Current Topics

Nuclear Magnetic Resonance in the Era of Structural Genomics[†]

J. H. Prestegard,* H. Valafar, J. Glushka, and F. Tian

Complex Carbohydrate Research Center, University of Georgia, Athens, Georgia 30602

Received January 31, 2001; Revised Manuscript Received June 7, 2001

ABSTRACT: Current interests in structural genomics, and the associated need for high through-put structure determination methods, offer an opportunity to examine new nuclear magnetic resonance (NMR) methodology and the impact this methodology can have on structure determination of proteins. The time required for structure determination by traditional NMR methods is currently long, but improved hardware, automation of analysis, and new sources of data such as residual dipolar couplings promise to change this. Greatly improved efficiency, coupled with an ability to characterize proteins that may not produce crystals suitable for investigation by X-ray diffraction, suggests that NMR will play an important role in structural genomics programs.

The enormous progress that has been made in the sequencing of DNA over the past few years is forcing much of the biochemistry and molecular biology community to reassess its approach to scientific investigation and discovery. Rather than focus on a single system with dogged testing of hypotheses related to mechanisms of action, there is a sense that new insight, or at least new hypotheses, can be generated by taking a broader view and systematically analyzing vast sets of data that have been accumulated in the absence of highly focused objectives. The human genome alone is expected to encode on the order of 30 000 proteins; add to this the genomes of agriculturally important plants and the genomes of pathogenic organisms, and the amount of information to be tapped is staggering (*1*). Most fruits of this new mode of investigation will not come, however, without additional effort, both in analysis and in information gathering. It is estimated that current analysis techniques can assign a function to no more than half of the proteins encoded by newly sequenced genes (*2*, *3*). To improve this situation, new efforts in functional genomics and proteomics have been initiated. Among the proteomics efforts are ones in structural genomics. It is the objective of this paper to discuss new ways that nuclear magnetic resonance (NMR) can contribute to the structural genomics effort. NMR has to date contributed just 17% of the structures currently deposited in the Protein Data Bank (PDB),[1] compared to 82% by X-ray crystallography (*4*), but broad applicability to proteins which fail to provide diffraction quality crystals promises to increase the importance of contributions by NMR in the future.

Structural Genomics

Structural genomics, more properly structural proteomics, refers to the attempt to provide three-dimensional structural information about a significant fraction of the proteins encoded by the genes sequenced in various genome projects. The attempt is driven in part by a belief that three-dimensional structure will provide a better basis for recognition of function than sequence similarity (*5*−*8*). This belief is, in fact, supported by numerous examples of proteins of similar function that have little sequence homology but recognizable geometric similarities once functionally important residues are attached to specific points in a backbone fold (*9*, *10*). Given the possible utility of folds in these applications, as well as their use as a basis for homology modeling, most structural genomics efforts will not target all proteins in a genome, but will attempt to produce representative structures in each protein fold family. With fold families estimated to number from several thousand to ten thousand, this is still a daunting task, but one that may allow emerging computational methods to produce structures on a more global basis (*11*).

Meeting the challenge of structural genomics is now a worldwide effort with a large program underway in Japan (*12*) and new programs starting in Europe (*13*). Recently, the National Institute of General Medical Sciences (NIGMS), a part of the National Institutes of Health, announced the establishment of seven pilot centers in the United States for testing and establishing protocols for the large-scale production of protein structures (*14*). These new centers will build on a number of preexisting structural genomics projects in North America (*15*) and mesh with efforts to standardize

[†] This work was supported by grants from the National Science Foundation (MCB-9707728) and the National Institutes of Health (GM-62407).

[1] Abbreviations: PDB, Protein Data Bank; HSQC, heteronuclear single-quantum coherence; CLEANEX, clean chemical exchange spectroscopy; rmsd, root-mean-square deviation; BFP, blue fluorescent protein; 1D, 2D, and 3D, one-, two-, and three-dimensional, respectively; NOE, nuclear Overhauser effect; NOESY, nuclear Overhauser effect spectroscopy; TROSY, transverse relaxation-optimized spectroscopy; COSY, coupling-correlated spectroscopy; DMPC, dimyristoylphosphatidylcholine; DHPC, dihexanoylphosphatidylcholine; DMPG, dimyristoylphosphatidylglycerol.

10.1021/bi0102095 CCC: $20.00
Published on Web 07/04/2001

procedures for capturing and storing both X-ray and NMR data (*16*−*18*).

Most of the new NIH centers will rely heavily on X-ray crystallography to produce structures. This is not surprising given the large percentage of the structures currently in the PDB that have been produced by X-ray crystallography. The advances in data collection efficiency at X-ray producing synchrotron sources will also ensure that X-ray crystallography will continue to be the major contributor to this database; however, one of the new centers does have NMR as a major component, and NMR is well represented in several others. NMR is also a major component in the Japanese and European efforts. As we discuss below, there is reason to believe that investment in this second contributor to protein structure determination will be productive.

NMR in Structural Genomics

There are several reasons to suggest that NMR will play a role in structural genomics activities. NMR can provide an important technique for selecting well-behaved proteins and optimizing conditions for structure determination, whether it be by NMR or X-ray crystallography. NMR can provide a means of identifying small ligands as well as macromolecular associations that may be essential for proper folding and function. And, NMR can provide an inherently complementary structure determination methodology. With respect to its complementarity, NMR structures are produced in solution, and there are a few documented cases where structural differences exist between solution and crystal forms of proteins (*19*−*21*). However, these cases are small in number, and differences are not likely to be significant at the level of structures for a protein fold database. A more important aspect is the possibility that NMR can provide structures for proteins that may be difficult to crystallize. For example, producing crystals of membrane proteins is still a challenge. Integral membrane proteins of the helical bundle class are estimated to represent 20−25% of most genomes (*22*). Yet, only ∼1% of the structures deposited in the current PDB are classified as membrane proteins (*4*). There are NMR structures of several small proteins with transmembrane helical segments, some produced in micelle media by high-resolution techniques and some produced in more extended membrane mimetics using solids NMR techniques (*23*, *24*). Also, there are new techniques on the horizon that promise to make membrane proteins in micelles more accessible to high-resolution NMR (*25*). Several of the structural genomics pilot centers plan to initially exclude membrane proteins from consideration because of the difficulties in structure production by either NMR or X-ray crystallography. But membrane proteins will eventually have to be tackled, and they may not be the only class of proteins that prove to be difficult to crystallize.

