Automated protein structure calculation from NMR data

Mike P Williamson and C Jeremy Craven

Department of Molecular Biology and Biotechnology, University of Sheffield, Firth Court, Western Bank, Sheffield S10 2TN, UK

Corresponding author: Mike P Williamson, email m.williamson@sheffield.ac.uk. Department of Molecular Biology and Biotechnology, University of Sheffield, Firth Court, Western Bank, Sheffield S10 2TN, UK. Fax +44 114 222 2800, Tel +44 114 222 4224.

Abstract Current software is almost at the stage to permit completely automatic structure determination of small proteins of < 15 kDa, from NMR spectra to structure validation with minimal user interaction. This goal is welcome, as it makes structure calculation more objective and therefore more easily validated, without any loss in the quality of the structures generated. Moreover, it releases expert spectroscopists to carry out research that cannot be automated. It should not take much further effort to extend automation to ca 20 kDa. However, there are technological barriers to further automation, of which the biggest are identified as: routines for peak picking; adoption and sharing of a common framework for structure calculation, including the assembly of an automated and trusted package for structure validation; and sample preparation, particularly for larger proteins. These barriers should be the main target for development of methodology for protein structure determination, particularly by structural genomics consortia.

Keywords NMR structure calculation of proteins; automation; structural genomics; expert; programs

Introduction

Approaches to complete automation of protein structure calculation by NMR have been reviewed (Altieri and Byrd 2004; Baran et al. 2004; Gronwald and Kalbitzer 2004; Güntert 2003; Huang et al. 2005), most recently in an authoritative review (Güntert 2008). This Perspective is not intended to be a review: our interest is to survey current practices and problems, and to stimulate a debate as to the best way forward, following an earlier Perspective on a related theme (Billeter et al. 2008). In particular, we aim to address the questions of what currently is automated, what can be automated and what should be automated. Does one need to be an 'expert' to calculate a protein structure from NMR data, and how much human intervention is best?

In any consideration of automated structure calculation by NMR, one must inevitably make comparisons with X-ray crystallography, where automated methods have been developed extensively over the last few years, and made a major difference to the practice of protein crystallography. Indeed, they are close to removing the need for 'expert' crystallographers at all in straightforward cases. The benefits are obvious. Structure calculation is on average faster and cheaper (Chandonia and Brenner 2006); and expert crystallographers are freed to concentrate on the more difficult and biologically important challenges. A potential problem is that each new structure gets less attention than in previous years, implying that novel or unusual features may get overlooked.[1] However, improvements in bioinformatics tools mean that structures get seized on and picked over as soon as they hit the public domain, so interesting features are unlikely to stay uncovered for long.

[1] Plus of course it makes it harder to publish a new structure.

Would the same be true for NMR structures? Is automation an obvious benefit? The answer has to be a resounding yes. As above, automated methods will free the expert from spending excessive time sorting out what should be trivial issues such as spectral referencing and folding, details of peak shapes and peak picking, dealing with unfriendly software, worries about overlapping or misaligned peaks, and so on. They will free both expert and novice from what one has to admit is the tedium of sorting peak lists, checking assignments, checking NOEs, and iteratively correcting errors in the input data, and thus make everyone's life a good deal more pleasant. All this is a much more persuasive argument for NMR than it is for crystallography. However, the most convincing argument in favor of automated methods is the elimination of subjective and irreproducible bias. Many bioinformaticians are reluctant to use NMR structures, partly because NMR structures are typically presented as a large number of structures, making the structure files much larger and raising the question of which single structure is the 'best'; but more importantly because they have doubts over the reliability of NMR structures. In crystal structures, there is a clear relationship between the structure and the experimental data, so that by comparing resolution, *R*-factor and free *R*-factor one can immediately form a reasonably reliable judgement as to the overall quality of the structure. In NMR, there is no such relationship. There is a qualitative relationship between restraint lists and structure, but there is such a lot of subjective processing between spectra and restraint lists that one cannot directly compare input data and structures in the same way as for crystallography.

There have been several analyses of the 'quality' of NMR structures. Although the global folds of protein structures produced by NMR and crystallography are closely

similar, it has been shown that on average there is a real difference on the local scale between crystal and NMR structures, which is unrelated to crystal packing (Andrec et al. 2007). Maybe this difference reflects a real difference between crystal and solution, but it remains to be proven. Even more worrying, NMR structures are typically much worse geometrically than crystal structures (Bhattacharya et al. 2007; Spronk et al. 2004 and refs therein), with geometrical parameters so far away from the expectation values that they correspond to crystal structures with resolution of 3 Å or worse (Figure 1). Structures produced by different software have reproducibly different peculiarities (Bhattacharya et al. 2007; Spronk et al. 2004). Furthermore, analyses suggest that a typical ensemble of NMR structures has a precision that is smaller (ie tighter) than its accuracy, meaning that where one can compare, the crystal structure is often significantly outside the cluster of structures represented by the NMR structures (Andrec et al. 2007; Snyder et al. 2005; Spronk et al. 2003). This again possibly reflects the fact that real structures in solution are genuinely different from the structure in the crystal (Andrec et al. 2007), but the point has been well made that most NMR structures probably have artificially over-tight precision, with the implication that structural statistics on NMR structures have at best a non-intuitive and non-standard meaning[2]. Many of the irregularities can be cleared up by the simple expedient of refinement in explicit solvent (Bhattacharya et al. 2007; Linge et al. 2003b; Nederveen et al. 2005; Ramelot et al. 2008), implying that this should be a standard procedure for all NMR structures (Spronk et al. 2004), but unfortunately it is not, and the problems are bad enough that it is an entirely

[2] The distribution of structures represented by an NMR ensemble is not (as one might expect) a representation of the true precision of the calculation. Rather, it is a representation of the reproducibility of the algorithm given the input data (Markwick et al. 2008) However, this is a subtlety that is easily overlooked.

reasonable position to have concerns over the reliability of NMR structures. Since NMR structures comprise some 15% of the Protein Data Bank, and a rather higher fraction of protein folds (Spronk et al. 2004), this is an unacceptable state.

Fully automated methods of structure calculation will not remove problems of structure quality. But they will not make them worse: analyses of structures produced by high-throughput methods suggest that the structures are at least as good as those determined manually (Snyder et al. 2005). The big benefit to be derived from automated methods is that the structures are produced by an objective repeatable method, and that the same result would be obtained by other research groups working with the same data, a situation that certainly is not true at present (Figure 1). This means that once a few (or more likely a few hundred) structures have been calculated with a given automated method, it would be entirely possible to analyse the reliability of the structures and their problems, and therefore retrospectively to correct them if needed, in a similar manner to the RECOORD project (Nederveen et al. 2005). Despite the best efforts of a large number of NMR spectroscopists, there is still no good way of telling if an NMR structure has been calculated correctly. Automation offers the best hope of solving this problem. And to answer a question posed at the start of this section, every stage of structure calculation will benefit from automation (though this of course does not mean that human intervention is not required, as discussed further below).

If this is true, why has automation not been addressed more urgently? There are a number of reasons. Protein crystallography has by and large a single and easily digitized type of data and a single ‘correct’ way to analyse the data: the methodology is not ‘owned’ by different research groups, and new methods are recognized as useful, adopted

and shared rapidly. The same could not be said to be true for NMR. This means that there are many competing software packages, each used by a subset of the NMR community, and each with its own benefits and problems (Jahnke 2007): we examine this in more detail below, particularly in the context of CCPN. NMR groups are of course often interested in many aspects of proteins other than just structure, implying that for many groups, structure calculation is just one aim among many (Billeter et al. 2008). The parameters and techniques of NMR continue to develop rapidly and 'divert' interest away from the possibly less glamorous task of calculating structures (Altieri and Byrd 2004).

For the crystallographic community, automation received a big boost with the advent of Structural Genomics (SG), which explicitly recognized the need for funding the development of automated methods. This is such an important issue that we survey SG in the following section, before turning to look at the software that currently exists.

