Acc. Chem. Res. **2003,** *36,* 207−214

NMR and Structural Genomics

DAVID STAUNTON, JO OWEN, AND IAIN D. CAMPBELL*

Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, U.K.

Received June 20, 2002

ABSTRACT

The role of NMR in structural genomics is outlined, with particular emphasis on using protein domains as targets. Strategies for domain expression, characterization, and labeling are presented.

Introduction

The goal of structural genomics is to accelerate protein structure determination, thus obtaining structural examples for all protein folds. In principle, this should allow a reasonable structural model to be constructed for any gene sequence, thus giving insight into gene function.

As will be apparent from other articles in this volume, the majority of currently funded projects and consortia (http://www.x12c.nsls.bnl.gov/StrGen.htm) concentrate on developing and refining techniques for high-throughput structure determination. Although the majority of these centers contain an NMR component, the main investment has been in X-ray crystallography. Existing high-throughput methodologies for seleno-labeling recombinant proteins, automated crystallization, data collection, and processing can yield structures very quickly, provided that protein and suitable crystals can be obtained. NMR is not expected to be able to compete with this productivity rate, especially for larger proteins. Surprisingly, perhaps, preliminary reports from structural genomics pilot studies indicate that the success rate of "gene to structure" is, so far, very similar for both methods.[1,2] With both techniques, the current success rate is low, with only about 20% of expressed genes producing structures. This success rate is likely to improve as the targets, constructs, and their expression conditions are refined.

In this article we will very briefly review the role of NMR in structural genomics projects and consider potential new technologies that should improve the rate of NMR structure determination. Such new technologies are likely to have a significant impact on a range of research efforts, not just on structural genomics. In addition to structure determination, there will be a growing need to relate structure to protein function.[3] For these studies, NMR will almost certainly play an increasingly important role, especially where protein dynamics and interactions between a variety of macromolecules and ligands are important.

Targets in Structural Genomics

A key issue in structural genomics is the choice of suitable target proteins for study. This is likely to develop from simple approaches that exploit readily expressed proteins from a particular genome ("low-lying fruit"), to systematic attempts to obtain examples of all structural families identified by sequence comparisons, regardless of the genome source.[4] The absolute number of structures determined will become less important than the "coverage" of protein fold space. Combined with homology modeling, good coverage will give low-resolution structures of most proteins in the various genomes. This is an attractive and achievable milestone on the road to a complete "structural map" but has limited value in the design of new drugs, where the high-resolution structure of a specific target, e.g., enzyme or receptor, is required.

Most current projects avoid proteins regarded as having little chance of success. The most obvious example of this category is intrinsic membrane proteins that comprise a large fraction of most genomes. Success rates in membrane protein structure determination have significantly improved in recent years.[5] One of the key problems is expression of sufficient quantities of these proteins for structural studies. The Japanese structural genomics project is addressing this particular problem by in vitro expression of membrane proteins in the presence of detergents, but such an approach requires extensive screening. It is worth noting that NMR can be used effectively for the structure determination of some transmembrane proteins in both detergent[6] and the solid phase.[7]

Since the first structural genomics projects were established, targeting has been refined from simply accumulating structures for all identified open reading frames (ORFs) to a more systematic approach. Most proteins are constructed from identifiable domains or modules (a subset of protein domains that have a contiguous amino acid sequence). These domains are tabulated in InterPro (http://www.ebi.ac.uk/interpro/index.html), a sophisticated domain database, constructed in 2001 from various earlier databases, including PFAM, SMART, PROSITE, TIGRFAMs, and SWISS-PROT. The current release of InterPro contains 5312 entries, representing 1177 domains in 4028 families (InterPro release 5.2, Sept 2002). Such a database is a convenient starting point for a coordinated, cross-genome approach to cover conformational space.[4] Of the SWISS-PROT protein sequences, 85% have one or more hits in InterPro, and it is estimated that domains currently identified account for around 25%

David Staunton received both of his degrees (M.A. in 1985 and D.Phil. in 1988) from the University of Oxford. He is currently a postdoctoral fellow in Iain Campbell's group at Oxford, developing new expression systems.

Jo Owen received a B.Sc. degree in biology from the University of York in 1995, and then moved to Oxford to work in the pharmaceutical industry on chemokine receptors. She is currently studying for a D.Phil. in biochemistry at the University of Oxford in Iain Campbell's group.

Iain Campbell received a B.Sc. (1963) and Ph.D. (1967) from the University of St. Andrews. He is currently Professor of Structural Biology at the University of Oxford.

10.1021/ar010119s CCC: $25.00
Published on Web 12/07/2002

of ORFs. Assuming that 30% sequence homology is required for a reasonable model to be generated from a structure, it has been calculated that 90% coverage of all the sequences in the PFAM database could be accurately modeled from 50 000 selected structures.[4] A simplified approach would be to generate structures for all of the domains identified so far, giving a database on which to build. New databases are being set up that will facilitate this approach. A recent example is SUPERFAMILY,[8] which is based on SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/), a structural domain-based classification of protein folds.

Sequence comparisons alone can fail to identify proteins that have the same folding patterns. This is likely to result in domain families being amalgamated as their structural characteristics are identified. There are numerous examples of this sort; a typical example is the ribosome anti-association factor IF6,[9] whose structure was recently generated by a structural genomics group. This was found to have a fold identical to an existing protein structure,[10] although no identity was found at the sequence level.[11] While this did not produce a novel fold, the new structure still provides the basis for homology modeling for related sequences; in other words, the databases of sequence and structure will become more and more powerful as new information is added.