It is widely recognized that heterogeneity in protein preparations (due to aggregation, for example) is detrimental to crystallization, and screening using light scattering is often used to avoid conditions that contribute to heterogeneous aggregation; both NMR and X-ray approaches suffer limitations due to aggregation. However, other factors are more detrimental to crystal-based approaches than NMR-based approaches to structure. For example, posttranslational modifications such as glycosylation seem to limit crystallizability, and disordered regions that may be integral parts of the protein structure seem to limit crystallizability (*26*). These latter factors do not normally prevent NMR-based structure determination of at least the well-structured regions.

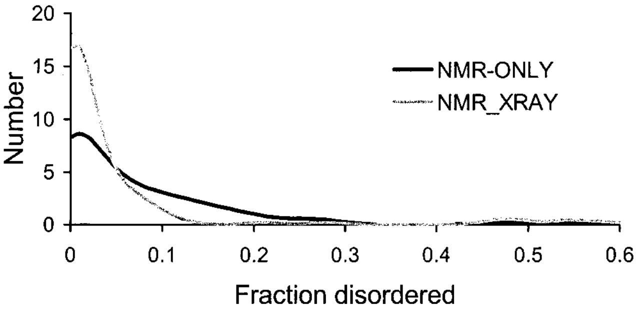

FIGURE 1: Influence of unstructured regions in proteins on the choice of structure determination approach. Data were taken from the July 1999 version of the Protein Data Bank.

It is possible to support an argument for the applicability of NMR in cases where crystallization has proved to be difficult by an examination of current versions of the PDB (*4*). There are now (January 2001) approximately 14 000 structures in the PDB; approximately 2200 have been determined by NMR methods. Most of these are deposited, not as a single structure, but as a set of 20 or more structures determined using random starting conditions in the structure determination protocol. Regions along the polypeptide backbone that are poorly defined can be located by calculating a root-mean-square deviation (rmsd) from the mean of the positions of backbone atoms residue by residue. We take an arbitrary limit of 4.0 Å as an indicator of poor structural definition. Poor definition can come simply from the absence of adequate NMR constraints, but the absence of constraints frequently correlates with motion or disorder in a region of the protein. Making this correlation, we can use large rmsds as an indicator of disordered regions in a protein.

Entries based on NMR (1504 from the July 1999 PDB) were filtered to produce a nonredundant set of proteins having more than 50 residues and adequate numbers of deposited structures (918). These entries were then evaluated on the basis of the fraction of residues showing an rmsd of >4.0 Å. The structures represented by these entries were divided into two sets on the basis of whether a corresponding X-ray structure existed. A search of all X-ray structures was conducted using a sequence identity search algorithm that allowed modest levels of amino acid substitution, deletion, or insertions (fewer than 12 residues different in size and fewer than 4 unmatched residues). On this basis, 182 were found to have a corresponding structure determined by X-ray crystallography. A search of the remaining NMR structures was conducted to select a set of comparable size having the least similarity to existing X-ray structures. A 163-member set having a BLAST (*27*) alignment homology score of less than 43 when compared to any X-ray structure was selected. The two subsets, NMR_ONLY and NMR_XRAY, were examined separately for the extent of disorder as given by the rmsd criterion applied to the NMR structure.

The analysis is presented graphically in Figure 1. For proteins with a low percentage of disorder, related X-ray structures are abundant, but for proteins with a higher percentage of disorder (more than 10%), the number of related X-ray structures drops off dramatically. NMR-based structures persist to a much higher percentage of disorder.

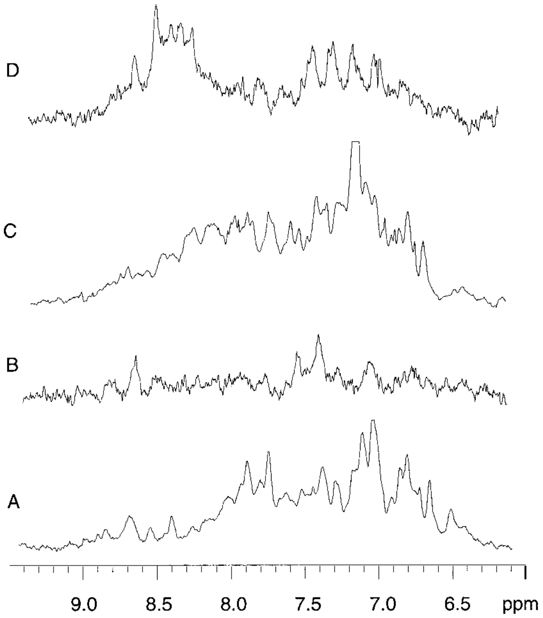

FIGURE 2: Screening for structured proteins using 1D amide proton exchange spectra. The sample is blue fluorescent protein at 0.5 mM: (A) 1D spectrum with Ca^{2+}, (B) CLEANEX spectrum with Ca^{2+}, (C) 1D spectrum without Ca^{2+}, and (D) CLEANEX spectrum without Ca^{2+}.

Such observations clearly support the suggestion that NMR-based structure determination is more applicable to proteins with disordered regions, possibly because these proteins may be difficult to crystallize.

NMR as a Screening Tool

Whether NMR is able to produce structures for disordered proteins, a simple ability to identify these proteins, or proteins that are heterogeneous because of aggregation or other conformational effects, would be a valuable contribution. Screening of expressed and purified proteins for sample conditions that are apt to promote crystallization or give good NMR samples is therefore a major activity of many of the pilot centers. We present in Figure 2 an illustration of the type of experiment that might be carried out on a very small amount of protein in a highly automated way. The experiment incorporates a simple test for rapidly exchanging amide protons. At pH 7, amide protons of unstructured regions of polypeptide chains undergo exchange with protons of water on time scales of tenths of seconds or less (*28*). In structured regions, amide protons are either buried in the hydrophobic interior of the protein or involved in hydrogen bonding. These amides exchange much more slowly (minutes to hours). A simple magnetization transfer experiment that uses magnetization associated with protons of water to provide detectable protein signal when they exchange into rapidly exchanging sites will selectively show amides in disordered regions. [In actual fact, a more sophisticated experiment that eliminates artifacts due to transfers from α-protons underlying the water resonance, CLEANEX, is used (*29*).]