Structural Genomics

Interest in SG first surfaced in the early 1990s, as a result of the obvious success of genomics. The question arose, now that methods for sequencing genomes are so rapid, can we do the same for the structures of proteins encoded by the genome? Groups began to get together to set up pilot studies, and to explore issues around high-throughput expression and purification, as well as technology for high-throughput structure calculation. In 1997, the Protein Folds project started in Japan, with the aim of solving enough structures to have at least one structure for each unique protein fold. [The estimate for the number of unique folds has ranged from less than 1000 to tens of thousands: it now appears that the number is somewhat less than 2000 (Levitt 2007)].

This project moved to the Genomic Sciences Center in Yokohama, Japan in 1998, with a large center dedicated to protein expression and NMR, of which more later (Yokoyama et al. 2000). Also in 1998, a pilot began in Ontario, Canada, concentrating on bacterial genomes (Yee et al. 2003). This project had a different aim, namely to determine structures for a single bacterial genome, and focused on high-throughput expression and screening for solubility and foldedness rather than calculational methodology, subsequently distributing well-behaved proteins around conventional structural groups (Yee et al. 2002). The pilot turned into the Ontario Center for Stuctural Proteomics and has solved around 300 structures, 20% of these by NMR (Yee et al. 2006).

1998 also saw the start of serious discussions on funding SG in the USA, which resulted in 2000 in the funding by NIH (in the form of the National Institute of General Medical Sciences) of nine SG centers based in the USA (Burley 2000; Terwilliger 2000). Between 2000 and 2005 these solved more than 1100 structures. In the second phase (from 2005), the Protein Structure Initiative (PSI) continued to fund four of the original centers plus several others (http://www.nigms.nih.gov/Initiatives/PSI/Centers/). Of these, the Northeast SG Consortium, based in Rutgers University, has the most significant NMR component. It has solved 580 structures, 240 by NMR. In addition, the Center for Eukaryotic SG, based in Madison, Wisconsin, also has a big NMR component, and has solved 113 structures, 40 by NMR. The Southeast Collaboratory also had an NMR component, though more for screening than for structure determination (Adams et al. 2003). Interestingly, the PSI was reviewed in 2007, concluding that structure output was good (in crystallography and NMR), but that dissemination of results, relevance to biology, and value for money were poor

(http://www.nigms.nih.gov/News/Reports/PSIAssessmentPanel2007.htm). The panel concluded that they were 'not enthusiastic about the benefit to biomedical research of the current large-scale, high-throughput structure determination effort', which does not bode well for future large-scale US funding.

In Europe, SG has been rather more fragmented, with 'little European coordination of SG' (Heinemann 2000). The EU decided to fund method development, largely in crystallography rather than NMR, focusing on protein complexes and benefits to human health, with high-throughput structure determination being left to individual countries. As a result, the first SG project, Spine (running 2002-2006), resulted in only about 300 structures, around 20% of these by NMR. There was no focus on automated methodology, and thus no major development in this area (AB et al. 2006). The Structural Genomics Consortium, based in Oxford, UK, has solved around 300 structures (5 by NMR), while the Protein Structure Factory, based in Berlin, has solved 17, 6 by NMR (http://www.proteinstrukturfabrik.de/public/PSF_Status_1.shtml).

In Japan, the initial project turned into Protein 3000, with the aim of determining 3000 protein structures over the period 2002-2007 (http://www.tanpaku.org/eng). By the end of December 2007, it had produced over 2000 crystal structures and 1400 NMR structures (http://p3krs.protein.osaka-u.ac.jp/p3kdb/status/s201_g_process_statistics.php), thus easily meeting its target, and making it the single most productive group worldwide, particularly for NMR structures. However, it received considerable criticism (eg (Cyranoski 2006) [energetically refuted by its director (Yokoyama et al. 2007)] on the grounds that many of the structures determined were closely related and therefore 'easy'.

The NMR-based SG centers are not yet carrying out fully automated structure calculations. Nevertheless, NMR structures have kept pace with crystal structures from SG centers and still dominate for proteins of less than 120 residues (Figure 2), although the proportion of structures larger than 20 kDa remains insignificant. Thus, even without full automation, NMR is keeping pace with crystallography in terms of number of structures. This reinforces the point that even now, NMR methods do not have a problem with *quantity*: their real problem is with *quality*.

In summary, the majority of SG funding has gone to fund crystallography, which has in general been very successful in meeting its aim of developing technology for high-throughput structure calculation. By contrast, most of the funding for NMR has been for small individual projects. Within the SG centers, methodology for automated NMR calculation took place mainly in two locations: in RIKEN, Japan and in the Northeast SG consortium (NESG) in the US. We discuss the outcomes from these and other developments below, after an analysis of the different stages required.

The stages of structure determination

The conventional method for determination of protein structures from NMR data has 6 stages: expression and purification of proteins; acquisition of spectra; processing of spectra and peak picking; resonance assignment; collection of structure restraints, for example NOE assignment; and structure calculation. (Note that we distinguish *resonance assignment* [assigning a chemical shift value to each nucleus] from *NOE assignment* [assigning an NOE crosspeak to its corresponding nuclei].) Although there are other

ways of doing structure calculation (discussed briefly later), this remains overwhelmingly the sequence of events. Where are the bottlenecks in each of these stages?

(i) Expression and purification of proteins

In many ways this stage is the most important and most overlooked of all, since it involves molecular biology rather than NMR, and is therefore a stage that many NMR practitioners do not see themselves as experts in, and would rather leave to someone else to worry about. Fortunately, the SG centers have devoted considerable effort here, and methods are becoming faster, cheaper and more predictable: see for example an excellent recent summary (Gräslund et al. 2008). There is no reason why the streamlined procedures adopted by crystallographic groups (Billeter et al. 2008) should not also work for NMR. SG consortia have records of which methods work for which proteins, and of course which do not (Christendat et al. 2000). One hopes that these records will be mined to produce useful guidelines.

Several of the SG centers have tried cell-free expression, because this potentially provides more routine and reliable expression: however, to date it is still far from routine. For NMR it also has the significant benefit that amino acids are incorporated intact and therefore the positions of isotopic labels do not get scrambled so much because of cellular metabolic processes. With this goes the disadvantage that one has to use isotopically labelled amino acids rather than minimal medium, which tends to make it a more expensive option.

In summary, the situation for protein expression is continually improving. For NMR, probably the most important question (particularly for larger proteins) is how to

increase solubility and reduce aggregation, since these problems limit the protein concentration and tend to increase the T_2 relaxation rate, thus degrading signal-to-noise, as discussed below.

(ii) Acquisition of spectra

Probably still the biggest drawback of NMR is its low sensitivity (Billeter et al. 2008). Clearly, higher fields and cryoprobes have helped enormously, but the fact still remains that NMR spectra have low signal-to-noise ratios (S/N). There has therefore been considerable work in this area. For automation, a very significant development is 'reduced dimensionality' spectra, such as the GFT method (Atreya and Szyperski 2004; Shen et al. 2005). In these spectra, the chemical shifts of several nuclei are measured simultaneously, as sums and differences of frequencies. Subsequent processing then untangles these to regenerate the original frequencies. The big advantage is that reduced dimensionality can enable the acquisition of 4D or 5D spectra in a few days, thus giving a major gain in speed of acquisition, at least in cases where protein concentration and linewidth is favorable. The ability to correlate a large number of resonances in one spectrum makes the task of resonance assignment simpler and more amenable to automated methods (Liu et al. 2005). Disadvantages of GFT are the rather longer pulse sequences needed and the increased complexity of spectra and processing, meaning that such methods are applicable only to small proteins: in practice this means a limit of about 15 kDa. Projection spectra are also a promising method (Hiller et al. 2007; Kupče and Freeman 2004).

The development of nonlinear sampling methods is another important development (Barna et al. 1987; Malmodin and Billeter 2005). The idea behind this method is that it is not necessary to acquire all the indirect time points in order to determine the indirect frequencies. One can therefore make a considerable saving in acquisition time by missing out a fraction of the indirect time points. Moreover, data acquired at short indirect acquisition times (eg, short t_1 times in a 2D experiment) have better S/N than data acquired at longer indirect acquisition times. Therefore to improve S/N one should spend more time acquiring data with short indirect acquisition times, and only acquire relatively few FIDs from the longer acquisition times, to provide the required resolution. In this way, total acquisition times can be reduced very significantly, by factors of 3-5 in 3D spectra (Malmodin and Billeter 2005). However, the drawback is that one can no longer process the data using a Fourier transform. Another development worth mentioning is the more rapid recycling of 'unused' non-amide proton magnetisation to permit more rapid pulsing (see references in AB et al. 2006). Such methods when used methodically have the potential to increase S/N by at least a factor of 2, a not insignificant gain, and have certainly not been adopted widely as yet.

