Quarterly Reviews of Biophysics **44**, 3 (2011), pp. 257–309. © Cambridge University Press 2011
doi:10.1017/S0033583510000326 Printed in the United States of America

Advances in automated NMR protein structure determination

Paul Guerry and Torsten Herrmann*

Centre Européen de RMN à très Hauts Champs, Université de Lyon/CNRS/ENS, Lyon/UCB Lyon 1, 5 rue de la Doua, 69100 Villeurbanne, France

Abstract. Around half of all protein structures solved nowadays using solution-state nuclear magnetic resonance (NMR) spectroscopy have been because of automated data analysis. The pervasiveness of computational approaches in general hides, however, a more nuanced view in which the full variety and richness of the field appears. This review is structured around a comparison of methods associated with three NMR observables: classical nuclear Overhauser effect (NOE) constraint gathering in contrast with more recent chemical shift and residual dipole coupling (RDC) based protocols. In each case, the emphasis is placed on the latest research, covering mainly the past 5 years. By describing both general concepts and representative programs, the objective is to map out a field in which – through the very profusion of approaches – it is all too easy to lose one's bearings.

1. Introduction 258

2. The classical approach 261
2.1. NOE assignment 262
2.1.1. Chemical shift ambiguity 262
2.1.2. Initial ranking of assignments 263
2.1.3. Network anchoring 263
2.1.4. Bayesian statistics 264
2.1.5. Ambiguous restraints 264
2.1.6. Using intermediate structures: violation analysis 267
2.1.7. Eliminating artefacts 268
2.1.8. Distance calibration 268
2.1.9. Recent developments 269
2.1.10. Structure calculation and energy refinement 269
2.2. Structure validation 270
2.3. Chemical shift assignment 272
2.3.1. Generalities 272
2.3.2. AutoAssign 273
2.3.3. MATCH 273
2.3.4. MARS 274
2.3.5. ABACUS 275

* Author for correspondence: Torsten Herrmann, Centre Européen de RMN à très Hauts Champs, Université de Lyon/CNRS/ENS, Lyon/UCB Lyon 1, 5 rue de la Doua, 69100 Villeurbanne, France.
Tel.: +33 (0)423 233882; Email: torsten.herrmann@ens-lyon.fr

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

2.4. Peak picking 276
2.4.1. AUTOPSY 276
2.4.2. ATNOS 276
2.4.3. Noise reduction 277
2.4.4. Using new experiments 278
2.4.5. Integration 278
2.5. Limitations 279

3. Alternatives 281
3.1. Chemical shift-based methods 281
3.1.1. Torsion angle constraints 281
3.1.2. From structures to chemical shifts 282
3.1.3. From chemical shifts to structures 283
3.1.3.1. CHESHIRE 283
3.1.3.2. CS-Rosetta 284
3.1.3.3. CS23D 284
3.1.3.4. Chemical shift restraints in MD simulations 284
3.2. RDC-based methods 285
3.2.1. Exploiting previously solved structures 286
3.2.1.1. Molecular fragment replacement (MFR) 286
3.2.1.2. RDC-Rosetta 287
3.2.2. Non-database methods 287
3.2.2.1. MECCANO 287
3.2.2.2. 3P 288
3.2.2.3. RDC-PANDA 288
3.2.3. Combined chemical shift assignment and structure determination 289
3.2.3.1. ITAS and FastNMR 289
3.3. RosettaNMR 289
3.3.1. CS-DP-Rosetta 290
3.3.2. CS-RDC-(NOE)-Rosetta 290
3.4. Structure calculation with unassigned chemical shifts 291
3.4.1. CLOUDS 291

4. Outlook 292
4.1. Towards more realistic structure models 292
4.2. New frontiers for automated protein structure determination 293

5. References 298

1. Introduction

Structural biologists should fear for their livelihoods. As soon as the monkeys finish re-writing Hamlet, their typewriters will be upgraded to computers running Foldit (Cooper *et al.* 2010) and a whole swathe of the scientific community will be made redundant. Come to think of it though, if the author's efforts (top score 8636, roughly equivalent to a Tetris score of 40) are anything to go by, that is not going to happen for a while … Jokes aside though, the fact that cutting-edge research has been translated into a video game (and generates interest!) stresses the aesthetic and intellectual beauty of protein structures. Another initiative from members of the Foldit group has been to run their program for structure prediction, Rosetta (Das & Baker, 2008), on BOINC

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

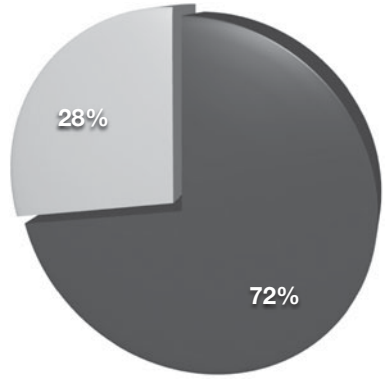


Fig. 1. Pie-chart representation of the percentage of structures in the PDB solved using solution-state NMR with entirely manual or interactive methods (black) or with some automation (grey), as of 27 August 2010. 'Some automation', entails that one or more program(s) are cited that perform unsupervised NMR data analysis. Exact search strings can be found in the Appendix, Table A4.

(http://boinc.bakerlab.org/rosetta/, http://boinc.berkeley.edu/). This immediately brings SETI (http://setiathome.berkeley.edu/) to mind, the platform of course, but mainly because it implies that proteins have the same power of popular fascination as extraterrestrial life.

Proteins govern all living systems. The structure, function, life and death of cells all rest upon structural, regulatory or defence proteins. Enzymes on the other hand catalyse the chemical reactions of an organism's metabolism. The evolution of different folds and protein families over millions of years underpins biology on a fundamental level. In terms of practical applications, the development of new drugs hinges on a better understanding of specific metabolic pathways and cellular processes encoded in protein structures.

Proteomics was revolutionised in the 1990s as mass spectrometry turned protein sequencing from a matter of days, to a matter of seconds (Wilm *et al.* 1996; Steen & Mann, 2004). This transformation has created (and continues to create) a host of new targets for structural investigation. The latest release of the RefSeq database for non-redundant protein sequences contained 10 640 515 entries (http://www.ncbi.nlm.nih.gov/refseq/, the latest release at the time of writing was dated 21 July 2010). In contrast, as of 14 September 2010, only 67 981 structures had been deposited in the PDB. The gap is not just huge but growing, as the number of known sequences doubles every 28 months (Levitt, 2009). While at present it is an unfeasible aim to determine all these structures experimentally, it maybe possible to map out fold space in a reasonable amount of time (Levitt, 2009).

This was the aim of a number of structural proteomics initiatives set up at the turn of the century (Yokoyama *et al.* 2000b; Terwilliger *et al.* 2009). The premise being that once a relatively small fraction of structures had been solved, any number of interesting targets would be accessible via homology modelling. The main techniques used were X-ray crystallography and NMR spectroscopy. The advantages of the NMR approach are well known: proteins do not need to be crystallised, experiments are performed in a (close to) physiological environment and are sensitive to protein dynamics. On the down side, however, NMR protein structure calculation is both slow and limited to relatively small targets (see Fig. 2). The protocol established in the 1980s (Wüthrich, 1986) has survived to this day with remarkably few changes. And at the time when structural genomics projects were being set up, while robust systems for automatic data analysis were emerging (Güntert, 1998; Moseley & Montelione, 1999), the main tasks were

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

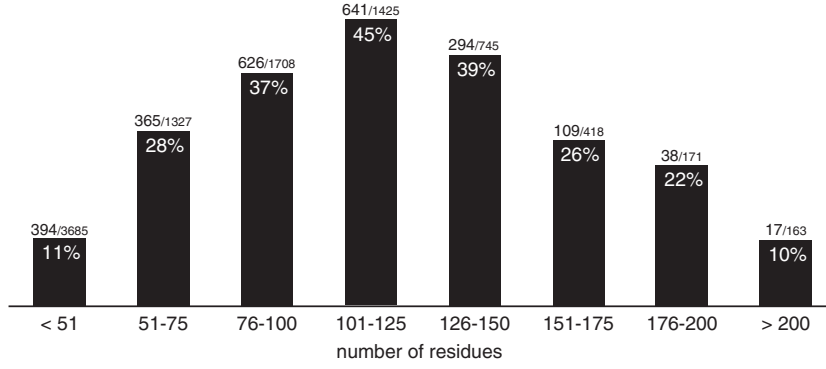


Fig. 2. Histograms representing, per size category (measured as the number of amino-acid residues), the percentage of solution-state NMR deposition in the PDB for which some automation was used. 'Some automation', entails that one or more program(s) are cited that perform unsupervised NMR data analysis. Exact search strings can be found in the Appendix, Table A5.

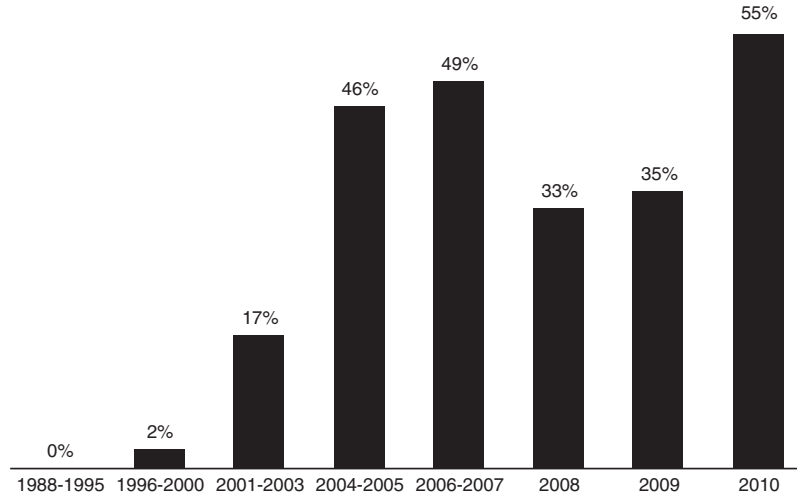


Fig. 3. Histograms representing, per period, the percentage of solution-state NMR deposition in the PDB for which some automation was used. 'Some automation', entails that one or more program(s) are cited that perform unsupervised NMR data analysis. Exact search strings can be found in the Appendix, Table A5.

overwhelmingly performed manually. For a couple of centres, some of the funding was allocated specifically to NMR software development (Williamson & Craven, 2009). Figure 3 highlights the influence of structural proteomics, in general, on the field. (One notes in particular the plunge that corresponds to the end of the RIKEN project.) We shall see that many of the programs discussed in this review are affiliated in some way with a structural genomics centre.

