[3] Molecular Fragment Replacement Approach to Protein Structure Determination by Chemical Shift and Dipolar Homology Database Mining

By GEORG KONTAXIS, FRANK DELAGLIO, and AD BAX

Abstract

A novel approach is described for determining backbone structures of proteins that is based on finding fragments in the protein data bank (PDB). For each fragment in the target protein, usually chosen to be 7–10 residues in length, PDB fragments are selected that best fit to experimentally determined one-bond heteronuclear dipolar couplings and that show agreement between chemical shifts predicted for the PDB fragment and experimental values for the target fragment. These fragments are subsequently refined by simulated annealing to improve agreement with the experimental data. If the lowest-energy refined fragments form a unique structural cluster, this structure is accepted and side chains are added on the basis of a conformational database potential. The sequential backbone assembly process extends the chain by translating an accepted fragment onto it. For several small proteins, with extensive sets of dipolar couplings measured in two alignment media, a unique final structure is obtained that agrees well with structures previously solved by conventional methods. With less dipolar input data, large, oriented fragments of each protein are obtained, but their relative positioning requires either a small set of translationally restraining nuclear Overhauser enhancements (NOEs) or a protocol that optimizes burial of hydrophobic groups and pairing of β-strands.

Introduction

With the completion of the sequencing of the human and many other genomes and the availability of an abundance of protein sequence data, there is a strong demand for rapid determination of tertiary protein structures. There are two main experimental avenues toward obtaining atomic resolution protein structures: X-ray crystallography and solution state nuclear magnetic resonance (NMR) spectroscopy. The process of structure determination by X-ray crystallography is already quite streamlined due to the availability of robotics for optimizing crystallization conditions, high-intensity synchrotron radiation sources, and standardized, semiautomated

0076-6879/05 $35.00

analysis software. Structure determination by NMR spectroscopy on the other hand is still a time-consuming and labor-intensive process with a turnaround time typically on the order of several months, which additionally requires ^{15}N, ^{13}C, and, for larger proteins, ^{2}H isotopic enrichment. Usually, an NMR structure determination project proceeds in several stages: assignment of backbone resonances using pairs of now standard triple-resonance experiments, assignment of side chain resonances using ^{13}C-, ^{1}H- or ^{15}N-mediated TOCSY- and COSY-type experiments, followed by assignment of NOE cross-peaks, and structure calculation.

The introduction of facile methods for weakly aligning proteins relative to the magnetic field now also allows measurement of residual dipolar couplings (RDCs). In favorable cases, the alignment necessary for a non-vanishing dipolar interaction can be imposed on the solute macromolecules directly by the magnetic field (Bothner-by *et al*., 1985; Kung *et al*., 1995; Tjandra *et al*., 1996; Tolman *et al*., 1995), but more commonly an anisotropic aqueous medium is used. Many such media are now available, including lyotropic liquid crystalline solutions of phospholipid bicelles (Tjandra and Bax, 1997), Pf1, *fd*, or TM phage particles (Clore *et al*., 1998b; Hansen *et al*., 1998), cellulose crystallites (Fleming *et al*., 2000), and polyethylene glycol (Ruckert and Otting, 2000) or cetylpyridinium halide-based bilayers (Barrientos *et al*., 2000; Prosser *et al*., 1998). Anisotropically compressed, low-density polyacrylamide gels (Chou *et al*., 2001a; Ishii *et al*., 2001; Meier *et al*., 2002; Sass *et al*., 2000; Tycko *et al*., 2000; Ulmer *et al*., 2003) and suspensions of magnetically oriented purple membrane fragments (Koenig *et al*., 1999; Sass *et al*., 1999) also have proven useful for this purpose.

RDCs are global parameters in the sense that they restrain the orientations of the corresponding dipolar interaction vectors all relative to a single reference frame, often referred to as the principal axis frame of the alignment tensor. In this respect, they differ in nature from NOEs and dihedral restraints derived from J couplings, which report on atomic positions relative to one another. Besides improving local geometry (Chou *et al*., 2001b; Tjandra *et al*., 1997), RDCs have been shown to be highly useful for determining the relative orientation of individual domains in multisubunit proteins, nucleic acids, and their complexes (Braddock *et al*., 2001; Clore, 2000; Lukavsky *et al*., 2003).

In the principal frame of the alignment tensor, the dipolar coupling is given by

$$D_{ij}(\theta,\phi) = D_{\mathrm{a}}[(3\cos^2\theta_{ij} - 1) + 3/2\, R \sin^2\theta_{ij}\cos^2\phi_{ij}] \qquad (1)$$

where θ and ϕ are the polar angles of the dipolar interaction vector, r_{ij}, in the alignment frame; D_{a} is the magnitude of the alignment tensor, which

includes constants related to the magnetogyric ratio and internuclear distance of nuclei i and j, and R is the rhombicity of the alignment tensor (Bax *et al.*, 2001). Clearly, with a single experimental $D_{ij}(\theta,\phi)$ value, and two variable parameters, in general an infinite number of (θ,ϕ) solutions exist. This degeneracy may be partly lifted if RDCs in a different alignment medium and with a different, independent alignment tensor are available (Ramirez and Bax, 1998). However, even in this case, a vector orientation can never be distinguished from its inverse because both orientations lead to the same dipolar coupling. Only once the ''handedness'' of a local element of structure that involves several dipolar interactions in at least two alignment frames is known, can the absolute orientation of such a fragment and thereby of its vectors be determined (Al-Hashimi *et al.*, 2000). As a consequence, when attempting to build a full protein structure that simultaneously satisfies Eq. (1) for all dipolar interactions, the number of false minima scales exponentially with this number of couplings. Solving this problem by means of a ''brute force'' simulated annealing or Monte Carlo program on a full protein has proven very difficult. However, when first assembling local substructures, this problem is no longer intractible, and a number of recently proposed approaches rely on this principle (Andrec *et al.*, 2001; Delaglio *et al.*, 2000; Hus *et al.*, 2000; Rohl and Baker, 2002).

Our present approach represents a much improved and more stable version of the molecular fragment replacement (MFR) method described earlier, which derived backbone torsion angles from searching the PDB for seven-residue peptide fragments that fit experimental dipolar couplings in a fragment of the target protein (Delaglio *et al.*, 2000). In the original procedure, a starting model was first built using these backbone torsion angles and subsequently refined by optimizing agreement between this model and the full set of dipolar couplings. Although the method results in reasonable structures, provided that a nearly complete set of dipolar coupling is available, convergence to a satisfactory final structure and accuracy of its local details remain limited by the quality of the fragments of the original search. However, in favorable cases, substantial regions of a protein can be assembled from such data, even in cases in which all dipolar couplings and assignments are derived from a single experiment (Zweckstetter and Bax, 2001).

Our MFR approach is related to work by Annila *et al.* (1999), who proposed to use dipolar couplings for finding structurally homologous proteins in the PDB. Instead of searching for complete proteins, the MFR program searches the PDB for structural homology for only 7–10 residues at a time. The idea of using small substructures from a database of representative protein structures as ''templates'' for building a structure was pioneered by Jones and co-workers and has been very successful in

X-ray crystallography (Jones and Thirup, 1986). The approach has also been applied to solving structures on the basis of NOEs, where it searches a database for substructures compatible with the experimental NOEs (Kraulis and Jones, 1987). However, because only short and medium range NOEs can be used in the search process, obtaining the correct tertiary fold remains very difficult with such a method. Other approaches relying on database substructures have also been described in recent years. Work by the Baker group (Rohl and Baker, 2002) relies on selecting a large number of database fragments that are roughly compatible with the experimental parameters measured for the corresponding target fragment and then using efficient Monte Carlo methods to assemble these fragment into a common structure with reasonable packing properties, where the fragments retain an orientation needed to satisfy dipolar coupling restraints. A method proposed by Andrec *et al.* (2001) is similar in spirit to our own MFR method but uses ''postprocessing'' to distinguish correct from incorrect fragments by comparing them with the overlapping region of an adjacent fragment. Using a so-called bounded-tree search, self-consistent sets of overlapping fragments can be identified relatively rapidly, resulting in a backbone structure.

A different approach to building complete protein backbone structures from dipolar couplings, which does not rely on a database for finding suitable substructures, has been proposed by Hus *et al.* (2000) and Giesen *et al.* (2003). It is conceptually somewhat similar to approaches pursued in determining polypeptide structure on the basis of ^{15}N–^{1}H dipolar couplings derived from solid-state NMR measurements (Brenneman and Cross, 1990; Marassi and Opella, 1998; Nishimura *et al.*, 2002; Wu *et al.*, 1995) and conducts a systematic search in Cartesian space when adding a peptide plane to the chain. The approach requires a very complete set of dipolar couplings when applied *de novo*, but it has other applications too. For example, it was shown to be particularly powerful for pinpointing the precise structural differences in the backbone at the active site of a 27-kDa enzyme [methionine sulfoxide reductase (MsrA) from *Erwinia chrysanthemi*] and its *Escherichia coli* homologue, for which an X-ray structure was available (Beraud *et al.*, 2002). Conceptually, Hus' method shares features with a method devised by Mueller *et al.* (2000), which determines the (usually 4-fold degenerate) peptide plane orientations compatible with experimental dipolar couplings prior to finding a chain compatible with these orientations. Finally, Fowler *et al.* (2000) demonstrated it is feasible to get information on the fold from ^{15}N–$^{1}H^{N}$, $^{1}H^{N}$–$^{1}H^{N}$, and $^{1}H^{N}$–$^{1}H^{\alpha}$ couplings without the need for ^{13}C enrichment.

Our improved MFR approach, which we refer to as MFR+, represents a much more versatile and stable version of the original MFR method. The

method differs from all previous database substructure methods by introduction of an intermediate step where the fragments are refined with respect to the experimental observables (shifts, couplings, and possibly short and medium range NOEs or torsion angle restraints) prior to their final selection and incorporation into a structure. It also utilizes the unique advantage of dipolar tensor parameters to maintain reasonable orientations for each fragment at all stages of the substructure assembly. The user has complete freedom to specify weighting factors used and to define the minimal criteria for deeming a selection to be ''reliable.'' With the default settings, and with dipolar couplings available from two different media, the program can rapidly generate backbone structures for the proteins ubiquitin and GB3 that are considerably less than 1 Å from their true structure. With less experimental data, accurate partial structures can be obtained, which subsequently can be used to assemble the structure either manually or by using docking algorithms (Clore, 2000; Clore and Schwieters, 2003).

Description of the MFR+ Method

The routines to conduct the homology search, visualize the results, and the assembly of a structure were written in the Tcl/Tk language and use ''NMRWish,'' an in-house version of the Tcl/Tk interpreter ''wish,'' which has been customized by the addition of routines to handle and manipulate tables, databases, and PDB format files. It can perform chemical shift (CS) and dipolar coupling (DC) simulations, carry out coordinate alignments, handle restraints (NOE, dihedral, CS, DC, J), and also includes facilities for structure calculation and a molecular dynamics engine, named DYNAMO (available through http://spin.niddk.nih.gov/bax).