Often overlooked, but stressed in an excellent review from the NESG, is the observation that when comparing data from different spectra (for example, for resonance assignment), the quality of the comparison is much better when the spectra have the same offsets, spectral widths, and broadly similar processing and digital resolutions (Baran et al. 2004). That is, for automation one requires a standardised data collection strategy. Such a procedure reduces the need for expert advice and makes subsequent processing and analysis much quicker, and importantly means that the chemical shift tolerance

needed when making NOE assignments can be reduced. Easy to say but less easy to achieve in practice is the crucial need to keep the temperature (as well as pH and buffer composition) the same among all spectra.

We conclude that a carefully considered choice of methods (most likely including reduced dimensionality) in the context of a standardised collection methodology is capable of producing high quality data in acceptably short times: in other words, for soluble monomeric proteins of up to 15 kDa, data acquisition should not be a bottleneck. (It is of course true that actually implementing methods such as GFT or non-linear processing poses problems even for specialist NMR labs.) Particularly for larger proteins, the situation is quite different (Figure 2). The solution that is most often offered here, both for acquisition and assignment, is selective deuteration, for example using the SAIL methodology in which well over half the protons in a protein are removed by selective deuteration (Takeda et al. 2007).

(iii) Processing of spectra and peak picking

The vast majority of NMR spectra are processed by fast Fourier transforms (FFT). It seems likely that non-linear sampling (previous section) provides such a major gain in acquisition rate, particularly for 3D spectra, that it should be widely adopted, implying that everyone working with multidimensional spectra of proteins should be routinely using processing methods such as maximum entropy or three-way decomposition (Malmodin and Billeter 2005). Nevertheless, the vast majority of NMR labs continue to use conventional acquisition and processing methods. Why is this? We suggest that it is mere inertia (and time pressure). If it was easy to acquire and process using non-linear

sampling, everyone would do it. However, time and effort are required to get it working and to integrate it with existing procedures, and it is therefore given low priority. A similar argument holds also for GFT and many of the other topics discussed below. The problem would clearly be solved if (as is roughly the case for the crystallographic community) there was one standard data format and processing package which everyone used, and into which one could simply 'slot in' a new method. But the fact that each program uses a different data format means that incorporating a new method requires time and effort in reformatting. We therefore suggest that a major bottleneck in automation is the multiplicity of competing software packages, many of which do roughly the same things in similar ways (Malmodin and Billeter 2005).

Having processed the time-domain data to obtain frequency-domain spectra, the next step is peak picking: that is, the creation of lists containing the frequencies of peaks within the spectrum. A popular program for this application is NMRView (Johnson and Blevins 1994). It is clear that peak picking remains one of the major problem areas, and an urgent target for improvement. The reason for this is that the picked peak lists form the basis for the subsequent steps of resonance assignment and NOE assignment. If the peak list is missing real peaks (for example, because it cannot pick out peaks from noise, or if peaks are missing because of motional and/or relaxation problems) then the resonance assignment will be incomplete, whereas if it contains incorrect peaks (eg noise, solvent, or artifacts) then both assignment and structure calculation suffer. Numerous studies (see for example Altieri and Byrd 2004; Baran et al. 2004; Güntert 2003) have shown that both automated resonance assignment and automated NOE assignment

become unreliable if the peak lists that they rely on start to contain too many incorrect peaks, as discussed in more detail below.

In principle automatic peak picking is simple: go through the spectra and find local maxima. However, there are three main problems: noise, artifacts and peak overlap, of which noise is the hardest and less tractable problem (Bartels et al. 1997; Slupsky et al. 2003; Zimmerman et al. 1997). Various methods have been used to distinguish noise from real peaks, including lineshape (crudely, a real peak will probably be several points wide while a noise peak may be a single-point spike; more sophisticated, one can match peaks to expected or real lineshapes), position (a real peak in a NOESY spectrum should have a corresponding diagonal peak; peaks too close to the water or the diagonal can be automatically eliminated), and frequency matching (NOE peaks will usually have symmetry-related partners; peak positions in different spectra should have matching shifts). To date there has been relatively little development on peak picking, but there are suggestions that effort could pay off handsomely. As an example, automated preprocessing of spectra using the program APART (which matches frequencies in multiple spectra, as well as carrying out a check for assignments that deviate too markedly from normal shift ranges) was shown to yield a significant improvement in the number of correct assignments carried out subsequently by Autoassign, particularly for noisy data but even for data with good S/N (Pawley et al. 2005). Peak picking is thus a crucially important area, and one that should yield relatively easily to a determined assault.

(iv) Resonance assignment

The assignment of backbone resonances for proteins < 15 kDa is clearly possible using several automated programs. Possibly the simplest method uses GFT spectra, where effectively all the information is contained in one or two 5D spectra and there are therefore no problems of interspectral alignment, but standard sets of 3D spectra also work well, even for rather larger proteins. The same cannot be said of sidechain assignment. Here there is a major problem, and both the two main SG groups note that manual intervention is usually required at this stage (Baran et al. 2004; Kobayashi et al. 2007). The difficulties with sidechain assignment stem from both missing peaks (incomplete ^{13}C TOCSY transfer) and overlapping peaks. However, this is largely a problem in pattern recognition, which is difficult but by no means impossible.

There are two complementary approaches being taken, which seem likely to make automated sidechain assignment possible in the near future. The first is improvements to current methods, for example using methods for predicting or matching expected chemical shift values (Fiorito et al. 2008; Hitchens et al. 2003; Malmodin et al. 2003; Moseley et al. 2004a), and combining *J*-coupling-based spectra with NOE-based spectra in conjunction with databases that give distance distributions (Kamisetty et al. 2006; Xiong et al. 2008). And the second is to combine resonance assignments with the subsequent stages of NOE assignment and structure calculation, and iteratively to improve the accuracy and completeness of each stage. In particular, it is claimed that the program FLYA is now capable of completely automated sidechain assignment by a combination of these approaches (López-Méndez and Güntert 2006). Thus, it appears that incremental improvements in current methods should permit fully automated assignments in the near future.

(v, vi) NOE assignment and structure calculation

Although the NOE assignment and structure calculation stages are conceptually different, in practice they have always gone together (Williamson et al. 1985). Typically a preliminary, incomplete and error-rich NOE assignment is used to calculate the first set of structures; this set is used to correct and expand the NOE restraint list, and a new calculation round is launched; and so on for several iterations. The question is, how much initial error can be tolerated. The main structure calculation packages note that they require the resonance assignments to be approximately 85% complete and correct for the subsequent NOE assignment to work properly, because it is the resonance assignment list that is used subsequently for the iterative assignment of ambiguous and unassigned NOEs (Baran et al. 2004; Herrmann et al. 2002; Jee and Güntert 2003). They differ in their requirements for completeness and accuracy of the initial NOE assignments, which appears to be much less critical. In the classical manual structure calculation, it is this stage of going through NOE spectra and trying to make unambiguous (and correct) NOE assignments that is the most tedious and time-consuming part. Automated packages have made major developments in this area, and can cope with an initial NOE peak list in which 50% or less of the peaks correspond to real NOEs (Kuszewski et al. 2004; López-Méndez and Güntert 2006). Systematic investigations of parameters used for NOE assignment, particularly the chemical shift tolerance and iteration methodology, have identified suitable strategies (Fossi et al. 2005a; Fossi et al. 2005b). These two stages are the areas that the automated calculations have tended to focus on, with the result that they are now reasonably robust and almost completely automated.