In Figure 2, we show the amide proton regions of one-dimensional proton NMR spectra of a 20 kDa protein, obelin. Obelin is a photoprotein that emits light on addition of Ca^{2+} in a manner similar to that of its close relative, aquorin. The crystal structure of obelin has recently been determined, but only in the pre-emission form (*30*); it has not so far been possible to produce diffraction quality crystals of obelin in its reacted form, a form also known as blue fluorescent protein (BFP). Figure 2A shows the amide region of a simple 1D spectrum of BFP in the presence of Ca^{2+}; signals from all amides as well as a few aromatic protons are present. In Figure 2B, we show a spectrum of the same region produced using the CLEANEX experiment. As only signals that derive magnetization from protons on water can be detected in this experiment, those seen are from amide protons in moderately rapid exchange (exchange times of fewer than a few seconds) with water. The intensity in the 7.0−7.8 ppm region is seen for most proteins and includes intensity from side chain amide protons such as those on glutamine and asparagine. The few signals in the 8.1−8.9 ppm region are typical of a well-folded protein with few unstructured or surface-exposed amides. The spectrum suggests that at least in the presence of Ca^{2+} production of quality crystals may be possible. Spectra C and D of Figure 2 show a similar pair, but from the protein with Ca^{2+} removed. The intensity in the 8.1−8.9 ppm region is now abnormally high and indicative of partial unfolding of the backbone. The data suggest that the protein would be difficult to crystallize in the absence of Ca^{2+}, but the protein may be sufficiently folded for an NMR study. Screening based on experiments such as the CLEANEX experiment can be made quite efficient, and even automated, using NMR flow probes and micro manipulator robots. The spectra that are shown were acquired in approximately 15 min each on 0.5 mM protein samples using a standard 600 MHz spectrometer.

There is of course more to be gained from screens based on NMR spectra. Chemical shifts contain information. The appearance of the transfer intensity near 8 ppm is an example. This is a region characteristic of amides in random coil configuration. α-Proton chemical shifts in the 4−5 ppm region can provide equally valuable indicators of α-helix or *β*-strand structures, as can chemical shifts of the backbone α-carbons (*31*). Screening based on various types of NMR experiments has been adopted by several structural genomics projects (*32*, *33*). Screening is often based on 2D $^{15}N-^{1}H$ NMR, rather than the simple one-dimensional proton experiments described above. Resolution is greatly improved in 2D experiments but requires ^{15}N-labeled protein (this can be accomplished by growing *Escherichia coli* in minimal medium supplemented with [^{15}N]ammonium chloride at modest additional cost). The experiment employed is a workhorse 2D experiment called a heteronuclear single-quantum coherence (HSQC) experiment. The resulting spectrum shows cross-peaks, approximately one for each residue, that correlate chemical shifts of an amide proton with that of the directly bonded nitrogen. The experiment is efficient, requiring on the order of 15 min on a 1 mM protein sample using a modern 600 MHz NMR spectrometer. Positions of peaks can suggest disordered regions, and once the peaks are assigned to sequential positions, the information could actually be used to engineer proteins with the unfavorable segments removed. Also, given the normal count of approximately one cross-peak per residue, the detection of a surplus of peaks can indicate conformational heterogeneity that may adversely affect structure determination. The efficiency of the experiment also allows its use along with variation of pH or ionic strength or addition of cofactors in identification of conditions that produce minimal heterogeneity as well as optimum resolution for subsequent NMR experiments.

3D Structures by NMR

With the seeming advantages of NMR, why stop at screening; why not routinely continue on to complete structures? Obviously, there must be reasons if only 15−20% of the deposited structures in the PDB are NMR-derived. One reason that NMR has not played a larger role in structure determination is the limitation on sizes of proteins to which NMR methods can be applied. The limit on size stems from the need to resolve and assign resonances to particular proton sites in the protein sequence. Cross-peaks between assigned resonances in nuclear Overhauser effect spectroscopy (NOESY) ultimately give rise to the pairwise distance constraints used in most NMR structure determinations (*34*). Resolution depends inversely on resonance line widths, and the widths depend in turn on effective molecular weight. Limits imposed by resolution have been pushed back over the years, first by use of isotope enrichment (^{13}C, ^{15}N, and ^{2}H) (*35*) and extension of spectra to three and four dimensions and more recently by transverse relaxation optimization techniques (TROSY) in combination with high magnetic fields (*25*). However, complete structure determination of monomeric proteins still has not pushed beyond a molecular mass limit of 50 kDa, and determinations not requiring deuteration of the protein still seem to be limited to 25−30 kDa (*33*). While this is a severe limitation, it is mitigated by the fact that the average domain size of encoded proteins appears to be only slightly greater than 150 amino acids, or 17 kDa (*3*).

A more severe limitation is that the time required for data acquisition and analysis is long, and sample preparation requires the use of isotopically labeled media (^{15}N- and ^{13}C-labeled proteins). There have been enormous strides made in the efficient production of proteins through expression in *E. coli* (*32*), and new cell-free production techniques pioneered in Japan promise more latitude in produced proteins and incorporated labels (*12*). However, the 4−6 weeks of acquisition and subsequent months-long periods required for assignment and structure determination is still a major obstacle (*33*). This time scale is not compatible with structural genomics objectives that would require 100−200 structures per year from each of the seven NIH-sponsored pilot centers (*14*).

Automation of NOE-Based Methods

One approach to reducing the time requirements relies on extensive automation to reduce the analysis time and improved instrument hardware to reduce data acquisition time. Isotopic labeling of proteins with both ^{13}C and ^{15}N has led to a very reliable and automatable assignment strategy. Primary experiments use one bond scalar coupling connectivities to walk along the peptide backbone. Extending connectivities out to the β carbons of the amino acid side chains allows assignment of many sequentially connected residues to amino acid types based on chemical shift correlations. Inclusion of experiments such as HCCH TOCSY that use isotropic mixing sequences to make multiple bond connectivities through the amino acid side chains allows assignment of most carbon and proton resonances. The program AUTOASSIGN developed by the Rutgers group is representative of programs that provide a very efficient automation of much of this general strategy (*36*). While AUTOASSIGN and most other automated assignment programs rely on ^{13}C and ^{15}N isotopic labeling (*37*), some can be used to aid assignment even in the absence of ^{13}C labeling (*38*−*40*). These alternate strategies may become important for proteins that are difficult to express in bacterial systems or cell-free systems.