A domain-targeted approach has obvious advantages for NMR, as the size of the protein targets can be reduced to individual modules of 100−200 amino acids, well within the size limitations of the technique.[12]

The Role of NMR

NMR relies on the property that some atomic nuclei, including ^{1}H, ^{15}N, and ^{13}C, can be made to resonate in a static magnetic field by applying radio frequency radiation. The NMR "resonances" are very sensitive to the electronic environment around the nuclei (a property known as chemical shift) and have fine structure or "spin−spin coupling" which arises from interactions between nuclei. Chemical shift and spin−spin coupling result in an NMR spectrum that is characteristic for a particular molecule. The method has high enough resolution to distinguish identical groups in a macromolecule; for example, a small protein with five alanines will give five different methyl side-chain resonances. By detecting dipolar interactions, or nuclear Overhauser effects (NOEs) between protons, distance information between specific atoms in a protein can be obtained. The main method used to determine structures of proteins in solution is to calculate families of structures consistent with a set of NOE restraints[13] (see also the section, "Assignment and Structure Calculation", below). NMR studies are limited by protein size, since the width of the resonances increases (and hence the resolution decreases) with molecular weight, and assignment of an observed resonance to a particular chemical group eventually becomes impossible due to spectral overlap. The main advantage of NMR over other structural methods is that the protein is studied in aqueous solution and avoids the requirement for crystals. For these reasons, NMR has a number of key roles to play in structural genomics.

Since NMR is a solution-state method, it is relatively easy to explore a range of solution conditions (pH, temperature, salts) to find *optimum conditions* for data collection. Early characterization of the folding and solution behavior of proteins reduces time and effort spent on constructs that will not give viable spectra. Circular dichroism and calorimetry have also been used to characterize the fold of protein constructs from high-throughput systems, but these results are not always consistent with their NMR spectra.[1] Some groups now routinely label with ^{15}N to facilitate the collection of 2D $^{15}N-^{1}H$ spectra. Such spectra are a powerful diagnostic tool for detecting a well-behaved, folded protein.[2]

Sensitive modern NMR spectrometers can allow a 2D spectrum to be obtained from a 50 μM sample in 10 min, giving an efficient screen for suitable samples.[14] Crystallization trials and initial diffraction data represent a comparable process, but this can take from hours to months before successful samples can be identified. Since one of the fundamental requirements for crystallization is good solution properties, this type of screen can also be used to assess suitable solution conditions for crystallization trials.

Another important role is the ability of NMR to screen for *flexible proteins or proteins with flexible regions*. Dunker and co-workers have analyzed several known genomic sequences. Using a computer algorithm, they predict that 7−33% of bacterial proteins are unfolded; for higher organisms, the same analysis yields the astonishing prediction that 36−63% of eukaryotic proteins are unfolded.[15] Dyson and Wright have pointed out that the coupling of protein folding with binding may have significant biological advantages, especially in multicellular organisms.[16] Significant fractions of expressed proteins in a structural genomic program may thus be unfolded, and therefore will not crystallize in the absence of their biological binding partner. As mentioned above, simple inspection of NMR spectra can determine if a protein contains extensive regions that are unfolded. More sophisticated analysis can be carried out using relaxation and heteronuclear NOE measurements to detect flexible regions.

The ability of NMR to *generate coordinates* of a family of protein structures consistent with the experimental data is now well established.[13] About 2500 NMR structures have been deposited in the Protein Data Bank. In the best cases, this can lead to protein structures with resolution comparable to X-ray structures, but generally the structures are of lower quality, although there is scope for improvement.[17] A recent analysis of protein structures showed that structures determined by NMR had a higher proportion of disorder than those determined by X-ray crystallography, suggesting that NMR is more applicable to these proteins due to their reluctance to crystallize. This is supported by the lack of corresponding high-resolution crystal structures for these NMR structures.[18]

As already mentioned, the size of macromolecule that can be studied by NMR is limited by spectral line width and overlap. Although structures up to 40 kDa can be determined by NMR,[19] relatively rapid de novo structure determination, which requires well-resolved spectra for easy assignment, is currently restricted to systems of less than about 20 kDa. This means that the NMR approach is restricted to a subset of genes encoding proteins of less than about 200 amino acids. If the targets were defined as the products of ORFs, NMR would thus be directly applicable to less than a quarter of microbial genomes and a much smaller proportion for higher organisms.

NMR has, however, an obvious role in the structural analysis of modules from proteins for which structure determination of the whole protein is unfeasible. Many of these modules are of suitable size for study, and this laboratory has used this approach for a number of years.[12,20] A recent example is the dumpy protein from *D. melanogaster*. This 2.5 MDa extracellular protein mostly consists of contiguous repeats of an EGF-DPY-EGF unit. While the structure of EGF domains is well characterized, that of the 21 amino acid DPY repeat was unknown. Structure determination by NMR revealed the module to have a compact fold consisting of a 3_{10} helix and an antiparallel β-sheet with two conserved disulfide bridges forming the module core. Knowledge of this structure allows a simple model of the EGF-DPY-EGF repeat to be constructed and the structure of the whole protein to be inferred as a fibrous molecule at least 0.8 μm in length.[21] This example illustrates the fact that, for small domains, NMR is an efficient route to obtaining coordinates and structural insight.