The MFR+ method uses chemical shifts and dipolar couplings as input for database mining and subsequent model building. As mentioned earlier, MFR is closely related to a method for protein fold recognition based on ''dipolar homology,'' where a database is searched for either a complete protein or a protein domain, compatible with a given set of dipolar couplings (Annila *et al.*, 1999). This ''protein fold recognition'' procedure typically requires that some degree of sequence homology is available, so that a sequence alignment can be performed, and it is known which residue in the database protein corresponds to which residue in the target protein. The MFR approach is not subject to this requirement and simply searches for all substructures in the PDB that are compatible with dipolar couplings and chemical shifts measured for a given fragment of the target protein. Provided the fragment is chosen to be small enough in size, invariably there are a large number of fragments in the database that exhibit a reasonable fit to the experimental parameters (Du *et al.*, 2003), but frequently they do

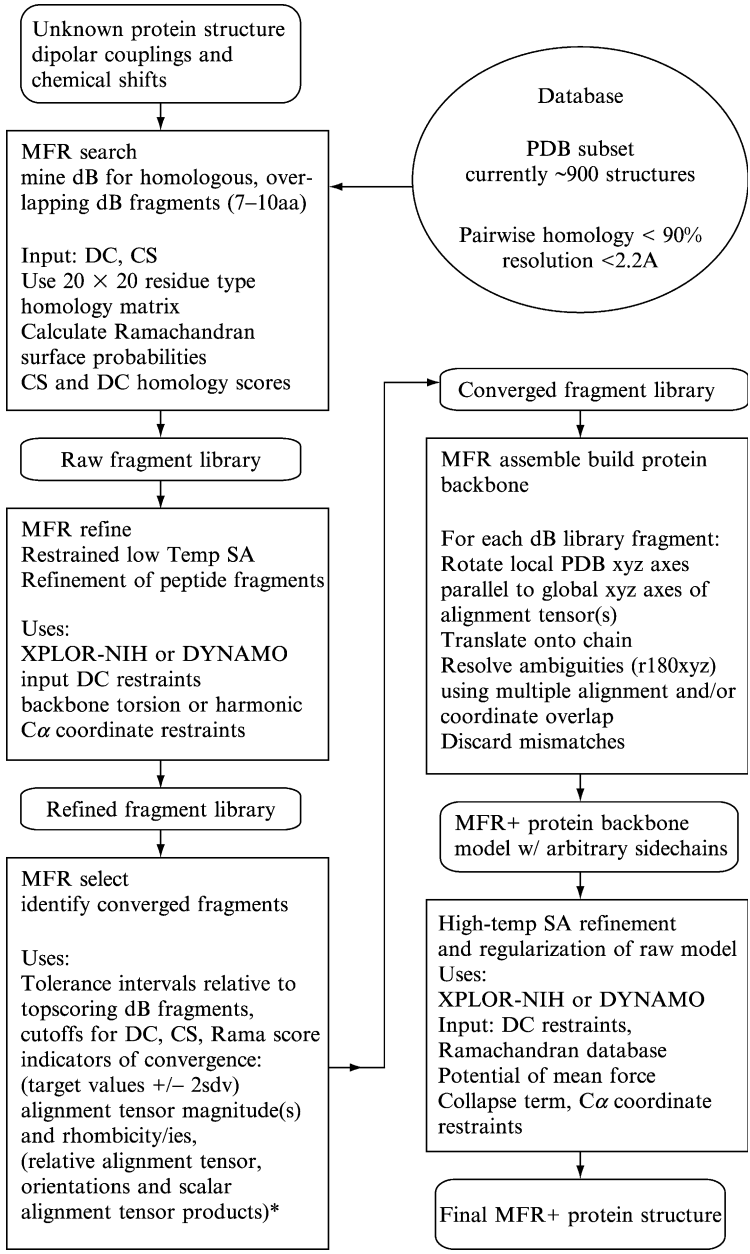
not represent a unique structural cluster. For such small fragment sizes, the experimental information therefore is less discriminating in defining the correct database substructures than when searching for full proteins. However, when choosing too large a fragment length, it is possible that no adequate hits may be found in the database. In practice, we find a fragment size in the 7–10 residue range to be optimal for MFR+, with larger fragments preferred if couplings are available in only a single medium, or if only relatively few couplings (two or less) are available per residue.

The MFR approach is based on the premise that if two fragments can be fit to the same set of dipolar couplings and exhibit similar secondary chemical shift (i.e., deviation from random coil shifts) patterns, they are likely to have similar structures. By comparing with one another the best ''hits'' found in the database for a given set of experimental dipolar couplings and secondary chemical shifts, a degree of certainty is obtained about how well the database substructures represent the corresponding fragment in the query protein.

Several stages can be distinguished in the MFR+ method:

1. Search of the database for fragments with dipolar, chemical shift, and sequence homology; optionally, other NMR observables based on local structure such as J couplings or NOEs can also be used.
2. Low temperature simulated annealing refinement of an initial set of best matches found in the search, followed by automated selection of a subset of these refined structures.
3. Assembly of a tertiary structure.

A flow diagram of this procedure is presented in Fig. 1, and a detailed discussion of the steps involved will be presented below. As mentioned above, for finding the structurally feasible peptide conformations, we rely on the PDB. Although for very short fragments, a small number of proteins is sufficient to represent all possible conformations found in nature (Jones and Thirup, 1986; Jones *et al*., 1991), for larger peptide fragments such as those used in our MFR+ approach this is no longer the case, and a database as large as possible is desirable. At the same time, structures with high sequence homology, or low-resolution structures with considerable uncertainty in their coordinates, carry little useful information. As a compromise, we have filtered the PDB to retain only X-ray–derived structures, solved at a resolution of $\leq$2.2 Å, and to discard structures that are more than $\sim$90% identical to an already selected protein. This final database then contains 893 entries (distributed as part of the MFR+ package). For testing the MFR+ program, which is carried out on proteins for which the structure already is available in this database (e.g., ubiquitin or GB3),


Unknown protein structure
dipolar couplings and
chemical shifts
Database
PDB subset
currently ~900 structures
Pairwise homology < 90%
resolution <2.2A
MFR search
mine dB for homologous, over-
lapping dB fragments (7–10aa)
Input: DC, CS
Use 20 × 20 residue type
homology matrix
Calculate Ramachandran
surface probabilities
CS and DC homology scores
Raw fragment library
MFR refine
Restrained low Temp SA
Refinement of peptide fragments
Uses:
XPLOR-NIH or DYNAMO
input DC restraints
backbone torsion or harmonic
Cα coordinate restraints
Refined fragment library
MFR select
identify converged fragments
Uses:
Tolerance intervals relative to
topscoring dB fragments,
cutoffs for DC, CS, Rama score
indicators of convergence:
(target values +/– 2sdv)
alignment tensor magnitude(s)
and rhombicity/ies,
(relative alignment tensor,
orientations and scalar
alignment tensor products)*
Converged fragment library
MFR assemble build protein
backbone
For each dB library fragment:
Rotate local PDB xyz axes
parallel to global xyz axes of
alignment tensor(s)
Translate onto chain
Resolve ambiguities (r180xyz)
using multiple alignment and/or
coordinate overlap
Discard mismatches
MFR+ protein backbone
model w/ arbitrary sidechains
High-temp SA refinement
and regularization of raw model
Uses:
XPLOR-NIH or DYNAMO
Input: DC restraints,
Ramachandran database
Potential of mean force
Collapse term, Cα coordinate
restraints
Final MFR+ protein structure

the relevant identical and/or highly homologous structures are excluded from the database search.

Search Procedure

As a first step, the protein under study (target protein) is broken up into small, overlapping peptide fragments, typically 7–10 residues in length. So, for a 100-residue protein, 91–94 such query fragments are generated. In a first, preliminary evaluation, singular value decomposition (SVD) (Losonczi *et al*., 1999; Sass *et al*., 1999) is used to determine how well these couplings fit to any of the roughly 180,000 substructures in the database. In practice, for reliable evaluation at least a dozen or more couplings per fragment are required, as there are five degrees of freedom (i.e., five independent elements of the alignment tensor) in the fitting process. In its simplest and fastest mode of operation, only the 10 or 20 best-fitting fragments are retained.

The SVD fitting procedure, while fast, does not take advantage of information that frequently is known prior to the start of the search, such as the magnitude and rhombicity of the alignment tensor, which can be obtained by inspection of the ''powder pattern'' (Bryce and Bax, 2004; Clore *et al*., 1998a). Alternatively, these two parameters frequently can be derived at good accuracy by fitting dipolar couplings, measured for a stretch of residues that carries a clear α-helical chemical shift signature, to an idealized model α-helix. Such information can be incorporated into nonlinear least-squares optimization techniques, but carrying out such an optimization many thousands of times per query fragment is exceedingly slow. As a compromise, an alternate strategy is to retain a larger subset of the best SVD results, e.g., 1000, and subject this subset to restrained least-squares minimization, using prior knowledge of magnitude and rhombicity. As an interesting side note, the MFR SVD search procedure itself provides an automated method for estimating tensor magnitude and rhombicty, before the structure is known, since the collection of best-fitting fragments found by SVD will have magnitudes and rhombicities that cluster around the true values for the intact target protein (Fig. 2). In practice, average values over the collection of fragments are weighted to give the highest

FIG. 1. Flow diagram of the MFR+ protein structure determination protocol. aa, amino acid; CS, chemical shift; dB, database; DC, dipolar coupling; PDB, Protein Data Bank; Rama, Ramachandran map; sdv, standard deviation; SA, simulated annealing. The selection criteria involving relative alignment tensor orientation, marked with an asterisk, applies only to the case in which multiple alignment media are used.

importance to regions where the fragments show the greatest structural consensus. Results of this method are shown in Fig. 2, which indicates that tensor parameters for 11 different samples could be predicted with a root mean square (RMS) error of less than 5%. For the case in which multiple media are used for a given protein, the same procedure can also be used for determining the relative orientation of the alignment tensors.

During the database search, a dipolar coupling score F_{DC} is used to measure the goodness of fit between a target fragment's set of N observed dipolar couplings and the values calculated for a database substructure. The score is a simple weighted root-mean-square deviation (RMSD):

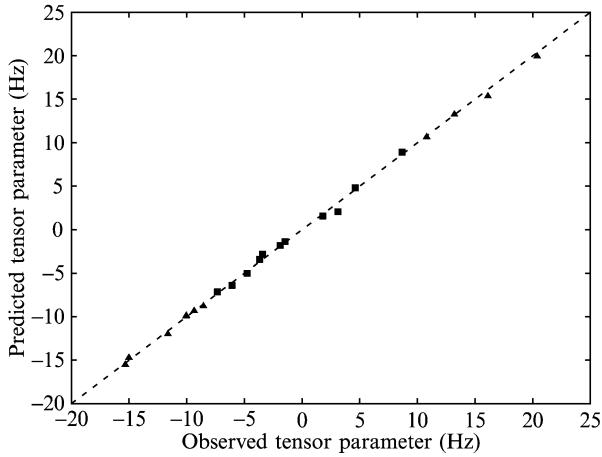


FIG. 2. Plot of dipolar coupling tensor parameters obtained from SVD fits of experimental dipolar couplings to known structures vs. values predicted automatically by the MFR search procedure, without prior knowledge of the structure. (▲) tensor magnitude D_a; (■) rhombicity, R, times tensor magnitude D_a. Values are shown for 11 different samples, comprising the proteins ubiquitin (two media), DinI (two media), Gb3 (five media), and γ-crystallin (two media; J. Wu, personal communication). All values are scaled relative to ${}^1D_{HN}$ couplings. The MFR tensor estimates are weighted averages of the SVD-derived tensor parameters obtained for database fragments selected by the MFR search. For calculating this weighted average, first for each set of selected database fragments, corresponding to target protein fragment j, linear averages, $<D_a(j)>$ and $<R(j)>$, are calculated. Then, the weighted average over the chain is calculated as $<D_a> = \sum_{j=1,\ldots,N} w_j <D_a(j)> / \sum_{j=1,\ldots,N} w_j$, and similarly for $<R>$, where the summation extends over all N fragments. The weighting factor is given by $w_j = \exp(-\alpha q_j)$, where $\alpha = 3$ Å^{-1}, and q_j is the backbone RMSD over the collection of selected database segments for fragment j. Typical values found for q are 0.1–0.3 Å for helical regions, 0.2–0.5 Å for well-defined β-sheet regions, and 0.8–2.0 Å for regions where MFR results are ambiguous. In all cases, tensor parameters are predicted to an accuracy of better than 0.5 Hz.

$$F_{\mathrm{DC}} = \left\{ (1/N) \sum_{ij} [w_{ij}(D_{ij}^{\mathrm{obs}} - D_{ij}^{\mathrm{calc}})]^2 \right\}^{1/2} \tag{2}$$

In practice, the weighting factor w_{ij} is selected to scale all types of couplings into a similar range, commonly as $(\gamma_{\mathrm{N}}\gamma_{\mathrm{H}}/r_{\mathrm{NH}}^3)/(\gamma_i\gamma_j/r_{ij}^3)$. The weighting factor w_{ij} can also be scaled to include adjustment for the estimated uncertainties in the measured couplings. In general, no such adjustment is needed if the estimated random error in a given type of coupling is less than about $\pm$ 15% of the applicable D_{a} value.

The chemical shift score for each fragment is calculated by comparing observed values with those expected for the database substructure. These latter values are derived using (ϕ,ψ)-dependent chemical shift surfaces, obtained from the TALOS chemical shift database (Cornilescu *et al.*, 1999).

For each (ϕ,ψ), the program looks up the average secondary chemical shift on this (ϕ,ψ) surface and its rms spread, $\sigma(\phi,\psi)$. The chemical shift score is then defined as

$$F_{\mathrm{CS}} = \left\{ \sum_i [(\mathrm{CS}_i^{\mathrm{obs}} - \mathrm{CS}_i^{\mathrm{calc}})/\sigma_i(\phi,\psi)]^2/N \right\}^{1/2} \tag{3}$$

Other terms can be used too in the search for suitable database fragments. Considering that the search does not account for residue type, but that certain residues in the database (e.g., Gly) have very different (ϕ,ψ) distributions from those seen for other residues, one such additional term is referred to as ''Ramachandran surface quality,'' or ''ϕ,ψ surface quality.'' It describes how well the trial fragment falls into the most favored region of the Ramachandran plot of the residue(s) in the target protein, onto which it is mapped. It is calculated as the normalized probability for a particular residue to assume a particular ϕ,ψ combination, and it is therefore a measure of how likely the target sequence can assume the conformation of the database substructure. For an N-residue fragment this surface score is defined as

$$F_{\mathrm{SURF}} = \sum_i - \ln[p_i(\phi_i,\psi_i)/p_{i,\max}]/N \tag{4}$$

where $p_i(\phi_i,\psi_i)$ represents the database population of the (ϕ_i,ψ_i) conformation for the target residue type at position i, and $p_{i,\max}$ is the population of the most favored conformation of that particular residue type.