Assignment of resonances, however, still falls short of assignment of peaks in NOESY spectra. This is a particularly challenging task because of the severe overlap and resulting ambiguities in assignment. Automated methods for this step fortunately also exist (*41*, *42*). They mesh with commonly used simulated annealing protocols for structure determination and allow iterative determinations for elimination of ambiguities in assignment. All of these automated assignment tools are still very much in the development stage. The number of structures produced with the aid of automation is at this point small, but structural genomics efforts will certainly require implementation of these techniques. In a recent example that used some of the automated tools that are mentioned, time from receipt of a plasmid containing a 90-residue soluble (2−3 mM) protein to production of a tertiary fold was reduced to <1 month (*43*).

Very recently, the promise of reducing actual data collection time has come from the introduction of NMR cryoprobes in which the receiver coil and associated electronic components are cooled to very low temperatures (*33*). Sensitivity improvements on protein samples of a factor of ∼3 can be achieved. This translates to a savings in time of a factor of 9 when substantial signal averaging is required. A recent example on a protein of 180 residues shows a reduction to 1.5 days for experiments required for backbone assignments (*44*).

Substantial savings can also be achieved in acquisition and analysis of experiments needed for side chain assignments and NOE cross-peak analysis by selective protonation of groups in otherwise fully deuterated proteins. The object is to reduce the NOE peaks to a set very rich in useful distance constraints, often a set containing peaks from side chains of residues concentrated at the hydrophobic core of proteins. This strategy was first well-illustrated in the work of Gardner and Kay where methyl groups were selectively protonated (*45*). The strategy has recently been extended by Fesik and co-workers and combined with the use of a cryoprobe to acquire sufficient data for the determination of the fold of a 180-residue protein in just 4 days (*44*). The procedure does, however, require preparation of several samples with different distributions of isotopic labels, and it does require the use of amino acids synthesized to have ^{13}C and/or protons in specific places.

Applications of NOE-Based Methods

It is useful to examine at least one recent example of using the triple-resonance assignment, NOE-based structure determination, approach in a structural genomics application. We choose an application to MHT538, a protein of approximately 110 amino acids from the thermophilic archaeon *Methanobacterium thermoautotrophicum* (*46*). Aside from efficiency issues, which are not discussed in detail, this application includes a nice illustration of additional ways that NMR might contribute in pushing past structure, to the point of functional characterization. The protein target was chosen without hypothesis as to function in an attempt to see what could be learned from the structure itself. A PSI-BLAST search (*27*) using the MHT538 sequence suggested a weak similarity (18−23% for a 95-residue segment) to members of a family of proteins annotated as putative ATPases or kinases (COG1618). As this level of similarity would normally be inadequate for prediction of any structural homology, an NMR-based structure determination proceeded using ^{13}C- and ^{15}N-labeled protein and a battery of experiments similar to that described above. A structure falling into the fold family $(\alpha/\beta)^5$ (minus the helix between strands 2 and 3) was determined with a 0.88 Å rmsd for backbone atoms of residues 4−108. This structure does not, in fact, bear a relationship to structures found for other members of the sequence family of putative ATPases or kinases. Using tools for finding structural homologies [SCOP (*47*) and DALI (*48*)], a relationship to either flavodoxins, or a receiver domain of two-component response regulator systems such as that of CheY, was found.

At this point, NMR screening methodology was used to draw a distinction between these two functional classification choices. As described previously, the simple two-dimensional ^{15}N−^{1}H heteronuclear single-quantum coherence (HSQC) spectrum provides a map of the protein backbone with approximately one cross-peak for each amino acid (each H−N amide pair). Movement of specific cross-peaks on binding of ligands to proteins is now used extensively in the pharmaceutical industry to identify binding sites on the protein surface (*49*). In the MTH538 study, this technique was used to explore possible flavin interactions, as expected if the protein were a flavodoxin, and possible Mg^{2+} binding, as expected for receiver domains; no flavin binding was detected, but Mg^{2+} binding to a site homologous to one in CheY was found. Although some other characteristics of receiver domains were missing, the close structural homology, along with the specific Mg^{2+} site, was used to suggest a possible receiver domain function.

Fold Determination Methods

Despite the advances in the efficiency of traditional NMR approaches to structure determination, and the successful illustrations of the contributions that these NMR approaches can make to structural genomics, the rate of data production still pales in comparison to that of synchrotron-based X-ray crystallography. This observation suggests that some reconsideration of both the objectives and approaches used in NMR applications to structural genomics projects might prove to be useful. First, in terms of an objective, is the traditional complete high-resolution structure appropriate for structural genomics applications? High-resolution structures will always be an important part of detailed mechanistic investigations, but functional classification of proteins may be a different matter. One of the principal hopes of structural genomics is that representative structures in each of a few thousand fold families will provide a basis upon which computational biologists can produce structures of related proteins and draw conclusions about function. When homology modeling is used on systems that have sequences that are as little as 25% identical, one clearly does not use the precise placement of either backbone or side chain atoms of representative structures in a fold library. Also, functional identification algorithms have been devised that work on the basis of approximate placement of backbone atoms as opposed to precise placement of side chain atoms (*6*). In the future, new computational methods for placing side chains into backbone structures promise to make backbone structures even more useful (*50*, *51*). In these cases, placement of backbone atoms with accuracy that is better than that which can be achieved by homology modeling appears to be necessary, and new experimental methods directed at accurate backbone structure determination will be useful.

If backbone structures become a primary objective, the traditional NOE-based NMR approach to structure can also be questioned. This approach is not optimal for direct investigation of backbone structures because the short-range character of the NOE, from which distance constraints are derived, dictates that side chain−side chain contacts in the hydrophobic core play an important part in determining a protein fold. However, a source of structural data that can constrain more directly backbone atom positions has come on the scene recently, residual dipolar couplings (*52*, *53*). The same dipolar interaction that gives rise to the NOE actually contains both angle- and distance-dependent terms. In the NOE, the angle-dependent part is responsible for modulation of the interaction as a protein tumbles in solution and is essential for making the distance dependence of NOEs measurable as a spin relaxation phenomenon. However, it normally makes no direct contribution to spin state energy differences that might be efficiently measured through variations in resonance positions. This is because the angular term rigorously averages to zero over the time course of molecular tumbling in isotropic solutions. Measurements of residual dipolar couplings depend on restoring small levels of directional order to proteins as they tumble in solution (of order one part in a thousand). This was first done for NMR observations on proteins using the inherent tendency of paramagnetic proteins to orient in high magnetic fields (*54*), but since that time, the use of aqueous liquid crystal media such as bicelles (*53*), phage (*55*, *56*), or cellulose microcrystals (*57*) has become standard practice.