Over the past decade, technical advances in NMR hardware and experiments have also been key in the drive towards high throughput. The advent of cryo-probes in particular has lead to stunning gains in sensitivity. NMR has also benefited from ever-increasing magnetic field strengths and instrument stability. Spreading atomic connectivity pathways over several dimensions is essential in order to reduce spectral crowding and assignment ambiguity. However, such methods rely on several stages of magnetisation transfer, each reducing the sensitivity and

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

correspondingly increasing the experimental time required to achieve an adequate signal-to-noise level. Put together, the aforesaid improvements mean for instance that multidimensional (3D +) spectra – the basis of most NMR protein structure calculations – are now routine to acquire. In parallel, a number of alternative schemes for data acquisition and processing have been developed to access even higher-dimensional correlations (Atreya & Szyperski, 2005; Malmodin & Billeter, 2005; Gryk *et al.* 2010).

It is clear that structural genomics projects have provided the main impetus for automatic NMR protein structure determination. However, interest in the matter does not end there. Crucial also are the quality and reproducibility of results. An advantage of computational approaches that is therefore just as important as their speed is their objectivity. Moreover, the availability of structure-calculation software opens the field to non-experts who may be interested in other aspects of protein science.

To date, three main approaches to automated protein structure determination have been proposed, each one based on a different NMR parameter. This review is structured likewise; we contrast the classical pathway based on the gathering of nuclear Overhauser effect (NOE) inter-proton distance constraints with more recent residual dipole coupling (RDC)- and chemical shift-based methods. The number of competing software packages is extremely disconcerting. We hope to provide a more organised view of 'software space' as it were, by opting for a component-by-component description of the various programs, aiming to shed light on similarities, differences, advantages and/or disadvantages where they exist.

An alternative take on the subject may be obtained from several sources, from which the evolution of the field over the past few years may be deduced (Güntert, 1998, 2003; Moseley & Montelione, 1999; Gronwald & Kalbitzer, 2004; Billeter *et al.* 2008; Markwick *et al.* 2008). The literature on structural proteomics initiatives is vast but instructive in terms of the ambitions and limits of NMR structural biology. We found the following papers to be particularly informative (Yokoyama *et al.* 2000a, 2007; Yee *et al.* 2002, 2006; Cyranoski, 2006; Terwilliger *et al.* 2009; Ledford, 2010). A number of reviews have newly been published, which treat the RDC (Blackledge, 2005; Donald & Martin, 2009) and chemical shift (Wishart, 2010) sections in more detail, NMR of proteins more generally speaking (Banci *et al.* 2010; Bermejo & Llinás, 2010). Finally, we only mention sample preparation techniques and new NMR experiments in the context of the automated protocol into which they have been incorporated, the reader interested in a more complete description is referred to Sundström *et al.* (2006), Yee *et al.* (2006) and Shin & Lee (2008) for sample preparation and Atreya & Szyperski (2005), Malmodin & Billeter (2005) and Gryk *et al.* (2010) for new experiments.

2. The classical approach

The standard approach for solving protein structures by NMR involves two stages of assignment. In the first, the chemical shifts of the NMR-active nuclei in the protein are identified according to their position in the amino-acid sequence, re-expressing the protein's primary structure in terms of an NMR-measurable parameter. In the second, a self-constrained picture of the protein fold is modelled through the evaluation of thousands of inter-proton contacts and a biophysical force field.

Magnetisation transfer through NOEs registers as signals in NOESY spectra. For two isolated protons, the NOESY peak volume is proportional to the inverse sixth power of the inter-nuclear

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

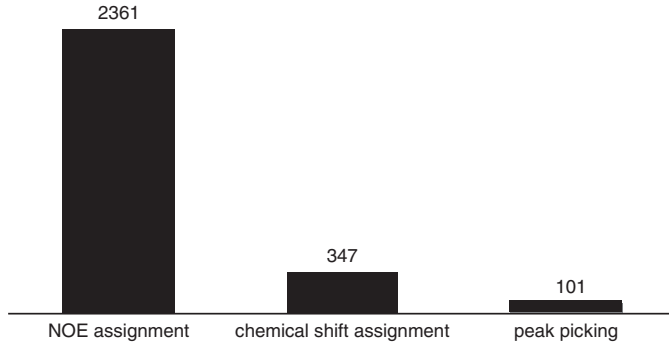


Fig. 4. The use of automation – in terms of PDB depositions – for the different stages of the traditional protocol for NMR protein structure determination. The histograms represent the number of structures returned when searching the PDB for one of the programs published for the respective stages. Exact search strings can be found in the Appendix (Tables A1, A2 and A3).

separation. Under the isolated two-spin approximation (Cavanagh, 2007):

$$V = C \times d^{-6}, \tag{1}$$

where C is a constant of proportionality linking the peak volume, V, to the inter-nuclear distance, d. Because of the effect of neighbouring spins, a somewhat looser relationship is assumed for proteins whereby the volume imposes an upper-distance limit on the inter-nuclear separation. Multi-spin effects, generally ignored, tend indeed to reduce the magnetisation transfer. For the same reason, the absence of an NOE signal is generally not taken as a lower-distance limit. Such conservative interpretation of the spectra relies on a dense network of upper-distance limits to reproduce the natural fold.

Because many of the tasks involved with this data analysis are repetitive, and involve a prodigious amount of book keeping, automation has always been a key focus. We shall cover the stages in reverse as it were, starting with NOE assignment before discussing algorithms for chemical shift assignment and peak picking. As Fig. 4 shows, this order follows how successfully the respective computer programs have established themselves in the structural biology community.

2.1 NOE assignment

2.1.1 Chemical shift ambiguity

NOE-assignment programs start from a list of peaks (compiled from all the NMR spectra) and a list of chemical shifts, each of which corresponds to a particular atom (or group of atoms) in the protein. It is common for the chemical shifts of many nuclei to be degenerate within the limits of the NOESY linewidth and the measurement precision. Indeed, it can be shown, using a sphere model to represent a biomolecule (Güntert, 2003) that the number of peaks assignable to a unique pair of protons decreases exponentially with the size of the protein, while the average initial ambiguity for a 3D ^{13}C-resolved NOESY spectrum is typically of the order of 10–20.

An expert spectroscopist will use one's trained intuition to discard incorrect assignments and artefacts, and will exploit preliminary structures once they become available. In this context, computer programs must find ways of transalating human reasoning into systematic procedures.

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

Table. 1. *Principle components of NOE-assignment algorithms*

	NOE-assignment programs using[a]	
ambiguous restraints		ALL
Bayesian assignment		KNOWNOE
constraint combination		CANDID
lower-distance constraints		PASD
network anchoring		ALL except KNOWNOE
partial assignment		ALL
relaxation matrix		ARIA
upper-distance constraints		ALL
violation analysis		ALL except PASD

[a]Individual elements are pattern-coded according to their function; in vertical stripes, methods for converting assignments to distance constraints; in dots, methods for reducing the assignment ambiguity and in vertical stripes, methods for eliminating artefacts.

And while it is relatively trivial to cyclically incorporate knowledge from (correct) intermediate 3D structures, more intricate techniques must be employed to reduce the assignment ambiguity at the very start of the structure-determination process.

2.1.2 Initial ranking of assignments

The simplest way of ranking the assignment possibilities of an individual peak is to compare how close the chemical shift coordinates match those of the peak. However, it is easy to overinterpret this ranking, since unaccounted-for effects from misaligned spectra, slight referencing offsets and variations in experimental conditions can lead to significant changes (actual or reported) in the resonance frequencies.

2.1.3 Network anchoring

Network anchoring (Herrmann *et al.* 2002b) measures how well a particular assignment is supported by the list of all other potential assignments compiled across all available spectra. Correct assignments form a self-consistent set and will score highly, whereas wrong assignments act incoherently in the network and do not support each other.

In CANDID (Herrmann *et al.* 2002b), the network anchoring score is calculated by finding all the atoms connected to both interacting protons either through another candidate assignment or via the covalent polypeptide structure. The procedure is cyclic in that individual scores depend

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

on all the others; by that means, connections through highly supported assignments are more influential.

A somewhat more formal expression of network anchoring is implemented in the BACUS algorithm, in which Bayes' theorem is used to convert the network support into assignment probabilities (Grishaev & Llinás, 2004; Lemak *et al.* 2008). The BACUS approach is original in that network anchoring is used without chemical shift assignment. Rather, the covalent information necessary to understand the mutual network support is extracted directly from additional assignment-type experiments (TOCSY, COSY, etc.) and database statistics.

The value of network anchoring has been demonstrated qualitatively – for ARIA (Bardiaux *et al.* 2009) and CANDID (Herrmann *et al.* 2002b) – and quantitatively, whereby for two test cases BACUS was able to reduce the initial chemical shift ambiguity by at least a factor of six.

2.1.4 Bayesian statistics

As an alternative to network anchoring, a Bayesian approach has been proposed to select the most probable assignment using information from a database of know structures (Gronwald *et al.* 2002; Hung & Samudrala, 2006).

In the context of NOE assignment, Bayes' theorem gives the probability of each putative assignment as a function of its prior probability and corrected for the overall reliability of the peak. More precisely,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{2}$$

where $P(A|B)$ is the probability of the candidate contact given the observation of the current peak and conversely for $P(B|A)$, whereas $P(B)$ is the prior probability of the peak. $P(A)$ finally is the prior probability of the contact, calculated as follows.

Different classes are constructed depending on the amino-acid type, the residue separation and the secondary structure. The latter only in the ProtInfo approach (Hung & Samudrala, 2006), where the prior probability for each class is then the proportion of such contacts in the database liable to interact via an NOE. KNOWNOE (Gronwald *et al.* 2002) on the other hand measures the distribution of distances from the database for each contact type and converts these to volume distributions. These are then used for each putative assignment in an attempt to identify a unique solution: one that has a high-enough probability to account for peak intensity in its entirety. While it is clear that such approaches are effective at ranking short-range above medium- and long-range assignments, the primary objective of ambiguity reduction is to select amongst several long-distance assignments. In this respect, the prior probabilities are not sufficiently discriminatory and no clear criterion for accepting or rejecting assignments emerges.

2.1.5 Ambiguous restraints

A key feature of an NOE-assignment protocol is its ability to deal with peaks with several assignment possibilities. Indeed, it is often the case – especially for larger proteins – that even the best filtering is unable to completely resolve all the ambiguities. The application of early algorithms (Güntert *et al.* 1991, 1993; Mumenthaler & Braun, 1995; Mumenthaler *et al.* 1997; Duggan *et al.* 2001) was thereby limited to cases where the number of unambiguous signals was sufficient to obtain an initial fold.