Automated programs and methods

As noted above, the most significant automated programs originate from the work of Güntert (who was affiliated with the ETH Zürich, Switzerland from 1987 to 2002, then with RIKEN in Japan until 2007, before joining the Goethe University in Frankfurt, Germany) and the NESG (led by Montelione). There are of course many other programs, at various stages of automation (Table 1). Güntert's approach has been evolutionary: each new program takes the good features of earlier ones, and he has not hesitated to take good ideas from others. A good example is the crucial idea of ambiguous NOEs, which came from the group of Nilges, the author of ARIA (Nilges and O'Donoghue 1998). It is very frequently found that more than one assignment can be made for a given NOE peak. The insight of Nilges was to recognise that the peak can be treated as the sum of all possible assignments, and can therefore be represented by a sum of restraints, each weighted by the inverse sixth power of the corresponding distance. This combined restraint will always be correct as long as the correct assignment is included within the list of possible restraints; the ambiguity merely makes it less powerful as a restraint. Other ideas are Güntert's own, in particular network anchoring (giving more weight to a restraint if it is supported by other spatially related restraints) and constraint combination (inclusion of restraints as pairs, only one of which needs to be satisfied, to reduce the potential distortions caused by inclusion of a genuinely incorrect restraint) (Herrmann et al. 2002). A collection of routines (eg CALIBA, GLOMSA, HABAS) were combined together with a torsional angle molecular dynamics routine to make the program DIANA, which then made use of further routines (REDAC, ASNO, the assignment program GARANT, and

molecular dynamics program OPAL) and evolved into DYANA (Güntert et al. 1997), which in turn absorbed further routines, in particular the NOE assignment module CANDID, and was packaged within an updated torsion angle molecular dynamics program to become the widely used structure calculation program CYANA (Güntert 2003). This program has more recently incorporated the peak picking routine AUTOPSY and become FLYA (López-Méndez and Güntert 2006). The key elements of FLYA are made up from NMRView, AUTOPSY, GARANT, CYANA, and OPALp: it is thus a combination of programs written at various times by various people though mainly by Güntert and colleagues, but assembled into a coherent package.

FLYA is the only package that claims, with some justification (Scott et al. 2006), to be able to calculate protein structures from NMR spectra (but not raw time-domain data) in a fully automatic way. It is based on its demonstrably successful predecessors DYANA and CANDID, as described above, and has so far only been tested on three small (< 16 kDa) proteins (López-Méndez and Güntert 2006). Güntert's summary of the program concludes that 'Fully automated structure determination of proteins up to 140 amino acid residues is possible now', a conclusion that seems fully justified, at least for well-behaved proteins.

Probably the key feature of FLYA that distinguishes it from many of its competitors (eg CYANA and AutoStructure) is that whereas CYANA and AutoStructure require approximately 85-90% complete resonance assignment before a successful structure calculation can begin, FLYA requires no initial assignment, but the outcome is an assignment over 95% complete and correct (or at least it was on the test set of 3 small proteins).

Güntert spent about 6 years working at the high-throughput NMR SG centre in RIKEN, which recently produced the program KUJIRA[3] (Kobayashi et al. 2007). KUJIRA has been used to calculate what is by a long way the largest number of semi-automatically calculated structures, and uses CYANA and NMRView together with a number of specifically written modules. KUJIRA is deliberately and explicitly not a fully automated package, with user intervention required in particular for sidechain assignment, and also for checking of NOE restraints via graphical interfaces. It and its predecessors were used to calculate approximately 800 structures, largely using a 'production line' method, using dedicated NMR data acquisition and structure calculation staff, with typically 3 weeks analysis time per structure. One can therefore conclude that both KUJIRA and CYANA have been thoroughly road-tested.

The approach taken by NESG has been rather different (Baran et al. 2004; Huang et al. 2005). The center took its high-throughput role seriously, and looked at each stage of the process, refining each stage and writing new software and procedures where required. The aim was to produce an automated package that would encompass all stages of structure determination with minimal user intervention. The group has therefore a clearly defined standardised data collection protocol; a database for archiving and organising NMR and structural data; packages for automatic processing of NMR data, peak picking, editing and checking; automated assignment (AutoAssign: Moseley et al. 2001); automated structure calculation (AutoStructure: Huang et al. 2005); and automated structure validation. As for FLYA, the packages include various externally written software (NMRPipe, Sparky, DYANA/CNS/XPLOR-NIH) assembled within a purpose-written core. The approach of NESG is also different from that of most others in

[3] Kujira is Japanese for whale: a reference to the size and complexity of the program?

that NOE restraints are incorporated and verified using a 'bottom-up' procedure that identifies and builds regular secondary structure first, in order to generate a low-resolution fold that can be used to accept or reject NOEs (as compared to the 'top-down' use of ambiguous restraints as described above, in which all possible assignments are included initially and the incorrect ones subsequently removed). User intervention is required particularly for sidechain assignment.

Numerous programs and packages have been written elsewhere. Among these we note particularly the comprehensive package Auremol, which has introduced a number of very useful steps, including an automated assessment of structure quality and a Bayesian NOE assignment program (Gronwald and Kalbitzer 2004); and ARIA (Nilges 1995), which has excellent automated procedures for steps (v) and (vi) above, a graphical interface and compatibility with the CCPN data model (see below) (Linge et al. 2003a; Rieping et al. 2007). As noted above, many of the ideas developed for ARIA have also been used by other programs.

Of particular significance to automation is the PASD (Probabilistic Assignment Algorithm for Structure Determination) program, for which the goal was to produce a program that can accept an automatically picked NOE list containing a high proportion of errors (up to 80% of incorrectly picked peaks!) and still be able to calculate structures correctly (Kuszewski et al. 2004). It achieved this using a combination of a linear NOE error function that is not dependent on size of violation and therefore does not penalize incorrect NOEs too severely; plus probabilistic NOE assignments, in which all possible assignments are allowed but only the correct ones 'work' because they tend to act

together to pull the structure towards the correct fold, whereas wrong assignments pull in random directions. More recent additions are a network anchoring preprocessing step, plus repulsive distance restraints ('non-NOEs') representing deductions from the network analysis on protons that cannot be close together, which act to speed up the calculation (Kuszewski et al. 2008).

It therefore appears that the major programs are gradually evolving more closely together, with similar ideas being used. This probably suggests that it is becoming less important which method and/or program was used to assign spectra, assign restraints and calculate the structure. This is still an issue, but one hopes that as validation methods become more useful (in particular, as they work out how to compare structures to input spectra rather than the restraints derived from them), then any differences in methodology that lead to differences in the structures produced will become more obvious.

We have noted that a major drawback to the more rapid adoption of new methods is the proliferation of programs, many with non-compatible formats, which therefore require significant efforts to transfer information from one to another. A very interesting development in this area is the Collaborative Computing Project for NMR (CCPN), based in Cambridge, UK, which has developed a 'data model' to act as a framework describing proteins and their associated NMR data, into which actual data can be loaded (Vranken et al. 2005). It is designed to facilitate the exchange and interconversion of data from different sources, as well as its deposition and archiving, so that different programs can use and transfer data essentially transparently, and all relevant data can be stored in one location. As part of the project, an analysis program CCPNmr Analysis has been written (Vranken et al. 2005), based on ANSIG and Sparky, that can exchange data with a wide

range of other programs. Thus effectively, CCPN aims to be an organizational structure within which other programs can be linked. CCPN has been linked to several other projects, including ARIA and the BioMagResBank (BMRB).

Future prospects and goals

In this final section, we look first at possible alternative approaches to those discussed above, and then try to summarise our conclusions as to what should be the next steps.

There have been numerous attempts to bypass the difficult step of resonance assignment, effectively by allowing the structure calculation process to make assignments as it goes. The most discussed of these is the CLOUDS/ABACUS method (Grishaev and Llinás 2002; Grishaev et al. 2005), though there are many others. It is an interesting approach and claims to be fully automatic, but so far has seen few applications.

The most important restraints for structure calculation have always been NOEs. It seems very unlikely that the need for NOEs will ever be removed, because of their uniquely powerful and direct structural information. However, a number of other parameters are being explored and have shown remarkable power. These include residual dipolar couplings (rdcs) and chemical shifts (Korukottu et al. 2007; Shen et al. 2008). It is clear that incorporation of these as restraints is useful both in providing additional and often complementary structural information, and in improving the accuracy of the crucial initial protein fold calculation. In particular, because chemical shift assignments are always generated either before or during the structure calculation and are therefore always available, we can look forward to their further incorporation as restraints.