Even when a protein structure is determined by X-ray crystallography, there may be regions and domains that remain undefined. An interesting example of a crystal structure of a protein complex containing a module of unknown structure is that of the extracellular segment of integrin $\alpha V\beta 3$.[22] The β chain contains a PSI domain (so called after three of the better-characterized families in which it exists: *p*lexins, *s*emaphorins, and *i*ntegrins) at its amino terminus. Most of the β subunit structure is well resolved, but the PSI region has no coordinates due to poorly defined electron density, which is suggestive of a relatively disordered domain. Therefore, such a module would be an ideal candidate for an NMR study, although the integrin PSI domain would probably not be selected, as it has an uneven number of cysteine residues which may result in incorrect disulfide bond formation and aggregation during refolding. However, the PSI domain homology is well documented,[23] and the domain in a molecule such as C21orf3 (a membrane protein of unknown function), which contains eight cysteine residues, would be a much better target for NMR.[24]

Another key role of NMR is in screening for *ligand interactions*. The method is now used widely in the pharmaceutical industry[25,26] to detect the binding of small molecules to macromolecular receptors. NMR is also good at detecting protein−protein interactions, and specific interaction sites can be mapped on the protein structure. Methods used for this mapping include induced chemical shift differences, reduction of hydrogen exchange rates, or, more recently, sophisticated cross-relaxation methods that allow large complexes to be studied using isotope labeling strategies including complete deuteration.[27,28]

NMR is also *complementary* to other coordinate-generating tools. It can be used to compare the solution-state behavior of proteins versus their behavior in the crystal. While there is no longer any doubt that the coordinates obtained from crystallography provide a valid starting point for understanding protein function, there are cases where crystal contacts can cause local distortions or restrictions in domain movements. An example is the ^{9}F3−^{10}F3 domain pair from fibronectin that has been studied both by crystallography[29] and by NMR,[30,31] where the integrin-binding RGD sequence is rigid in the crystal structure but flexible in solution. Another recent elegant example is a study of calmodulin using residual dipolar coupling measurements. NMR clearly showed significant conformational flexibility within each of the two Ca^{2+} binding domains and deviations of the average structure from that detected by crystallography.[32] NMR thus has a role to play in complementing the crystallographic view of protein structure.

These examples also illustrate another advantage of NMR in its ability to investigate local and global *dynamics* in a protein. Relaxation methods can be used to detect motion over a wide range of time scales. These methods are largely independent of structural information. It is becoming increasingly clear that the dynamic properties of proteins are important for a wide range of functions, including catalysis, binding, and protein stability.[33−35] In cases where proteins are constructed from a number of domains, relative changes in domain orientation may play an essential role in regulation of protein function and mechanism. NMR is again likely to play a key role in detecting and characterizing domain movements relative to one another.[18,36]

What Needs To Be Made Better?

The prospects are that structural genomics initiatives will lead to significantly improved methodology. Some areas where improvements are expected are outlined below.

(a) NMR Methodology. The technical limitations of NMR arise mainly from lack of resolution and poor signal-to-noise ratio. Since NMR spectra contain many thousands of peaks, there is also a problem assigning these peaks to individual chemical groups in the protein. The first commercial instruments in the 1950s operated at a ^{1}H frequency of around 40 MHz. Since then, there has been a steady increase in the strength of the static field (B_o) in which the sample is placed, and 900 MHz instruments have recently been introduced. Higher fields lead directly to better resolution: the line width stays the same to first order, but the chemical shift increases; the signal-to-noise ratio also increases with B_o, roughly following a $(B_o)^{7/4}$ law.[37] Other interesting effects arise at high fields, including an increase in chemical shift anisotropy. Line

width contributions (relaxation) generally arise from both dipolar (D) and chemical shift anisotropy (CSA) effects. In a doublet, e.g., arising from $^{15}N-^{1}H$ spin−spin coupling, the line width of one of the two components is dominated by the sum of D and CSA, while in the other it is dominated by the difference in D and CSA. This results in one of the doublet components being relatively narrow at high fields, because of cancellation; the optimum effect on resolution is predicted to occur around 1000 MHz. Powerful tools, such as TROSY, have been developed to exploit this effect;[38] these can be very useful, especially in large complexes where one component (e.g., a protein module) is selectively labeled with isotopes. Other methods being explored to enhance resolution include the encapsulation of proteins in inverted micelles in low-viscosity solvents.[39] This leads to relatively rapid tumbling in solution and narrower lines.

One of the major problems with NMR is low sensitivity (signal-to-noise ratio). This means that data collection can take several days. Spectrometer sensitivity has improved greatly over the years, with increasing the field strength and improved electronics and resonators. A relatively new development is to cool the resonator in the sample probes to very low temperatures (~10 K) (the sample is still studied at ambient temperature, unlike cryogenic crystallography studies). This approach was first applied in the 1980s,[40] but cryoprobes are now commercially available, and their use can lead to a significant reduction in data collection time.[41]

(b) Assignment and Structure Calculation. Spectral assignment is usually the rate-limiting step in NMR structure determination. Sophisticated pulse sequences tailored to study isotopically labeled samples (^{15}N, ^{13}C, and ^{2}H) have greatly enhanced the resolving power and the ability to assign backbone and side chain resonances, since scalar coupling pathways can be traced through bonds.[42] Many groups have sought to automate the assignment process.[43,44] While there are still some problems with difficult spectra, these automation procedures are expected to become increasingly effective, especially under the pressure of global structural genomics efforts, where the number of spectra that require assignment will overwhelm traditional "hands-on" methodology.