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

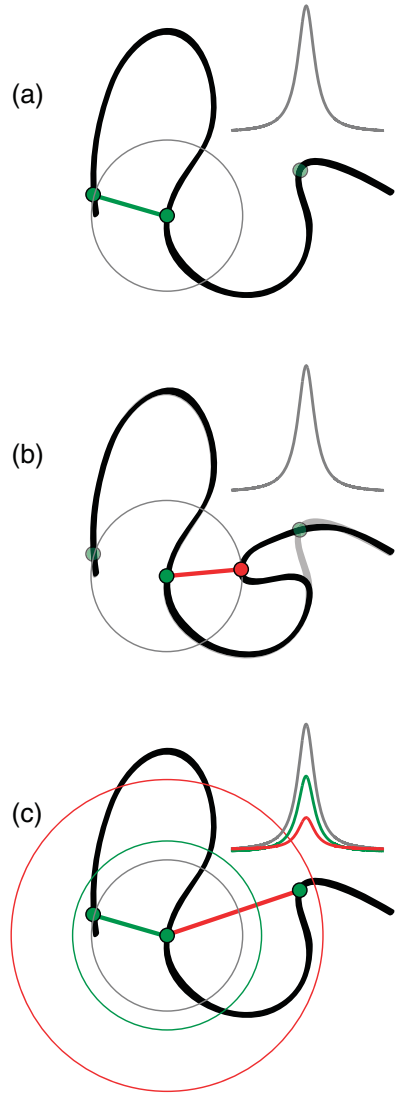


Fig. 5. Illustration of the principle of ambiguous restraints for the case of two-fold assignment ambiguity. (*a*) The peak is unambiguously assigned to the correct candidate constraint. (*b*) The peak is unambiguously assigned to the wrong restraint, distorting the structure. (*c*) The peak volume is assumed to originate from both candidate contacts. An ambiguous restraint is applied such that the summed contribution from both assignments equals the experimental volume. Since constraints are applied as upper-distance limits, the correct structure is always fulfilled as long as the correct assignment is included.

To overcome this obstacle, Michael Nilges (Nilges, 1993, 1995) proposed treating ambiguous peaks as an overlap of signals from all the n possible assignments (which, in fact, is almost never the case, but we shall see that this does not matter) and to set the constraint on the summed volume rather than on one of the individual components.

$$V_A = \sum_{k=1}^{n} V_k \tag{3}$$

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

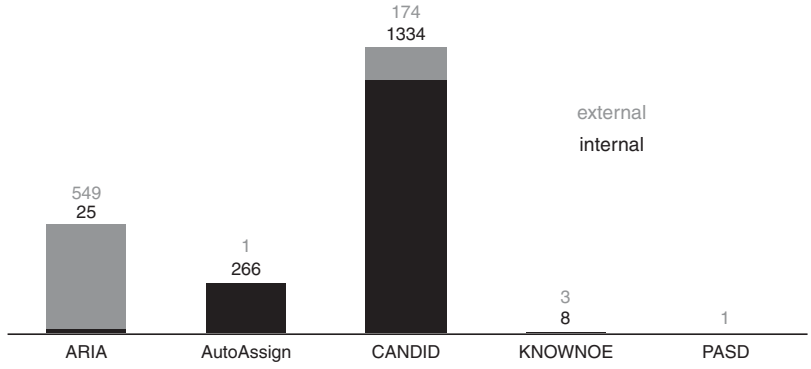


Fig. 6. Histogram plot of the number of citations from PDB depositions for NOE assignment programs as of 27 August 2010. Citations from internal depositions are represented in black whereas those from external use of the program are shown in grey. In this context, we define internal depositions as those for which one or more of the developers appear as a structure author, or which originate from the same structural genomics project in which the program was developed. Exact search strings can be found in the Appendix, Table A2.

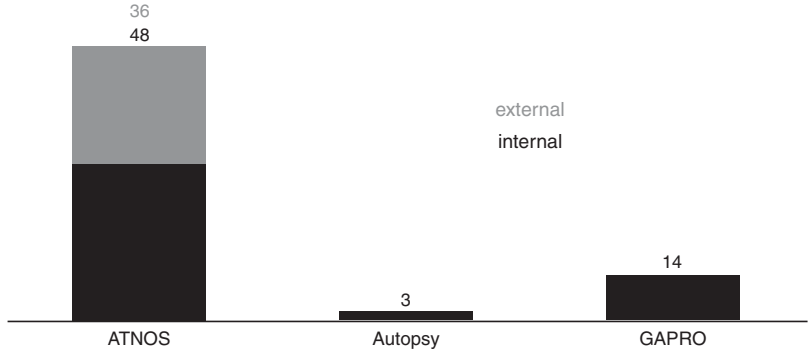


Fig. 7. Histograms representing the number of citations from PDB depositions for peak picking algorithms as of 27 August 2010. Citations from internal depositions are represented in black, whereas those from external use of the program are shown in grey. In this context, we define internal depositions as those for which one or more of the developers appear as a structure author, or which originate from the same structural genomics project in which the program was developed. Exact search strings can be found in the Appendix, Table A3.

where V_A is the total (ambiguous) peak volume and V_k is the volume associated with a particular assignment. In terms of distances, this becomes

$$d_A = \left(\sum_{k=1}^{n} d_k^{-6} \right)^{-\frac{1}{6}} \tag{4}$$

As illustrated in Fig. 5, because the intensity of the signal is inversely related to separation, ambiguous restraints are always laxer than those corresponding to the native structure. This means that the information is somewhat blurred out but more importantly, it guarantees that ambiguous constraints never distort a structure as long as the correct assignment is present. In fact, the concerted action of many ambiguous restraints during the molecular dynamics (MD)

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

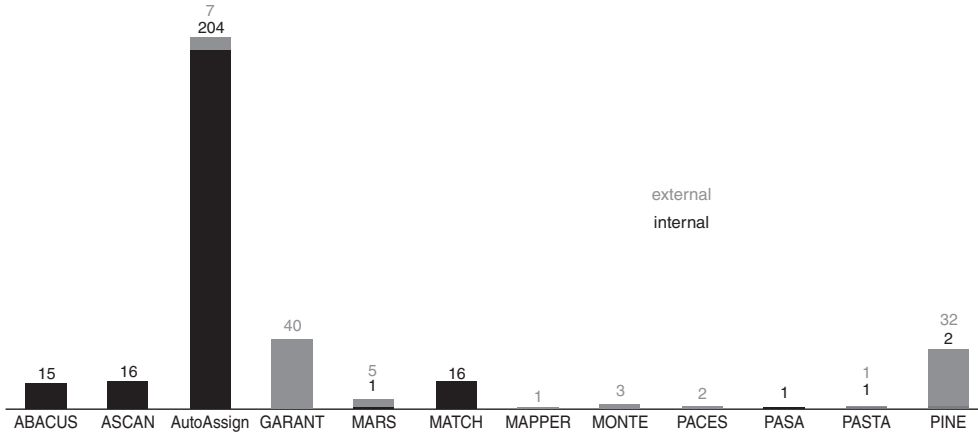


Fig. 8. Histogram plot of the number of citations from PDB depositions for selected chemical shift assignment programs as of 27 August 2010. Only those programs are listed for which at least one citation was found. Citations from internal depositions are represented in black, whereas those from external use of the program are shown in grey. In this context, we define internal depositions as those for which one or more of the developers appear as a structure author, or which originate from the same structural genomics project in which the program was developed. Note that (Bermejo & Llinás, 2010) count 31 depositions for ABACUS. Exact search strings can be found in the Appendix, Table A1.

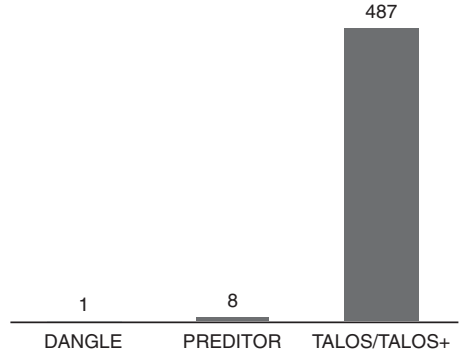


Fig. 9. Histogram plot of the number of citations – as of 7 September 2010 – from PDB depositions for torsion angle-prediction software. Exact search strings can be found in the Appendix, Table A6.

calculation forces each one to converge towards the correct unambiguous assignment. However, it is clear that this action becomes less and less effective as the average ambiguity increases. It is therefore vital that incorrect assignments be weeded out as much as possible prior to the constraints being set.

2.1.6 Using intermediate structures: violation analysis

Once a preliminary structural model has been built, this can be used to re-evaluate the experimental data and filter assignments based on their agreement with the intermediate fold. Not only does this significantly reduce the initial chemical shift-based assignment ambiguity but it also identifies artefacts as the peaks with no structure-supported assignment. The difficulty, however, lies in identifying true violations from those caused by insufficient convergence or distortions

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

caused by other artefacts. Violation analysis is therefore a double-edged sword; on the one hand, it ensures that the algorithms converge, but it makes it extremely hard to recover the native fold if this has not already been obtained after the first cycle. The algorithms are therefore heavily reliant on efficient, *ab initio* analysis of the experimental data.

As described above, ambiguous restraints are a means to let the MD algorithm decide which of the potential assignments should take precedent. Partial NOE assignment (Nilges & O'Donoghue, 1998; Linge *et al.* 2001; Kuszewski *et al.* 2004) aims at retrieving this information from the intermediate structures and may be used to supplement or replace (Kuszewski *et al.* 2004) violation analysis in reducing the assignment ambiguity.

For a given peak, the relative volume contributions from the list of candidate assignments are calculated according to the separation of the corresponding atoms in the latest structure. Usually, a cycle-dependent cut-off is imposed whereby initially, only very small contributions (long distances) are discarded with the selection criterion becoming more stringent after each cycle.

2.1.7 Eliminating artefacts

Experimental NMR spectra contain a substantial number of artefacts. As described above, intra-peak ambiguous restraints never distort a structure as long as the correct assignment is included. Artefacts clearly violate this rule and will lead to distortions. The risk being that these are carried through when intermediate structures are used to reassign the spectra. When working with unrefined peak lists, it is therefore critical to minimise the effect of artefacts even before the first structure is calculated.