Structure calculation, particularly by NMR, has always used prior information that was not derived from the NMR data: for example, bond lengths, van der Waals radii and hydrogen bond geometry. It is widely accepted (though by no means universally) that for the 'best' structure, one should use all the available information. There have been a number of interesting studies using homology-based modeling and structures from databases, which hold out the promise to predict both global fold and local structure with good accuracy. Of particular interest is the use of structure prediction algorithms such as the program Rosetta: as an aid in both fold generation and structure refinement, this appears a very promising avenue (Korukottu et al. 2007; Meiler and Baker 2003; Ramelot et al. 2008).

A key stage in structure calculation is validation of the quality of the structures produced. We have not discussed this question here in any detail: not because it is not important, but because correctly assembled automated packages should do the validation routinely, as for example discussed by the NESG consortium (Moseley et al. 2004b). Several authors have commented that as yet there is no agreed method for validation. Clearly this is a big problem, and one that is vital for the success of any truly objective calculation.

In conclusion, it appears that fully automated structure calculation is already possible, at least for proteins up to 15 kDa. However, an important question is whether full automation is actually desirable. A crystallographic colleague commented to us that although fully automated crystal structures are now possible, 'only a fool would trust one': in other words, some human intervention is always needed, particularly to check

that results look sensible. This must surely be much more the case for NMR than it is for crystallography, because the input data are more varied and prone to artifacts. (We note a highly relevant remark in Güntert (2003): '*If used sensibly*, automated NOESY assignment with CANDID has no disadvantage compared to the conventional, interactive approach but is a lot faster, and more objective' [our italics].) We suggest that some steps that currently require human intervention, such as sidechain assignment and NOE assignment, can and should be run in an essentially fully automated manner; but that humans are needed to check at least the input and output, aided of course by automated quality reports.

What then does our analysis suggest as the most important steps in the future? In a rough order of importance (which is approximately in agreement with Güntert (2008)), we can conclude that effective automation requires:

1. An efficient and effective peak picking routine, that incorporates features shown to work already such as comparisons between different spectra, and that works in an integrated and dynamic manner (ie, peak lists get updated during the calculation) with other parts of the package.
2. An integrated and user-friendly package, incorporating the best ideas from existing packages in a modular way. Our analysis suggests that it would have a FLYA-like core, with data acquisition, data management and structure validation (ie the beginning and end parts) based on NESG procedures, all within a CCPN data model (and therefore allowing users to use alternative programs if preferred). In particular, automation of the validation process into a reliable and routine package is a key goal

(Billeter et al. 2008). In the medium term, it is probably structure validation where developments are most urgently required, and where human intervention is (and will remain) most necessary.

3. Improved methods for dealing with proteins larger than 15 kDa (clearly a problem, as shown by Figure 2). Some likely targets here are: selective deuteration schemes such as SAIL, in conjunction with cell-free expression; improved methods for predicting domain boundaries; methods for looking at one domain in the context of an intact protein, for which inteins look increasingly interesting (Skrisovska and Allain 2008; Yagi et al. 2004); improved sensitivity, possibly by making better use of 'unused' magnetisation; and methods for improving solubility, reducing aggregation and reducing the rotational correlation time. Such methods will of course also be useful for the more 'difficult' proteins less than 15 kDa.

Our analysis suggests that it is in the interests of the whole NMR community for this to happen as soon as possible, primarily to increase the 'respectability' of NMR as a structural tool. In particular, we suggest that NMR groups within SG centers (and their funding bodies) should put these goals at the top of their list.

Acknowledgements

We thank members of the Editorial Board of *JBNMR* for their helpful comments.

References

AB E, Atkinson AR, Banci L, Bertini I, Ciofi-Baffoni S, Brunner K, Diercks T, Dötsch V, Engelke F, Folkers GE, Griesinger C, Gronwald W, Günther U, Habeck M, de Jong RN, Kalbitzer HR, Kieffer B, Leeflang BR, Loss S, Luchinat C, Marquardsen T, Moskau D, Neidig KP, Nilges M, Piccioli M, Pierattelli R, Rieping W, Schippmann T, Schwalbe H, Travé G, Trenner J, Wöhnert J, Zweckstetter M, Kaptein R (2006) NMR in the SPINE structural proteomics project. Acta Cryst. Sect. D 62:1150-1161

Adams MWW, Dailey HA, Delucas LJ, Luo M, Prestegard JH, Rose JP, Wang BC (2003) The Southeast Collaboratory for Structural Genomics: A high-throughput gene to structure factory. Accounts Chem. Res. 36:191-198

Altieri AS, Byrd RA (2004) Automation of NMR structure determination of proteins. Curr. Op. Struct. Biol. 14:547-553

Andrec M, Snyder DA, Zhou ZY, Young J, Montelione GT, Levy RM (2007) A large data set comparison of protein structures determined by crystallography and NMR: Statistical test for structural differences and the effect of crystal packing. Proteins: Struct. Funct. Bioinf. 69:449-465

Atreya HS, Szyperski T (2004) G-matrix Fourier transform NMR spectroscopy for complete protein resonance assignment. Proc. Natl Acad. Sci. USA 101:9642-9647

Baran MC, Huang YJ, Moseley HNB, Montelione GT (2004) Automated analysis of protein NMR assignments and structures. Chem. Rev. 104:3541-3555

Barna JCJ, Laue ED, Mayger MR, Skilling J, Worrall SJP (1987) Exponential sampling, an alternative method for sampling in two-dimensional NMR experiments. J. Magn. Reson. 73:69-77

Bartels C, Xia TH, Billeter M, Güntert P, Wüthrich K (1995) The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. J. Biomol. NMR 6:1-10

Bartels C, Güntert P, Billeter M, Wüthrich K (1997) GARANT - A general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. J. Comp. Chem. 18:139-149

Bhattacharya A, Tejero R, Montelione GT (2007) Evaluating protein structures determined by structural genomics consortia. Proteins: Struct. Funct. Bioinf. 66:778-795

Billeter M, Wagner G, Wüthrich K (2008) Solution NMR determination of proteins revisited. J. Biomol. NMR 42:155-158

Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL

(1998) Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Cryst. D 54:905-921

Burley SK (2000) An overview of structural genomics. Nature Struct. Biol. 7:932-934

Chandonia JM, Brenner SE (2006) The impact of structural genomics: Expectations and outcomes. Science 311:347-351

Chen L, Oughtred R, Berman HM, Westbrook J (2004) TargetDB: a target registration database for structural genomics projects. Bioinformatics 20:2860-2862

Christendat D, Yee A, Dharamsi A, Kluger Y, Savchenko A, Cort JR, Booth V, Mackereth CD, Saridakis V, Ekiel I, Kozlov G, Maxwell KL, Wu N, McIntosh LP, Gehring K, Kennedy MA, Davidson AR, Pai EF, Gerstein M, Edwards AM, Arrowsmith CH (2000) Structural proteomics of an archaeon. Nature Struct. Biol. 7:903-909

Cyranoski D (2006) 'Big science' protein project under fire. Nature 443:382-382

Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A (1995) NMRPipe: A multidimensional spectral processing system based on Unix pipe. J. Biomol. NMR 6:277-293

Fiorito F, Herrmann T, Damberger FF, Wüthrich K (2008) Automated amino acid side-chain NMR assignment of proteins using ^{13}C- and ^{15}N-resolved 3D [^{1}H, ^{1}H]-NOESY. J Biomol NMR 42:23-33

Fossi M, Linge J, Labudde D, Leitner D, Nilges M, Oschkinat H (2005a) Influence of chemical shift tolerances on NMR structure calculations using ARIA protocols for assigning NOE data. J. Biomol. NMR 31:21-34

Fossi M, Oschkinat H, Nilges M, Ball LJ (2005b) Quantitative study of the effects of chemical shift tolerances and rates of SA cooling on structure calculation from automatically assigned NOE data. J. Magn. Reson. 175:92-102

