would expect that the greater the number of these per residue the better defined the final model would be. Table IV shows the correlation matrix for statistics given in Tables I and II. From this it can be seen that there is no correlation between the distance constraints per residue and the percentage of residues which occupy the most highly favored regions of the Ramachandran plot ($r=0.21$). If, however, ϕ,ψ constraints are considered there is a much better correlation with the Ramachandran plot data (0.87). While this value must be treated with caution due to the reduced degrees of freedom (6) it does point to the greater influence of the specific local information which defines these angles rather than long-range NOE-derived constraints which play little part in their definition. As expected, the ϕ angle circular variance correlates strongly with that of the ψ angles ($r=0.94$). They both show a weaker association with the variance of the χ_1 angles ($r=0.80$ and 0.84) reflecting the differential between the more ordered interior and the extensive conformational heterogeneity observed in surface exposed loops. The correlation is slightly stronger between the circular variance of ψ and that of χ_1 ($r=0.84$). Both have in common a greater degree of uncertainty in the data leading to lower precision. Also the ψ variance has a better correlation with total distance constraints/residue ($r=-0.78$) than does the ϕ variance ($r=-0.72$). This probably reflects the greater dependence of the ψ angle calculation on sequential NOE distances. For the ϕ angle, $^3J_{HN\alpha}$ coupling constants provide the major contribution, and its determination is not as dependent on NOE distance constraints.

As expected there is a strong positive correlation between the ϕ,ψ circular variances and rmsd of the ensemble backbone atoms ($r=0.72$ and 0.84, respectively). However, the averaged overall rmsd tells us little about the stereochemistry and conformational status of the individual members of the ensemble. It provides an indication of variability between the different members, all of whose coordinates are consistent with the experimental observations and applied constraints. A large overall rmsd may represent a more comprehensive sampling of conformational space for a structure which may exhibit a greater than usual degree of disorder. As can be seen from Table IV there is no correlation between rmsd and the tightness of the Ramachandran plot distribution within the most favored regions ($r=-0.34$). Within the β region, for example, considerable variability of ϕ,ψ values across the ensemble is possible for a given residue, while a residue falling outside the favored regions could well have a similar location in every model.

Side Chains

The correlation of χ_1 circular variance to total number of constraints per residue at -0.71 is similar to that observed for the ϕ angles. If the percentage of unique rotameric occupancy is considered instead of χ_1 circular variance the correlation is slightly better ($r=0.77$), and is improved still further by inclusion of the χ_1 constraints and stereospecific assignments per residue. As these are progressively weighted relative to the distance constraints the correlation coefficient peaks at a maximum of 0.88. This is shown as a scatterplot in Figure 12.

Unlike the 360° variability there appears to be no correlation between any of the experimental parameters and the tightness of clustering *within* the χ_1 individual wells, whether these are uniquely defined across the ensemble or not. The mean over all ensembles (19.7°) of the pooled standard deviations for the uniquely defined χ_1 angles is typical of what might be observed in crystal structures of about 2.3 Å resolution.

α-Helix ϕ,ψ Angles

In good quality crystal structures the ϕ,ψ values of helical residues are tightly clustered. Not all of the structures in the NMR dataset have α-helices. Details of the ones which do are shown in Table V where the averages and standard deviations are given for ensemble ϕ,ψ angles in α-helices; the two helical terminal residues are excluded in the calcu-

TABLE IV. Matrix of Correlation Coefficients for Experimental and Derived Data

	1	2	3	4	5	6	7	8	9
1 (ϕ circ. variance)	1.00	0.94	0.80	0.80	0.10	−0.24	0.72	−0.72	0.25
2 (ψ circ. variance)	0.94	1.00	0.84	0.81	0.13	−0.42	0.84	−0.78	0.14
3 (χ_1 circ. variance)	0.80	0.84	1.00	0.92	−0.11	−0.20	0.75	−0.71	0.03
4 (% mult. rot. occ.)*	0.80	0.81	0.92	1.00	−0.01	−0.24	0.72	−0.77	−0.11
5 (pooled χ_1 sdev)†	0.10	0.13	−0.11	−0.01	1.00	−0.52	0.08	−0.02	0.32
6 (% in Ram. CORE)‡	−0.24	−0.42	−0.20	−0.24	−0.52	1.00	−0.34	0.21	0.87
7 (rmsd)	0.72	0.84	0.75	0.72	0.08	−0.34	1.00	−0.62	−0.26
8 (dist. constraints)§	−0.72	−0.78	−0.71	−0.77	−0.02	0.21	−0.62	1.00	−0.19
9 (ϕ,ψ constraints)**	0.25	0.14	0.03	−0.11	0.32	0.87	−0.26	−0.19	1.00

*Percentage multiple rotameric occupancy of χ_1 angles.
†Pooled standard deviations of the uniquely defined χ_1 angles.
‡Percentage of residues in the most favored regions of the Ramachandran plot.
§Number of distance constraints per residue.
**Number of ϕ,ψ constraints per residue.