Constraint combination (Herrmann *et al.* 2002b) involves the implementation of inter-peak ambiguous restraints, which ensures a significant reduction in the number of artefacts. In the first stage, long-distance peaks are ranked according to their network anchoring and split into four categories from the most reliable quarter to the least. For each set of four peaks, one from each category, the most reliable is paired once with each of the three others, while the two middle-ranking peaks are combined to form the fourth ambiguous constraint. Thereby, a combined constraint will only distort the structure if both peaks are artefacts. The extent to which artefacts are suppressed depends on the initial proportion thereof and how well the network anchoring score reflects the true error probability of the peaks. For instance, the analysis of Herrmann *et al.* (2002b) shows that for a set of 1000 peaks of which 100 are artefacts and for which the top-ranked quarter is, respectively, twice and three times more reliable than the middle two and bottom-ranked quarters, only seven artefacts remain after constraint combination.

2.1.8 Distance calibration

In NMR protein structure determination, the peak volumes (or intensities) measured experimentally are converted to distance constraints under the two-spin approximation. As mentioned above, this simplification is rather drastic for proteins. In addition, it is actually quite difficult to accurately measure a peak volume in the presence of many overlapping lines. However, several factors contribute to the robustness of this approach. To begin with, because of the inverse power of six relationships, the NOE method is not sensitive to errors in the volume assessment. Furthermore, the distance constraints need not be precise as long as there are many and as long as conservative estimates (upper-distance limits) are taken. The remaining task is then to determine C, the spectrum-dependent constant, for which several methods have been proposed.

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

CANDID applies a continuous calibration whereby initially, peaks assigned to covalent contacts (for which the polypeptide structure imposes a maximal spatial separation such that an NOE is always expected) are used to calculate the calibration constant. In later cycles, the calibration constant is set such that no more than a given percentage (15 then 10%) of constraints violate the intermediate structure. AutoStructure (Moseley & Montelione, 1999; Huang *et al.* 2006) on the other hand, classes peaks according to their relative intensities before deriving upper-distance limits for each class.

A more precise interpretation of the information contained in the peak volumes may be achieved by back-calculating the experimental spectrum using the intermediate structure. Bardiaux *et al.* (2009) have shown that the relaxation matrix (Ernst *et al.* 1987) approach employed in ARIA (Linge *et al.* 2004) effectively corrects for spin diffusion effects and results in tighter error bounds for the distance restraints. Understandably though, this more intricate approach is hungrier for computer resources.

2.1.9 Recent developments

The NOE-assignment programs mentioned above (see also Table 1 and Fig. 6), released roughly 10 years ago, are still in active development. A clear objective of the field in general is to make such tools more accessible to non-specialists. In this perspective, most programs now come complete with graphical user interfaces (GUIs). In fact, though the interpretation of the results is still the domain of experts, it reflects well on the maturity of the field that a virtual novice can set up and run a calculation in a matter of minutes.

To mention but a few examples, KUJIRA (Kobayashi *et al.* 2007) was developed as part of the RIKEN structural proteomics project (Yokoyama *et al.* 2000a) to provide a graphical front end for CANDID with CYANA (Güntert, 2004) and to facilitate data management and analysis in the context of high-throughput structure calculations. Similarly, CANDID as part of UNIO and also ARIA now provide a number of graphical reports at the end of a structure calculation to facilitate the evaluation of the result. In addition, ARIA automatically performs energy refinement and reports on the quality of the structure obtained using several validation tools (see Section 2.2), as does AutoStructure with the recall, precision and F-measure (RPF) score (Huang *et al.* 2005). Finally, the software package AUREMOL (Gronwald & Kalbitzer, 2004) associates KNOWNOE with graphical tools for NMR data processing and analysis.

2.1.10 Structure calculation and energy refinement

Being virtually unachievable manually or interactively, the building of atomic-resolution structure models was the first task to be successfully automated and indeed, this particular problem has been considered solved for the past 10 years or so. There is not much to add therefore to Peter Güntert's comprehensive review (Güntert, 1998) and the main programs (DYANA (Güntert *et al.* 1997), CNS (Brunger *et al.* 1998), X-PLOR (Brunger, 1992; Schwieters *et al.* 2003)) have not changed significantly since then. Briefly though, in CYANA for example, the structure is obtained starting from an initially random fold by restrained MD simulations using a hybrid energy function that incorporates physical interactions and experimental constraints. For reasons of efficiency, the physical component in the target function is over-simplified, and therefore the final round of structure calculation should be followed by refinement using a physically realistic potential in explicit solvent (Güntert, 1998; Spronk *et al.* 2004; Nabuurs *et al.* 2004a; Nederveen

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

et al. 2005). This is usually performed using force fields adapted from standard MD packages (Linge & Nilges, 1999; Koradi *et al.* 2000; Linge *et al.* 2003; Case *et al.* 2005; Christen *et al.* 2005; Phillips *et al.* 2005; Schwieters *et al.* 2006) but an alternative approach has recently been proposed, based on Rosetta structure prediction (Qian *et al.* 2007; Das & Baker, 2008; Ramelot *et al.* 2009). Protocols for energy refinement are constantly being updated and this improves structures past, present and future. A couple of recent studies have indeed demonstrated the value of amending prior depositions using the latest software (Nabuurs *et al.* 2004a; Nederveen *et al.* 2005): an excellent reason – if another were needed – to upload experimental restraints as well as the structure!

2.2 Structure validation

How confident can one be of the correctness of the structure obtained? There are two indirect criteria through which a structural biologist can address this most pertinent issue. First, does the structure truthfully represent the experimental data? Second, is the answer physically reasonable? Defining an NMR equivalent to the crystallographic R factor is problematic – low-quality data (e.g. few peaks) can easily lead to good scores, and therefore the latter question has made a more auspicious target (Spronk *et al.* 2004; Nabuurs *et al.* 2004b; Snyder *et al.* 2005).

As outlined in the previous section, a common source of problems is an incorrect balancing of experimental and electro-static potentials during structure calculation. Over fitting to experimental data can lead to unphysical structures, whereas under fitting leads to any number of unfolded random configurations. In general, force field tends to be biased towards experimental data notably through a simplification of the physico–chemical interactions. Over fitting is therefore much more common (Spronk *et al.* 2004; Nabuurs *et al.* 2004b), which explains the requirement and power of refinement in explicit solvent (Nabuurs *et al.* 2004a; Nederveen *et al.* 2005). On the other hand, problems stemming from incorrect interpretation of experimental data are much more insidious, as even the most drastic errors prove difficult to characterise unambiguously (Nabuurs *et al.* 2006; Saccenti & Rosato, 2008). Two recent studies are particularly informative on the ins and outs of validation in the context of automated protein structure determination.

In Nabuurs *et al.* (2006), an investigation of two abnormally differing depositions for homologous proteins leads to a search for similar anomalies in the PDB guided by commonly available validation software. (Note that structures deposited by structural genomics projects – where automation is prevalent – performed as well (or as badly) here as those from individual research groups: a cause for concern or celebration?) The authors use PROCHECK (Laskowski *et al.* 1993, 1996) and WHAT IF (Vriend, 1990; Hooft *et al.* 1996) to compare the local geometrical quality/normality of the homologues. Furthermore, the consistency of the experimental restraints is analysed by counting those that are violated while the QUEEN software (Nabuurs *et al.* 2003) measures the information content of individual constraints. No single parameter unambiguously identified the erroneous structures. Rather, evidence was accumulated based on a residue-by-residue analysis of several quality scores and a constraint-by-constraint dissection of the experimental data.

In contrast, Edoardo Saccenti and Antonio Rosato (Saccenti & Rosato, 2008) are interested in how well different programs for structure validation (FRST (Tosatto, 2005), HARMONY (Pugalenthi *et al.* 2006), HOPPscore (Sims & Kim, 2006), MolProbity (Lovell *et al.* 2003), PROCHECK, ProsaII (Sippl, 1993; Wiederstein & Sippl, 2007), TAP (Tosatto & Battistutta,

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

Table 2. *Software for the validation of protein structures solved using NMR data*

program	physico-chemical quality	experimental data
AQUA[a,b]	×	✓
FRST	✓	×
HARMONY	✓	×
HOPPscore	✓	×
MolProbity[c]	✓	×
PDB validation[c,d]	✓	×
PROCHECK[b,c]	✓	×
Prosall[c]	✓	×
QUEEN[b]	×	✓
RefDens[e]	✓	×
RFAC[f]	×	✓
RPF[c,g]	×	✓
TAP	✓	×
Verify3D[c]	✓	×
WHAT IF/WHAT CHECK[b]	✓	×

[a]Laskowski *et al.* (1996).
[b]Included in the CING (pronounced 'king') suite (http://nmr.cmbi.ru.nl/cing/Home.html).
[c]Included in the PSVS suite (Bhattacharya *et al.* 2007).
[d]Berman *et al.* (2000).
[e]Ginzinger *et al.* (2010).
[f]Gronwald *et al.* (2000).
[g]Huang *et al.* (2005).

2007), WHAT IF, all of which scrutinise physico-chemical quality) catch the increasingly serious mistakes (miss-calibrated and miss-assigned NOEs) they slip into the original set of experimental constraints. For a set of model proteins, the perturbed structures deviate by 2–3 Å from the reference without displaying obvious pathologies. In keeping with the results related above, the authors find that none of the tools reliably identify the distorted structures; the differences in particular scores between the erroneous and reference set are no larger than the variations within the latter.

It is well known that the tightness – or precision – of NMR structure bundles often overestimate their accuracy (Spronk *et al.* 2004). Nonetheless, this parameter is often assumed to be as meaningful as the X-ray resolution and is used as the primary gauge of structural quality. Saccenti and Rosato, however, find no correlation (positive or negative) between errors and precision, emphasising the perniciousness of this misconception.

The main conclusion to be drawn here from these studies is that as of yet, no simple protocol has been discovered for structure validation (see Table 2 for a list of available programs). On the contrary, a series of clues and pieces of circumstantial evidence must be gathered up painstakingly from all sources available. To this end, the software suites PSVS (Bhattacharya *et al.* 2007) and CING (http://nmr.cmbi.ru.nl/cing/Home.html) assemble several existing tools within a single framework. The layperson may, however, be repelled by the somewhat bewildering array of scores, tests, reports and checks. In light of this, some programs output a single quality score. However, such summarising is obviously not conducive to the detailed inspection advocated above.

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

Solving the same structure twice using different sets of experimental data appears to be the fail-safe option, be this through cross-validation or by referring to a homologous protein. It is indeed exceedingly unlikely that the same mistakes are made twice. Unfortunately though, even a homology model does not allay the need for manual interpretation (Nabuurs *et al.* 2006) and the experimental cost of complete cross-validation (RDCs are the preferred complement to NOEs) is often prohibitive.