Gräslund S, Nordlund P, Weigelt J, Bray J, Hallberg BM, Gileadi O, Knapp S, Oppermann U, Arrowsmith C, Hui R, Ming J, Dhe-Paganon S, Park HW, Savchenko A, Yee A, Edwards A, Vincentelli R, Cambillau C, Kim R, Kim SH, Rao Z, Shi Y, Terwilliger TC, Kim CY, Hung LW, Waldo GS, Peleg Y, Albeck S, Unger T, Dym O, Prilusky J, Sussman JL, Stevens RC, Lesley SA, Wilson IA, Joachimiak A, Collart F, Dementieva I, Donnelly MI, Eschenfeldt WH, Kim Y, Stols L, Wu R, Zhou M, Burley SK, Emtage JS, Sauder JM, Thompson D, Bain K, Luz J, Gheyi T, Zhang F, Atwell S, Almo SC, Bonanno JB, Fiser A, Swaminathan S, Studier FW, Chance MR, Sali A, Acton TB, Xiao R, Zhao L, Ma LC, Hunt JF, Tong L, Cunningham K, Inouye M, Anderson S, Janjua H, Shastry R, Ho CK, Wang DY, Wang H, Jiang M, Montelione GT, Stuart DI,

Owens RJ, Daenke S, Schütz A, Heinemann U, Yokoyama S, Büssow K, Gunsalus KC (2008) Protein production and purification. Nature Methods 5:135-146

Grishaev A, Llinás M (2002) CLOUDS, a protocol for deriving a molecular proton density via NMR. Proc. Natl Acad. Sci. USA 99:6707-6712

Grishaev A, Steren CA, Wu B, Pineda-Lucena A, Arrowsmith C, Llinás M (2005) ABACUS, a direct method for protein NMR structure computation via assembly of fragments. Proteins: Struct. Funct. Bioinf. 61:36-43

Gronwald W, Kalbitzer HR (2004) Automated structure determination of proteins by NMR spectroscopy. Progr. Nucl. Magn. Reson. Spectrosc. 44:33-96

Güntert P, Mumenthaler C, Wüthrich K (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. J. Mol. Biol. 273:283-298

Güntert P (2003) Automated NMR protein structure calculation. Progr. NMR Spectrosc. 43:105-125

Güntert P (2008) Automated structure determination from NMR spectra. Eur. Biophys. J. 38:in press

Heinemann U (2000) Structural genomics in Europe: Slow start, strong finish? Nature Struct. Biol. 7:940-942

Herrmann T, Güntert P, Wüthrich K (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. J. Mol. Biol. 319:209-227

Hiller S, Wasmer C, Wider G, Wüthrich K (2007) Sequence-specific resonance assignment of soluble nonglobular proteins by 7D APSY-NMR Spectroscopy. J. Am. Chem. Soc. 129:10823-10828

Hitchens TK, Lukin JA, Zhan YP, McCallum SA, Rule GS (2003) MONTE: An automated Monte Carlo based approach to nuclear magnetic resonance assignment of proteins. J. Biomol. NMR 25:1-9

Huang YPJ, Moseley HNB, Baran MC, Arrowsmith C, Powers R, Tejero R, Szyperski T, Montelione GT (2005) An integrated platform for automated analysis of protein NMR structures. Methods Enzymol. 394:111-141

Jahnke W (2007) Perspectives of biomolecular NMR in drug discovery: the blessing and curse of versatility. J. Biomol. NMR 39:87-90

Jee J, Güntert P (2003) Influence of the completeness of chemical shift assignments on NMR structures obtained with automated NOE assignment. J. Struct. Funct. Genomics 4:179-189

Johnson BA, Blevins RA (1994) NMRView: A computer program for the visualization and analysis of NMR data. J. Biomol. NMR 4:603-614

Kamisetty H, Bailey-Kellogg C, Pandurangan G (2006) An efficient randomized algorithm for contact-based NMR backbone resonance assignment. Bioinformatics 22:172-180

Kobayashi N, Iwahara J, Koshiba S, Tomizawa T, Tochio N, Güntert P, Kigawa T, Yokoyama S (2007) KUJIRA, a package of integrated modules for systematic and interactive analysis of NMR data directed to high-throughput NMR structure studies. J. Biomol. NMR 39:31-52

Korukottu J, Bayrhuber M, Montaville P, Vijayan V, Jung YS, Becker S, Zweckstetter M (2007) Fast high-resolution protein structure determination by using unassigned NMR data. Angewandte Chemie Int. Ed. 46:1176-1179

Kupče E, Freeman R (2004) Projection-reconstruction technique for speeding up multidimensional NMR spectroscopy. J. Am. Chem. Soc. 126:6429-6440

Kuszewski J, Schwieters CD, Garrett DS, Byrd RA, Tjandra N, Clore GM (2004) Completely automated, highly error-tolerant macromolecular structure determination from multidimensional nuclear Overhauser enhancement spectra and chemical shift assignments. J. Am. Chem. Soc. 126:6258-6273

Kuszewski JJ, Thottungal RA, Clore GM, Schwieters CD (2008) Automated error-tolerant macromolecular structure determination from multidimensional nuclear Overhauser enhancement spectra and chemical shift assignments: improved robustness and performance of the PASD algorithm. J. Biomol. NMR 41:221-239

Levitt M (2007) Growth of novel protein structural data. Proc. Natl Acad. Sci. USA 104:3183-3188

Linge JP, Habeck M, Rieping W, Nilges M (2003a) ARIA: automated NOE assignment and NMR structure calculation. Bioinformatics 19:315-316

Linge JP, Williams MA, Spronk CAEM, Bonvin AMJJ, Nilges M (2003b) Refinement of protein structures in explicit solvent. Proteins: Struct. Funct. Bioinf. 50:496-506

Liu GH, Shen Y, Atreya HS, Parish D, Shao Y, Sukumaran DK, Xiao R, Yee A, Lemak A, Bhattacharya A, Acton TA, Arrowsmith CH, Montelione GT, Szyperski T (2005) NMR data collection and analysis protocol for high-throughput protein structure determination. Proc. Natl Acad. Sci. USA 102:10487-10492

López-Méndez B, Güntert P (2006) Automated protein structure determination from NMR spectra. J. Am. Chem. Soc. 128:13112-13122

Malmodin D, Papavoine CHM, Billeter M (2003) Fully automated sequence-specific resonance assignments of heteronuclear protein spectra. J. Biomol. NMR 27:69-79

Malmodin D, Billeter M (2005) High-throughput analysis of protein NMR spectra. Progr. NMR Spectrosc. 46:109-129

Markwick PRL, Malliavin T, Nilges M (2008) Structural biology by NMR: Structure, dynamics and interactions. PLOS Comp. Biol. 4:e1000168

Meiler J, Baker D (2003) Rapid protein fold determination using unassigned NMR data. Proc. Natl Acad. Sci. USA 100:15404-15409

Moseley HNB, Monleon D, Montelione GT (2001) Automatic determination of protein backbone resonance assignments from triple resonance nuclear magnetic resonance data. Methods Enzymol. 339:91-108

Moseley HNB, Riaz N, Aramini JM, Szyperski T, Montelione GT (2004a) A generalized approach to automated NMR peak list editing: application to reduced dimensionality triple resonance spectra. J. Magn. Reson. 170:263-277

Moseley HNB, Sahota G, Montelione GT (2004b) Assignment validation software suite for the evaluation and presentation of protein resonance assignment data. J. Biomol. NMR 28:341-355

Nabuurs SB, Nederveen AJ, Vranken W, Doreleijers JF, Bonvin AMJJ, Vuister GW, Vriend G, Spronk CAEM (2004) DRESS: a database of REfined solution NMR structures. Proteins: Struct. Funct. Bioinf. 55:483-486

Nederveen AJ, Doreleijers JF, Vranken W, Miller Z, Spronk CAEM, Nabuurs SB, Güntert P, Livny M, Markley JL, Nilges M, Ulrich EL, Kaptein R, Bonvin AMJJ (2005) RECOORD: A recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank. Proteins: Struct. Funct. Bioinf. 59:662-672

Nilges M (1995) Calculation of protein structures with ambiguous distance restraints: Automated assignment of ambiguous NOE crosspeaks and disulfide connectivities. J. Mol. Biol. 245:645-660

Nilges M, O'Donoghue SI (1998) Ambiguous NOEs and automated NOE assignment. Progr. NMR Spectrosc. 32:107-139

Pawley NH, Gans JD, Michalczyk R (2005) APART: Automated preprocessing for NMR assignments with reduced tedium. Bioinformatics 21:680-682