Sequence homology information can also be used in the fragment scoring process. A homology score is generated to penalize for ''mutations'' between the target sequence and the sequence of a database substructure, according to

$$F_{\mathrm{HOMO}} = \left(\sum_i h_{jk}(i)^2 / N \right)^{1/2} \tag{5}$$

where $h_{jk}(i)$ are the elements of a residue-type similarity matrix between residue type j at position i in the target protein fragment and residue type k in the corresponding position in the database substructure. There are several possible schemes for homology scoring, and in practice a given scheme is selected by substituting the appropriate table corresponding to the desired homology matrix. In the present work, the matrix used is derived from the similarity in the residue-specific Ramachandran map distributions, $p(\phi,\psi)$:

$$h_{jk} = A \left\{ \sum_{\phi,\psi=0^\circ,\ldots,359^\circ} ([p_k(\phi,\psi) - p_j(\phi,\psi)]\, p_j(\phi,\psi))^2 \right\}^{1/2} \tag{6}$$

where A is an arbitrary factor, used to scale the results to a convenient range, and each $p(\phi,\psi)$ distribution is normalized such that its maximum value is unity. Equation (6) results in an asymmetric score matrix (especially for substitutions involving Gly residues) reflecting the fact that, for example, for a Gly residue in the target fragment, the presence of a non-Gly residue in the database fragment biases the selected fragment toward negative ϕ angles; this is more of a concern than the inverse scenario. (Note that a Gly residue in the trial fragment, with a bias toward positive ϕ, usually is already penalized by an increased Ramachandran surface quality score.) To illustrate the weight factors for residue similarity identified in this manner, Table I provides a condensed form of the 20×20 matrix. However, it is the full matrix that is used by the software.

Additional information, such as sequential or medium range backbone NOEs, or dihedral restraints also can be introduced at this stage of the fragment search, but this latter information was not needed in the application to small proteins with relatively complete sets of dipolar couplings, discussed in this chapter.

The final score for a particular fragment is a weighted sum of the individual terms:

$$F_{\mathrm{TOTAL}} = \sum_i c_i F_i \tag{7}$$

where i = CS, DC, surface (SURF), homology (HOMO). Empirically, a set with $c_{\mathrm{DC}} = 1.0$, $c_{\mathrm{CS}} = 0.2$, $c_{\mathrm{SURF}} = 0.2$, $c_{\mathrm{HOMO}} = 0.2$ was found to give close to optimal results. These coefficients were chosen such that the dipolar term dominates, but the other terms remain significant. It can be useful to also evaluate which fragments are selected if the DC term is scaled down, and the

TABLE I
HOMOLOGY FACTORS, h_{jk}, DERIVED FROM EQ. (6)[a]

Residue	A	RK	N	D	CST	QE	HFWY	G	IV	LM	P
A	0.0	0.6	1.8	0.9	2.1	0.4	1.5	8.7	2.8	0.6	5.1
RK	0.5	0.0	1.6	0.6	1.7	0.4	1.1	8.6	2.3	0.5	5.0
N	0.7	0.7	0.0	0.4	1.4	0.5	1.4	8.2	2.4	0.7	5.0
D	0.5	0.4	1.1	0.0	1.7	0.3	1.4	8.5	2.6	0.6	4.9
CST	0.9	0.9	1.5	0.8	0.0	0.8	1.3	8.7	1.7	0.9	4.8
QE	0.4	0.4	1.6	0.7	1.8	0.0	1.3	8.6	2.5	0.4	5.2
HFWY	0.9	0.9	1.9	1.1	1.7	1.0	0.0	8.6	1.9	0.8	5.0
G	0.4	0.4	1.5	0.7	1.9	0.5	1.2	0.0	2.8	0.5	4.9
IV	0.8	1.1	2.0	1.1	1.8	0.8	1.3	8.7	0.0	0.8	5.3
LM	0.5	0.6	1.8	0.8	1.8	0.4	1.2	8.7	2.2	0.0	5.2
P	4.0	4.3	4.9	4.6	5.0	4.4	4.8	9.7	5.5	4.3	0.0

[a] The full matrix used by MFR+ consists of 20 × 20 elements. The grouping above reflects the high degree of similarity of the coefficients pertaining to any given group.

CS term is given a high weight. A set of coefficients we commonly use for this purpose is $c_{DC} = 0.1$, $c_{CS} = 1.0$, $c_{SURF} = 0.1$, $c_{HOMO} = 0.1$. This alternate weighting scheme is particularly useful when in a given region of the polypeptide backbone the number of observed dipolar couplings is low. If both searches are conducted, the results are pooled together, for example, by retaining the 10 best fragments of each search. In a subsequent refinement (see below), database fragments that are close to their true structure will better converge to the correct solution, and the presence of lower quality fragments, usually selected on the basis of their chemical shifts, therefore does not pose a problem. Typical search parameters are summarized in Table II.

A convenient way to visualize the search results displays the collection of fragments as a ''Ramachandran flight path'' of the peptide fragments, which connects the (ϕ_i,ψ_i) position in the Ramachandran map of residue i to the (ϕ_{i+1},ψ_{i+1}) and (ϕ_{i-1},ψ_{i-1}) position of residue $i + 1$ and $i - 1$ (Fig. 3). The ''true'' structure needs to be represented by an unbroken path, and outliers can be identified very easily this way. Figure 3A shows that at this stage the ϕ, ψ angles of the selected substructures still exhibit a considerable spread. However, as described below, the quality of these fragments and the width of their ϕ, ψ distribution can be vastly improved by refining these fragments with respect to their dipolar couplings.

Fragment Refinement

Next, the backbone angles from the substructures resulting from the search are used to build fragments of the target protein, with the correct residue types, but with initial side chain orientations being random. As the

TABLE II
FILES AND PARAMETERS USED DURING THE MFR+ DATABASE HOMOLOGY SEARCH AND THEIR TYPICAL NAMES OR VALUES

Input file or parameter	Name or typical value
Dipolar coupling table(s)[a]	dObs[*].tab
Backbone chemical shift table[a]	csObs.tab
Reference structure[b]	ref.pdb
Name of output fragment table	mfr.tab
Location of PDB files	$PDBH_DIR[c]
List of PDB files to search	$PDBH_TAB[c]
csW/dcW/surfW/homoW[d]	0.2/1.0/0.2/0.2 or 1.0/0.1/0.1/0.1
segLength[e]	7 (default)–10
scoreCount[f]	10 (default)–20
csThresh[g]	2.2 (default)
undefFrac[h]	0.9 (default)
Additional parameters, calculated for each saved fragment in the output fragment table	
da[i]	From fit
dr[i]	From fit
scalarProduct[j]	−1 to 1
cosXX, cosYY, cosZZ[k]	0 to 1

[a] Dipolar coupling and backbone chemical shift input tables are PALES and TALOS input formats, respectively.
[b] Covalent template of the target protein, either in extended or randomized conformation, or as a model derived from other sources, in PDB format. The program can display agreement between selected fragments and the template (to be used for test purposes, when the true structure is known).
[c] UNIX environment variables.
[d] Weighting factors for F_{DC} [Eq. (2)], F_{CS} [Eq. (3)], F_{SURF} [Eq. (4)], and F_{HOMO} [Eq. (5)] in the total fragment score [Eq. (7)].
[e] Number of residues per segment.
[f] Number of database fragments retained after first round of MFR+ search.
[g] Chemical shift cut-off value. No fragment with F_{CS} greater than this value is retained.
[h] If a fraction smaller than undefFrac of the experimental data can be mapped onto the database fragment (e.g., due to a missing N–H vector for Pro in a database fragment), the database fragment is not considered.
[i] Axial and rhombic components of local best fit alignment tensor (obtained from SVD fit), calculated for each alignment medium used.
[j] Normalized generalized scalar product of local best fit (SVD) alignment tensors; this is a number in the [−1.0 to 1.0] range.
[k] cosXX, cosYY, and cosZZ specify the relative orientation of two alignment tensor axis frames (OXYZ and O′X′Y′Z′). cosXX = | *cos*[angle (OX O′X′)] |, etc.

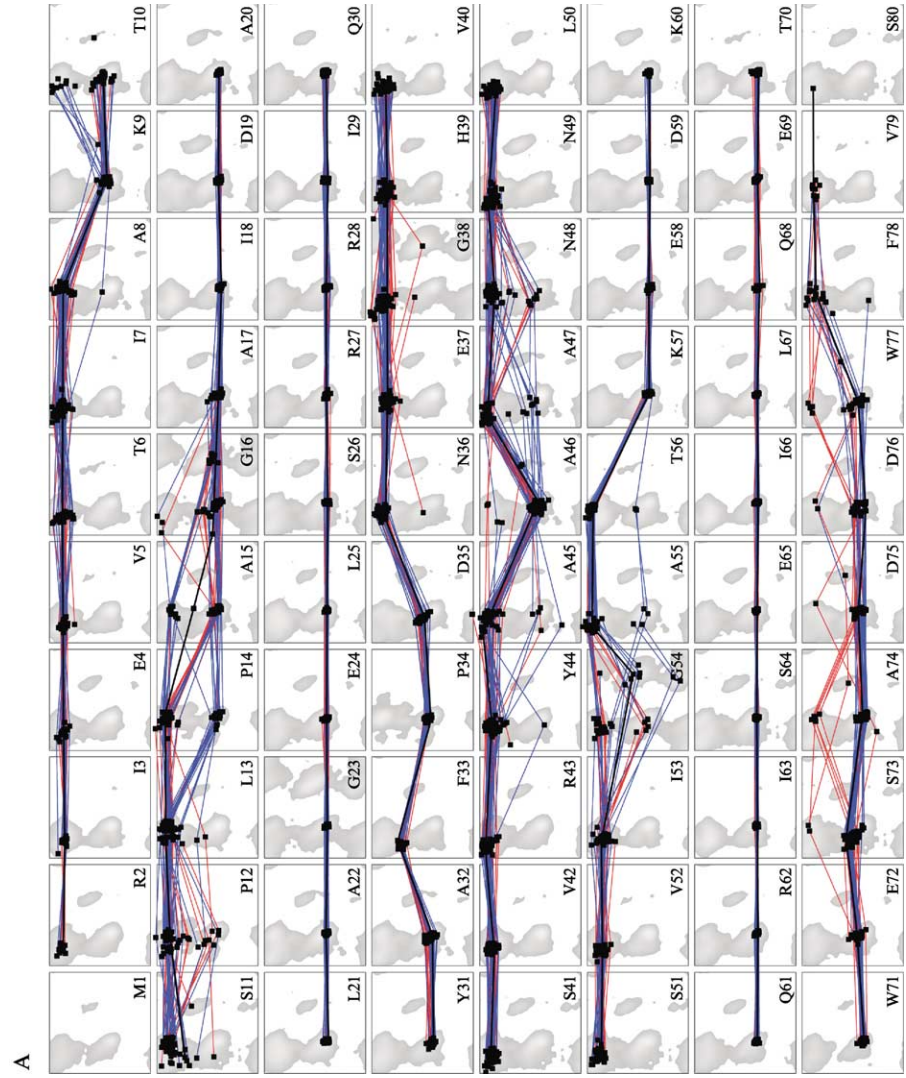
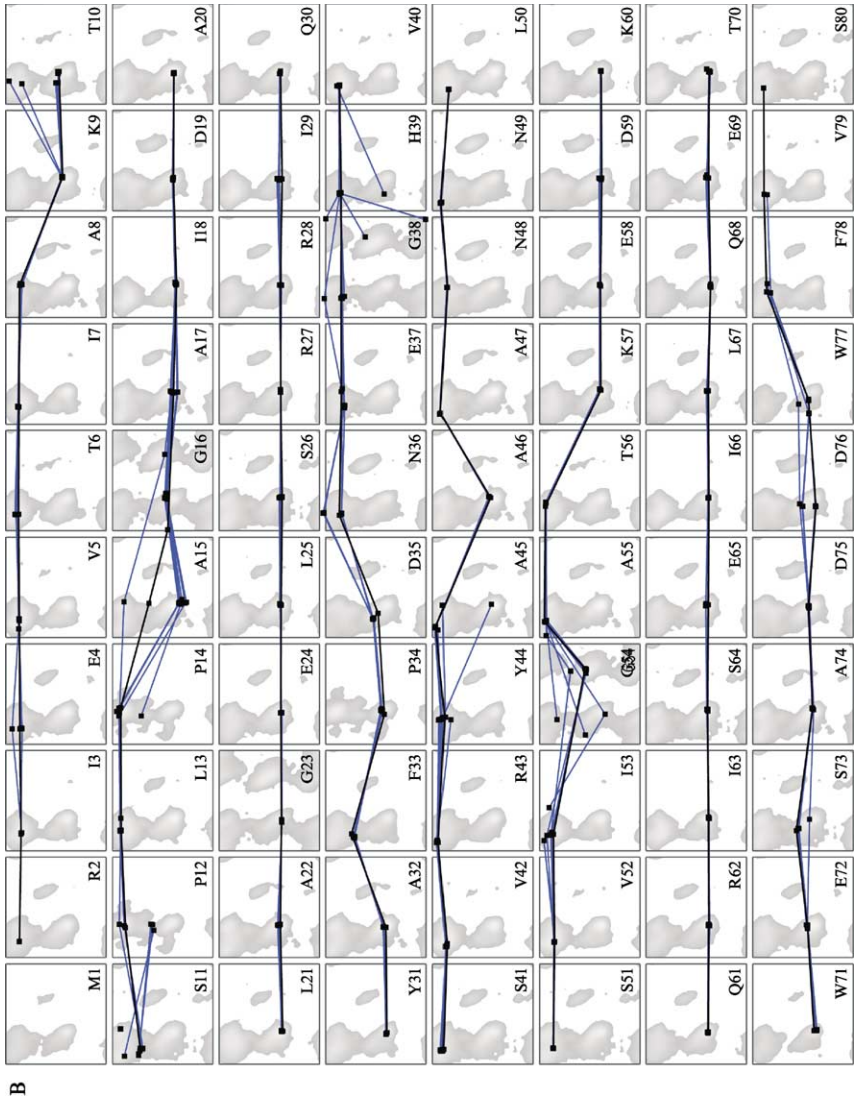