2.3 Chemical shift assignment

All the algorithms for NOE assignment described above require close to 100% backbone and side-chain chemical shift assignment for optimal performance. The shear number of methods published for automated chemical shift assignment is testament to the interest that this topic continues to attract. Perhaps the main driving force has been that even after 15 years or so of research, chemical shift assignment is seldom performed automatically as the tasks of manual assignment have proved difficult to implement systematically. The absence of a consensus procedure (notably in terms of the optimal set of experiments) highlights the difficulties in obtaining complete and reliable coverage. While scouring the literature in the context of writing this review, we have counted not less than 44 (more or less) novel automated procedures for backbone and/or side-chain assignment, published within the last 15 years or so. (The interested reader will find a list of these – undoubtedly not comprehensive – in the Appendix, and will find alternative compilations in the following reviews: Moseley & Montelione, 1999; Baran *et al.* 2004; Gronwald & Kalbitzer, 2004; Bermejo & Llinás, 2010) In light of this, we will concentrate on certain representative algorithms to sum up the different approaches. Many ideas are indeed common to several programs, but expressed or implemented in a different way.

2.3.1 Generalities

Automated chemical shift assignment is usually performed in four stages with the aim first, to group resonances into spin systems; second, to determine the amino-acid type of each spin system; third, to create fragments from collections of spin systems that match sequentially; fourth, to associate the said fragments with a particular position in the amino-acid sequence. Usually at this stage, only the backbone and Cβ atoms are assigned (in the context of chemical shift assignment, the Cβ atoms are generally treated with the backbone atoms, rather than the side chain. In the following therefore, we use 'backbone assignment' to mean 'backbone and Cβ assignment') and indeed, the assignment of the side-chain atoms has proved even more problematic to complete automatically. A (smaller) number of algorithms have nonetheless been developed that include the side-chain atoms in the initial spin systems (stage one) (Bartels *et al.* 1996, 1997; Moseley & Montelione, 1999; Baran *et al.* 2004; Gronwald & Kalbitzer, 2004). Alternatively, the two problems may be treated entirely separately with an algorithm dedicated to the side chains being applied (Fiorito *et al.* 2008) once the backbone assignment has been completed.

The defining feature of most programs is their approach to mapping fragments onto the primary sequence (stage four), which involves finding the optimal combination amongst myriad possibilities. Most programs can thereby be classified accordingly as exhaustive search, best-first, Monte-Carlo or genetic algorithms.

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

2.3.2 AutoAssign

The best-first approach employed in AutoAssign (Zimmerman *et al.* 1997; Moseley *et al.* 2001) is perhaps the simplest to understand conceptually, if only because its description is jargon-free and does not require the assignment problem to be recast in terms of a more general theory. AutoAssign performs backbone assignment starting from a set of peak lists (five to eight) from traditional assignment experiments and information on amino-acid-specific chemical shift distributions extracted from the Biological Magnetic Resonance Bank (BMRB) (Doreleijers *et al.* 2005). Following the approach outlined above, resonances from different spectra are aligned and grouped into dipeptide spin systems (stage one). At this stage, three classes are defined – distinct, overlapped and incomplete – for the spin systems depending on the ambiguity and/or completeness in terms of constituent resonances. The three subsequent steps are applied initially to the distinct class before exploiting the results to interpret and assign the overlapped and then incomplete spin systems. First, the amino-acid-type probabilities are determined for each dipeptide group (stage two) by comparing experimental chemical shifts with BMRB statistics. Then, spin systems are linked using a best-first approach. For each dipeptide, candidate neighbours are ranked according to the number of common resonances that match, and then according to the closeness of the match; the highest-ranking candidate is chosen and removed from subsequent consideration. Finally, the fragments of varying length are mapped to a unique position on the amino-acid sequence.

2.3.3 MATCH

In MATCH, the problem of sequence-specific backbone assignment is expressed in terms of the evolution of a set of individuals. The initial population is a set of sequence-specific resonance assignments (the individuals), each one built from fragments assembled from one or several APSY (Hiller *et al.* 2005) spectra and arranged in a random sequence. In MATCH, a spin system consists of up to 18 frequencies (6 intra- and 12 inter-residue) linking two or three residues. The consolidation of resonances into spin systems is facilitated greatly by the numerous correlations already apparent in the high-dimensional APSY peak lists.

Starting with the longest fragments but otherwise randomly, these are removed from the buffer to the best-matching available position on the current individual. This procedure is repeated until a heterogeneous population is obtained, the size of which is determined automatically according to the ambiguity of the input data.

The evolutionary cycle proceeds through the independent mutation of each individual. Randomly, a fragment is chosen along with a new position where a swap is performed if at least one extremity of the candidate fragment fits with its neighbour. Furthermore, an assignment is made if the sequence-specific score – the chemical shift matching score between the experimental and predicted values, evaluated for the fragment at its new location, is above a given threshold. Conversely, if no compatible location is found after a given number of attempts, the process is re-initialised for a different candidate. However, those fragments that have already been assigned are locked in their positions.

This local optimisation is followed by an evaluation of assignments across the whole population. Temporary assignments are associated with a particular individual, whereas a permanent assignment is defined if a given temporary assignment is found in more than a predetermined fraction of the population.

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

Each cycle is punctuated by selective breeding whereby the genetic information from the fittest individuals forms a gene pool from which descendants are constructed. Specifically: in the first stage, the parent individuals are ranked according to their sequence-specific scores. This evaluates how well each fragment fits at its current position and the overall quality of the assignment. A buffer is then created using fragments from the highest-ranked individuals. As for the initial genesis, descendant individuals are constructed fragment by fragment from the gene pool, only now, the position of the fragments is inherited from the previous cycle.

The optimisation ends once all the individuals in the population are identical. The resulting assignment is then stored in a buffer of elite individuals and the whole procedure is repeated a predetermined number of times. An assignment is considered correct if it appears in more than 50% of the final elite population.

2.3.4 MARS

The emphasis so far has been placed on chemical shift assignment for *de-novo* NMR protein structure determination. In many cases, however, for instance when protein dynamics are of interest, a structural model already exists. The MARS procedure is original in that it can be augmented with a structural model and RDCs (Jung & Zweckstetter, 2004a, b). Starting with the more conventional 'chemical shift-only' approach – which also forms the core of the 'structure & RDC' method – MARS relies on chemical shift and secondary structure prediction along with cumulative exhaustive searches to arrive at the correct assignment.

More precisely, a pseudo-energy is calculated for assigning a spin system to a particular position, based on how well the experimental chemical shifts match those expected from the primary structure. In MARS, the relevance of the comparison is enhanced by predicting the secondary structure of the protein using the program PSIRED (McGuffin *et al.* 2000) and then using this information along with corrections for neighbouring residues to refine the chemical shift statistics from the BMRB. It should be noted that MARS requires spin systems rather than peak lists as input; i.e. stage one must be performed externally (manually or otherwise).

The MARS algorithm is constructed around an exhaustive search procedure initiated by allocating to all spin systems a random position in the amino-acid sequence. A spin-system is selected randomly and a forward search is performed along the sequence until all possible five fragments (fragments of five spin systems) have been constructed. The procedure for establishing sequential connectivities is tolerant for missing or imprecise data in that a putative link is only rejected if both common resonances are present and are in disagreement above a certain threshold. In contrast to best-first approaches (e.g. AutoAssign described above), candidate fragments are not ranked according to the quality of sequential match and all possibilities are considered at each stage of the exhaustive search. Each fragment is then placed at every possible position in the amino-acid list, the pseudo energy is calculated and a ranking is established. The lowest-energy five-fragment with its associated position identifies the best position for the optimal five-fragment starting from the initially selected spin system. Now, starting from the last residue of the highest-ranked fragment, the procedure is repeated in reverse. If again, the same five-fragment comes out on top associated with the same position, the initial spin system is relocated. This back and forth search is repeated until all spin systems have been used once as the root residue before re-initialising the cycle with four, three and two fragments. However, moving previously assigned spin systems comes with an extra energy cost. This ensures that correctly assigned residues are not repeatedly (and incorrectly) shifted around.

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

At this stage, a comparison is made between the fragment-derived assignments and the lowest pseudo-energy position for each spin system, analysed individually. Assignments are only conserved if they are found repeatedly in multiple runs. Indeed, the MARS procedure involves performing both assignment schemes 100 times during which Gaussian noise is added to the predicted shifts according to an empirically devised strategy. Assignments that are conserved in the first 20 pairs are kept fixed for subsequent rounds, then similarly for consistent assignments in the next 30, then again for the next 20 and the final 30. At each stage, the variance of the noise is reduced such that the most reliable assignments constrain the search for more uncertain attributions.

When RDCs and a structural model are provided, MARS proceeds in the same way except for an extra term in the pseudo energy that evaluates the match between RDCs measured experimentally and those back calculated from the structure. Since RDCs are very sensitive to small structural changes – that can occur for example due to dynamics, the RDC matching weighs less than the chemical shift. In addition, the alignment tensor is initially calculated using one of four possible methods and then the estimate is refined gradually as more assignments are made using singular value decomposition (SVD).

2.3.5 ABACUS

ABACUS (Lemak *et al.* 2008) performs backbone and side-chain assignment using a traditional set of assignment experiments, minus those for establishing sequential connectivities. Indeed, in ABACUS, this information is extracted directly from the NOE spectra used subsequently for the structure calculation.

To begin with, NOEs are assigned to specific fragments using BACUS (Grishaev & Llinás, 2004; Lemak *et al.* 2008) and the relative proximities of the later are determined from the resulting distance information according to the CLOUDS procedure (Grishaev & Llinás, 2002a).

ABACUS employs a multi-canonical MC algorithm (Berg & Neuhaus, 1991; Berg & Celik, 1992) to find the minimum energy in the assignment space. The pseudo energy of the system has two components, the first of which accounts for the matching between the chemical shifts of the current fragment and those for the amino acid type at the candidate position, while the second evaluates whether the current fragment is close enough to its putative neighbour to form a sequential link. Thereby, the closer the two fragments are in the clouds (constructed in the previous cycle), the greater the number of NOEs connecting the pair (assigned by BACUS) and the higher the chemical shift similarity at the candidate position, the lower the energy.