NMR structure determination methods that were developed in the 1980s depend on short-range distance information.[13] The "traditional" method mainly involves assigning a large number of $^{1}H-^{1}H$ NOEs and calculating families of structures consistent with restraints that include NOEs, torsion angles, and inferred H-bonds. Additional restraints have been added in recent years, including those derived from chemical shifts and conformational database analysis, automated iterative assignment of NOEs, and long-range H-bonds. Particularly important recent additions are experiments that provide information about relatively long-range order in a macromolecule.[45] This information can be derived from analysis, either of residual dipolar couplings (RDCs), introduced by partial alignment of the protein studied,[46] or by measurement of T_1/T_2 ratios of isotopically labeled resonances.[47] Long-range information is particularly valuable when dealing with modular proteins[36] because the short-range information between domains is often minimal and interdomain flexibility is common.

In addition to experimental restraints, there is scope for improving procedures for calculating structures from these restraints. One of the most powerful methods is simulated annealing, which finds a global minimum of a target function. Energy barriers are overcome by raising the temperature of the system, followed by slow cooling while exploring conformational space using molecular dynamics (MD). The MD simulations can be carried out in Cartesian coordinates space or torsion angle space. These procedures will continue to be improved.[48]

Since traditional NMR structure determination methods remain relatively slow and tedious, alternative approaches are being sought to obtain information rapidly. Part of the rationale is that a relatively low-resolution view of a protein fold would still be valuable for some aspects of structural genomics (see, e.g., ref 18). One recent example involved collection of a single set of triple-resonance data from ^{15}N,^{13}C-labeled samples in isotropic and liquid crystalline media. This yielded not only assignments but also structural information that included $^{13}C-^{1}H$ dipolar couplings and deviations from random coil shifts.[49] This allowed good representations of the substructures of ubiquitin and calmoldulin to be obtained. There seems to be little doubt that this kind of approach can be extended significantly.

(c) Protein Expression. The requirement for stable proteins with good solution behavior is common to both X-ray crystallography and NMR techniques. NMR has the additional requirements for growth in minimal media to allow efficient and economical isotope labeling throughout the protein molecule. In practical terms, this limits the protein expression systems to *Escherichia coli* and yeast. For high throughput, the homologous recombination required for the yeast expression (e.g., *Pichia pastoris*) makes it relatively slow, although its success with disulfide bond-containing proteins makes it an attractive fall-back system for targets that fail during the first round of expression. Through the use of these systems, sample production and purification is no longer the rate-limiting step in NMR experiments, but we still rely on finding a few suitably behaved samples from a large number of initial targets. The low success rate going from constructs to well-formed crystals and NMR samples in some of the preliminary structural genomics projects suggests that this is common, even when the targets have been selected for stability (thermophiles), with difficult ones excluded. While acceptable in a research environment where projects require only a few well-behaved samples, this is unacceptable when such selection could result in unrepresentative families of proteins becoming the majority of structures determined. This may only become apparent as the structures from such projects become available.

Optimization of Constructs. Once a sample has been generated, it often requires fine-tuning, involving changes in protein sequence and length before an optimum can

be reached. This requires a repetition of the subcloning, expression, and purification, which is tedious and time-consuming but often essential for success. It is possible to use proteolytic degradation and mass spectrometry to identify the appropriate domain termini of NMR samples, but this is still a reiteration process that does not translate easily to a high-throughput format.

Our experience with domains suggests that inter-domain interactions can make significant contributions to module stability and hence expression.[50] Since it is very important to choose the correct domain boundary for a given target module, multiple overlapping constructs should be made as a matter of course to provide a library from which proteins can be selected, using various screens, such as NMR. An example of this approach is the analysis of EGF domains of fibulin-1, an extracellular matrix protein, where binding sites for fibronectin and other ligands are localized to the C-terminal array of EGF modules.[51] EGF domains 4−7 were expressed as a range of domain combinations in order to select the best fragments for study. This proved fruitful, as fragments EGF 4/5 and 5/6 produced 1D NMR indicative of folded protein with β-strand components as expected, but EGF 4/5/6 exhibited a poorly dispersed 1D spectrum suggesting an unfolded structure (Figure 1). An additional benefit is that the bank of overlapping fragments can be analyzed for functional properties, such as ligand binding, allowing binding sites to be mapped to particular domains.

Development of in Vitro Expression. Another major problem with high-throughput protein expression is the jump from the microliter volumes of the DNA/construct level to the liters required for expression fermentations. This produces a physical bottleneck, where the 96-well plate format has to be sacrificed to cope with the 1000-fold increase in volume early in the procedure. In vitro or cell-free expression systems have been used for over 40 years to identify the protein products of mRNA and DNA. However, their usefulness was limited by the depletion of essential components, leading to a halt in protein production after 1−2 h. In 1988, Spirin and co-workers described a bioreactor in which a supply of amino acids and high-energy substrates increased reaction times and yields,[52] but it was not until recently that these yields were developed to levels (mg/mL) suitable for structural studies.[53]

The main attraction of in vitro expression for high-throughput screening is that constructs and conditions can be explored in batch reactions of less than 100 μL, allowing a 96-well format to be used. The yields of both soluble and insoluble protein are followed by [^{35}S]methionine or [^{14}C]leucine incorporation, until optimized. The successful batch reactions can then be scaled up to the milliliter level in a semicontinuous batch reaction, in which the expression system is supplied with reagents from a reservoir through a dialysis membrane.[54] This extends the life of the reaction to more than 12 h and the protein yield to the milligrams required for NMR sample preparation or crystallization trials. Another advantage of this open system is that the amino acids are chosen by