FIG. 3. *(continued)*


B
M1 R2 I3 E4 V5 T6 I7 A8 K9 T10
S11 P12 L13 P14 A15 G16 A17 I18 D19 A20
L21 A22 G23 E24 L25 S26 R27 R28 I29 Q30
Y31 A32 F33 P34 D35 N36 E37 G38 H39 V40
S41 V42 R43 Y44 A45 A46 A47 N48 N49 L50
S51 V52 I53 G54 A55 T56 K57 E58 D59 K60
Q61 R62 I63 S64 E65 I66 L67 Q68 E69 T70
W71 E72 S73 A74 D75 D76 W77 F78 V79 S80

dipolar couplings are extremely sensitive to internuclear vector orientations, even relatively small variations in these orientations can result in a large dipolar residual and a wrong (usually too small) magnitude and orientation of the best-fit alignment tensor (Zweckstetter and Bax, 2002). This can render the long-range information content of the residual dipolar couplings difficult to utilize and also can make it difficult to distinguish good database substructures from false hits, which accidentally yield some above average agreement with the dipolar couplings.

Both these problems can be addressed by refining the selected database fragments by means of a gentle, low-temperature simulated annealing protocol against the experimental RDCs (and other restraints, if available), using either an in-house modified version of X-PLOR (Brunger, 1993; Schwieters *et al*., 2003) (referred to as XPLOR-NIH, available through http://nmr.cit.nih.gov) or the program DYNAMO (available through http://spin.niddk.nih.gov/bax). Parameters used during this refinement are listed in Table III. Refinement in the presence of dipolar couplings is carried out using an artificial, harmonic coordinate restraint term (similar to a noncrystallographic symmetry term), which keeps backbone C^{α} atoms of the refined fragment reasonably close ($< \sim 1$ Å) relative to the initial database substructure but allows small adjustments in the orientations of the individual peptide planes, while usually increasing the agreement with the dipolar couplings considerably.

Incorporation of dipolar coupling restraints into X-PLOR structure calculation and refinement has been described previously (Schwieters *et al*., 2003; Tjandra *et al*., 1997), and relies on the use of a tetraatomic orthonormal ''pseudomolecule,'' OXYZ, to represent the principal axis system of the alignment tensor, whose orientation is allowed to float. Both XPLOR-NIH and DYNAMO software packages now allow the magnitude and rhombicity of the alignment tensor to float during refinement, but this option is not used in the current evaluation of the performance of the

FIG. 3. Summary of MFR search results for DinI, (A) before and (B) after refinement of the fragments with respect to dipolar restraints. Each subpanel displays the (ϕ,ψ) backbone angles found for a given residue in the MFR search, with gray regions marking the most occupied region in the database for the particular residue type. Solid lines connect (ϕ,ψ) pairs found for adjacent residues in the same fragment, and ''dead ends'' such as observed for P12 in (A) correspond to the fragment where this residue is the last in the selected stretch. In (A), blue (dark gray) lines correspond to MFR search results heavily weighted toward dipolar couplings (using parameters of Table II); red (light gray) lines correspond to MFR results when emphasizing chemical shifts. After refinement, using parameters of Table III and convergence criteria of Table IV, the majority of residues display unique (ϕ,ψ) backbone angles (B). The black line connects the (ϕ,ψ) pairs seen in the previously determined NMR structure (PDB entry 1GHH).

TABLE III
TYPICAL PARAMETERS FOR LOW-TEMPERATURE SIMULATED ANNEALING REFINEMENT OF DATABASE FRAGMENTS USING EITHER XPLOR-NIH OR DYNAMO

Parameter	Simulated annealing protocol	
	Fragment refinement	Structure regularization
Temperature (K)[a]	500 → 1	1000 → 1
Temperature step (K)	10	10
Number of steps	3000	20000
Timestep (fs)	3	3
Masses (amu)	100	100
Force parameters[b]		
k_{bond} (kcal mol^{-1} Å^{-2})	1000	1000
k_{angle} (kcal mol^{-1} rad^{-2})	400 → 1000	400 → 1000
$k_{improper}$ (kcal mol^{-1} rad^{-2})	100 → 1000	100 → 1000
k_{vdw} (kcal mol^{-1} Å^{-4})	0.01 → 1	0.01 → 1
Repel[c]	0.8	0.8
k_{rama} (kcal mol^{-1})	0.002 → 1	0.002 → 1
k_{bor} (kcal mol^{-1} rad^{-2})[d]	1 → 40	1 → 40
k_{cen} (kcal mol^{-1} rad^{-2})[d]	1 → 10	1 → 10
k_{dipo} (kcal mol^{-1} Hz^{-2})	0.01 → 1	0.01 → 1
$k_{collapse}$ (kcal mol^{-1} Å^{-2})	—	50
k_{harm} (kcal mol^{-1} Å^{-2})[e]	10 → 0.1	10 → 0.1
k_{cdih} (kcal mol^{-1} rad^{-2})[f]	—	300 → 1

[a] Temperature control is achieved by coupling to a heat bath with a coupling constant of 10 fs.
[b] Weighting factors for the energy terms (if any) are included in the force constants.
[c] Scale factor for van der Waals radius.
[d] Force constants for vector angle restraint energy term (border and center exclusion) as defined in Meiler *et al.* (2000).
[e] Harmonic coordinate restraint, applied to C^{α} only.
[f] Force constant for dihedral restraints (if available).

MFR+ approach. We find that whenever reasonably reliable values for the alignment parameters can be extracted, either from the histogram of observed dipolar couplings (Bryce and Bax, 2004; Clore *et al.*, 1998a), or from the first round of the database fragment search (Fig. 2), it is better to leave these values fixed during refinement.

The dipolar energy term is of the form $E_{dipo} = k_{dipo}\,(D^{obs} - D^{calc})^2$. To avoid large initial erratic forces, the program starts with the alignment frame in an optimal orientation, determined by best fitting the experimental dipolar couplings to the coordinates of the database peptide. Using additional a priori information about the relative orientations of the alignment tensors, applicable in cases in which measurements have been carried

out in multiple media, is advantageous at this point. As shown for the magnitude and rhombicity of the alignment tensors (Fig. 2), the relative tensor orientations can also be obtained from a weighted average of the results of the first round of the MFR database search. In practice, we define the relative orientation of the alignment tensors by so-called ''vector angle restraints'' (Meiler *et al.*, 2000) between the axes of the two pseudomolecules (OXYZ and O′X′Y′Z′), representing the two alignment tensors.

The refinement protocol also takes advantage of a very weak, database-derived dihedral energy term or potential of mean force, which is commonly referred to as a ''Ramachandran term.'' This term disfavors conformations that are not or very sparsely represented in the PDB and can improve the local quality of structures (Kuszewski *et al.*, 1997). Another important benefit of this potential energy term is that it positions the side chains in the orientations found to be most likely for the corresponding backbone torsion angles. The parameters for the refinement protocol are summarized in Table III.

After refinement, it is usually fairly straightforward to separate the correctly refined substructures from those that involve ''false positives'' in the initial database search. In this process, the refined structures are ranked according to their DC residual, CS agreement, as well as their ''surface quality,'' calculated using Eqs. (2)–(4). Fragments beyond an adjustable cut-off threshold for F_{DC}, F_{CS}, and F_{SURF} are discarded.

Even though the best-fitting substructures were selected from the database and subsequently refined, if even these best fits are relatively poor, these substructures frequently are structurally quite diverse and the precise value of the dipolar residual becomes a less discriminating factor. This is illustrated by a plot of backbone coordinate RMSD vs. normalized dipolar score, using $[(0.8D_a)^2 + (0.6D_r)^2]^{-1/2}$ for normalization (Clore and Garrett, 1999) (Fig. 4). For fragments for which the RMSD between best-fit and observed dipolar couplings exceeds ca. $0.2D_a$, the spread relative to the true structure rapidly increases, and above $0.3D_a$ essentially no correlation between the conformation of the database substructure and the target fragment remains. Therefore, such fragments are discarded at this stage.

For the converged fragments, the ''energies'' typically fall within a small margin (typically 10–20%) of the lowest value observed for that fragment, and only those fragments that are within this margin are accepted. Typical tolerance values, tolDC, tolCS, and tolSurf, for the selection of converged fragments are listed in Table IV.

Additional criteria for validation include the requirement that the magnitude and rhombicity of the local ''best fit'' alignment tensor, as determined by SVD, falls within an adjustable fraction (typically ±20%) of the values used during refinement (Table IV). This latter selection can eliminate ''false positives'' that did not converge properly in the course of

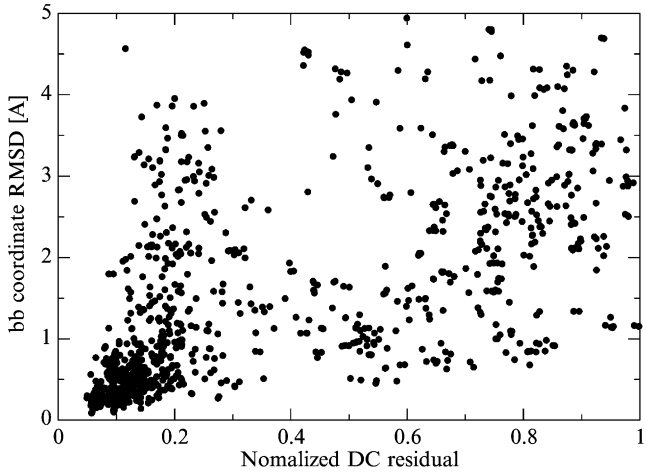


FIG. 4. Plot of backbone coordinate RMSD when best fitting the refined nine-residue fragment library to the X-ray structure (PDB entry 1UBQ) vs. normalized dipolar coupling RMSD. Prior to calculating the dipolar coupling RMSD, couplings are normalized by dividing each difference between observed and best-fitted coupling by $[(0.8\ \mathrm{Da})^2 + (0.6\ \mathrm{Dr})^2]^{1/2}$. The normalized dipolar residual used for making the plot corresponds to the average of the two different data sets available from the two alignment media.

the dipolar coupling refinement. Similarly, if data from more than one alignment medium are available, a requirement is that the relative orientation and normalized scalar product (Sass *et al.*, 1999) of the tensors fall within a specified margin, typically two standard deviations from what is observed for the full set of fragments.

The results of the refinement and the selection of accepted fragments again are inspected by viewing the Ramachandran flight path (Fig. 3B). As expected, the refinement of the individual substructures yields considerably smaller spreads in ϕ,ψ angles and improved agreement relative to the reference structure. Typically, when inspecting the full protein in this manner, distinctly different solutions are observed at several locations, as seen, for example, for residues T10, P14, and G54 in DinI (Fig. 3B). However, as discussed below, many of these ambiguities are resolved in an automated manner at the assembly stage.