During each MC run, fragment swaps are performed randomly and accepted if they lower the energy, or if they do not, with a probability

$$P_i = e^{-S_i(\mathrm{E})} \tag{5}$$

where E is the total energy of the system. Following a multi-canonical approach, the weight function $S(E)$ is updated after each MC cycle such that

$$P_{i+1} = e^{-(S_i(E)+H_i(E))} \tag{6}$$

where $H_i(E)$ is the energy histogram from MC run i. In addition, an absolute energy limit is imposed, which is reduced after each cycle; the idea in both cases being to push the sampling of assignment space towards low-energy regions.

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

2.4 Peak picking

Peak lists are the most common starting point for automated chemical shift or NOE assignment algorithms. It may seem surprising that the procedure of scanning a spectrum for maxima is still nowadays overwhelmingly performed manually. However, it turns out that NMR-based methods have extremely stringent requirements in terms of the accuracy of the picked peaks. Furthermore, it is extremely difficult to systematically distinguish between weak signals (which are often the most important) and artefacts. Consequently, and maybe also because of the convenience and intuitiveness of graphical interactive programs (Johnson & Blevins, 1994; Bartels *et al.* 1995; Neidig *et al.* 1995; Goddard & Kneller, 2008), the peak-picking algorithms described in the following sections have not (yet?) become familiar tools for most spectroscopists. While an intimate coupling between peak picking and peak prediction has proved so far to be the most effective approach (Herrmann *et al.* 2002b), a promising recent development has been to incorporate peak picking in the design of new NMR experiments (Korzhneva *et al.* 2001; Orekhov *et al.* 2001; Hiller *et al.* 2005; Eghbalnia *et al.* 2005a).

2.4.1 AUTOPSY

AUTOPSY (Koradi *et al.* 1998) is a stand-alone method for automatic peak picking in the sense that it does not use any additional information to that which is contained in the spectrum. The main features of the program are an algorithm for local noise estimation, the segmentation of the spectrum into more-manageable chunks, the use of spectrum-specific peak templates and the exploitation of the intrinsic symmetry of magnetisation transfers.

Invariably, due to noise bands, water lines, artefacts, etc., the noise in NMR spectra is not evenly spread. The first stage of the AUTOPSY program involves calculating the average noise level for each 1D slice of the spectrum. The spectrum is then simplified by setting the intensity of pixels to zero if their value is below the local threshold. Next, the spectrum is split into connected regions using a flood-filling algorithm (Foley, 1996). This ensures that even the most bulky multidimensional NMR spectra are treated efficiently. A set of reference line shapes is created by comparing the selections of well-resolved peaks from different regions. These templates are then used to resolve regions of strong overlap. Thereby, a peak – or rather a peak component – is accepted as a signal if a corresponding model can be found within the reference set. In addition, the presence of a symmetric peak, mirrored with respect to the spectrum diagonal is taken as an indication of the validity of a particular candidate.

2.4.2 ATNOS

The algorithm ATNOS was developed initially for combined NOE signal identification and protein structure calculation (Herrmann *et al.* 2002b). It has subsequently been adapted for side-chain chemical shift assignment in combination with ASCAN (Fiorito *et al.* 2008) and is used in APSY for the identification of projection peaks (Hiller *et al.* 2005). ATNOS is designed to interface cyclically with assignment algorithms such that the selection criteria – and therefore the resulting peak lists – evolve as new assignments or more accurate structures become available. Updating the peak lists in this way ensures that the likelihood of an artefact being fed repeatedly to the assignment algorithm is much lower than if the peak picking were performed once only, as is common when signal identification is performed manually.

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

NOE assignment and structure calculation programs require essentially artefact-free peak lists. Indeed, it has been shown that when trading between accuracy and completeness, the former is much more valuable (Güntert, 2003; Jee & Güntert, 2003). To this end, ATNOS applies a series of increasingly stringent criteria for candidate resonances. Notably, and in contrast with other approaches, ATNOS makes extensive use of the information contained in the chemical shift assignments.

As a first step, all peaks whose intensities are no more than three times the AUTOPSY-calculated local noise level are removed. The covalent polypeptide structure imposes upper-distance limits on inter-nuclear distances. In ATNOS, these are calculated analytically for all proton pairs separated by fewer than two torsion angles. Correlations are expected for distances less than 5·5 Å. Candidate peaks whose coordinates match those for one of the predicted covalent signals form a reference set from which stricter area and intensity thresholds are defined.

The remaining peaks are divided into three classes: the reference or covalent set, peaks which are close to the diagonal and the rest. If no structural model is available, the 'diagonal' set is ignored. Peaks in the covalent set are accepted so long as they do not appear in a noise band, defined as an area in which the local noise level is more than three times the spectrum average. The remaining peaks must satisfy three additional criteria. First, the candidate peak must be well removed from others in the vicinity; if an area of noise does not separate it from its neighbours then the candidate peak must either be the most intense of the local group or be detached by a sufficiently deep intensity valley. Second, the candidate peak must have a symmetric double. Third, the candidate peak should be aligned with at least two peaks from the same plane (i.e. well anchored in the network of signals) and its coordinates should match at least one combination from the list of assigned atoms. If on the other hand, a preliminary structure is available, the grid of predicted signals is extended to all atoms within 5·5 Å and all peaks appearing close to an intersection are accepted. Those that do not satisfy this criterion are analysed as described above.

For ASCAN side-chain assignment, since the objective is to discover hitherto unknown assignments, rather than only accepting peaks at grid intersections where the *n*D coordinates are prescribed by the *n* interacting atoms, whole lines are considered as valid locations for peaks whereby the $n-1$ previously assigned chemical shifts provide the anchor points.

2.4.3 Noise reduction

Low signal-to-noise ratios are a clear obstacle to automated routines. A number of recently published methods aim at removing (some or most of) the noise in order to facilitate peak picking and render the process more robust.

The MICE (Serban, 2007) protocol involves de-noising spectra using wavelet reconstruction (Daubechies, 1992; Donoho & Johnstone, 1994, 1995; Mallat, 1999). Local maxima are then accepted above a certain threshold that is set relative to the average noise level in the spectrum. All line shapes are assumed to be Lorentzian and the average line width calculated from well-resolved peaks is then used to fit crowded regions.

Dancea & Gunther (2005) have investigated the effect of different de-noising algorithms on a 3D ^{15}N-NOESY spectrum. The optimal schemes were then applied for the automated structure calculation of the polysulfide-sulfur transferase (Sud) protein (a homodimer with 2×137 residues), by which automatically picked peaks were fed into ARIA. Perhaps unsurprisingly, the

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

more the noise level is reduced, the more spectral fine structure is lost. Consequently, two different de-noising schemes were employed during the structure calculation process. The most aggressive scheme was used to obtain a preliminary fold during five ARIA cycles and a spectrum that retained more weak features was analysed for four additional iterations. The final four cycles exploited a peak list extracted from the original spectrum. In contrast to the ATNOS protocol described above, each spectrum was analysed once only and no structure-based filters were used.

Within this framework, local maxima are accepted if they satisfy three criteria. First, they must be more intense than the local noise level (calculated as in AUTOPSY). Second, they must be supported by at least one different non-diagonal peak that is assignable to the same two residues (cf. network anchoring). And third, they must be supported by a symmetric double the other side of the diagonal.

An alternative approach has been proposed with PICKY, to remove noise using SVD (Alipanahi *et al.* 2009b). Simple selection rules are then applied whereby the remaining noise is assumed to be evenly spread across the spectrum and maxima are accepted based on a minimum signal-to-noise threshold. Any additional spectra available are analysed in parallel such that depending on the type of experiment, signals originating from common atoms support each other.

2.4.4 Using new experiments

With many types of recently developed experiments, such as those based on reduced dimensionality or projections from higher-order dimensions (Malmodin & Billeter, 2005; Gryk *et al.* 2010), signals appear in standard patterns that can be analysed systematically in order to obtain peak lists. MUNIN (Korzhneva *et al.* 2001; Orekhov *et al.* 2001) exploits the multi-dimensional decomposition (MDD) spectrum reconstruction procedure through which multi-dimensional (3D or higher) spectra are reconstructed from a set of model peaks. Peak positions are thereby obtained as a by-product, as it were, of the MDD process. APSY (Hiller *et al.* 2005) experiments share the similarity with HIFI (Eghbalnia *et al.* 2005a) that peak positions are extracted directly from projected spectra without reconstructing the multi-dimensional parent. In APSY, a simple peak picker is included to analyse 2D projections such that all maxima are accepted as long as they satisfy the local threshold criterion described above. Then, GAPRO repeatedly selects $n-1$ of the peak lists and back calculates peak positions in the n-dimensional spectrum. The resulting triangulation of individual peaks is highly redundant and therefore extremely precise. Similarly, although artefacts are picked in individual projections, these do not correlate during the GAPRO reconstruction, whereas signals do. Such approaches should prove particularly useful for providing peak lists for chemical shift assignments, for which high dimensionality is particularly valuable at tracing out sequential connectivities, as demonstrated for instance in the association of APSY with the MATCH algorithm for backbone assignment (Volk *et al.* 2008).

2.4.5 Integration

A number of proposals have also been made to interface peak picking of traditional experiments directly with chemical shift assignment algorithms. ATNOS/ASCAN for instance performs automated side-chain resonance assignment directly from NOESY or TOCSY spectra (Fiorito *et al.* 2008). For backbone assignment, AUTOPSY has been interfaced directly with GARANT (Bartels *et al.* 1997; Malmodin *et al.* 2003). Once the chemical shift assignment is complete,

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

ATNOS/CANDID (Herrmann *et al.* 2002a) is well established as a robust algorithm for constraint gathering directly from NOE spectra and a similar approach has been proposed by Dancea and Gunther with ARIA in the context of wavelet de-noised NOE spectra (Dancea & Gunther, 2005).

Integration of the stages outlined in this section is a key objective in the ongoing quest for higher throughput and objectivity. At this stage, it seems that the mashups described above strike the optimal balance between efficiency and accuracy, and between time spent on data acquisition and analysis (see below). Indeed, a major obstacle for fully integrative approaches – in this context, those that read spectra and output a structure – is the cumulative propagation of errors.

The principle of fully automated NMR protein structure determination was first proven with the FLYA algorithm (Lopez-Mendez & Güntert, 2006); for three 12–16 kDa proteins the FLYA approach performed to a similar standard as traditional step-by-step automation. FLYA interleaves a series of well-established programs: peaks are picked automatically using NMRView (Johnson & Blevins, 1994) or AUTOPSY; GARANT performs the chemical shift assignment while the overall process is controlled through the CYANA (Güntert, 2004) suite for NOE assignment and structure calculation; finally, energy minimisation is performed using OPALp (Koradi *et al.* 2000) with the AMBER (Case *et al.* 2005) force field.