The introduction of order leads to a residual contribution to the splitting of resonance multiplets, $D^{\text{res}}{}_{ij}$. In the case of a directly bonded pair of spin $^{1}/_{2}$ nuclei (^{1}H−^{15}N, ^{1}H−^{13}C, etc.), where the internuclear distance is known, the contribution to splitting becomes a simple function of order and orientation adopted by the system. The effect of order is in turn conveniently expressed in terms of elements of an order tensor (S_{ij}) and direction cosines relating the internuclear vector to a principal order frame [$\cos(\alpha_k)$] (*58*):

$$D^{\text{res}}{}_{ij} = -\left(\frac{\mu_0}{4\pi}\right)\frac{\gamma_i\gamma_j h}{2\pi^2 r_{ij}{}^3}\sum_{kl} S_{kl}\cos(\alpha_k)\cos(\alpha_l) \qquad (1)$$

Despite the apparent complexity of the equation, it has only five independent parameters; these can alternately be described in terms of three Euler angles relating the orientation of a molecular fragment to a principle alignment frame and principal and rhombic alignment parameters (*53*). The latter two apply to all parts of a rigid protein molecule. The first

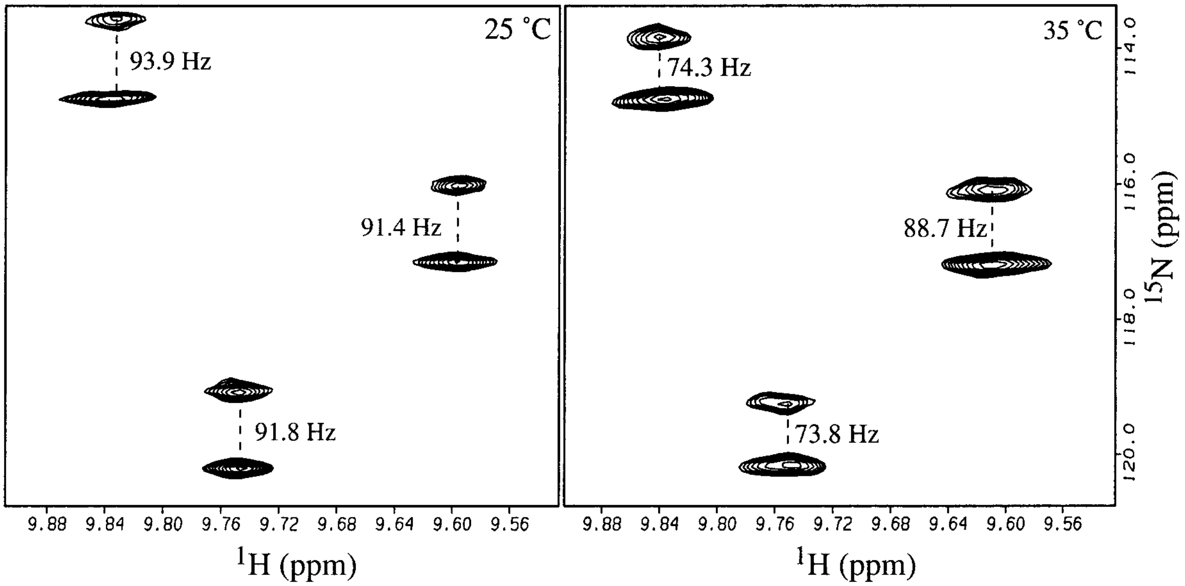

FIGURE 3: Measurement of $^{15}N-^{1}H$ residual dipolar couplings in amide groups using coupled HSQC spectra. The sample is 0.3 mM ADP ribosylating factor (ARF) in a 7% bicelle preparation (2.8:1 DMPC/DHPC or 15:1 DMPC/DMPG) at pH 6.5 (100 mM NaCl, 1 mM $MgCl_2$, and 10 mM NaH_2PO_4). The sample is isotropic at 25 °C and ordered at 35 °C.

three can be determined independently for any semirigid fragment in a protein, and can serve to relate orientations of identifiable structural elements.

Part of the appeal of residual dipolar couplings measurements for structural genomics applications is the efficiency with which these data can be acquired. Measurement of $^{15}N-^{1}H$ dipolar couplings of directly bonded amide pairs is based on the very efficient HSQC experiment described above. Modifications necessary for the measurement of residual dipolar couplings can be as simple as removing the normal 180 pulse used for refocusing proton scalar coupling during the indirect detection period. When this is done, the single cross-peak for each amide turns into a doublet with a splitting at the sum of one-bond scalar and dipolar contributions. The dipolar part is usually extracted by acquiring spectra under both isotropic and oriented conditions. As illustrated in Figure 3 for an ^{15}N-labeled sample of a 20 kDa protein, ADP ribosylating factor, this can be simply done using bicelle systems in which isotropic and ordered states can be produced by changing the temperature from 25 to 35 °C. At 25 °C, only scalar couplings contribute, and splittings are all of a similar magnitude. At 35 °C, the bicelles, which are simply discoidal pieces of lipid bilayer, order with normals perpendicular to the magnetic field. Occasional collisions of proteins with the ordered surfaces impart a small degree of order; the dipolar contributions to couplings between ^{1}H and ^{15}N nuclei fail to average to zero, and splittings increase or decrease depending on the angles that a particular internuclear vector make with the axes of the order frame. There are of course a number of more sophisticated pulse sequences that allow measurement of $^{1}H-^{15}N$ and $^{1}H-^{13}C$ couplings from either intensity variation or frequency variation of cross-peaks (*59, 60*). A primary advantage of the intensity-based sequences is that they encode couplings in single cross-peaks that can have the same positions as the cross-peaks in standard HSQC and triple-resonance experiments. There have also been 3D experiments devised that allow measurement of several couplings at once (*61*). For backbone couplings ($^{1}H-^{15}N$ amide, $^{1}H-^{13}C_{\alpha}$, $^{1}HN-^{13}CO$, $^{13}C_{\alpha}-^{15}N$, $^{13}CO-^{15}N$, etc.), it is important to realize that these involve atoms that are among the most easily assigned using standard triple-resonance experiments and automated assignment programs. Moreover, orientational relationships of fragments containing these atoms can be defined without close approach of the fragments. Hence, definition of a backbone structure without simultaneous placement of side chains is possible.