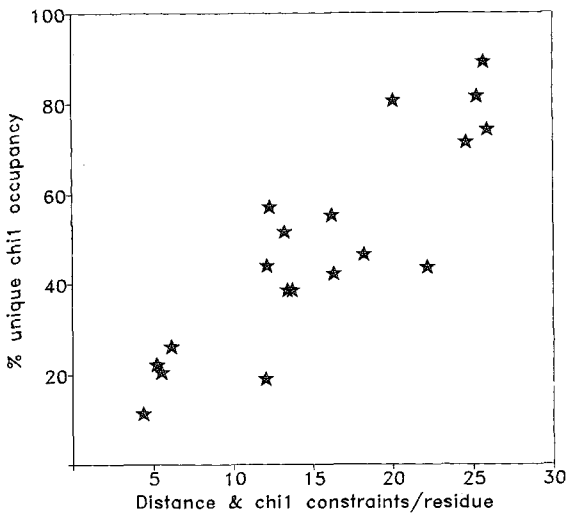


Fig. 12. Correlation of combined distance and χ_1 angle constraints per residue with the percentage of uniquely defined χ_1 rotameric occupancy. Each point identifies an individual ensemble in ascending order of constraints/residue as follows: 1c5a, 2sh1, 1bus, 1znf, 1aps, 1trx, 2znf, 1atx, 2hir, 4hir, 2bds, 2ait, 1egr, 2hoa, 4znf, 2cbh, 1gb1, 4trx. For this plot a best correlation coefficient (r=0.88) was obtained when the combined χ_1 constraints and stereospecific assignments were given six times the weighting of the distance constraints.

lations as these frequently have ϕ,ψ values which are not typical of the α region. The results compare favorably with those observed in well-determined crystal structures i.e. very tight. The helix geometry is well determined by NMR due to the multiple NOE constraints obtainable. In addition, once a region is defined as helical, the ϕ,ψ values and $O_i \cdots N_{i+4}$ separations are often restricted by the NMR experimental data within standard limits.

ENERGY MINIMIZED STRUCTURES

Despite the questionable validity of a single set of coordinates ever being a geniune portrayal of a protein structure in solution we nevertheless deemed it instructive to examine the models which were obtained by energy minimization of the averaged ensemble coordinates. Most crystal structures are conventionally presented in this way although it is now acknowledged that many display evidence of conformational heterogeneity.

Data

Of the 22 ensembles described above (see Table I), 12 have entries in the Protein Data Bank representing the energy minimized averaged coordinates. In addition there are nine[43–48] energy minimized sets for which the ensemble coordinates are unavailable. All 21 are listed in Table VI which also shows the percentages for the Ramachandran plot most favored regions and side chain χ_1 pooled standard deviations.

Bond Lengths and Bond Angles

The bond and angle terms in the target function are the most important of the stereochemical restraints. In a molecular dynamics simulation the force constants for the bond, angle, planarity, and α-carbon chirality are maintained at uniformly high levels in order to ensure near perfect stereochemistry. Typical values would be 600kcal/mol/Å^2, 500kcal/mol/rad^2, and 500kcal/mol/rad^2, respectively. These strong restraints are reflected in the tight clustering about ideal values of bonds and angles as observed by Laskowski et al.[66] An equally tight clustering about mean values is found in both the C_α improper torsion angles (average standard deviation = 1.1°) and the peptide bond ω angles (average standard deviation = 3.7°) which are subject to the same force constant. In interpreting these figures it must be remembered that in the methods employing distance geometry in torsion angle space these parameters are fixed at their standard values.

TABLE V. α-Helix ϕ,ψ Angles in NMR Ensembles

PDB code	No. ϕ,ψ	ϕ	SD	ψ	SD
1aps	102	−57.5	11.4	−47.4	12.1
1bus	46	−59.3	11.7	−59.8	13.4
1c5a	1,863	−64.5	9.5	−40.2	13.5
1ego	699	−61.5	22.0	−43.8	17.5
1egr	681	−60.6	18.1	−44.6	14.8
1gb1	811	−64.4	10.9	−40.0	12.6
4trx	1,166	−69.5	17.3	−34.6	18.8
4znf	368	−61.6	11.0	−42.8	11.0
Overall	5,736	−64.3	13.3	−40.4	14.0