Ramelot TA, Raman S, Kuzin AP, Xiao R, Ma LC, Acton TB, Hunt JF, Montelione GT, Baker D, Kennedy MA (2009) Improving NMR protein structure quality by Rosetta refinement: A molecular replacement study. Proteins: Struct. Funct. Bioinf. in press:

Rieping W, Habeck M, Bardiaux B, Bernard A, Malliavin TE, Nilges M (2007) ARIA2: Automated NOE assignment and data integration in NMR structure calculation. Bioinformatics 23:381-382

Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM (2003) The XPLOR-NIH NMR molecular structure determination package. J. Magn. Reson. 160:65-73

Scott A, López-Méndez B, Güntert P (2006) Fully automated structure determinations of the Fes SH2 domain using different sets of NMR spectra. Magn. Reson. Chem. 44:S83-S88

Shen Y, Atreya HS, Liu GH, Szyperski T (2005) G-matrix Fourier transform NOESY-based protocol for high-quality protein structure determination. J. Am. Chem. Soc. 127:9085-9099

Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu GH, Eletsky A, Wu YB, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT,

Baker D, Bax A (2008) Consistent blind protein structure generation from NMR chemical shift data. Proc. Natl Acad. Sci. USA 105:4685-4690

Skrisovska L, Allain FHT (2008) Improved segmental isotope labeling methods for the NMR study of multidomain or large proteins: Application to the RRMs of Npl3p and hnRNP L. J. Mol. Biol. 375:151-164

Slupsky CM, Boyko RF, Booth VK, Sykes BD (2003) Smartnotebook: A semi-automated approach to protein sequential NMR resonance assignments. J. Biomol. NMR 27:313-321

Snyder DA, Bhattacharya A, Huang YPJ, Montelione GT (2005) Assessing precision and accuracy of protein structures derived from NMR data. Proteins: Struct. Funct. Bioinf. 59:655-661

Spronk CAEM, Nabuurs SB, Bonvin AMJJ, Krieger E, Vuister GW, Vriend G (2003) The precision of NMR structure ensembles revisited. J. Biomol. NMR 25:225-234

Spronk CAEM, Nabuurs SB, Krieger E, Vriend G, Vuister GW (2004) Validation of protein structures derived by NMR spectroscopy. Progr. NMR Spectrosc. 45:315-337

Takeda M, Ikeya T, Güntert P, Kainosho M (2007) Automated structure determination of proteins with the SAIL-FLYA NMR method. Nature Protocols 2:2896-2902

Terwilliger TC (2000) Structural genomics in North America. Nature Struct. Biol. 7:935-939

Vranken WF, Boucher W, Stevens TJ, Fogh RH, Pajon A, Llinás P, Ulrich EL, Markley JL, Ionides J, Laue ED (2005) The CCPN data model for NMR spectroscopy: Development of a software pipeline. Proteins: Struct. Funct. Bioinf. 59:687-696

Williamson MP, Havel TF, Wüthrich K (1985) Solution conformation of proteinase inhibitor IIA from bull seminal plasma by ^{1}H nuclear magnetic resonance and distance geometry. J. Mol. Biol. 182:295-315

Xiong F, Pandurangan G, Bailey-Kellogg C (2008) Contact replacement for NMR resonance assignment. Bioinformatics 24:I205-I213

Yagi H, Tsujimoto T, Yamazaki T, Yoshida M, Akutsu H (2004) Conformational change of H^{+}-ATPase β monomer revealed on segmental isotope labeling NMR spectroscopy. J. Am. Chem. Soc. 126:16632-16638

Yee A, Chang XQ, Pineda-Lucena A, Wu B, Semesi A, Le B, Ramelot T, Lee GM, Bhattacharyya S, Gutierrez P, Denisov A, Lee CH, Cort JR, Kozlov G, Liao J, Finak G, Chen L, Wishart D, Lee W, McIntosh LP, Gehring K, Kennedy MA, Edwards AM,

Arrowsmith CH (2002) An NMR approach to structural proteomics. Proc. Natl Acad. Sci. USA 99:1825-1830

Yee A, Pardee K, Christendat D, Savchenko A, Edwards AM, Arrowsmith CH (2003) Structural proteomics: Toward high-throughput structural biology as a tool in functional genomics. Accounts Chem. Res. 36:183-189

Yee A, Gutmanas A, Arrowsmith CH (2006) Solution NMR in structural genomics. Curr. Opin. Struct. Biol. 16:611-617

Yokoyama S, Hirota H, Kigawa T, Yabuki T, Shlrouzu M, Terada T, Ito Y, Matsuo Y, Kuroda Y, Nishimura Y, Kyogoku Y, Miki K, Masui R, Kuramitsu S (2000) Structural genomics projects in Japan. Nature Struct. Biol. 7:943-945

Yokoyama S, Terwilliger TC, Kuramitsu S, Moras D, Sussman JL, Comm IE (2007) RIKEN aids international structural genomics efforts. Nature 445:21-21

Zimmerman DE, Kulikowski CA, Huang YP, Feng WQ, Tashiro M, Shimotakahara S, Chien CY, Powers R, Montelione GT (1997) Automated analysis of protein NMR assignments using methods from artificial intelligence. J. Mol. Biol. 269:592-610

Table 1 Citations of software in PDB files submitted September 2005 – September 2008

Program	Function	# PDB entries citing	Year of introduction
NMRPipe	Processing, display and peak picking	1340	1995
CYANA	Structure calculation	1160	2003
XWinNMR/Topspin	Bruker programs for acquisition and processing	1043	1997
NMRView	Viewing spectra; peak picking; analysis	910	1994
KUJIRA	Semi-automated processing and structure calc	736	2007
Sparky	Assignment, integration	365	1999
VNMR	Varian programs for acquisition and processing	317	1989
CNS	Structure calculation	242	1998
XPLOR-NIH	Structure calculation	153	2003
XEASY	Semi-automated analysis and assignment	130	1995
ARIA	NOE assignment and structure calculation	122	1995
DYANA	Structure calculation	114	1997

Autostructure	Structure calculation	103	2003
Autoassign	Assignment	82	2001
XPLOR	Structure calculation	75	1992
CcpNmr Analysis	Viewing, analysis, assignment	18	2004
Aurelia/Auremol	Semi-automated processing and structure calc	17	2004
ABACUS/CLOUDS	Structure calculation without assignments	4	2002
FLYA	Fully automated structure calculation	3	2006

A much more comprehensive discussion of programs can be found in Gronwald and Kalbitzer (2004). This is a selective list and programs listed here are not necessarily the most cited. References to software that are not given in the text: NMRPipe (Delaglio et al. 1995); Sparky T. D. Goddard and D. G. Kneller, SPARKY 3,University of California, San Francisco, http://www.cgl.ucsf.edu/home/sparky/; CNS (Brunger et al. 1998); XPLOR-NIH (Schwieters et al. 2003); XEASY (Bartels et al. 1995); XPLOR A.T. Brünger, X-PLOR Version 3.1, Yale University Press, NewHaven/London, 1992.

Figure legends

Fig 1 RMS *Z*-scores for parameters describing the local geometry of protein structures refined using a selection of different programs, together with structures re-refined in explicit solvent taken from the DRESS database (Nabuurs et al. 2004). The Figure also shows *Z*-scores for a random selection of high-resolution (< 1 Å) X-ray structures, as a comparison to what the distribution for 'high-quality' structures might be expected to look like. The RMS *Z*-score should be around 1: values lower than 1 indicate a narrower range of values than expected (typically indicating that restraints are too tight), while values larger than 1 indicate a wider range than expected. The local parameters measured are bond angles, bond lengths, improper dihedral angles, ω angles and sidechain planarity. The bars indicate the range of the data (with outliers drawn as circles), and boxes indicate the quartiles and the median. Redrawn using data given in Spronk et al. (2004)

Fig 2 Size distribution of protein structures determined by structural genomics centers and in the entire protein data bank. The top two panels show data for all structures determined by SG centers in TargetDB at 16 March 2004 (top) and 14 September 2008 (middle), taken from the TargetDB database http://targetdb.pdb.org/ (Chen et al. 2004). The bottom panel shows data for the entire PDB from January 2004 to November 2008, and is thus roughly comparable in time to the middle panel. Note the small but significant number of structures with > 200 residues in the bottom panel. There was no attempt to remove duplicates and close duplicates. Data on proteins of less than 50 residues in length are not shown. Crystal structures are shown in black and NMR structures in red.