Similarly, the FAPSY (Shin & Lee, 2008) approach associates AutoAssign (Zimmerman *et al.* 1997; Moseley *et al.* 2001) for chemical shift assignment with TALOS (Cornilescu *et al.* 1999) and CYANA for constraint gathering and structure calculation. Applied to two proteins of 90 and 111 residues, the FAPSY protocol calculated structures that deviated from the reference by 3·2 and 3·0 Å, respectively.

There is a clear disadvantage with these methods, however, – if high-throughput is a concern – in terms of the number of spectra required and hence the experimental expenditure. In the original FLYA implementation, for instance, 16 spectra were used. Subsequent implementations of the FLYA algorithm for SAIL-labelled proteins (Kainosho *et al.* 2006) suggest that peak picking is the limiting step in the procedure; with this task being performed interactively, nine spectra were sufficient to calculate the N- and C-terminus of calmodulin and five for ubiquitin (Takeda *et al.* 2008; Ikeya *et al.* 2009).

UNIO attempts to alleviate problems of cumulative errors by promoting user-intervention at the completion of each step. UNIO is composed of the assignment modules MATCH (Volk *et al.* 2008) and ATNOS/ASCAN (Fiorito *et al.* 2008), respectively for the backbone and side-chain atoms, along with ATNOS/CANDID (Herrmann *et al.* 2002a) for NOE constraint gathering and structure calculation (with CYANA, CNS or X-PLOR). The standard input requirements are three APSY (Hiller *et al.* 2005) and three NOE spectra although the assignment process with MATCH has already been demonstrated with a single 6D APSY spectrum (Volk *et al.* 2008). As part of the JCSG's proteins structure venture, the UNIO protocol has led to the deposition of 19 structures (67–198 residues) in the past two years.

2.5 Limitations

A common criticism of the classical NMR approach to structure calculation is that it leads to poor-quality structures. (In this context, we insist on the distinction between a good-quality structure (i.e. in terms of the covalent geometry) and one that exhibits the correct fold.) Or rather, that it does not perform as well as X-ray diffraction in terms of physico-chemical standards

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

(Doreleijers *et al.* 1998, 1999; Ratnaparkhi *et al.* 1998; Spronk *et al.* 2002; Fan & Mark, 2003; Laskowski, 2003). In fact, most of these problems stem from the simplified physical energy terms used during the structure calculation steps and are efficiently redressed by appropriate energy refinement in explicit solvent (Güntert, 1998; Spronk *et al.* 2004). Such criticism is therefore outdated and efforts are ongoing to rectify previously deposited data (Nabuurs *et al.* 2004a; Nederveen *et al.* 2005). This issue emphasises, however, the modelling aspect of NMR protein structure determination. The results obtained should always be interpreted within the framework of the underlying model and with its limitations in mind.

In this context, it should be pointed out that all NMR data interpretation is performed – and the resulting structures interpreted – on the basis of a single rigid conformation. This is perhaps a surprising simplification and in a sense, a pity, since NMR as a technique is uniquely suited to the characterisation of both protein structure and dynamics (and thereby function). This problem is well known in the community (Kim & Prestegard, 1989; Torda *et al.* 1990; Lindorff-Larsen *et al.* 2005; Markwick *et al.* 2008), but so far, progress towards a unified description has mostly been made, as it were, from the dynamics side (Korzhnev *et al.* 2004, 2010; Bouvignies *et al.* 2007; Korzhnev & Kay, 2008; Baldwin & Kay, 2009; Markwick *et al.* 2009). Alternatively, one may rely more heavily on computer modelling (Case, 2002; Mackerell, 2004; Markwick *et al.* 2005; Hornak *et al.* 2006), but as of yet, the simulation of long time-scale (>1 μs) motions in proteins remains a challenging prospect (Hamelberg *et al.* 2004; Elber, 2005; Markwick *et al.* 2008; van der Kamp *et al.* 2008; Chiang *et al.* 2010).

The scope of the traditional NMR method is also limited, both in terms of its efficiency and its applicability to larger systems. Even with full-automation (or perhaps even more so) chemical shift assignment typically requires upwards of ten spectra and notwithstanding cryo-probes and high magnetic fields, NOE spectra are still relatively weak. A balance must therefore be struck between expediting the experimental step and facilitating robust data analysis, although progress has been made towards solving both problems at once (Hiller *et al.* 2005, 2008; Fiorito *et al.* 2008; Volk *et al.* 2008). Systematic referencing across a large set of spectra – or rather the lack thereof – is also a problem. And while the internal calibration routines employed by most programs deal effectively with small discrepancies, meticulous manual referencing and alignment are prerequisites.

A more fundamental limit is the loss of resolution that follows the growth of targets – and *ipso facto* the number of peaks – as the corresponding slowing of protein motions leads to broader lines (Wüthrich, 1986; Neuhaus & Williamson, 2000). To date, automated NOE assignment has been used to solve just one structure longer than 250 amino-acid residues (de Oliveira *et al.* 2010) and more generally, structures heavier than 20 kDa are rarely targeted. In this instance, interactive data analysis is of little assistance (cf. Fig. 2). Perdeuteration is the traditional avenue towards accessing larger proteins and a number of selective labelling schemes have recently been proposed (Yee *et al.* 2006; Ohki & Kainosho, 2008; Shin & Lee, 2008) – including notably the SAIL protocol (Ikeya *et al.* 2006; Kainosho *et al.* 2006; Takeda *et al.* 2007b; Kainosho & Güntert, 2009), designed specifically with automation in mind – that promise to extend the range of classical NMR methods.

Of the steps involved, peak picking has proved to be the most troublesome to automate reliably. This is also the most critical stage as the dependability of the subsequent steps is underwritten by the accuracy of the peak lists provided. Indeed, the reliability of algorithms for chemical shift and NOE assignment deteriorates rapidly if more than a few artefacts are included. The latter also requires virtually 100% complete chemical shift assignment – of all

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

NOE-interacting protons and their attached heavy atoms; i.e. a full record in terms of the NOE spectra rather than the primary structure (Güntert, 2003; Jee & Güntert, 2003), rarely available from the outset. Rather, the preliminary fold obtained from a subset of reliable signals is used to recursively complete chemical shift and then NOE attributions. This is possible because the network of NOE constraints is highly redundant meaning that this approach is robust with respect to peaks missed (or ignored).

3. Alternatives

NMR structural biology aspires to high throughput and objectivity. Automation of the different steps of the classical approach has gone some way to achieve these twin aims. As we have seen, though, some problems remain unsolved. A few radical approaches suggest that circumventing rather than overcoming the difficulties is the way forward.

Eliminating the need for chemical shift assignment is perhaps the most obvious short cut: this step of the traditional approach has no biological relevance and typically requires a huge investment in terms of experiments and data analysis. On the other hand, problems stemming from weak NOE spectra are the simplest to avoid as a number of NMR observables – notably chemical shifts and RDCs – are readily related to structural parameters. The methods described in this section follow this general philosophy although as we shall see, the border between classical and non-classical methods is not as clear-cut as might suggest the paradigm shift outlined above.

3.1 Chemical shift-based methods

The correlation between chemical shifts in proteins and their structure is well established, but complicated (Szilagyi, 1995; Wishart & Case, 2001; Mielke & Krishnan, 2009; Mulder & Filatov, 2010; Wishart, 2010). In contrast to NOEs or RDCs, there is no single equation to link this observable to a pairwise orientation or distance (with the exception of paramagnetic pseudocontact shifts; Bertini *et al.* 2001). Moreover, chemical shifts vary primarily with each nucleus' local environment and depend only weakly on the non-covalent structure. Only recently have algorithms appeared that derive statistical predictions from ever-growing databases of structures and their corresponding chemical shifts (Wishart, 2010). (So the first stage of the traditional procedure is useful after all!) Two main classes of programs exist for which the output is either a set of torsion angles – to be used in conjunction with other restraints in subsequent structure calculation – or the query protein itself.

3.1.1 Torsion angle constraints

The programs for torsion angle constraints TALOS (Cornilescu *et al.* 1999), TALOS+ (Shen *et al.* 2009a), PREDITOR (Berjanskii *et al.* 2006) and DANGLE (Cheung *et al.* 2010) match sections of a target protein to those from a structural database based on their sequence and chemical shift similarity. The underlying idea is that a particular chemical shift pattern is indicative of a specific set of torsion angles. Thereby, if such a pattern from a query fragment can be matched to one in a database, so the corresponding torsion angles must also agree.

The overall strategy for each program being essentially identical, differences reside only in its implementation. A similarity score is calculated for each position in the amino-acid sequence,

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

measuring the level of agreement between the chemical shifts of the query triplet (quintuplet in the case of DANGLE) and fragments of the same length from the database. A residue difference correction is applied, which separates primary and secondary structural influences on the chemical shifts. A parallel search is conducted based solely on sequence homology, which exploits the structural information intrinsic to the amino-acid sequence. The best results are then combined and the prediction is output along with a confidence rating or error estimate. TALOS+ and DANGLE also classify results based on favoured regions in Ramachandran plots (Ramachandran *et al.* 1963).

To a first approximation, the larger the database, the more likely one is to find a match for a particular fragment and the more reliable are the predictions (Shen *et al.* 2009a). However, great care must be taken in constructing these databases such that the referencing for all chemical shifts is consistent. TALOS+ uses the 200-protein database constructed and calibrated as part of the SPARTA chemical shift prediction project (Shen & Bax, 2007). A subset of this database (186 structures) forms the basis of the DANGLE approach for which Cheung *et al.* (2010) have investigated the effect of systematic miss-referencing in the query protein. For PREDITOR, the reference-correction module ensures consistency between experimental shifts and those in the associated database of 141 proteins.

In terms of output, the TALOSs and DANGLE predict the backbone ϕ/ψ angles only, whereas PREDITOR also includes estimates of the side chain χ_1 and ω angles, thanks to a sequence homology analysis. Note in this context, the recent release of the PROMEGA software (Shen & Bax, 2010), which employs a TALOS-like approach to predict Xaa-Pro peptide bond conformations (pre-proline ω angles).

In terms of performance, Cheung *et al.* (2010) report similar prediction accuracy – ca. 70–80% within 30° – for all methods with their DANGLE algorithm providing more realistic error estimates. David Wishart has recently sought an objective ranking by running each program for the same set of 33 proteins. PREDITOR made accurate (ϕ and ψ within 30°) predictions for approximately 10% more residues than DANGLE or TALOS (94% vs. ∼85%).