Notional Resolution

Shown in Figure 13 is the Ramachandran plot assessment for all 21 energy-minimized averaged structures superimposed along the crystallographic regression line. There is a large spread in the percentage of residues in the highly favored ϕ,ψ regions with an overall mean value comparable to what might be expected of a crystal structure determined to about 2.9 Å resolution, i.e., an average of 67.5% of ϕ,ψ angles are located in the most favored regions. However, the Ramachandran plot for all residues in the set shows that very few non-glycine residues fall in the disallowed region. Rather, the bulk are diffusely scattered within the area commonly accepted as being allowed, with an increasing concentration toward the most highly favored regions. This is probably a reflection of the constraints placed on the ϕ,ψ values during the calculations. The ϕ angles are sometimes restrained to ranges of $-160° < \phi < -80°$ and $-90° < \phi < -40°$ on the basis of a geometry which corresponds to $^3J_{HN\alpha}$ coupling constant values of >8.0 and <7.0 Hz, respectively.[67] While the scalar connectivity $^3J_{HN\alpha}$ helps to define the ϕ angle there is no such through-bond coupling for the ψ angle. There are three potential short-range sequential ^1H–^1H distances available from NOESY (d_{NN}, $d_{\alpha N}$, and $d_{\beta N}$) but the possibility of peptide bond flipping in exposed locations can render interpretation more difficult. It is therefore not easy to improve the definition of the main chain conformation to help attain the rather stringent target value seen in high-resolution crystal structures, i.e., >90% of ϕ,ψ values in the most favored regions illustrated in Figure 7.

Examination of the χ_1 pooled standard deviations for the 21 structures reveals a similar situation (Fig. 14). Once again there is considerable spread (23.2°) corresponding to a notional crystallographic resolution of around 2.7 Å.

Correlation of χ_1 Standard Deviation With Accessibility

In X-ray crystal structures a greater spread of χ_1 values about their rotamer means is observed with increasing exposure. This is also seen in the NMR energy minimized average structures, as shown in Figure 15. Given for comparison are plots for the crystallographic data. While the correlation with accessibility for the NMR structure is much less regular due to the small dataset it is clear that for all levels of exposure the values are more weakly clustered about their means than they are for the highest quality X-ray structures. Rather it more closely resembles the correlation for the entire Brookhaven dataset of crystallographic structures which range in resolution from 1.0 to 3.5 Å.

Correlations With Experimental Data

Backbone

In the case of energy minimized structures a better correlation than that for the ensembles (Table IV) is observed between the concentration of ϕ,ψ values in the most favored regions of the Ramachandran plot and the total number of constraints per residue, though it is still a modest one ($r = 0.66$). As Figure 16 for the entire dataset indicates there is a suggestion of a trend toward improvement with increasing number of experimental constraints per residue. The poor correlation may be partly due to the small size of the dataset and, in addition, the overall quality depends not simply on the total number of constraints but also on their distribution which can vary widely with location and residue type.

Side chains

One might have expected that smaller numbers of distance constraints would be reflected in a greater uncertainty in the positions of the χ_1 angles within the energy wells (g^+, t, or g^-), as measured by their pooled standard deviations. No such correlation is observed ($r = -0.24$). The χ_1 pooled standard deviations do, however, correlate weakly with the number of χ_1 constraints per residue ($r = -0.62$). It is likely that any possible correlation between uncertainty of side chain position within their energy wells and the total number of distance constraints per residue is masked by the large number of sequential NOEs which play no direct part in determining the side chain conformation.

When compared with high-resolution well-refined crystal structures the NMR averaged energy minimized structures show a greater spread of values about their means within the three energy wells (g^-, t, or g^+). The overall pooled standard deviation for the set of NMR energy minimized average structures is 23.2° compared to 15.4° for the set of high-resolution crystal structures. There are also significant differences in detail for individual residue types, e.g., χ_1 pooled for cysteine (NMR 24.4° crystal 11.9°), serine (NMR 26.6° crystal 17.6°), and aspartate (NMR 24.3° crystal 14.6°).

TABLE VI. Backbone and Side Chain-Derived Data for the 21 Averaged Energy Minimized Coordinate Sets

PDB code	% Ram. CORE	χ_1 SD	PDB code	% Ram. CORE	χ_1 SD	PDB code	% Ram. CORE	χ_1 SD
1bds	72.5	33.6	1mrt	30.9	29.1	2mrt	56.0	24.7
1cbh	73.1	25.1	1sh1	60.0	16.1	3ait	79.1	22.8
1cti	60.0	18.5	2bus	58.7	27.2	3trx	73.7	20.1
1il8	82.5	22.9	2eti	61.8	31.4	3znf	42.3	20.4
1mca	81.0	15.6	2gb1	82.0	21.8	4ait	77.4	25.9
1mhu	52.0	23.8	2mhu	36.0	22.3	5hir	51.4	28.8
1mrb	37.5	32.3	2mrb	44.4	21.2	6hir	54.1	33.7
Overall							67.5	23.2