Assembly of Structure from Fragments

All the manipulations for model building are carried out in the Tcl/Tk script language, using the DYNAMO class of routines within ''NMRWish.'' A key routine in this process, dynAlign, is used to best-fit

TABLE IV
SELECTION CRITERIA AND THEIR TYPICAL SETTINGS FOR MFR+ SELECTION OF CONVERGED FRAGMENTS AFTER REFINEMENT[a]

Parameter	Typical value
tolDC[b]	10–33%
tolCS[b]	10–33%
tolSurf[b]	10–33%
daTol[c]	2 SD[d]
rTol[c]	2 SD[d]
scalarTol[e]	2 SD[d]
tolXX/YY/ZZ[e]	2 SD[d]

[a] Various "quality checks" are applied to remove fragments that result from accidental "false positives" in the MFR+ search stage. Knowledge of alignment tensor magnitude(s) and rhombicity(ies) (cf. Fig. 2) and relative orientations (if more than one alignment medium is employed) is used to define tolerance limits. Peptide fragments with deviations greater than the specified tolerance margins are discarded at the "quality check" stage.
[b] Maximum allowed deviation of F_{DC} [Eq. (2)], F_{CS} [Eq. (3)], and F_{Surf} [Eq. (4)] relative to the best of the refined database fragment in that particular residue range.
[c] Maximum allowed deviation relative to the average da and dr values (see Table II), when considering the entire ensemble of selected database fragments.
[d] SD, standard deviation.
[e] Maximum deviation of generalized scalar product and cosXX, cosYY, and cosZZ (see Table II) relative to the averaged value, when considering the entire ensemble of selected database fragments.

a new fragment onto the growing chain, by minimizing the backbone coordinate RMSD for the overlapping residues. In principle, there are three translational and three rotational degrees of freedom in this process. However, as the orientation of the fragment relative to the alignment frame is already known uniquely at this stage (or with 4-fold ambiguity in case of only a single alignment tensor), dynAlign has the ability to freeze the rotational degrees of freedom while optimizing the translational parameters only. In case of a single, axially symmetric or nearly axially symmetric alignment tensor, rotation about the z-axis is also allowed as an adjustable parameter.

The procedure for the assembly of the structure is visualized in Fig. 5. In a first step, the refined fragments are rotated such that their coordinate frames coincide with the principal frame(s) of their local alignment

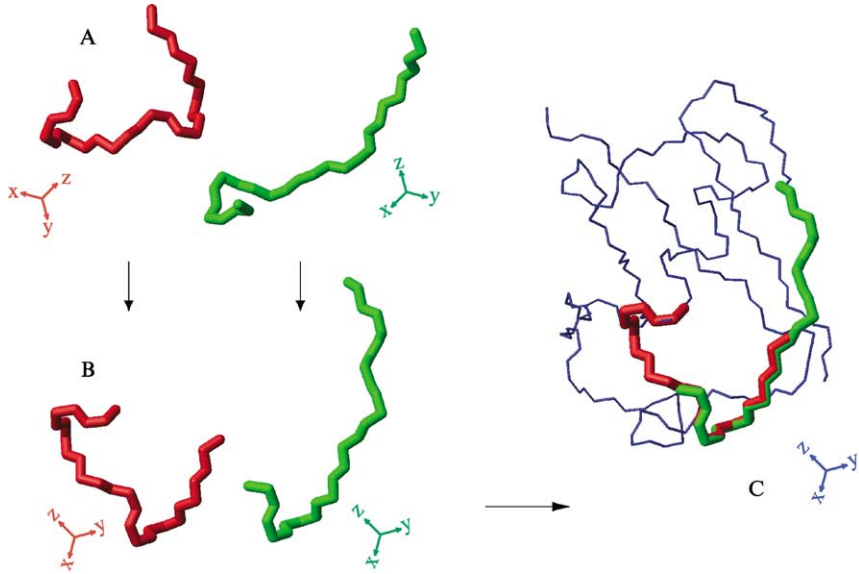


FIG. 5. Pictorial representation of the MFR+ assembly process, illustrated for ubiquitin. (A) Backbone representations of MFR+-derived fragments for residues 13–21 (red, light gray) and 17–25 (green, dark gray) fragments in arbitrary relative orientation, together with their corresponding alignment tensor frames. (B) After rotation such that their respective local alignment tensors have identical orientations. (C) Fragments are translated (with fixed orientation) such that the coordinate RMSD relative to the previously assembled chain is minimized. The figure was generated using the program MOLMOL (Koradi *et al.*, 1996).

tensor(s). When dipolar couplings from only a single medium are available, there is a 4-fold ambiguity in the orientation of the fragment. If dipolar couplings from multiple alignment media are available, the 4-fold degeneracy can easily be resolved (Al-Hashimi *et al.*, 2000) prior to assembly, but this is by no means a prerequisite. In practice, a computationally convenient solution generates all four orientations for each alignment frame and compares at the assembly stage, discussed below, which of the possible pairwise combinations yields the correct, compatible orientation for the fragment relative to the prebuilt fraction of the chain, as judged from the lowest coordinate RMSD, when optimizing their relative translation while maintaining a consistent relative alignment tensor orientation (Al-Hashimi *et al.*, 2000).

In the assembly step, the N- and C-terminal residue of each fragment are discarded, as they typically are less well defined by the data. Subsequently, each of the accepted, shortened fragments is translated onto the

growing chain by minimizing the coordinate RMSD relative to the previously built fraction of the chain (Fig. 5). All fragments that fit within an adjustable threshold (typically requiring a coordinate RMSD, max RMSDpRes $\leq 0.2 \times N$ Å, where N is the number of overlapping residues) are retained for generating the final structure. As mentioned above, when data from only a single medium are available, four distinctly different fragment orientations agree equally well with the experimental data (Al-Hashimi *et al.*, 2000), but usually at most one of these will yield a reasonable coordinate fit. This chain extension process is repeated until the assembly of the protein structure is complete. Parameters for the assembly process are presented in Table V. It is worth noting that in the case of

TABLE V
PARAMETERS AND TYPICAL VALUES FOR MFR+ CHAIN ASSEMBLY

Parameter	Typical value
Table with converged fragments	frag.tab
Dipolar coupling table(s)	dObs[*].tab
dadrFlag[a]	0 or 1
da, dr (for each tensor)[b]	From Table II
MaxRMSD[c]	[0.0–1.0 Å]
maxRMSDpRes[d]	[0.1–0.3 Å/residue]
minOVLPCount[e]	[1 – fragLength] = 2 (default)
skipFirst[f]	0 or 1

[a] Flag to select whether SVD (0) or nonlinear Powell minimization is used to fit a database fragment to experimental dipolar couplings.
[b] Determined from the fragment search (Fig. 2). Parameters da and dr are used only if dadrFlag $\neq 0$.
[c] Maximum acceptable backbone (N, Cα, C^1) coordinate RMSD between a new refined database fragment and the overlapping region of the previously assembled protein backbone. If the current backbone coordinate RMSD is smaller, the fragment is accepted; if it is larger, it not used in model building. When maxRMSD is set to zero, the parameter maxRMSDpRes is used instead to decide if a fragment should be accepted or rejected.
[d] Used only if maxRMSD = 0. maxRMSDpRes is the maximum acceptable coordinate RMSD per residue overlap between a new refined database fragment and the previously assembled protein backbone. Above this threshold the fragment will be discarded.
[e] Minimum number of residues required to overlap between a new fragment and a previously assembled protein chain.
[f] Flag to decide whether to discard (>0) or retain (0) N- and C-terminal residues of a database fragment in the assembly of the MFR model.

axially symmetric alignment, particularly when only data in a single alignment medium are available, the rotation about the z-axis of the alignment tensor is undefined and in that case it must be treated as an additional degree of freedom when building a fragment onto the chain. This additional degeneracy can pose significant problems in the assembly process and may require that additional data, such as sparse NOEs, are available.

In the final step, coordinates of all accepted fragments are averaged, and the resulting structure, which may have local nonphysical geometry, is regularized by simulated annealing using a protocol similar to the one used for refinement of individual fragments (Table III), but using 20,000 instead of 3000 steps, and a 2-fold (1000 K vs. 500 K) higher starting temperature.

Application to Model Proteins

Application of the MFR+ method is demonstrated for three proteins for which extensive sets of experimental backbone dipolar couplings were available, ubiquitin, GB3, and DinI. Crystallographically determined structures are available for ubiquitin and GB3, and NMR structures are available for all three. The method has also been applied to several slightly larger proteins for which dipolar couplings were simulated, including thioredoxin, profilin, and interleukin-1β. In all these applications, standard parameters (Table II) were used, with a query fragment length of nine residues, but nearly identical results were obtained using seven- or eight-residue fragments. For ubiquitin, GB3, and DinI, relatively complete sets of $^1D_{NH}$, $^1D_{NC'}$, $^1D_{C\alpha H\alpha}$, $^1D_{C\alpha C'}$, and $^2D_{C'HN}$, reported previously, were used (Ottiger and Bax, 1998; Ramirez *et al.*, 2000; Ulmer *et al.*, 2003). For the other three proteins, these couplings were generated with the program PALES, for alignment tensors predicted by PALES for media of bicelles and Pf1 (Zweckstetter and Bax, 2000; Zweckstetter *et al.*, 2004). The number of dipolar coupling and chemical shift restraints for each protein is listed in Table VI.

Ubiquitin

Ubiquitin has served as a test case for several programs that determine the NMR structure by nonconventional methods. Using a method similar to MFR+, but using only consistent sets of overlapping fragments and best fitting the coordinates of these, Andrec *et al.* (2001) obtained a model with a backbone coordinate RMSD of 2.4 Å relative to the X-ray structure. Using the same input data, the ROSETTA method of the Baker group (Rohl and Baker, 2002), operating in torsion angle space, resulted in models that deviate from the X-ray reference structure by only 1.03–1.17 Å.

TABLE VI
RESULTS OF THE MFR+ METHOD APPLIED TO MODEL PROTEINS[a]

Protein	Number of residues	PDB entry	N_{CS}[b]	N_{DC} medium A[b]	N_{DC} medium B[b]	Backbone RMSD to PDB (Å)	All atom RMSD to PDB (Å)	Angular ϕ/ψ RMSD (°)[c]
Ubiquitin	76	1UBQ	378	333	325	0.70 (0.81)[d]	1.66 (1.79)[d]	7.5/9.1
DinI	81	1GHH	389	343	209	1.51 (0.79)[e]	2.34 (1.92)[e]	9.6/7.6
GB3	56	2IGD	273	211	231	0.68 (1.12)[d]	1.49 (2.01)[d]	11.7/11.7
Thioredoxin	105	1ERT	416	501	501	0.83	1.67	8.0/7.7
Profilin	125	1ACF	598	588	588	0.44	0.85	6.0/8.9
Interleukin-1β	153	4ILB	574	707	707	1.95 (1.41)[f]	2.74 (2.41)[f]	10.4/9.9

[a] Experimental data are used for ubiquitin, DinI, and GB3; input RDC data for thioredoxin, profilin, and interleukin-1β are simulated using PALES software, but chemical shifts are experimental. Unless otherwise noted, reported coordinate RMSD values refer to residues 2–72 (ubiquitin), 2–77 (DinI), 2–55 (GB3), 2–104 (thioredoxin), 2–124 (profilin), and 2–152 (interleukin-1β).

[b] N_{CS} is the number of chemical shifts; N_{DC} (A) and N_{DC} (B) are the numbers of dipolar couplings available in media A (bicelles) and B (charged bicelles for ubiquitin; Pf1 for all others).

[c] Excludes regions where large crankshaft errors have occurred in model building and refinement (N52/G53 in ubiquitin, G58/G59 in profilin, S21/G22 in interleukin-1β).

[d] Using dipolar data from a single medium (bicelles) only.

[e] For residues 2–53.

[f] For residues 4–134.

ROSETTA proved particularly tolerant to incomplete RDC data. For example, using only D_{NH} input values, the fold could still be predicted reliably (backbone RMSD 2.75 Å) The original version of our MFR program yielded a backbone structure that differed by 0.88 Å from the X-ray structure (Delaglio *et al.*, 2000). The sequential chain building method of Hus *et al.* (2000) yielded comparable results (RMSD 1.0 Å), using the same sets of residual dipolar couplings. The present version of MFR+ yields a backbone structure that differs by 0.70 Å from the crystal structure (PDB entry 1UBQ) (Vijay-Kumar *et al.*, 1987) shown in Fig. 6A, 0.72 Å from the NMR structure (PDB entry 1D3Z) (Cornilescu *et al.*, 1998). This latter number increases to 0.81 Å, if data from only one alignment medium are used. The RMSD for all nonhydrogen atoms relative to the X-ray (1.66 Å) or lowest energy NMR structure (1.68 Å) is considerably larger, resulting from the lack of experimental data restraining the side chain orientations.