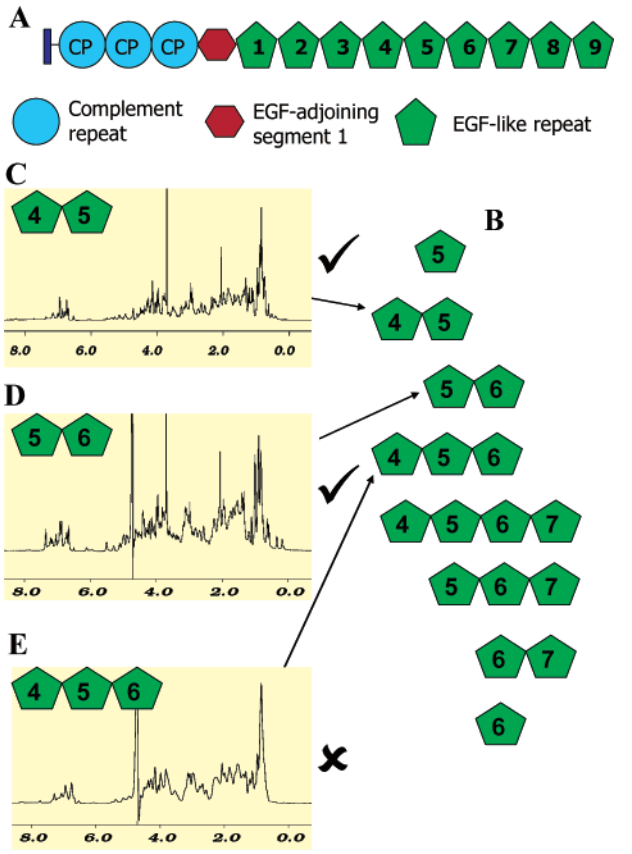

FIGURE 1. Example of protein construct screening to identify samples suitable for NMR. (A) Schematic of fibulin-1A, the shortest of four alternatively spliced isoforms comprising of an N-terminal signal sequence (dark blue), three complement repeat domains (light blue), and an EGF-adjoining segment (red) followed by nine consecutive EGF-like domains (green). (B) The complementary range of eight EGF domain constructs that were designed. (C−E) 1D NMR of EGF 4/5 fragment (C), EGF 5/6 (D), and EGF 4/5/6 (E) at pH 6.8, 30 °C. EGF 4/5 and 5/6 were selected as suitable for structural studies on the basis of dispersion, upfield shifts, and aliphatic region peaks indicative of β-sheets.

the researcher; isotopic labeling can thus be specific to particular amino acids. Examples where this approach has distinct advantages are the incorporation of Leu and Val with protonated methyl groups into an otherwise completely deuterated protein [55] or incorporation of amino acids synthesized with specific isotope labels.[56] Artificially charged suppressor tRNAs can also be used to label specific positions in the protein sequence.[54,57] This approach will become increasingly valuable as targets expand to include proteins with post-translational modifications (e.g., γ-carboxyglutamic acid), since these can be introduced during translation. The system does not readily allow the formation of disulfide bonds, but there is the potential to make rapid progress in this area.

Intein Technology. Although modern NMR methods can be used to study protein complexes of 50 kDa or more, the interpretation of these data is still limited by the complexity of the spectra. One way of reducing complexity would be to limit the isotopic labeling to a fragment of the protein under study. The utilization of inteins for protein ligation now allows isotopically labeled peptides

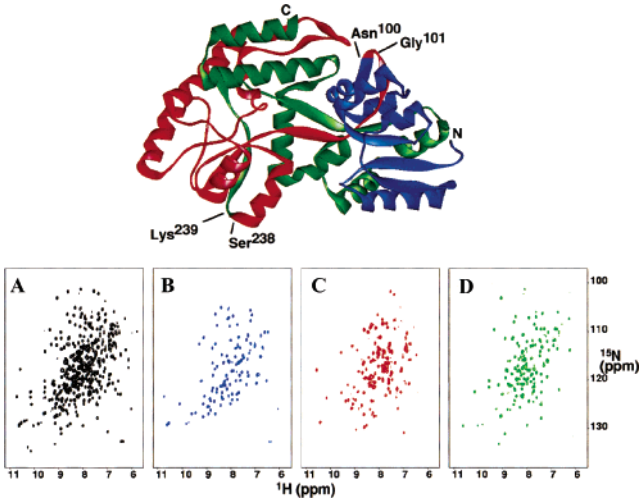

FIGURE 2. The use of intein-mediated protein ligation and isotope labeling to simplify protein NMR spectra. Segments of maltose-binding protein (MBP) were constructed by intein ligation and their HSQC spectra recorded. The N-terminal, central, and C-terminal segments (blue, red, and green, respectively) and their junction sites are illustrated on the MBP structure. $^{15}N-^{1}H$ HSQC spectra of (A) uniformly ^{15}N-labeled wild-type MBP/maltose complex and those of MBP segmentally labeled with ^{15}N; (B) N-terminal segment, (C) central segment, and (D) C-terminal segment (adapted from Otoma et al., 1999, with permission).

and proteins to be joined through peptide bonds to unlabeled proteins. Inteins are naturally occurring motifs that are excised from precursor proteins, and the two remaining fragments or exteins are then ligated through a normal peptide bond. In this way, a protein can be rebuilt to have an isotopically labeled fragment at either end of the sequence.[58] Two approaches have been taken for this methodology. In the first, a fragmented or split intein is used to trans-splice peptides or proteins expressed as fusions with N- or C-terminal portions of the intein.[59] In the second, an intein fusion protein is used in conjunction with a suitable thiol to generate a stable thioester of the recombinant protein. This thioester can then react with a suitable peptide or protein. A limitation of both techniques is that the proteins must be able to withstand prolonged exposure to extreme conditions, denaturing in the first and reducing in the second.