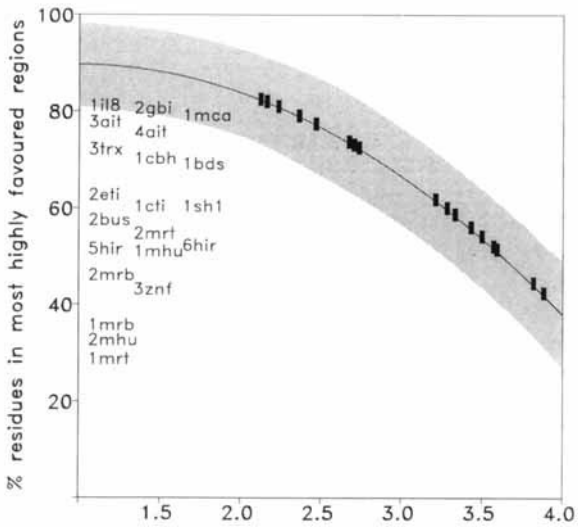


Fig. 13. Ramachandran plot distribution assessment for the 21 structures represented in PDB by single coordinate sets obtained by averaging and energy minimization. As described in Figure 6a the percentage of residues found in the most favored Ramachandran regions for each structure is matched with the regression line from crystallographic data which relates the percentages to resolution.

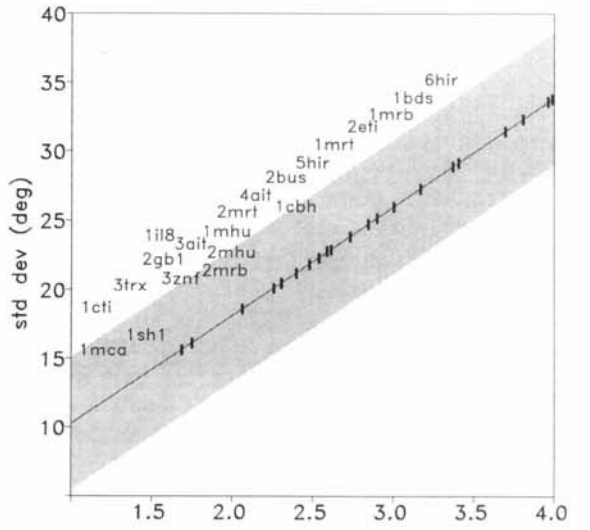


Fig. 14. Plot relating the χ_1 pooled standard deviations (Table VI) of the individual averaged structures to the crystal structure regression line. The captions identify the points by following an imaginary horizontal line.

CONCLUSIONS

In the above analysis the approach which had previously been applied to X-ray crystal structures was applied to those determined by NMR spectroscopy. In the course of the study this approach had to be modified to account for the presentation of the latter as an ensemble of models having multiple conformations. In considering NMR-derived structures it is vital to take into account the fact that parts of the structure usually on the surface of the protein are not well determined, either through inherent flexibility or lack of data. In analyzing the conformation and stereochemical quality, a single global measure can provide only a very rough guide and it is more informative to look at each residue separately, or at the core and flexible segments of chain separately, especially considering the correlation with accessibility. Taking such factors into account, the data provide evidence that the detailed conformation of the main chain (especially the peptide linkage) is difficult to determine accurately by NMR, although *J*-coupling constant data and stereospecific assignments provide better constraints. A major problem until recently has been how to apportion this heterogeneity between lack of experimental information and actual mobility. Now with increasing use of heteronuclear techniques there is the possibility of gaining a measure of inherent local mobility from the use of ^{15}N–^{1}H relaxation times.

Increased numbers of stereospecific assignments will inevitably be an important contributory factor to the achievement of a more precise definition of the structure, particularly side chain conformations. In this connection we have observed that determinations using database libraries of different torsion angle combinations give a degree of unique occupancy of side chain rotamers comparable to that ob-

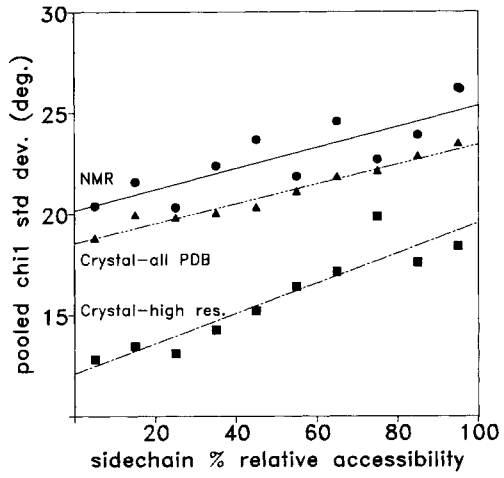


Fig. 15. Plots of pooled standard deviations of χ_1 rotamers (g^-,t,g^+) versus side chain relative percentage accessibilities. The data from the 21 NMR structures and two crystal datasets are shown. At all levels of accessibility the high-resolution (≤2.0 Å), well-refined (*R*-factor≤0.20) crystal dataset of 30 structures exhibits a very tight clustering of their χ_1 angles within their energy wells. The NMR structures display a pattern which more closely resembles that shown by the crystal dataset of 330 structures, which range in resolution from 1.0 to 3.5Å and vary greatly in degree of refinement.