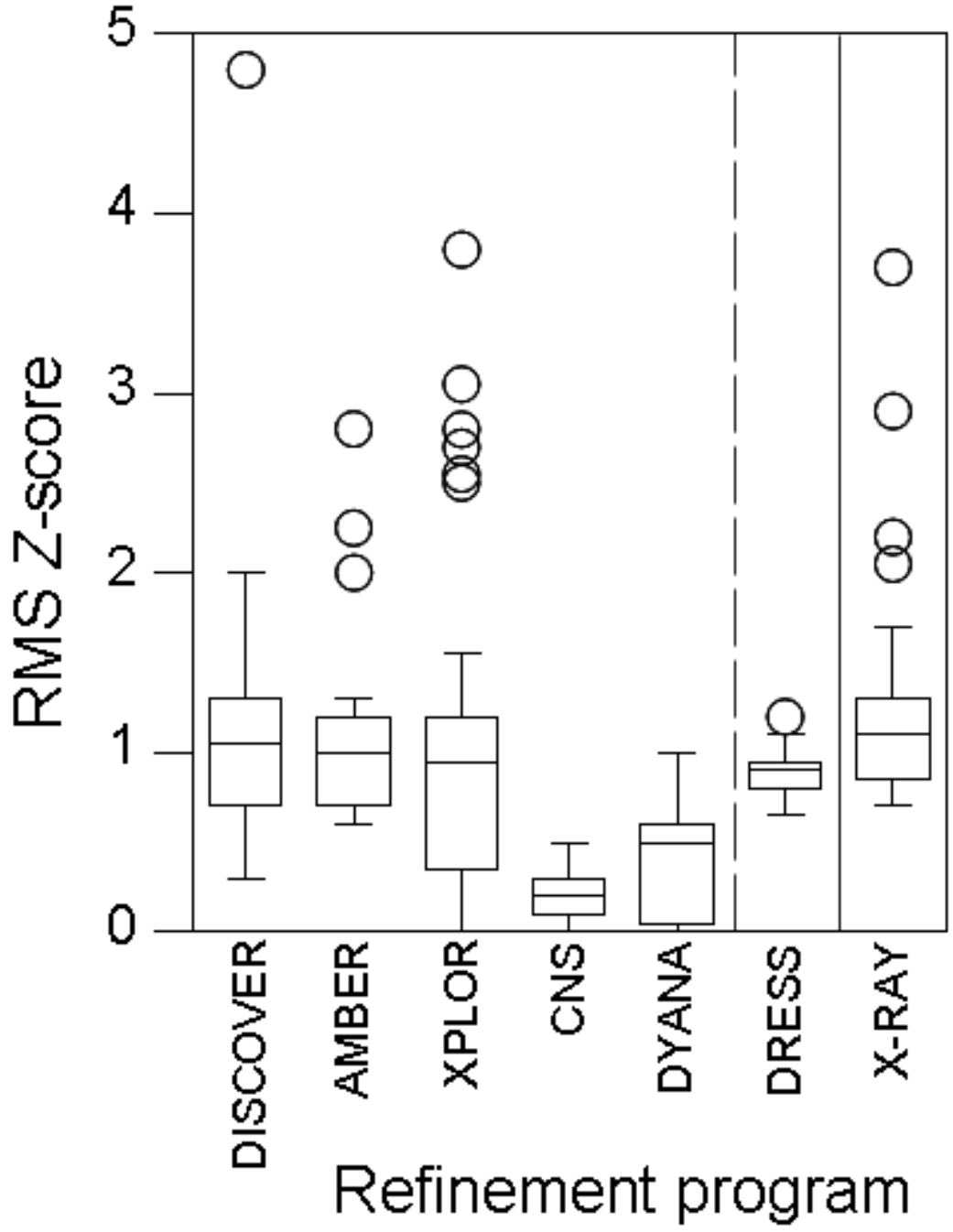
RMS Z-score
5
4
3
2
1
0
DISCOVER
AMBER
XPLOR
CNS
DYANA
DRESS
X-RAY
Refinement program

Figure 1

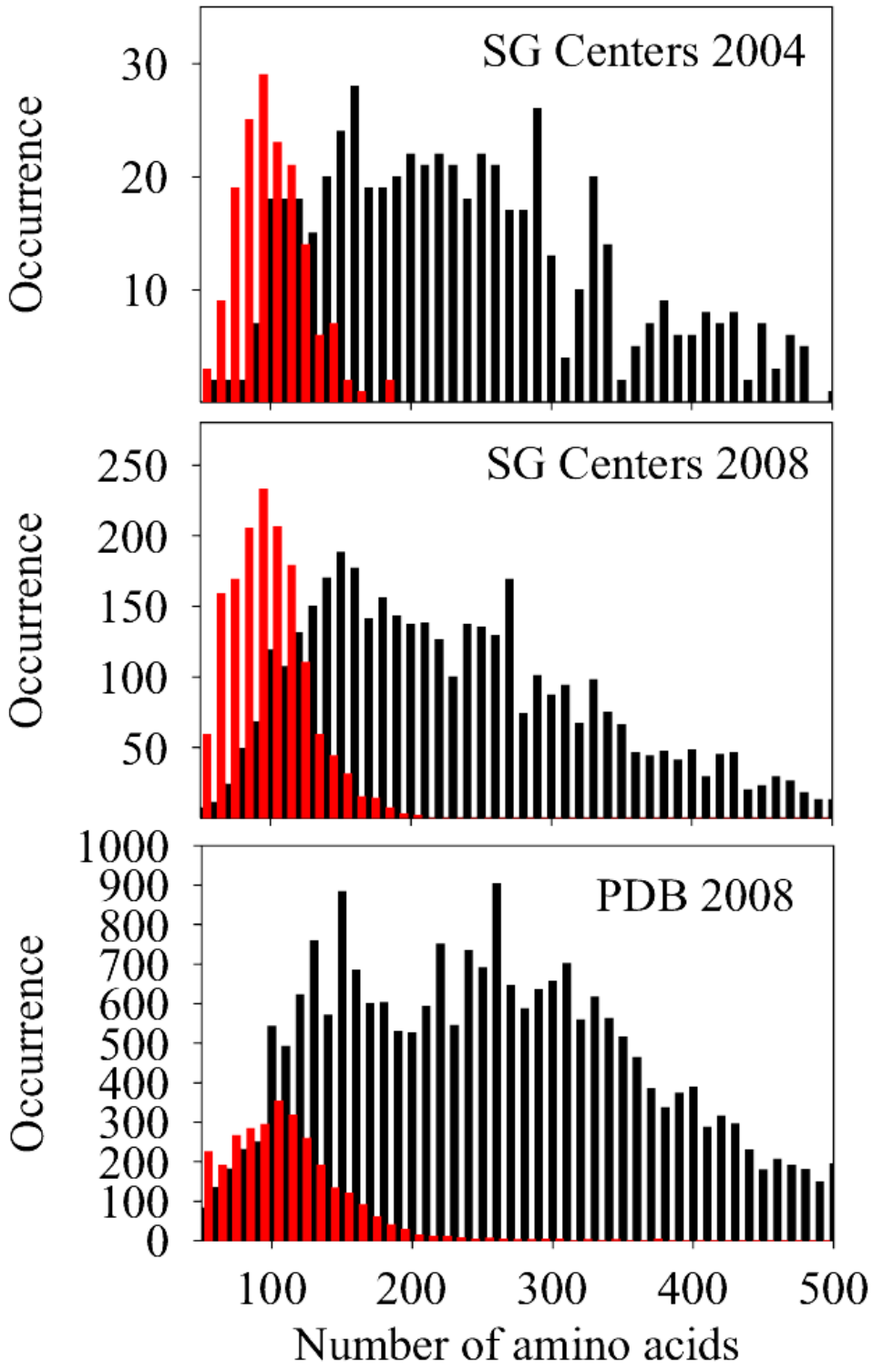
SG Centers 2004
Occurrence
30
20
10
SG Centers 2008
Occurrence
250
200
150
100
50
PDB 2008
Occurrence
1000
900
800
700
600
500
400
300
200
100
0
100
200
300
400
500
Number of amino acids

Figure 2