Despite these limitations, both techniques have been used to generate contiguous proteins from three polypeptide segments.[60,61] Otomo and co-workers used two inteins, PI-*pfu*I and PI-*pfu*II from *Pyrococcus furiosus*, to create maltose-binding protein (370 residues) from three segments (residues 1−100, 101−238, and 239−370). The ligation junctions were selected for flexibility and exposure and were therefore located in exposed loops. By using individually ^{15}N-labeled segments, $^{15}N-^{1}H$ HSQC spectra were obtained for each of the segments in the whole MBP. Comparison of uniformly labeled MBP with the signals from the segmental labeled proteins indicated very little perturbation of the spectra, except for those due to extra residues inserted at the joints (Figure 2). By combining intein ligation and the "library" approach described earlier, it may be possible to analyze proteins in 100 overlapping labeled amino acid segments in a relatively routine way. This would greatly simplify the resulting spectra and their automated assignment.

Conclusion

The quality of NMR spectrometers, like X-ray crystallography technology, has improved significantly in the past two decades. To meet the challenges of post-genomic research, this trend will continue; improvements in associated technologies, such as automated spectral assignment and structure calculation procedures, are also expected. Structural genomics is encouraging the NMR community to reassess techniques and procedures in order to speed up structure determination and functional analysis. It is likely that new protocols for engineering and producing samples, for example, in vitro and intein technology specifically tailored to the requirements of NMR experiments, will have a major impact.

While the main emphasis of projects around the world is on structure determination by X-ray crystallography, the complementary properties of NMR will make a contribution in a variety of ways. The established role of NMR in the structure determination of small proteins and modules and those that are unsuitable for crystallography will continue, especially since the proportion of proteins in the genomes that do not crystallize, due to flexibility, may be quite high. Sample screening by NMR has significant advantages for the early selection of constructs suitable for structure determination. Information obtained about protein dynamics and ligand-binding partners from NMR will also be increasingly important as the research emphasis moves from structure determination to function.

We acknowledge support from the Wellcome Trust and BBSRC (J.O.).

References

(1) 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.; Arrowsmith, C. H. Structural proteomics of an archaeon. *Nat. Struct. Biol.* **2000**, *7*, 903−909.

(2) Yee, A.; Chang, X. Q.; Pineda-Lucena, A.; Wu, B.; Semesi, A.; Le, B.; Ramelot, T.; Lee, G. M.; Bhattacharyya, S.; Gutierrez, P.; Denisov, A.; Lee, C. H.; Cort, J. R.; Kozlov, G.; Liao, J.; Finak, G.; Chen, L.; Wishart, D.; Lee, W.; McIntosh, L. P.; Gehring, K.; Kennedy, M. A.; Edwards, A. M.; Arrowsmith, C. H. An NMR approach to structural proteomics. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 1825−1830.

(3) Teichmann, S. A.; Murzin, A. G.; Chothia, C. Determination of protein function, evolution and interactions by structural genomics. *Curr. Opin. Struct. Biol.* **2001**, *11*, 354−363.

(4) Vitkup, D.; Melamud, E.; Moult, J.; Sander, C. Completeness in structural genomics. *Nat. Struct. Biol.* **2001**, *8*, 559−566.

(5) Dutzler, R.; Campbell, E.; Cadene, M.; Chait, B.; MacKinnon, R. X-ray structure of a ClC chloride channel at 3.0 A reveals the molecular basis of anion selectivity. *Nature* **2002**, *415*, 287−94.

(6) Arora, A.; Abildgaard, F.; Bushweller, J. H.; Tamm, L. K. NMR solution structure and dynamics of the outer membrane protein A transmembrane domain in dodecylphosphocholine micelles. *Biophys. J.* **2002**, *82*, 2512.

(7) Pauli, J.; Baldus, M.; van Rossum, B.; de Groot, H.; Oschkinat, H. Backbone and side-chain C-13 and N-15 signal assignments of the alpha-spectrin SH3 domain by magic angle spinning solid-state NMR at 17.6 T. *ChemBioChem* **2001**, *2*, 272−281.

(8) Gough, J.; Chothia, C. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.* **2002**, *30*, 268−272.

(9) Groft, C. M.; Beckmann, R.; Sali, A.; Burley, S. K. Crystal structures of ribosome anti-association factor IF6. *Nat. Struct. Biol.* **2000**, *7*, 1156−1164.

(10) Paoli, M. An elusive propeller-like fold. *Nat. Struct. Biol.* **2001**, *8*, 744−744.

(11) Groft, C. M.; Beckmann, R.; Sali, A.; Burley, S. K. Response to Paoli. *Nat. Struct. Biol.* **2001**, *8*, 745−745.

(12) Baron, M.; Norman, D.; Campbell, I. Protein Modules. *Trends Biol. Sci.* **1991**, *16*, 13−17.

(13) Wuthrich, K. The way to NMR structures of proteins. *Nat. Struct. Biol.* **2001**, *8*, 923−925.