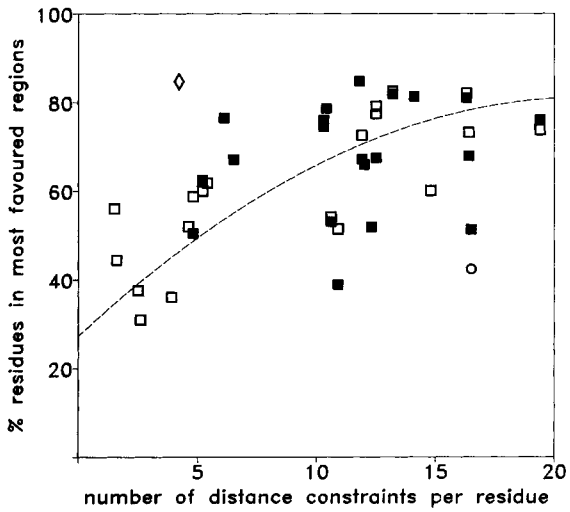


Fig. 16. Plot of percentage residues in Ramachandran map CORE regions versus the number of distance constraints per residue for the dataset listed in Table I. 1mca was excluded since the structure was derived by homology model building based on the 1il8 coordinates. The ensembles are shown by a solid square (■) and the minimized average structures by an open square (□). The complement factor (1c5a) which was homology modeled on the crystal structure of 1c3a is shown by a diamond (◇). The zinc finger (3znf) shows anomalous behavior and is marked by an open circle (○). Many of its ϕ,ψ angles appear close to the CORE periphery and none is in the disallowed regions.

served in crystal structures. In these, the few multiple occupancies such as do occur are mainly found in exposed surface residues.

It has to be noted that the dataset studied was a small one and included many "first generation" structures. Nevertheless the backbone is generally seen to be well traced, and indeed one of the NMR structures exposed serious error in an earlier crystal structure.[8] The detailed ϕ,ψ values, however, show much greater variation overall and less clustering than they do in crystal structures, although in secondary structure elements they are better defined. This is in part because additional restraints are often added once α-helices and β-sheets have been identified from the NMR data.

Both main chain and side chains are better defined in the interior, although considerable variation exists across the dataset. In general, for the protein core the result is comparable to a medium resolution (2.0–2.3 Å) crystal structure, although disorder toward the surface is often much more pronounced in NMR than in X-ray. Surface disorder is always likely to be greater in solution than in the crystal and will also depend on the nature of the molecule. The ability of NMR to study proteins in solution is a major advantage, and we can expect the improved techniques to reveal details about the conformational flexibility of the surface of the protein. This is of fundamental importance in increasing our understanding of molecular recognition between molecules, and the role of flexibility in this recognition.

This study of many protein structures determined by NMR has highlighted recent improvements in methodology, which provide ever more constraints to help determine the 3D structure more accurately. A careful analysis of geometry and local conformation, as presented here and mindful of conformational heterogeneity, is a useful tool to highlight possible errors, or regions of unusual conformation which may be of biological interest.

We are currently developing a program suite PROCHECK-NMR which will implement some of the checks for an ensemble of structures.

ACKNOWLEDGMENTS

M.W.M. is supported by a Science and Engineering Research Council Studentship. Our thanks to Gail Hutchinson, Roman Laskowski, and Mike Sutcliffe for help with programs, and to Jim Feeney, Mark Carr, Paul Driscoll, Iain Campbell, and Kurt Wüthrich for helpful discussions.

REFERENCES

1. Wagner, G., Hyberts, S.G., Havel, T.F. NMR structure determination in solution—a critique and comparison with X-ray crystallography. Annu. Rev. Biophys. Biomol. Struct. 21:167–198, 1992.
2. Billeter, M. Comparison of protein structures determined by NMR in solution and by X-ray diffraction in single crystals. Quart. Rev. Biophys. 25:325–377, 1992.
3. Havel, T., Wüthrich, K. A distance geometry program for determining the structures of small proteins and other macromolecules from nuclear magnetic resonance mea-