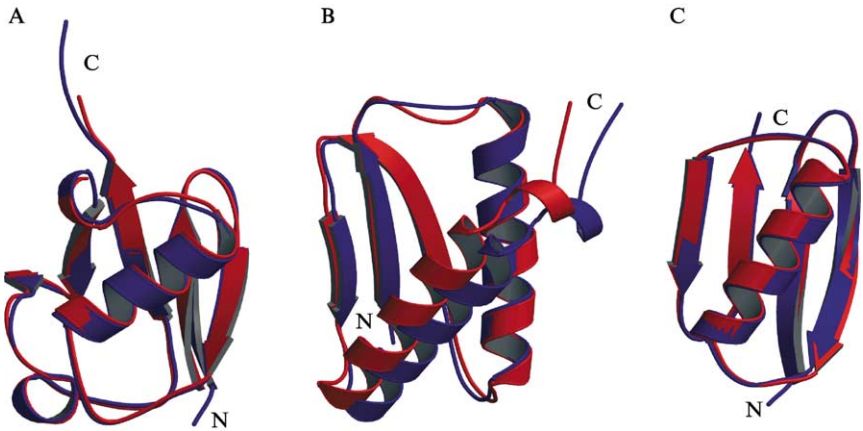


FIG. 6. Comparison of MFR+-derived structures and previously solved X-ray (A and C) and NMR (B) structures of (A) ubiquitin, (B) DinI, and (C) GB3. Blue (light gray) structures correspond to the PDB reference coordinates; red (dark gray) represents the MFR+-derived ribbon. For ubiquitin (A), the disordered C-terminus (residues 74–76) could not be built by the MFR+ method and is not shown. For DinI (B), the structure could not be assembled uniquely from dipolar couplings acquired in a single medium. Even with data from two media, some ambiguity remains (Fig. 3B) and is responsible for the erroneous lateral displacement of helix $\alpha 2$ by ca. 3.7 Å relative to the reference structure. For clarity, superposition of the MFR+ and reference structure is optimized for residues 2–53, highlighting the displacement of helix 2. Reference structures correspond to PDB entries 1UBQ, 1GHH, and 1IGD. Figures were generated using the programs Molscript (Kraulis, 1991) and Raster 3D (Merritt and Murphy, 1994).

For the four C-terminal residues, the MFR+ program was unable to find a unique solution. This is not surprising, considering that these residues are dynamically highly disordered (Tjandra *et al*., 1995; Wand *et al*., 1996). The inability to define a unique structure in the presence of such extensive motion suggests that the MFR+ automatically recognizes such regions. However, for other regions in the protein where increased dynamics is known to take place but is less extreme (e.g., around residues G10, K11, I23, E24, K48, and G53), and where several amide resonances and their corresponding dipolar couplings are unobservable as a result of conformational exchange, MFR+ can faithfully define the backbone structure. For these cases, the increased internal dynamics affects at most only a few sequential residues, which does not significantly perturb the search for optimally fitting nine-residue substructures. Clearly, the final NMR model calculated with MFR+ no longer carries a signature of this increased internal dynamics. However, the increased dynamics will generally be evident from the raw spectra, which exhibit either exchange-broadened weak resonances or motionally narrowed, intense resonances.

DinI

DinI is a small globular protein of 81 residues, implicated in DNA repair, for which an NMR structure has been obtained (PDB entry 1GHH) from both NOE and extensive dipolar coupling restraints (Ramirez *et al*., 2000). Compared to ubiquitin, this proved to be a more challenging test case. NMR data were collected at lower concentrations, resulting in lower signal-to-noise ratios and a less complete set of dipolar couplings. Two sets of RDCs were available, one nearly complete set, acquired in bicelles, and a somewhat less complete set, due to stronger homonuclear ^{1}H–^{1}H dipolar broadening resulting from overalignment, obtained in a Pf1 solution.

For an extended loop region, K9 to G16, several of the fragments did not find satisfactory substructures in the database that met the standard cut-off criteria. In fact, comparing the NMR structure for this region with all fragments in our database indicates that although the search indeed selects the best fitting fragment, none of the database fragments found by the MFR+ search agrees to better than 0.8 Å. The dipolar search is affected significantly by structural differences of this magnitude. A subsequent evaluation, searching simply for database fragments with the closest backbone RMSD relative to this fragment of the previously determined DinI structure, confirms that no fragments that are closer than 0.8 Å in backbone structure are present in the database.

A number of false positives in the MFR+ search, which differed substantially in structure but accidentally yielded better than random RMSDs between observed and best-fitted dipolar couplings, was also obtained in the search for the K9–G16 fragment. This problem was compounded by the presence of two Pro residues at positions 12 and 14, which resulted in far fewer dipolar couplings for the fragments that include these residues. As illustrated in Fig. 3, it was not possible to unambiguously extract reliable fragments with the standard protocol. However, because at the chain building stage none of the erroneously identified fragments yielded a suitable match to the previously built chain, even the large degeneracy encountered in this loop region did not prevent successful chain extension.

For DinI, the chemical shifts and dipolar coupling fragment searches are clearly indicative of two long α-helices, stretching from G16 to A32 and from K57 to W77 (with a kink at S73). Such stretches of helix, which typically result in very good hits when searching the database, are very helpful in accurately defining the relative orientations as well as the magnitudes and rhombicities of the two alignment tensors. Using this additional information, the search hits were then screened for the correct magnitude and relative orientation of the alignment tensors. Several fragments failed to converge to a unique structure in the course of their refinement, despite having almost indistinguishable scores and energies, e.g., fragments around T10–A15, A45–N48, and G54. Nevertheless, there was sufficient overlap between converged fragments that the chain could be built. As mentioned above, the requirement that a fragment must give a reasonable backbone coordinate match to the previous one is key in resolving such remaining ambiguities.

The final backbone differs rather substantially, by 1.51 Å, from the previously determined NMR structure (PDB entry 1GHH). As illustrated in Fig. 6B, a comparison of the two structures reveals that this high RMSD results from a lateral translational error in the position of the second long helix: If residues 2–53 are superimposed, helix α2 is shifted by 3.7 Å from that seen in 1 GHH, apparently caused by an incorrect formation of the reverse turn centered at G54. For residues R2–I53, the backbone RMSD is only 0.79 Å, and the C-terminal kinked helix, K57–W77, agrees with the previously determined solution structure to within 0.44 Å.

GB3

A very high-resolution X-ray structure, solved at 1.1 Å resolution (PDB entry 1IGD) (Derrick and Wigley, 1994), is available for this protein, and the solution structure (PDB entries 1P7E/1P7F) (Ulmer *et al.*, 2003) agrees with this structure to within 0.3 Å.

Using dipolar couplings from only one alignment medium (bicelles), the standard protocol resulted in a model that differed by a backbone RMSD of 1.12 Å from the X-ray structure (1IGD). When using dipolar couplings from both bicelle and phage media, this RMSD decreased to 0.68 Å. The recently reported solution structure and the two structures derived with MFR+ are superimposed on one another in Fig. 6C. The backbone RMSD of the MFR+ model relative to the NMR structure (1P7E) is smaller than found relative to the X-ray structure, both when using data from only one alignment medium (bicelles; 0.89 Å) and when using data measured in bicelles and phage media (0.49 Å). This smaller difference reflects a previously noted small change in the twist of the β-sheet, which is constrained by intermolecular hydrogen bonds in the crystalline lattice (Derrick and Wigley, 1994).

Tests Using Simulated Data

To further test the MFR method on larger systems, residual dipolar couplings were simulated for three proteins, using alignment tensors predicted on the basis of their three-dimensional structure. The targets chosen were proteins from the TALOS database (Cornilescu *et al*., 1999), for which nearly complete backbone chemical shift assignments and high-resolution X-ray crystal structures were available. Data for thioredoxin, profilin, and interleukin-1β were simulated using a version of the program PALES (Zweckstetter and Bax, 2000), which has been modified to include the effects of electrostatic alignment as appropriate for phage (Zweckstetter *et al*., 2004). Two different sets of residual dipolar couplings, corresponding to neutral bicelles and Pf1 phage media, were simulated for each protein. Couplings that are generally not measured with the standard methods (residues preceding Pro and residues with missing shifts) were removed from the coupling tables. Noise was added to the simulated, normalized couplings (1.0 Hz for thioredoxin and profilin; 1.5 Hz for interleukin-1β). Increasing the rms error in the simulated data up to 30% of the applicable D_a value has little effect on the regions of the target protein that yield unique fragments in the database search at low noise levels. However, it tends to increase the width of the selected fragment distribution for regions that exhibit ambiguity when smaller errors in the simulated data are used.

For thioredoxin [105 residues, X-ray crystal structure PDB entry 1ERT (Weichsel et al., 1996)] the chemical shifts were taken from Qin *et al*. (1996). For residues 2–104, the resulting MFR+ model exhibits a backbone RMSD of 0.83 Å relative to the X-ray structure. This difference (Fig. 7A) results mainly from small translational displacements of secondary structure

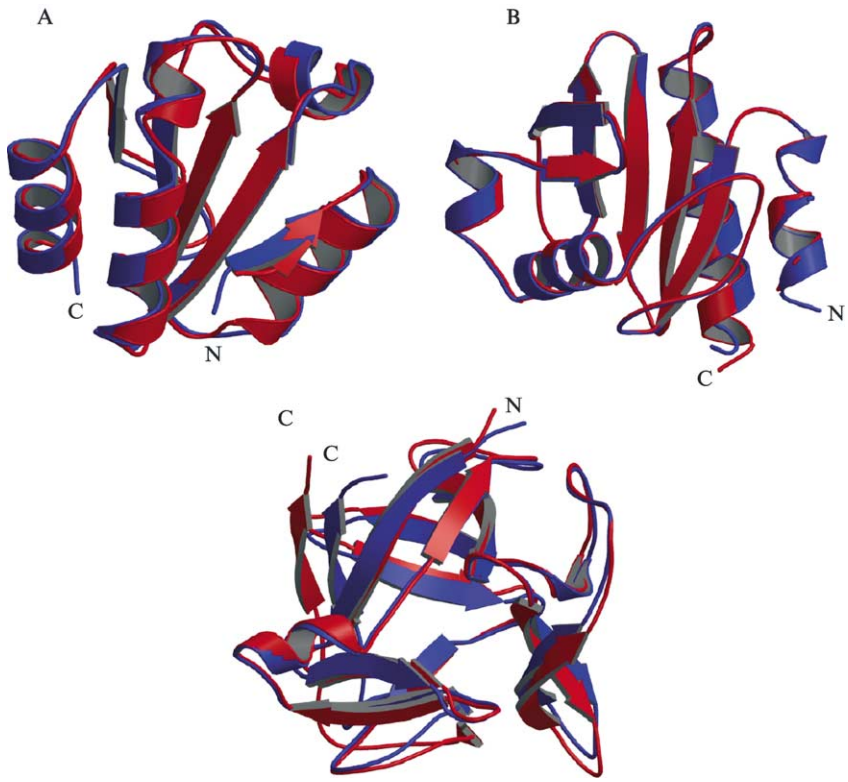


FIG. 7. Comparison of MFR+ structures [red (dark gray)] and X-ray reference structures [blue (light gray)] for (A) reduced human thioredoxin (PDB entry 1ERT), (B) profilin (PDB entry 1ACF), and (C) interleukin-1β (PDB entry 4ILB). MFR+ structures were derived from experimental chemical shifts and using two sets of dipolar couplings, simulated for the X-ray structures, for media containing 50 mg/ml bicelles and 15 mg/ml Pf1. Figures were generated using the programs Molscript (Kraulis, 1991) and Raster 3D (Merritt and Murphy, 1994).

elements, which presumably could be corrected by the use of a very minimal set of NOEs. Thioredoxin had one problematic loop region around M74–P75. At M74, in the region connecting the C-cap of α3 to β4, the backbone adopts an unusual conformation ($\phi = -80°$, $\psi = -107°$), slightly outside the allowed region of the Ramachandran plot. The presence of the Pro residue reduces the number of dipolar couplings and resulted in a ''crankshaft'' difference in this critical region, while retaining the correct orientation for adjacent residues. As a consequence, the whole region comprising strands β4, β5, and α4 shows a small lateral

displacement relative to the X-ray structure. Other interesting features, such as a kink in helix $\alpha 2$ at position I38 and a twist of strand $\beta 5$ at residue V86, are correctly recognized by MFR+.

Application to profilin [125 residues, X-ray crystal structure PDB entry 1ACF (Fedorov *et al.*, 1994); NMR structure PDB entry 2PRF (Archer *et al.*, 1994; Vinson *et al.*, 1993)] presented an interesting case due to its relatively high Gly contents (17 Gly residues). Gly residues can adopt unusual geometries that are less well represented in the PDB. However, with a backbone RMSD of only 0.44 Å, the final MFR+ model agrees remarkably well with the 1ACF structure that was used to generate the dipolar couplings (Fig. 7B).