(14) Hajduk, P. J.; Gerfin, T.; Boehlen, J. M.; Haberli, M.; Marek, D.; Fesik, S. W. High-throughput nuclear magnetic resonance-based screening. *J. Med. Chem.* **1999**, *42*, 2315−2317.

(15) Dunker, A. K.; Obradovic, Z. The protein trinity—linking function and disorder. *Nat. Biotechnol.* **2001**, *19*, 805−806.

(16) Dyson, H. J.; Wright, P. E. Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.* **2002**, *12*, 54−60.

(17) Spronk, C.; Linge, J. P.; Hilbers, C. W.; Vuister, G. W. Improving the quality of protein structures derived by NMR spectroscopy. *J. Biomol. NMR* **2002**, *22*, 281−289.

(18) Prestegard, J. H.; Valafar, H.; Glushka, J.; Tian, F. Nuclear magnetic resonance in the era of structural genomics. *Biochemistry* **2001**, *40*, 8677−8685.

(19) Choy, W. Y.; Tollinger, M.; Mueller, G. A.; Kay, L. E. Direct structure refinement of high molecular weight proteins against residual dipolar couplings and carbonyl chemical shift changes upon alignment: an application to maltose binding protein. *J. Biomol. NMR* **2001**, *21*, 31−40.

(20) Campbell, I. D.; Downing, A. K. NMR of modular proteins. *Nature Struct. Biol.* **1998**, *5*, 496−499.

(21) Wilkin, M. B.; Becker, M. N.; Mulvey, D.; Phan, I.; Chao, A.; Cooper, K.; Chung, H. J.; Campbell, I. D.; Baron, M.; MacIntyre, R. Drosophila Dumpy is a gigantic extracellular protein required to maintain tension at epidermal-cuticle attachment sites. *Curr. Biol.* **2000**, *10*, 559−567.

(22) Xiong, J. P.; Stehle, T.; Zhang, R. G.; Joachimiak, A.; Frech, M.; Goodman, S. L.; Aranout, M. A. Crystal structure of the extracellular segment of integrin alpha V beta 3 in complex with an Arg-Gly-Asp ligand. *Science* **2002**, *296*, 151−155.

(23) Bork, P.; Doerks, T.; Springer, T. A.; Snel, B. Domains in plexins: links to integrins and transcription factors. *Trends Biochem. Sci.* **1999**, *24*, 261−263.

(24) Yaspo, M. L.; Aaltonen, J.; Horelli-Kuitunen, N.; Peltonen, L.; Lehrach, H. Cloning of a novel human putative type Ia integral membrane protein mapping to 21q22.3. *Genomics* **1998**, *49*, 133−136.

(25) Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. NMR-based screening in drug discovery. *Q. Rev. Biophys.* **1999**, *32*, 211−240.

(26) Hicks, R. P. Recent advances in NMR: Expanding its role in rational drug design. *Curr. Med. Chem.* **2001**, *8*, 627−650.

(27) Wuthrich, K. Protein recognition by NMR. *Nat. Struct. Biol.* **2000**, *7*, 188−189.

(28) Takahashi, H.; Nakanishi, T.; Kami, K.; Arata, Y.; Shimada, I. A novel NMR method for determining the interfaces of large protein−protein complexes. *Nat. Struct. Biol.* **2000**, *7*, 220−223.

(29) Leahy, D. J.; Aukhil, I.; Erickson, H. P. 2.0 angstrom crystal structure of a four-domain segment of human fibronectin encompassing the RGD loop and synergy region. *Cell* **1996**, *84*, 155−164.

(30) Spitzfaden, C.; Grant, R. P.; Mardon, H. J.; Campbell, I. D. Module−module interactions in the cell binding region of fibronectin: Stability, flexibility and specificity. *J. Mol. Biol.* **1997**, *265*, 565−579.

(31) Copie, V.; Tomita, Y.; Akiyama, S. K.; Aota, S.; Yamada, K. M.; Venable, R. M.; Pastor, R. W.; Krueger, S.; Torchia, D. A. Solution structure and dynamics of linked cell attachment modules of mouse fibronectin containing the RGD and synergy regions: Comparison with the human fibronectin crystal structure. *J. Mol. Biol.* **1998**, *277*, 663−682.

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

(33) Palmer, A. G. NMR probes of molecular dynamics: Overview and comparison with other techniques. *Annu. Rev. Biophys. Biomol. Struct.* **2001**, *30*, 129−155.

(34) Spyracopoulos, L.; Sykes, B. D. Thermodynamic insights into proteins from NMR spin relaxation studies. *Curr. Opin. Struct. Biol.* **2001**, *11*, 555−559.

(35) Wand, A. J. Dynamic activation of protein function: A view emerging from NMR spectroscopy. *Nat. Struct. Biol.* **2001**, *8*, 926−931.

(36) Hashimoto, Y.; Smith, S. P.; Pickford, A. R.; Bocquier, A. A.; Campbell, I. D.; Werner, J. M. The relative orientation of the fibronectin (6)F1(1)F2 module pair: A N-15 NMR relaxation study. *J. Biomol. NMR* **2000**, *17*, 203−214.

(37) Boyd, J.; Soffe, N.; Campbell, I. Nmr At Very High Fields. *Structure* **1994**, *2*, 253−255.

(38) Riek, R.; Pervushin, K.; Wuthrich, K. TROSY and CRINEPT: NMR with large molecular and supramolecular structures in solution. *Trends Biochem. Sci.* **2000**, *25*, 462−468.