There has been a wide range of applications of residual dipolar couplings to protein structure determination, including their use to refine structures that have been produced using substantial amounts of data from NOE distance restraints (*62*). However, the appeal in structural genomics results from the possibility of more direct use in structure recognition or structure determination. The fact that $^{1}H-^{15}N$ dipolar couplings could be used in properly classifying a structure using a fold library was clearly illustrated in work by Annila et al. (*63*). Here a protein called calerythrin (~180 residues) was labeled with ^{15}N and ^{13}C to assign backbone resonances, and $^{15}N-^{1}H$ dipolar couplings for amide bonds were measured. Experimental couplings were then compared to couplings predicted for a small set of known structures. The measured sequence was threaded into each structure using secondary structure information from backbone carbon chemical shifts to improve threading, and coupling sets were generated for a grid of possible order frame alignments. The set of structures included two sarcoplasmic calcium binding proteins, one from sandworm and one from an amphioxus, whose sequences were 27 and 15% identical, respectively. These proved to be the only structures giving a good match to measured dipolar coupling patterns. Hence, a strong argument for an ability to place new proteins into existing fold classes strictly on the basis of easily acquired residual dipolar coupling data resulted. Procedures for searching databases for homologous folds based on dipolar data, secondary structure information, and pseudocontact shifts accessible for paramagnetic proteins have been efficiently programmed by other groups. The Griesinger group reports success with three test proteins using a more extended, but still small, fold database having 125 members (*64*).

A variation of the above procedure that can extend structure prediction to proteins representing new folds has also been described (*65*). The method, termed molecular fragment replacement, does not require that an entire fold be represented in a database, but only short segments of structure (approximately seven residues in the application described). Residual dipolar data on the small protein ubiquitin (76 residues) that included $^{1}H-^{15}N$ amide, $^{1}H-^{13}C_{\alpha}$,$^{13}CO-^{15}N^{1}$, and $^{1}HN-^{13}CO$ couplings obtained in two different aligned media were used. Seven-residue segments were threaded trough structures in a reduced PDB (1560 proteins), back calculating couplings from optimized alignment parameters at each step. Matching couplings produced local structures that could be assembled, and the resulting complete structure could be refined against the residual dipolar and chemical shift data. A backbone model, excluding a flexible C-terminus, matches the crystal structure of ubiquitin to within 0.88 Å. Again, this procedure employs only data that can be easily and efficiently acquired.

Direct calculation of backbone structures, without the use of a database, and using primarily residual dipolar data can also be accomplished. In our own work, two small proteins were used as test cases, acyl carrier protein (ACP) from *E. coli*, a protein of 77 amino acids whose structure had previously been characterized by traditional NMR methods, and NodF, a 92-amino acid product of a *Rhizobium leguminosarum* gene required for the synthesis of molecules that stimulate symbiotic root nodule formation (*66*). The acyl carrier protein test was carried out without the benefit of ^{13}C isotopic labeling (only ^{15}N labeling was used). Because ^{15}N-enriched media can be prepared at a fraction of the cost of preparing ^{15}N- and ^{13}C-enriched media, this may be important for gene products that give poor yields when expressed using *E. coli*, or require expression using other organisms. Elements of secondary structure were identified from a combination of backbone NOE patterns and three-bond ($^{3}J_{HN-HA}$) scalar couplings. These elements, three α-helices, were modeled as polyalanine helices of ideal geometry, and the orientation of each helix relative to a global orienting frame was determined by inverting equations similar to eq 1. The residual dipolar data came from $^{1}H-^{15}N$ couplings measured in HSQC-based experiments and $^{1}H-^{1}H$ couplings measured in a ^{15}N-edited constant time COSY experiment. A very small number of easily assigned NOEs connecting to backbone nuclei (five) were used to help restrict translational degrees of freedom for the helices. The resulting structure agreed with the previously determined NMR structure to within 3 Å for backbone atoms of the secondary structure elements (see Figure 4). This level of resolution is not representative of the precision of the measurements but is limited by the attempt to fit real helices with idealized models. Allowing helices to bend greatly improves the fit to data and presumably the accuracy of the structures (*66*). More importantly for structural genomics applications, the total data acquisition time was estimated to be <1 week on a conventional 500 MHz spectrometer. With the benefit of cryoprobes and high-field spectrometers, this should be reduced to approximately 1 day.

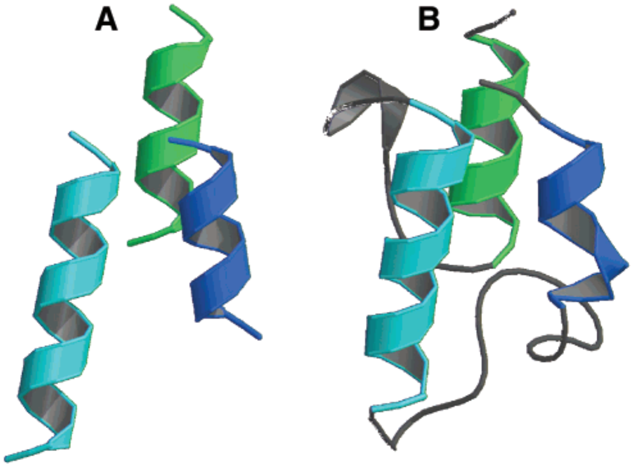

FIGURE 4: Structure of the acyl carrier protein (ACP) determined from residual dipolar couplings (A) and from NOEs (B).

The NodF protein was doubly labeled and, hence, benefited from a more efficient assignment strategy and the measurement of additional residual dipolar couplings ($^{1}H-^{13}C$, $^{13}C'-^{1}HN$, and $^{13}C'-^{15}N$ couplings, in addition to the couplings used in the ACP case). The additional couplings allowed modeling with helices assembled from several canonical segments (to allow bending). Data acquisition was actually shorter than for the ^{15}N-labeled ACP sample, approximately 5 days. No other structure is available for this protein, so no evaluation of the accuracy of the structure could be made.

Other procedures for more directly building backbone structures of protein from residual dipolar data are being developed by other groups (*67*, *68*). Some of these strategies are based on an ability to treat peptide planes as independent fragments and orient them one plane at a time. They use an extensive set of residual dipolar data from doubly labeled proteins, but no or little NOE data. Data acquisition time is again short, and will be further reduced with the aid of cryogenic probes.