For interleukin-1β [150 residues, X-ray crystal structure PDB entry 4ILB (Veerapandian *et al.*, 1992)] the chemical shifts were taken from Clore *et al.* (1990). Due to its complicated topology, this protein proved to be the most challenging test case for MFR+, largely because it consists almost exclusively of β-sheet. Although for conventional NOE-based structure determination this is generally beneficial, the large number of reverse turns and the possibility for the accumulation of error when the chain gets longer pose significant challenges to the MFR+ procedure. Moreover, due to the absence of a clear α-helical segment in interleukin-1β, it is less straightforward to extract accurate alignment tensor parameters. The inherently higher structural variability of β-strands makes these less suitable for such a purpose. Therefore, the relative orientation of the two alignment tensors could not be established a priori and could not be used as a restraint during the initial MFR search.

The presence of two adjacent Gly residues in the last loop (G139 and G140), which adopt an unusual conformation in a sparsely occupied region of the Ramachandran map, resulted in the lack of properly matching fragments in the database, and no prediction could be made for this region. Because the relative orientation of the last β-strand relative to the rest of the protein could be inferred from the two sets of RDCs, coordinates for G139 and G140 were added to the model by adding an extended Gly–Gly dipeptide to bridge the gap prior to subjecting the full protein to another cycle of regularization and refinement. However, because no NOEs were used in this process, the local geometry resulting from this procedure is only very approximate, resulting in a displacement by over 4 Å of the last β-strand relative to the crystal structure (Fig. 7C). This misplacement increased the overall backbone RMSD to 1.95 Å, but a considerably smaller difference (RMSD 1.41 Å) is obtained when considering only residues 5–134, excluding the last β-strand and its preceding loop. Other turns for which MFR+ did not yield accurate conformations include those between $\beta 2$ and $\beta 3$ and between $\beta 6$ and $\beta 7$.

Concluding Remarks

The MFR+ approach provides a remarkably direct way to determine solution NMR structures from protein backbone RDC data, either without or with inclusion of a small set of local backbone NOE data. The approach utilizes only protein backbone data and thereby bypasses the side chain and NOE assignment step. However, resulting structures have limitations that are distinct from those encountered in conventional, NOE-based structural studies. The most significant limitation of the MFR+ method in its application to full backbone structures is its requirement for relatively complete sets of residual dipolar couplings, preferably in two media. Short gaps of one residue at a time, which may result from exchange broadening, rapid solvent exchange, or the presence of a Pro residue, are typically easily bridged. Gaps longer than two residues at a time, especially in loop regions, often make it impossible to define the local structure uniquely. In such cases, only pieces or subdomains of a protein can be built reliably, and additional information such as NOE contacts or hydrophobic packing-based modeling is required to assemble these pieces correctly. Regions of the protein for which no well-matching substructures can be found in the database tend to be more problematic when applying the MFR+ procedure. This problem is compounded by the fact that the absence of tightly fitting fragments in the database occurs almost exclusively outside regions of well-defined secondary structure. If the best database hits are still relatively poor and structurally diverse, this can result in additional complications in the assembly process.

Owing to the symmetry of the dipolar interaction, if a backbone bond is parallel to any of the three principal axes of the alignment tensor, all couplings will be invariant to a 180° rotation about this bond, and data from a second alignment medium are required to resolve such an ambiguity. If the dataset for the other alignment medium happens to be incomplete for this region, it may be impossible to distinguish the two cases. In practice, it is usually the torsion angle ψ that causes such problems, as for many values of ψ, $\psi + 180°$ also falls in the allowed region of the Ramachandran map. With the exception of Gly residues, most 180° changes in ϕ result in severe steric clashes and then are easily filtered out.

As mentioned earlier, problems also arise when attempting to build extended loop regions that are not very well represented in the PDB. Particularly if the density of RDCs is low in such a region, it may become impossible to uniquely define matching substructures in the database. Such problematic areas usually can be spotted at an early stage of the fragment search by divergence in the ''Ramachandran flight map patterns,'' or at a later stage by inspection of the corresponding flight maps for the selected, refined fragments (Fig. 3).

Although our data demonstrate that reliable backbone models for small and medium sized proteins can be built on the basis of quite complete dipolar coupling and chemical shift data, application of the MFR+ program is not limited to these cases. Zweckstetter previously has shown that in favorable cases, fragments of a structure can be derived from chemical shifts together with as few as two dipolar couplings per residue (Zweckstetter and Bax, 2001). It is primarily at the assembly phase that additional data are needed. In its simplest form, tight backbone torsion angle restraints of the uniquely defined fragments could be used in conjunction with a limited number of backbone–backbone NOEs as restraints in a regular simulated annealing protocol to determine full structures in these cases. Alternatively, more sophisticated "docking" procedures based on rigid body refinement (Clore and Schwieters, 2002; Schwieters and Clore, 2001; Schwieters *et al.*, 2003) or Monte Carlo–based "shuffling" approaches (Rohl and Baker, 2002) may be used for assembling these fragments into a final structure. Note that the known orientations of each fragment (except for a 4-fold degeneracy) provide important additional restraints during such an assembly procedure.

Axially symmetric alignment tensors yield data that are less discriminating when building a structure from dipolar couplings than highly asymmetric (rhombic) alignment tensors (Delaglio *et al.*, 2000). In the axially symmetric case, only the unique axis (z-axis) of the local alignment frame is defined, and a rotational degree of freedom around this axis remains in the placement of fragments.

Our results demonstrate that it is possible for relatively small proteins to completely determine the backbone structures from backbone dipolar couplings. However, the MFR+ allows for convenient use of local NOEs too, when searching the database. The sequential $d_{\mathrm{H}\alpha\mathrm{HN}}(i, i+1)$ connectivity can be particularly useful for excluding fragments with the wrong ψ angle. Long-range NOEs are incorporated most easily after the initial model has been built, using a simulated annealing refinement protocol, where the backbone torsion angles are restrained relatively tightly to the values obtained from the initial model (excepting residues that proved uncertain after the fragment refinement procedure), and the dipolar coupling and NOE restraints are incorporated in the usual manner.

In its present implementation, the MFR+ method focuses only on building protein backbone structures, and side chains are essentially positioned according to their most likely conformation for the corresponding ϕ,ψ angles, using a database-derived empirical energy term (Dunbrack and Karplus, 1994; Kuszewski *et al.*, 1997). Our results indicate that this rather crude approach to side chain modeling yields reasonable results, with total increases in RMSD between the all-heavy-atom model and the corresponding reference structure typically being less than 1 Å. This is not much worse than found

for many medium resolution NMR structures in the PDB, if corresponding X-ray structures are taken as the reference. Not surprisingly, however, a substantial subset of side chains is poorly positioned with such an approach, but these tend to be easily identified by their very poor van der Waals contacts, using programs such as AQUA or PROCHECK (Laskowski *et al.*, 1996).

Although the MFR+ method here has been presented as a method for building protein structures without recourse to NOEs, it is likely that it will become most valuable in a hybrid approach where it is used to define structures of smaller fragments that subsequently require few NOEs for assembling them into a structure. Such an approach will be much less demanding in terms of completeness of the dipolar coupling data and will take advantage of the subset of NOEs that is frequently identified very easily, including H^N–H^N interactions.

References

Al-Hashimi, H. M., Valafar, H., Terrell, M., Zartler, E. R., Eidsness, M. K., and Prestegard, J. H. (2000). Variation of molecular alignment as a means of resolving orientational ambiguities in protein structures from dipolar couplings. *J. Magn. Reson.* **143,** 402–406.

Andrec, M., Du, P. C., and Levy, R. M. (2001). Protein backbone structure determination using only residual dipolar couplings from one ordering medium. *J. Biomol. NMR* **21,** 335–347.

Annila, A., Aitio, H., Thulin, E., and Drakenberg, T. (1999). Recognition of protein folds via dipolar couplings. *J. Biomol. NMR* **14,** 223–230.

Archer, S. J., Vinson, V. K., Pollard, T. D., and Torchia, D. A. (1994). Elucidation of the poly-L-proline binding-site in Acanthamoeba profilin-I by NMR-spectroscopy. *FEBS Lett.* **337,** 145–151.

Barrientos, L. G., Dolan, C., and Gronenborn, A. M. (2000). Characterization of surfactant liquid crystal phases suitable for molecular alignment and measurement of dipolar couplings. *J. Biomol. NMR* **16,** 329–337.

Bax, A., Kontaxis, G., and Tjandra, N. (2001). Dipolar couplings in macromolecular structure determination. *Methods Enzymol.* **339,** 127–174.

Beraud, S., Bersch, B., Brutscher, B., Gans, P., Barras, F., and Blackledge, M. (2002). Direct structure determination using residual dipolar couplings: Reaction-site conformation of methionine sulfoxide reductase in solution. *J. Am. Chem. Soc.* **124,** 13709–13715.

Bothner-by, A. A., Gayathri, C., Vanzijl, P. C. M., Maclean, C., Lai, J. J., and Smith, K. M. (1985). High-field orientation effects in the high-resolution proton NMR-spectra of diverse porphyrins. *Magn. Reson. Chem.* **23,** 935–938.

Braddock, D. T., Cai, M. L., Baber, J. L., Huang, Y., and Clore, G. M. (2001). Rapid identification of medium- to large-scale interdomain motion in modular proteins using dipolar couplings. *J. Am. Chem. Soc.* **123,** 8634–8635.

Brenneman, M. T., and Cross, T. A. (1990). A method for the analytic determination of polypeptide structure using solid-state nuclear magnetic-resonance—the metric method. *J. Chem. Phys.* **92,** 1483–1494.

Brunger, A. T. (1993). ''XPLOR: A System for X-ray Crystallography and NMR, 3.1 Ed.'' Yale University Press, New Haven, CT.

Bryce, D. L., and Bax, A. (2004). Application of correlated residual dipolar couplings to the determination of the molecular alignment tensor magnitude of oriented proteins and nucleic acids. *J. Biomol. NMR* **28,** 273–287.

Chou, J. J., Gaemers, S., Howder, B., Louis, J. M., and Bax, A. (2001a). A simple apparatus for generating stretched polyacrylamide gels, yielding uniform alignment of proteins and detergent micelles. *J. Biomol. NMR* **21,** 377–382.

Chou, J. J., Li, S. P., Klee, C. B., and Bax, A. (2001b). Solution structure of Ca2+-calmodulin reveals flexible hand-like properties of its domains. *Nat. Struct. Biol.* **8,** 990–997.

Clore, G. M. (2000). Accurate and rapid docking of protein-protein complexes on the basis of intermolecular nuclear Overhauser enhancement data and dipolar couplings by rigid body minimization. *Proc. Natl. Acad. Sci. USA* **97,** 9021–9025.

Clore, G. M., and Garrett, D. S. (1999). R-factor, free R, and complete cross-validation for dipolar coupling refinement of NMR structures. *J. Am. Chem. Soc.* **121,** 9008–9012.

Clore, G. M., and Schwieters, C. D. (2002). Theoretical and computational advances in biomolecular NMR spectroscopy. *Curr. Opin. Struct. Biol.* **12,** 146–153.

Clore, G. M., and Schwieters, C. D. (2003). Docking of protein-protein complexes on the basis of highly ambiguous intermolecular distance restraints derived from H-1(N)/N-15 chemical shift mapping and backbone N-15-H-1 residual dipolar couplings using conjoined rigid body/torsion angle dynamics. *J. Am. Chem. Soc.* **125,** 2902–2912.

Clore, G. M., Bax, A., Driscoll, P. C., Wingfield, P. T., and Gronenborn, A. M. (1990). Assignment of the side-chain H-1 and C-13 resonances of interleukin-1-beta using double-resonance and triple-resonance heteronuclear 3-dimensional NMR-spectroscopy. *Biochemistry* **29,** 8172–8184.

Clore, G. M., Gronenborn, A. M., and Bax, A. (1998a). A robust method for determining the magnitude of the fully asymmetric alignment tensor of oriented macromolecules in the absence of structural information. *J. Magn. Reson.* **133,** 216–221.

Clore, G. M., Starich, M. R., and Gronenborn, A. M. (1998b). Measurement of residual dipolar couplings of macromolecules aligned in the nematic phase of a colloidal suspension of rod-shaped viruses. *J. Am. Chem. Soc.* **120,** 10571–10572.

Cornilescu, G., Marquardt, J. L., Ottiger, M., and Bax, A. (1998). Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. *J. Am. Chem. Soc.* **120,** 6836–6837.

Cornilescu, G., Delaglio, F., and Bax, A. (1999). Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR* **13,** 289–302.

Delaglio, F., Kontaxis, G., and Bax, A. (2000). Protein structure determination using molecular fragment replacement and NMR dipolar couplings. *J. Am. Chem. Soc.* **122,** 2142–2143.