(39) Wand, A. J.; Ehrhardt, M. R.; Flynn, P. F. High-resolution NMR of encapsulated proteins dissolved in low-viscosity fluids. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 15299−15302.

(40) Styles, P.; Soffe, N.; Scott, C. An improved cryogenically cooled probe for high-resolution NMR. *J. Magn. Reson.* **1989**, *84*, 376−378.

(41) Russell, D. J.; Hadden, C. E.; Martin, C. E.; Gibson, A. A.; Zens, A. P.; Carolan, J. L. A comparison of inverse-detected heteronuclear NMR performance: Conventional vs cryogenic microprobe performance. *J. Nat. Prod.* **2000**, *63*, 1047−1049.

(42) Gardner, K. H.; Kay, L. E. The use of H-2, C-13, N-15 multidimensional NMR to study the structure and dynamics of proteins. *Annu. Rev. Biophys. Biomol. Struct.* **1998**, *27*, 357−406.

(43) Moseley, H. N. B.; Monleon, D.; Montelione, G. T. Automatic determination of protein backbone resonance assignments from triple resonance nuclear magnetic resonance data. *Methods Enzymol.* **2001**, *339*, 91−108.

(44) Linge, J. P.; O'Donoghue, S. I.; Nilges, M. Automated assignment of ambiguous nuclear Overhauser effects with ARIA. *Methods Enzymol.* **2001**, *339*, 71−90.

(45) Clore, G. M.; Gronenborn, A. M. New methods of structure refinement for macromolecular structure determination by NMR. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 5891−5898.

(46) Tjandra, N.; Bax, A. Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium (vol. 278, pg 1111, 1997). *Science* **1997**, *278*, 1697−1697.

(47) Tjandra, N.; Garrett, D. S.; Gronenborn, A. M.; Bax, A.; Clore, G. M. Defining long range order in NMR structure determination from the dependence of heteronuclear relaxation times on rotational diffusion anisotropy. *Nat. Struct. Biol.* **1997**, *4*, 443−449.

(48) Clore, G.; Schweiters, C. Theoretical and computational advances in biomolecular NMR spectroscopy. *Curr. Opin. Struct. Biol.* **2002**, *12*, 146−153.

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

(50) Pickford, A. R.; Smith, S. P.; Staunton, D.; Boyd, J.; Campbell, I. D. The hairpin structure of the (6)F1(1)F2(2)F2 fragment from human fibronectin enhances gelatin binding. *Embo J.* **2001**, *20*, 1519−1529.

(51) Tran, H.; VanDusen, W. J.; Argraves, W. S. The self-association and fibronectin-binding sites of fibulin-1 map to calcium-binding epidermal growth factor-like domains. *J. Biol. Chem.* **1997**, *272*, 22600−22606.

(52) Spirin, A. S.; Baranov, V. I.; Ryabova, L. A.; Ovodov, S. Y.; Alakhov, Y. B. A Continuous Cell-Free Translation System Capable of Producing Polypeptides in High-Yield. *Science* **1988**, *242*, 1162−1164.

(53) Kim, D. M.; Kigawa, T.; Choi, C. Y.; Yokoyama, S. A highly efficient cell-free protein synthesis system from Escherichia coli. *Eur. J. Biochem.* **1996**, *239*, 881−886.

(54) Kigawa, T.; Yabuki, T.; Yoshida, Y.; Tsutsui, M.; Ito, Y.; Shibata, T.; Yokoyama, S. Cell-free production and stable-isotope labeling of milligram quantities of proteins. *FEBS Lett.* **1999**, *442*, 15−19.

(55) Gardner, K. H.; Rosen, M. K.; Kay, L. E. Global folds of highly deuterated, methyl-protonated proteins by multidimensional NMR. *Biochemistry* **1997**, *36*, 1389−1401.

(56) Oba, M.; Kobayashi, M.; Oikawa, F.; Nishiyama, K.; Kainosho, M. Synthesis of C-13/D doubly labeled L-leucines: Probes for conformational analysis of the leucine side-chain. *J. Org. Chem.* **2001**, *66*, 5919−5922.

(57) Yabuki, T.; Kigawa, T.; Dohmae, N.; Takio, K.; Terada, T.; Ito, Y.; Laue, E. D.; Cooper, J. A.; Kainosho, M.; Yokoyama, S. Dual amino acid-selective and site-directed stable-isotope labeling of the human c-Ha-Ras protein by cell-free synthesis. *J. Biomol. NMR* **1998**, *11*, 295−306.

(58) Xu, R.; Ayers, B.; Cowburn, D.; Muir, T. W. Chemical ligation of folded recombinant proteins: Segmental isotopic labeling of domains for NMR studies. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 388−393.

(59) Otomo, T.; Teruya, K.; Uegaki, K.; Yamazaki, T.; Kyogoku, Y. Improved segmental isotope labeling of proteins and application to a larger protein. *J. Biomol. NMR* **1999**, *14*, 105−114.
(60) Otomo, T.; Ito, N.; Kyogoku, Y.; Yamazaki, T. NMR observation of selected segments in a larger protein: Central-segment isotope labeling through intein-mediated ligation. *Biochemistry* **1999**, *38*, 16040−16044.
(61) Blaschke, U. K.; Cotton, G. J.; Muir, T. W. Synthesis of multidomain proteins using expressed protein ligation: Strategies for segmental isotopic labeling of internal regions. *Tetrahedron* **2000**, *56*, 9461−9470.

AR010119S