surements of intramolecular ^1H–^1H proximities in solution. Bull. Math. Biol. 46:673–698, 1984.
4. Clore, G.M., Brünger, A.T., Karplus, M., Gronenborn, A.M. Application of molecular-dynamics with interproton distance restraints to 3-dimensional protein structure determination—a model study of crambin. J. Mol. Biol. 191: 523–551, 1986.
5. Billeter, M., Kline, A.D., Braun, W., Huber, R., Wüthrich, K. Comparison of the high-resolution structures of the α-amylase inhibitor Tendamistat determined by nuclear magnetic resonance in solution and by X-ray diffraction in single crystals. J. Mol. Biol. 206:677–687, 1989.
6. Clore, G.M., Gronenborn, A.M. Comparison of the solution nuclear magnetic resonance and X-ray crystal structures of human recombinant interleukin-1β. J. Mol. Biol. 221: 47–53, 1991.
7. Berndt, K.D., Güntert, P., Orbons, L.P.M., Wüthrich, K. Determination of a high-quality nuclear magnetic resonance solution structure of the bovine pancreatic trypsin inhibitor and comparison with three crystal structures. J. Mol. Biol. 227:757–775, 1992.
8. Braun, W., Vašák, M., Robbins, A.J., Stout, C.D., Wagner, G., Kägi, J.H.R., Wüthrich, K. Comparison of the NMR solution structure and the X-ray crystal-structure of rat metallothionein-2. Proc. Natl Acad. Sci. U.S.A. 89:10124–10128, 1992.
9. van Nuland, N.A.J., van Dijk, A.A., Dijkstra, K., van Hoesel, F.H.J., Scheek, R.M., Robillard, G.T. Three dimensional ^{15}N—^1H—^1H and ^{15}N—^{13}C—^1H nuclear-magnetic resonance studies of HPr a central component of the phospho*enol*pyruvate-dependent phosphotransferase system from *Escherichia coli*. Eur. J. Biochem. 203:483–491, 1992.
10. El-Kabbani, O.A.L., Waygood, E.B., Delbaere, L.T.J. Tertiary structure of histidine-containing protein of the phosphoenolpyruvate:sugar phhosphotransferase system of *Escherichia coli*. J. Biol. Chem. 262:12927–12929, 1987.
11. Clore, G.M., Gronenborn, A.M. Comparison of the solution nuclear magnetic resonance and crystal structures of interleukin-8. J. Mol. Biol. 217:611–620, 1991.
12. Williams, R.J.P. Protein dynamics studied by NMR. Eur. Biophys. J. 21:393–401, 1993.
13. Huber, R., Scholze, H., Pasques, E.P., Deisenhofer, J. Crystal structure analysis and molecular model of human C3a anaphylatoxin. Hoppe Seyler's Z. Phys. Chem. 361:1389–1399, 1980.
14. Zuiderweg, E.R.P., Nettesheim, D.G., Fesik, S.W., Olejniczak, E.T., Mandecki, W., Mollison, K.W., Greer, J., Carter, G.W. Studies of the 3D structure of complement protein C5a mutants by 2D and 3D NMR. In: "Frontiers of NMR in Molecular Biology." Live, D., Armitage, I.M., Patel, D., eds. New York: Wiley-Liss, 1990: 75–87.
15. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The Protein Data Bank: A computer based archival file for macromolecular structures. J. Mol. Biol. 122:535–542, 1977.
16. Morris, A.L., MacArthur, M.W., Hutchinson, E.G., Thornton, J.M. Stereochemical quality of protein structure coordinates. Proteins 12:345–364, 1992.
17. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
18. Laskowski, R.A., MacArthur, M.W., Moss, D.S., Thornton, J.M. PROCHECK: A program to check the stereochemical quality of protein structures. J. Appl. Crystallogr. 26:283–290, 1993.
19. Lee, B., Richards, F.M. The interpretation of protein structures: Estimation of static accessibility. J. Mol. Biol. 55: 379–400, 1971.
20. Hubbard, S.J. Analysis of protein-protein molecular recognition. Ph.D. thesis, 1992.
21. Chothia, C. The nature of the accessible and buried surfaces in proteins. J. Mol. Biol. 105:1–14, 1976.
22. Pastore, A., Saudek, V., Ramponi, G., Williams, R.J.P. Three-dimensional structure of acylphosphatase. Refinement and structure analysis. J. Mol. Biol. 224:427–440, 1992.
23. Widmer, H., Billeter, M., Wüthrich, K. Three-dimensional structure of the neurotoxin ATX 1a from *Anemonia sulcata* in squeous solution determined by nuclear magnetic resonance spectroscopy. Proteins 6:357–371, 1989.