Derrick, J. P., and Wigley, D. B. (1994). The 3rd igg-binding domain from streptococcal protein-G—an analysis by X-ray crystallography of the structure alone and in a complex with Fab. *J. Mol. Biol.* **243,** 906–918.

Du, P. C., Andrec, M., and Levy, R. M. (2003). Have we seen all structures corresponding to short protein fragments in the Protein Data Bank? An update. *Protein Eng.* **16,** 407–414.

Dunbrack, R. L., and Karplus, M. (1994). Conformational-analysis of the backbone-dependent rotamer preferences of protein side-chains. *Nat. Struct. Biol.* **1,** 334–340.

Fedorov, A. A., Magnus, K. A., Graupe, M. H., Lattman, E. E., Pollard, T. D., and Almo, S. C. (1994). X-ray structures of isoforms of the actin-binding protein profilin that differ in their affinity for phosphatidylinositol phosphates. *Proc. Natl. Acad. Sci. USA* **91,** 8636–8640.

Fleming, K., Gray, D., Prasannan, S., and Matthews, S. (2000). Cellulose crystallites: A new and robust liquid crystalline medium for the measurement of residual dipolar couplings. *J. Am. Chem. Soc.* **122,** 5224–5225.

Fowler, C. A., Tian, F., and Prestegard, J. H. (2000). An NMR method for the rapid determination of protein folds using dipolar couplings. *Biophys. J.* **78,** 2827.

Giesen, A. W., Homans, S. W., and Brown, J. M. (2003). Determination of protein global folds using backbone residual dipolar coupling and long-range NOE restraints. *J. Biomol. NMR* **25,** 63–71.

Hansen, M. R., Mueller, L., and Pardi, A. (1998). Tunable alignment of macromolecules by filamentous phage yields dipolar coupling interactions. *Nat. Struct. Biol.* **5,** 1065–1074.

Hus, J. C., Marion, D., and Blackledge, M. (2000). *De novo* determination of protein structure by NMR using orientational and long-range order restraints. *J. Mol. Biol.* **298,** 927–936.

Ishii, Y., Markus, M. A., and Tycko, R. (2001). Controlling residual dipolar couplings in high-resolution NMR of proteins by strain induced alignment in a gel. *J. Biomol. NMR* **21,** 141–151.

Jones, T. A., and Thirup, S. (1986). Using known substructures in protein model-building and crystallography. *EMBO J.* **5,** 819–822.

Jones, T. A., Zou, J., Cowan, S. W., and Kjeldgaard, M. (1991). Improved methods for building protein models in electron density maps and location of errors in these models. *Acta Crystallogr. A.* **47,** 110–119.

Koenig, B. W., Hu, J. S., Ottiger, M., Bose, S., Hendler, R. W., and Bax, A. (1999). NMR measurement of dipolar couplings in proteins aligned by transient binding to purple membrane fragments. *J. Am. Chem. Soc.* **121,** 1385–1386.

Koradi, R., Billeter, M., and Wuthrich, K. (1996). MOLMOL: A program for display and analysis of macromolecular structures. *J. Mol. Graph* **14,** 51–55.

Kraulis, P. J. (1991). MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24,** 946–950.

Kraulis, P. J., and Jones, T. A. (1987). Determination of 3-dimensional protein structures from nuclear magnetic-resonance data using fragments of known structures. *Proteins* **2,** 188–201.

Kung, H. C., Wang, K. Y., Goljer, I., and Bolton, P. H. (1995). Magnetic alignment of duplex and quadruplex DNAs. *J. Magn. Reson. Ser. B* **109,** 323–325.

Kuszewski, J., Gronenborn, A. M., and Clore, G. M. (1997). Improvements and extensions in the conformational database potential for the refinement of NMR and X-ray structures of proteins and nucleic acids. *J. Magn. Reson.* **125,** 171–177.

Laskowski, R. A., Rullmann, J. A. C., MacArthur, M. W., Kaptein, R., and Thornton, J. M. (1996). AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* **8,** 477–486.

Losonczi, J. A., Andrec, M., Fischer, M. W. F., and Prestegard, J. H. (1999). Order matrix analysis of residual dipolar couplings using singular value decomposition. *J. Magn. Reson.* **138,** 334–342.

Lukavsky, P. J., Kim, I., Otto, G. A., and Puglisi, J. D. (2003). Structure of HCVIRES domain II determined by NMR. *Nat. Struct. Biol.* **10,** 1033–1038.

Marassi, F. M., and Opella, S. J. (1998). NMR structural studies of membrane proteins. *Curr. Opin. Struct. Biol.* **8,** 640–648.

Meier, S., Haussinger, D., and Grzesiek, S. (2002). Charged acrylamide copolymer gels as media for weak alignment. *J. Biomol. NMR* **24,** 351–356.

Meiler, J., Blomberg, N., Nilges, M., and Griesinger, C. (2000). A new approach for applying residual dipolar couplings as restraints in structure elucidation. *J. Biomol. NMR* **16,** 245–252.

Merritt, E. A., and Murphy, M. E. P. (1994). Raster3d Version-2.0—a program for photorealistic molecular graphics. *Acta Crystallogr. Sect. D-Biol. Crystallogr.* **50,** 869–873.

Mueller, G. A., Choy, W. Y., Skrynnikov, N. R., and Kay, L. E. (2000). A method for incorporating dipolar couplings into structure calculations in cases of (near) axial symmetry of alignment. *J. Biomol. NMR* **18,** 183–188.

Nishimura, K., Kim, S. G., Zhang, L., and Cross, T. A. (2002). The closed state of a H+ channel helical bundle combining precise orientational and distance restraints from solid state NMR-1. *Biochemistry* **41,** 13170–13177.

Ottiger, M., and Bax, A. (1998). Determination of relative N-H-N N-C′, C-alpha-C′, and C(alpha)-H-alpha effective bond lengths in a protein by NMR in a dilute liquid crystalline phase. *J. Am. Chem. Soc.* **120,** 12334–12341.

Prosser, R. S., Losonczi, J. A., and Shiyanovskaya, I. V. (1998). Use of a novel aqueous liquid crystalline medium for high-resolution NMR of macromolecules in solution. *J. Am. Chem. Soc.* **120,** 11010–11011.

Qin, J., Clore, G. M., and Gronenborn, A. M. (1996). Ionization equilibria for side-chain carboxyl groups in oxidized and reduced human thioredoxin and in the complex with its target peptide from the transcription factor NF kappa B. *Biochemistry* **35,** 7–13.

Ramirez, B. E., and Bax, A. (1998). Modulation of the alignment tensor of macromolecules dissolved in a dilute liquid crystalline medium. *J. Am. Chem. Soc.* **120,** 9106–9107.

Ramirez, B. E., Voloshin, O. N., Camerini-Otero, R. D., and Bax, A. (2000). Solution structure of DinI provides insight into its mode of RecA inactivation. *Protein Sci.* **9,** 2161–2169.

Rohl, C. A., and Baker, D. (2002). *De novo* determination of protein backbone structure from residual dipolar couplings using rosetta. *J. Am. Chem. Soc.* **124,** 2723–2729.

Ruckert, M., and Otting, G. (2000). Alignment of biological macromolecules in novel nonionic liquid crystalline media for NMR experiments. *J. Am. Chem. Soc.* **122,** 7793–7797.

Sass, H. J., Musco, G., Stahl, S. J., Wingfield, P. T., and Grzesiek, S. (2000). Solution NMR of proteins within polyacrylamide gels: Diffusional properties and residual alignment by mechanical stress or embedding of oriented purple membranes. *J. Biomol. NMR* **18,** 303–309.

Sass, J., Cordier, F., Hoffmann, A., Rogowski, M., Cousin, A., Omichinski, J. G., Lowen, H., and Grzesiek, S. (1999). Purple membrane induced alignment of biological macromolecules in the magnetic field. *J. Am. Chem. Soc.* **121,** 2047–2055.

Schwieters, C. D., and Clore, G. M. (2001). Internal coordinates for molecular dynamics and minimization in structure determination and refinement. *J. Magn. Reson.* **152,** 288–302.

Schwieters, C. D., Kuszewski, J. J., Tjandra, N., and Clore, G. M. (2003). The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson.* **160,** 65–73.

Tjandra, N., and Bax, A. (1997). Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium [see comments]. *Science* **278,** 1111–1114.

Tjandra, N., Feller, S. E., Pastor, R. W., and Bax, A. (1995). Rotational diffusion anisotropy of human ubiquitin from N-15 NMR relaxation. *J. Am. Chem. Soc.* **117,** 12562–12566.

Tjandra, N., Grzesiek, S., and Bax, A. (1996). Magnetic field dependence of nitrogen-proton J splittings in N-15-enriched human ubiquitin resulting from relaxation interference and residual dipolar coupling. *J. Am. Chem. Soc.* **118,** 6264–6272.

Tjandra, N., Omichinski, J. G., Gronenborn, A. M., Clore, G. M., and Bax, A. (1997). Use of dipolar H1-N15 and H1-C13 couplings in the structure determination of magnetically oriented macromolecules in solution. *Nat. Struct. Biol.* **4,** 732–738.

Tolman, J. R., Flanagan, J. M., Kennedy, M. A., and Prestegard, J. H. (1995). Nuclear magnetic dipole interactions in field-oriented proteins—information for structure determination in solution. *Proc. Natl. Acad. Sci. USA* **92,** 9279–9283.

Tycko, R., Blanco, F. J., and Ishii, Y. (2000). Alignment of biopolymers in strained gels: A new way to create detectable dipole-dipole couplings in high-resolution biomolecular NMR. *J. Am. Chem. Soc.* **122,** 9340–9341.

Ulmer, T. S., Ramirez, B. E., Delaglio, F., and Bax, A. (2003). Evaluation of backbone proton positions and dynamics in a small protein by liquid crystal NMR spectroscopy. *J. Am. Chem. Soc.* **125,** 9179–9191.

Veerapandian, B., Gilliland, G. L., Raag, R., Svensson, A. L., Masui, Y., Hirai, Y., and Poulos, T. L. (1992). Functional implications of interleukin-1-beta based on the 3-dimensional structure. *Proteins* **12,** 10–23.

Vijay-Kumar, S., Bugg, C. E., and Cook, W. J. (1987). Structure of ubiquitin refined at 1.8 A resolution. *J. Mol. Biol.* **194,** 531–544.

Vinson, V. K., Archer, S. J., Lattman, E. E., Pollard, T. D., and Torchia, D. A. (1993). 3-Dimensional solution structure of Acanthamoeba profilin-I. *J. Cell Biol.* **122,** 1277–1283.

Wand, A. J., Urbauer, J. L., McEvoy, R. P., and Bieber, R. J. (1996). Internal dynamics of human ubiquitin revealed by C-13-relaxation studies of randomly fractionally labeled protein. *Biochemistry* **35,** 6116–6125.

Weichsel, A., Gasdaska, J. R., Powis, G., and Montfort, W. R. (1996). Crystal structures of reduced, oxidized, and mutated human thioredoxins: Evidence for a regulatory homodimer. *Structure* **4,** 735–751.

Wu, C. H., Ramamoorthy, A., Gierasch, L. M., and Opella, S. J. (1995). Simultaneous characterization of the amide H-1 chemical shift, H-1-N-15 dipolar, and N-15 chemical-shift interaction tensors in a peptide-bond by 3-dimensional solid-state NMR-spectroscopy. *J. Am. Chem. Soc.* **117,** 6148–6149.

Zweckstetter, M., and Bax, A. (2000). Prediction of sterically induced alignment in a dilute liquid crystalline phase: Aid to protein structure determination by NMR. *J. Am. Chem. Soc.* **122,** 3791–3792.

Zweckstetter, M., and Bax, A. (2001). Single-step determination of protein substructures using dipolar couplings: Aid to structural genomics. *J. Am. Chem. Soc.* **123,** 9490–9491.

Zweckstetter, M., and Bax, A. (2002). Evaluation of uncertainty in alignment tensors obtained from dipolar couplings. *J. Biomol. NMR* **23,** 127–137.

Zweckstetter, M., Hummer, G., and Bax, A. (2004). Prediction of charge-induced molecular alignment of biomolecules dissolved in dilute liquid-crystalline phases. *Biophys. J.* **86,** 3444–3460.

[4] Rapid NMR Data Collection

By HANUDATTA S. ATREYA and THOMAS SZYPERSKI

Abstract

Rapid data collection is an area of intense research in biomolecular NMR spectroscopy, in particular for high-throughput structure determination in structural genomics. NMR data acquisition and processing protocols for rapidly obtaining high-dimensional spectral information aim at avoiding sampling limited data collection and are reviewed here with emphasis on G-matrix Fourier transform NMR spectroscopy.

0076-6879/05 $35.00