Future of NMR in Structural Genomics

With advances such as the ones described above, we can remain optimistic about the role NMR will play in structural genomics. It can provide structures for proteins that may not easily crystallize, and it can provide these structures rapidly, especially if specification of backbone atoms in a protein fold is the primary objective. It is also useful to look beyond completion of an appropriate fold database to functional characterization of proteins. NMR will play a role here as well. We have already mentioned the potential of HSQC spectra in screens for ligand binding and drug design (*49*). It is important to realize that these experiments can also be used in screens for protein−protein interactions (*69*, *70*). It is now clear that relatively few proteins act in isolation, and understanding function may require looking at combinations of proteins. Going beyond simple detection of interaction, perturbations of specific peaks in HSQC experiments can identify regions of contact. The relative orientation of proteins in contact can also be determined using the residual dipolar methods mentioned above. Entire proteins of known structure can be treated as orientable fragments, just as we did for secondary structure elements or individual peptide planes in the work described above. In fact, there are now several examples of the determination of the relative orientation of loosely connected domains in multiple domain proteins using residual dipole methods (*71*−*73*). In one of these examples, binding of ligands between domains makes the relative orientations in the presence and absence of ligand an integral part of understanding the mechanism (*72*). In a more hypothetical case, cooperative binding of carbohydrate ligands on the surfaces of membranes may be modulated by the relative orientation of subunits in multimeric lectins (*74*). In short, prospects for contributions beyond the initial focus of structural genomics initiatives on domain structures are bright. We look forward to an ability to report progress in this direction in a few years.

ACKNOWLEDGMENT

We thank Dr. C. A. Fowler for his analysis of the ACP structure and Drs. J. W. Lee and E. Vysotski for their collaboration on the obelin samples.

REFERENCES

1. Sali, A. (1998) *Nat. Struct. Biol. 5*, 1029−1032.
2. Blundell, T. L., and Mizuguchi, K. (2000) *Prog. Biophys. Mol. Biol. 73*, 289−295.
3. Burley, S. K. (2000) *Nat. Struct. Biol. 7*, 932−934.
4. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) *Nucleic Acids Res. 28*, 235−242.
5. Orengo, C. A., Todd, A. E., and Thornton, J. M. (1999) *Curr. Opin. Struct. Biol. 9*, 374−382.
6. Skolnick, J., and Fetrow, J. S. (2000) *Trends Biotechnol. 18*, 34−39.
7. Sanchez, R., Pieper, U., Melo, F., Eswar, N., Marti-Renom, M. A., Madhusudhan, M. S., Mirkovic, N., and Sali, A. (2000) *Nat. Struct. Biol. 7*, 986−990.
8. Thornton, J. M., Todd, A. E., Milburn, D., Borkakoti, N., and Orengo, C. A. (2000) *Nat. Struct. Biol. 7*, 991−994.
9. Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J. R., Booth, V., Mackereth, C. D., Saridakis, V., Ekiel, I., Kozlov, G., Maxwell, K. L., Wu, N., McIntosh, L. P., Gehring, K., Kennedy, M. A., Davidson, A. R., Pai, E. F., Gerstein, M., Edwards, A. M., and Arrowsmith, C. H. (2000) *Nat. Struct. Biol. 7*, 903−909.
10. Skolnick, J., Fetrow, J. S., and Kolinski, A. (2000) *Nat. Biotechnol. 18*, 283−287.
11. Gerstein, M. (2000) *Nat. Struct. Biol. 7*, 960−963.
12. Yokoyama, S., Matsuo, Y., Hirota, H., Kigawa, T., Shirouzu, M., Kuroda, Y., Kurumizaka, H., Kawaguchi, S., Ito, Y., Shibata, T., Kainosho, M., Nishimura, Y., Inoue, Y., and Kuramitsu, S. (2000) *Prog. Biophys. Mol. Biol. 73*, 363−376.
13. Heinemann, U. (2000) *Nat. Struct. Biol. 7*, 940−942.
14. Norvell, J. C., and Machalek, A. Z. (2000) *Nat. Struct. Biol. 7*, 931.
15. Terwilliger, T. C. (2000) *Nat. Struct. Biol. 7*, 935−939.
16. Berman, H. M., Bhat, T. N., Bourne, P. E., Feng, Z., Gilliland, G., Weissig, H., and Westbrook, J. (2000) *Nat. Struct. Biol. 7*, 957−959.
17. Brunger, A. T., and Laue, E. D. (2000) *Curr. Opin. Struct. Biol. 10*, 557.
18. Markley, J. L., Bax, A., Arata, Y., Hilbers, C. W., Kaptein, R., Sykes, B. D., Wright, P. E., and Wuthrich, K. (1998) *Eur. J. Biochem. 256*, 1−15.
19. Wuthrich, K. (1995) *Acta Crystallogr. D51*, 249−270.
20. Clore, G. M., Bax, A., Ikura, M., and Gronenborn, A. M. (1993) *Curr. Opin. Struct. Biol. 3*, 838−845.
21. Engh, R. A., Dieckmann, T., Bode, W., Auerswald, E. A., Turk, V., Huber, R., and Oschkinat, H. (1993) *J. Mol. Biol. 234*, 1060−1069.
22. von Heijne, G. (1999) *Q. Rev. Biophys. 32*, 285−307.
23. Sanders, C. R., and Oxenoid, K. (2000) *Biochim. Biophys. Acta 1508*, 129−145.
24. de Groot, H. J. M. (2000) *Curr. Opin. Struct. Biol. 10*, 593−600.
25. Riek, R., Pervushin, K., and Wuthrich, K. (2000) *Trends Biochem. Sci. 25*, 462−468.
26. Kwong, P. D., Wyatt, R., Desjardins, E., Robinson, J., Culp, J. S., Hellmig, B. D., Sweet, R. W., Sodroski, J., and Hendrickson, W. A. (1999) *J. Biol. Chem. 274*, 4115−4123.
27. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W., and Lipman, D. J. (1997) *Nucleic Acids Res. 25*, 3389−3402.
28. Englander, S. W., Mayne, L., Bai, Y., and Sosnick, T. R. (1997) *Protein Sci. 6*, 1101−1109.
29. Hwang, T. L., Mori, S., Shaka, A. J., and vanZijl, P. C. M. (1997) *J. Am. Chem. Soc. 119*, 6203−6204.
30. Vysotski, E. S., Liu, Z. J., Rose, J., Wang, B. C., and Lee, J. (1999) *Acta Crystallogr. D55*, 1965−1966.
31. Wishart, D. S., and Sykes, B. D. (1994) *Methods Enzymol. 239*, 363−392.
32. Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Gerstein, M., Arrowsmith, C. H., and Edwards, A. M. (2000) *Prog. Biophys. Mol. Biol. 73*, 339−345.
33. Montelione, G. T., Zheng, D., Huang, Y. J., Gunsalus, K. C., and Szyperski, T. (2000) *Nat. Struct. Biol. 7*, 982−985.
34. Wider, G., and Wuthrich, K. (1999) *Curr. Opin. Struct. Biol. 9*, 594−601.
35. Gardner, K. H., and Kay, L. E. (1998) *Annu. Rev. Biophys. Biomol. Struct. 27*, 357−406.
36. Zimmerman, D. E., Kulikowski, C. A., Huang, Y. P., Feng, W. Q., Tashiro, M., Shimotakahara, S., Chien, C. Y., Powers, R., and Montelione, G. T. (1997) *J. Mol. Biol. 269*, 592−610.
37. Moseley, H. N. B., and Montelione, G. T. (1999) *Curr. Opin. Struct. Biol. 9*, 635−642.
38. Oschkinat, H., and Croft, D. (1994) *Methods Enzymol. 239*, 308−318.
39. Bartels, C., Guntert, P., Billeter, M., and Wuthrich, K. (1997) *J. Comput. Chem. 18*, 139−149.
40. Gronwald, W., Kirchhofer, R., Gorler, A., Kremer, W., Ganslmeier, B., Neidig, K. P., and Kalbitzer, H. R. (2000) *J. Biomol. NMR 17*, 137−151.
41. Xu, Y., Wu, J., Gorenstein, D., and Braun, W. (1999) *J. Magn. Reson. 136*, 76−85.
42. Nilges, M., Macias, M. J., Odonoghue, S. I., and Oschkinat, H. (1997) *J. Mol. Biol. 269*, 408−422.
43. Kozlov, G., Ekiel, I., Beglova, N., Yee, A., Dharamsi, A., Engel, A., Siddiqui, N., Nong, A., and Gehring, K. (2000) *J. Biomol. NMR 17*, 187−194.
44. Medek, A., Olejniczak, E. T., Meadows, R. P., and Fesik, S. W. (2000) *J. Biomol. NMR 18*, 229−238.
45. Gardner, K. H., Rosen, M. K., and Kay, L. E. (1997) *Biochemistry 36*, 1389−1401.
46. Cort, J. R., Yee, A., Edwards, A. M., Arrowsmith, C. H., and Kennedy, M. A. (2000) *J. Mol. Biol. 302*, 189−203.
47. Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) *J. Mol. Biol. 247*, 536−540.
48. Holm, L., and Sander, C. (1999) *Nucleic Acids Res. 27*, 244−247.
49. Hajduk, P. J., Meadows, R. P., and Fesik, S. W. (1999) *Q. Rev. Biophys. 32*, 211−240.
50. Moult, J., and Melamud, E. (2000) *Curr. Opin. Struct. Biol. 10*, 384−389.
51. Looger, L. L., and Hellinga, H. W. (2001) *J. Mol. Biol.* (submitted for publication).
52. Prestegard, J. H. (1998) *Nat. Struct. Biol. 5*, 517−522.
53. Tjandra, N., and Bax, A. (1997) *Science 278*, 1111−1114.
54. Tolman, J. R., Flanagan, J. M., Kennedy, M. A., and Prestegard, J. H. (1995) *Proc. Natl. Acad. Sci. U.S.A. 92*, 9279−9283.
55. Hansen, M. R., Mueller, L., and Pardi, A. (1998) *Nat. Struct. Biol. 5*, 1065−1074.
56. Clore, G. M., Starich, M. R., and Gronenborn, A. M. (1998) *J. Am. Chem. Soc. 120*, 10571−10572.