24. Williamson, M.P., Madison, V.S. Three-dimensional structure of porcine $C5a_{desArg}$ from ^1H nuclear magnetic resonance data. Biochemistry 29:2895–2905, 1990.
25. Xia, T.-H., Bushweller, J.H., Sodano, P., Billeter, M., Björnberg, O., Holmgren, A., Wüthrich, K. NMR structure of oxidized *Escherichia coli* glutaredoxin—comparison with reduced *E. coli* glutaredoxin and functionally related proteins. Protein Sci. 3:310–321, 1992.
26. Sodano, P., Xia, T.-H., Bushweller, J.H., Björnberg, O., Holmgren, A., Billeter, M., Wüthrich, K. Sequence-specific ^1H n.m.r. assignments and determination of the three-dimensional structure of reduced *Escherichia coli* glutaredoxin. J. Mol. Biol. 221:1311–1324, 1991.
27. Billeter, M., Qian, Y., Otting, G., Müller, M., Gehring, W.J., Wüthrich, K. Determination of the three-dimensional structure of the *Antennapedia* homeodomain from *Drosophila* in solution by ^1H nuclear magnetic resonance spectroscopy. J. Mol. Biol. 214:183–197, 1990.
28. Dyson, H.J., Gippert, G.P., Case, D.A., Holmgren, A., Wright, P.E. Three-dimensional solution structure of the reduced form of *Escherichia coli* thioredoxin determined by nuclear magnetic resonance spectroscopy. Biochemistry 29:4129–4136, 1990.
29. Lee, M.S., Gippert, G.P., Soman, K.V., Case, D.A., Wright, P.E. Three-dimensional solution structure of a single zinc finger DNA-binding domain. Science 245:635–637, 1989.
30. Güntert, P., Qian, Y.Q., Otting, G., Müller, M., Gehring, W., Wüthrich, K. Structure determination of the *Antp*(C39→S) homeodomain from nuclear magnetic resonance data in solution using a novel strategy for the structure calculation with the programs DIANA, CALIBA, HABAS, and GLOMSA. J. Mol. Biol. 217:531–540, 1991.
31. Summers, M.F., South, T.L., Kim, B., Hare, D.R. High-resolution structure of HIV zinc fingerlike domain via a new NMR-based distance geometry approach. Biochemistry 29:329–340, 1990.
32. Kline, T.P., Brown, F.K., Brown, S.C., Jeffs, P.W., Kopple, K.D., Mueller, L. Solution structures of human transforming growth factor-α derived from ^1H NMR data. Biochemistry 29:7805–7813, 1990.
33. Williamson, M.P., Havel, T.F., Wüthrich, K. Solution conformation of proteinase inhibitor IIA from bull seminal plasma by ^1H nuclear magnetic resonance and distance geometry. J. Mol. Biol. 182:295–315, 1985.
34. Gronenborn, A.M., Filpula, D.R., Essig, N.Z., Achari, A., Whitlow, M., Wingfield, P.T., Clore, G.M. A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein-G. Science 253:657–661, 1991.
35. Kline, A.D., Braun, W. Determination of the complete three-dimensional structure of the α-amylase inhibitor tendamistat in aqueous solution by nuclear magnetic resonance. J. Mol. Biol. 204:675–724, 1988.
36. Driscoll, P.C., Gronenborn, A.M., Beress, L., Clore, G.M. Determination of the three-dimensional solution structure of the antihypertensive and antiviral protein BDS-I from the sea anemone *Anemonia sulcata:* A study using nuclear magnetic resonance and hybrid distance geometry—dynamical simulated annealing. Biochemistry 28:2188–2198, 1989.
37. Kraulis, P.J., Clore, G.M., Nilges, M., Jones, T.A., Pettersson, G., Knowles, J., Gronenborn, A.M. Determination of the three-dimensional solution structure of the C-terminal domain of cellobiohydrolase I from *Trichoderma reesei*. A study using nuclear magnetic resonance and hybrid distance geometry—dynamical simulated annealing. Biochemistry 28:7241–7257, 1989.
38. Folkers, P.J.M., Clore, G.M., Driscoll, P.C., Dodt, J., Köhler, S., Gronenborn, A.M. Solution structure of recombinant hirudin and the Lys-47 → Glu mutant: A nuclear magnetic resonance and hybrid distance geometry—dynamical simulated annealing study. Biochemistry 28: 2601–2617, 1989.
39. Clore, G.M., Appella, E., Yamada, M., Matsushima, K., Gronenborn, A.M. Three-dimensional structure of interleukin 8 in solution. Biochemistry 29:1689–1696, 1990.
40. Fogh, R.H., Kem, W.R., Norton, R.S. Solution structure of neurotoxin I from the sea anemone *Stichodactyla helianthus* J. Biol. Chem. 265:13016–13028, 1990.

41. Forman-Kay, J.D., Clore, G.M., Wingfield, P.T., Gronenborn, A.M. High-resolution three-dimensional structure of reduced recombinant human thioredoxin in solution. Biochemistry 30:2685–2698, 1991.
42. Omichinski, J.G., Clore, G.M., Appella, E., Sakaguchi, K., Gronenborn, A.M. High-resolution three-dimensional structure of a single zinc finger from a human enhancer binding protein in solution. Biochemistry 29:9324–9334, 1990.
43. Holak, T.A., Gondol, D., Otlewski, J., Wilusz, T. Determination of the complete 3-dimensional structure of the trypsin inhibitor from squash seeds in aqueous solution by nuclear magnetic resonance and a combination of distance geometry and dynamical simulated annealing. J. Mol. Biol. 210:635–648, 1989.
44. Gronenborn, A.M., Clore, G.M. Modeling the three-dimensional structure of the monocyte chemo-attractant and activating protein MCAF/MCP-1 on the basis of the solution structure of interleukin-8. Protein Eng. 4:263–269, 1991.
45. Messerle, B.A., Schäffer, A., Vašák, M., Kägi, J.H.R., Wüthrich, K. Three-dimensional structure of human [$^{113}Cd_7$] methallothionein-2 in solution determined by nuclear magnetic resonance spectroscopy. J. Mol. Biol. 214: 765–779, 1990.
46. Arseniev, A., Schultze, P., Wörgötter, E., Braun, W., Wagner, G., Vašák, M., Kägi, J.H.R., Wüthrich, K. Three-dimensional structure of rabbit liver [Cd_7] metallothionein-2a in aqueous solution determined by nuclear magnetic resonance. J. Mol. Biol. 201:637–657, 1988.
47. Schultze, P., Wörgötter, E., Braun, W., Wagner, G., Vašák, M., Kägi, J.H.R., Wüthrich, K. Conformation of [Cd_7]-methallothionein-2 from rat liver in aqueous solution determined by nuclear magnetic resonance spectroscopy. J. Mol. Biol. 203:251–268, 1988.
48. Chiche, L., Gaboriaud, C., Heitz, A., Mornon, J.-P., Castro, B., Kollman, P.A. Use of restrained molecular dynamics in water to determine three-dimensional protein structure: Prediction of the three-dimensional structure of *Ecballium elaterium* trypsin inhibitor II. Proteins 6:405–417, 1989.
49. Braun, W., Gō, N. Calculation of protein conformations by proton-proton distance constraints. A new efficient algorithm. J. Mol. Biol. 186:611–626, 1985.
50. Weiner, P.K., Kollman, P.A. AMBER: Assisted model building with energy refinement—a general program for modelling molecules and their interactions. J. Comp. Chem. 2:287–303, 1981.
51. Brooks, B.R., Bruccoleri, R.E., Olafson, B.O., States, D.J., Swaminathan, S., Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J. Comp. Chem. 4:187–217, 1983.
52. Brünger, A.T., Kuriyan, J., Karplus, M. Crystallographic R factor refinement by molecular dynamics. Science 235: 458–460, 1987.
53. Havel, T.F., Wüthrich, K. A distance geometry program for determining the structures of small proteins and other macromolecules from nuclear magnetic resonance measurements of intramolecular ^1H-^1H proximities in solution. Bull. Math. Biol. 46:673–698, 1984.
54. Metzler, W.J., Hare, D.R., Pardi, A. Limited sampling of conformational space by the distance geometry algorithm: Implications for structures generated from NMR data. Biochem. 28:7045–7052, 1989.
55. Banks, K.M., Hare, D.R., Reid, B.R. Three-dimensional solution structure of a DNA duplex containing the *BclI* restriction sequence: two-dimensional NMR studies, distance geometry calculations, and refinement by back-calculation of the NOESY spectrum. Biochemistry 28:6996–7010, 1989.
56. Billeter, M., Engeli, M., Wüthrich, K. Interactive program for investigation of protein structures based on ^1H NMR experiments. J. Mol. Graph. 3:79, 1985.
57. Nilges, M., Clore, G.M., Gronenborn, A.M. ^1H NMR stereospecific assignments by conformational database searches. Biopolymers 29:813–822, 1990.
58. Schaumann, T., Braun, W., Wüthrich, K. The program FANTOM for energy refinement of polypeptides and proteins using a Newton-Raphson minimizer in torsion angle space. Biopolymers 29:679–694, 1990.
59. Gros, P., Fujinaga, M., Dijkstra, B.W., Kalk, K.H., Hol, W.G.J. Crystallographic refinement by incorporation of molecular dynamics: Thermostable serine protease thermitase complexed with eglin c. Acta Crystallogr. B45:488–499, 1989.
60. Mardia, K.V. "Statistics of Directional Data." London: Academic Press, 1972.
61. Smith, J.L., Hendrickson, W.A., Honzatko, R.B., Sheriff, S. Structural heterogeneity in protein crystals. Biochemistry 25:5018–5027, 1986.
62. Thornton, J.M., Disulphide bridges in globular proteins. J. Mol. Biol. 151:261–287, 1981.
63. Richardson, J.S. The anatomy and taxonomy of protein structure. Adv. Protein Chem. 34:167–339, 1981.
64. Wagner, G., Nirmala, N.R., Montelione, G.T., Hyberts, S. Static and dynamic aspects of protein structure. In: "Frontiers of NMR in Molecular Biology." Live, D., Armitage, I.M., Patel, D., eds. New York: Wiley-Liss, 1990: 129–143.
65. Sutcliffe, M.J. Representing an ensemble of NMR-derived protein structures by a single structure. Protein Sci. 2:936–944, 1993.
66. Laskowski, R.A., Moss, D.S., Thornton, J.M. Main-chain bond lengths and bond angles in protein structures. J. Mol. Biol. 231:1049–1067, 1993.
67. Wüthrich, K. "NMR of Proteins and Nucleic Acids." New York: John Wiley & Sons, 1986.