57. Fleming, K., Gray, D., Prasannan, S., and Matthews, S. (2000) *J. Am. Chem. Soc. 122*, 5224−5225.
58. Prestegard, J. H., Al-Hashimi, H. M., and Tolman, J. R. (2001) *Q. Rev. Biophys.* (in press).
59. Ottiger, M., and Bax, A. (1998) *J. Biomol. NMR 12*, 361−372.
60. Tolman, J. R., and Prestegard, J. H. (1996) *J. Magn. Reson., Ser. B 112*, 245−252.
61. Wang, Y. X., Marquardt, J. L., Wingfield, P., Stahl, S. J., Lee-Huang, S., Torchia, D., and Bax, A. (1998) *J. Am. Chem. Soc. 120*, 7385−7386.
62. Clore, G. M., Starich, M. R., Bewley, C. A., Cai, M. L., and Kuszewski, J. (1999) *J. Am. Chem. Soc. 121*, 6513−6514.
63. Annila, A., Aitio, H., Thulin, E., and Drakenberg, T. (1999) *J. Biomol. NMR 14*, 223−230.
64. Meiler, J., Peti, W., and Griesinger, C. (2000) *J. Biomol. NMR 17*, 283−294.
65. Delaglio, F., Kontaxis, G., and Bax, A. (2000) *J. Am. Chem. Soc. 122*, 2142−2143.
66. Fowler, C. A., Tian, F., Al-Hashimi, H. M., and Prestegard, J. H. (2000) *J. Mol. Biol. 304*, 447−460.
67. Mueller, G. A., Choy, W. Y., Yang, D. W., Forman-Kay, J. D., Venters, R. A., and Kay, L. E. (2000) *J. Mol. Biol. 300*, 197−212.
68. Hus, J. C., Marion, D., and Blackledge, M. (2001) *J. Am. Chem. Soc. 123*, 1541−1542.
69. Sette, M., Spurio, R., Van Tilborg, P., Gualerzi, C. O., and Boelens, R. (1999) *RNA 5*, 82−92.
70. Rajagopal, P., Waygood, E. B., Reizer, J., Saier, M. H., and Klevit, R. E. (1997) *Protein Sci. 6*, 2624−2627.
71. Bewley, C. A., and Clore, G. M. (2000) *J. Am. Chem. Soc. 122*, 6009−6016.
72. Skrynnikov, N. R., Goto, N. K., Yang, D. W., Choy, W. Y., Tolman, J. R., Mueller, G. A., and Kay, L. E. (2000) *J. Mol. Biol. 295*, 1265−1273.
73. Fischer, M. W. F., Losonczi, J. A., Weaver, J. L., and Prestegard, J. H. (1999) *Biochemistry 38*, 9013−9022.
74. Imberty, A., and Drickamer, K. (1999) *Curr. Opin. Struct. Biol. 9*, 547−548.

BI0102095