The backbone restraints obtained using these programs are approximate and local. It is therefore not possible to reconstruct the full protein fold using these constraints alone and indeed, in situations where a dense network of NOE restraints is available, the information gained from torsion angles adds little to the precision. However, in situations where inter-atomic distances are harder to come by – for large proteins for example, or for solid-state NMR – these approaches come to the fore. To bypass NOE-type experiments completely though, different recipes are required of which angle predictions are nonetheless one of their main ingredients.

3.1.2 From structures to chemical shifts

The fundamentals of chemical shift prediction from structure are not within the scope of this review. Moreover, a thorough description of this topic has newly been published (Wishart, 2010). For the sake of protein structure determination, the reverse problem of predicting chemical shifts from existing structures is not a concern *per se*. However, they do allow much larger X-ray databases to be translated. This expansion of protein fold space turns out to be critical for the accuracy of algorithms for (full) structure calculation based (primarily) on chemical shift data (see the following section). Briefly therefore, programs for chemical shift prediction (we have censured 10, namely 4DSPOT (Lehtivarjo *et al.* 2009), CamShift (Kohlhoff *et al.* 2009), PROSHIFT (Meiler, 2003), PRSI (Wang, 2004), SHIFTCALC (Iwadate *et al.* 1999),

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

SHIFTS (Xu & Case, 2001), SPARTA (Shen & Bax, 2007), and SHIFTX/SHIFTX2 (Neal *et al.* 2003; Wishart, 2010)) construct nucleus-specific contour maps that translate torsion angles into chemical shifts. In general, these structure-shift tables are assembled from a set of high-resolution X-ray structures for which chemical shifts are also available. These relationships are generally supplemented by semi-classical or quantum-mechanical calculations of physico-chemical effects (e.g. solvent exposure, peptide anisotropy, motional averaging, pH, temperature, etc.). Corrections for protein dynamics can also be made (Lehtivarjo *et al.* 2009) and some programs make direct use of sequence homology (Meiler, 2003; Shen & Bax, 2007). Thereafter, variations mainly concern the size and content of the reference database, the number and nature (which nuclei) of the predictions for each residue and of course, the implementation of the prediction algorithm.

As for torsion angles, David Wishart has measured the correlation between experimental chemical shifts and those predicted by the algorithms SHIFTX, SHIFTX2, CamShift, Shifts and ProShifts for a set of 61 high-resolution X-ray structures (Wishart, 2010). In terms of the different nuclei, Cβ predictions score highest on average (correlation coefficient $\sim$0·99) and HN lowest ($\sim$0·69). In terms of the different programs, the correlation coefficients obtained with SHIFTX predictions are the highest on average ($\sim$0·95), and lowest with SHIFTS ($\sim$0·79).

3.1.3 From chemical shifts to structures

A structure defines a single chemical shift pattern, no unique solution exists to the reverse problem as many different global conformations can satisfy specific local arrangements. The recent development of accurate chemical shift prediction algorithms (see the preceding section) points towards an iterative solution to the problem whereby candidate structures can be ranked according to their chemical shift agreement with the query protein. To this date three methods, namely CHESHIRE (Cavalli *et al.* 2007), CS-Rosetta (Shen *et al.* 2008b) and CS23D (Wishart *et al.* 2008) have been published although a general pathway to low-resolution structures was first published by Gong *et al.* (2007) and already in 2001, the structure of ubiquitin was computed using chemical shift data exclusively (Wishart & Case, 2001). The idea moreover of supplementing Rosetta structure prediction with experimental data dates from the turn of the century (Bowers *et al.* 2000).

Chemical shifts are again used to scan a database of solved structures but compared to the programs for torsion angle prediction, alternative approaches are implemented for the subsequent reconstruction of the 3D fold. CS-Rosetta simply uses the fragments as a restricted (but much more relevant) database for a standard Rosetta (Das & Baker, 2008) calculation; CHESHIRE uses molecular fragment replacement based on Monte-Carlo substitution of candidate fragments; CS23D meanwhile assembles fragments based on simple matching criteria. Chemical shift matching along with a standard energy functional then guides the refinement of the coarse-grain CHESHIRE and CS23D models. Specific details of these protocols may be found in the following.

3.1.3.1 CHESHIRE

The first stage of the CHESHIRE protocol involves predicting the secondary structure of three and nine-residue fragments from the query protein. Bayesian inference is used to account for the propensity of certain groups of amino acids to form specific conformations. This complements a

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

more conventional chemical shift matching test. The second stage involves collecting dihedral angle constraints in a TALOS-type approach, albeit with a larger database. The backbone fold is then reconstructed from the selected fragments using a Monte-Carlo swapping routine whereby the Metropolis condition (Metropolis *et al.* 1953) is applied according to the level of agreement between SHIFTX-predicted and experimental chemical shifts. Side chains are then added to the 20000-strong group of models that is whittled down to 500 during two stages of energy/chemical shift refinement. A converged calculation is one for which the energy vs. root mean square deviation (RMSD) (to lowest-energy structure) landscape is strongly funnelled.

3.1.3.2 CS-Rosetta

As the name suggests, CS-Rosetta employs chemical shifts (experimental from the query protein and SPARTA-predicted from the structural database) to construct a database of relevant fragments that form the basis for the well-established Rosetta protocol for Monte-Carlo assembly based on sequence homology followed by high-resolution energy minimisation (Das & Baker, 2008). During the generation of the refined database, optimal accuracy is ensured on the one hand by comparing secondary chemical shifts (i.e. removing amino acid type bias), and on the other hand by accounting for uncertainties associated with the SPARTA predictions. Shen *et al.* (2008b) show that these high-quality fragments lead to convergence – which is tested using the familiar 'funnelled landscape' criterion mentioned above – in some of the cases where Rosetta (without chemical shifts) fails.

3.1.3.3 CS23D

Rosetta is also included in the CS23D protocol, but only as a 'fail-safe' alternative and with a restricted set of structures generated (500 vs. 10–20 000), the emphasis being placed on efficiency. In the hands of Wishart *et al.* (2008), CS23D solved the same structures ca. 10 000 times faster than CS-Rosetta or CHESHIRE notably through extensive homology modelling. Indeed, CS23D avoids time-consuming Monte-Carlo structure reconstructions by comparing and directly selecting outputs of three different database searches. Candidate fragments are found based on sequence homology and chemical shift matching (using SHIFTX-prediction for the structural database), but also according to torsion-angle homology after conversion of PREDITOR predictions to one of nine dihedral-angle labels. The backbone of the query protein may then be assembled by directly stitching fragments (with no limit on length) from one or other of the searches according to their relative success for a given stretch. The algorithm reverts to the restricted Rosetta run should the previous approach fail. In both cases, the procedure is completed by energy refinement in torsion angle space and a minimisation of chemical shift discrepancies.

3.1.3.4 Chemical shift restraints in MD simulations

The feasibility of the fragment assembly chemical shift methods was initially proven for ~10–20 small proteins (Cavalli *et al.* 2007; Wishart *et al.* 2008; Shen *et al.* 2008a). Since then, the application of CS-Rosetta and CHESHIRE protocols has been extended to situations where the chemical shift data for the query protein are incomplete (Robustelli *et al.* 2008; Shen *et al.* 2009b). Owing to their database approach, however, all these approaches struggle to reproduce novel folds (Shen *et al.* 2008a) and in this respect the CS23D suffers most having sacrificed flexibility for efficiency (Wishart *et al.* 2008). In this case, why not use chemical shifts directly? Why not add a chemical shift term to an MD force field?

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

This is perhaps a more intuitive implementation of chemical shift restraints and efforts to incorporate these in the structure calculation process can be traced back to the early 1990s (Ösapay & Case, 1991; Ösapay *et al.* 1994; Celda *et al.* 1995; Szilagyi, 1995; Kuszewski *et al.* 1995a, b, c; Clore & Gronenborn, 1998). However, these initial investigations were limited to refining structures that had already converged to the native state. Indeed, all but the tiniest conformational changes incur large (and often erroneous) energy penalties and imprecisions in chemical shift predictions only exacerbate this problem. Therefore, the large structural rearrangements required for a uniform sampling of conformation space are formidably difficult to provoke on the basis of chemical shift pseudo energies.

Recent improvements in the accuracy of the predictions have revived interest in the subject. Robustelli *et al.* (2009) combined SHIFTX prediction with Monte-Carlo rearrangement governed by a physical potential with an extra term to penalise chemical shift discrepancies. While this approach suffered less from the limitations outlined above, still more than 90% of the random swaps were rejected, severely restricting the sampling of conformation space.

Contrary to other chemical shift prediction software, the functions output by CamShift are differentiable with respect to atomic coordinates which means that MC sampling can be replaced with MD simulations. Lately thereby, Robustelli *et al.* (2010) have upgraded their approach and successfully identified the native fold of 11 out of 12 small proteins starting from partially unfolded structures (56–108 residues, RMSD bias to reference 0·8–2·2 Å). *De novo* structure calculation is still not possible, however, as the MD simulations become trapped in local minima when starting from random structures. Loop regions are also problematic as errors in the CamShift predictions prevent otherwise valid conformational changes. At this stage, therefore, the authors suggest that CS-MD should be used in conjunction with other restraints or methods and indeed, they demonstrate the potential of CS-RDC-MD for the excited state peptide bound form of the Abp1p SH3 domain, with restraints obtained from relaxation dispersion experiments (Korzhnev *et al.* 2004; Korzhnev & Kay, 2008; Baldwin & Kay, 2009).

3.2 RDC-based methods

RDCs can be measured for proteins by placing samples in a liquid crystalline material, thereby limiting the tumbling that usually cancels out dipolar interactions in solution. RDCs constrain the dipole vectors of all atoms with respect to the same alignment tensor. As for NOEs, there is a direct theoretical link between experimental values and structural parameters (Bax *et al.* 2001; Prestegard *et al.* 2004; Blackledge, 2005):

$$D_{ij} = -S\mu_0 h \frac{\gamma_i \gamma_j}{16\pi^3 \gamma_{ij}^3} \left(A_a (3\cos^2\theta - 1) + \frac{3}{2} A_r \sin^2\theta \cos 2\phi \right) \tag{7}$$

The order parameter, S, is assumed constant; r_{ij} is the inter-nuclear distance oriented (θ, ϕ) with respect to the alignment tensor with axial and rhombic components A_a and A_r, respectively.

RDCs are, however, highly degenerate with respect to the alignment tensor (Ramirez & Bax, 1998; Bax *et al.* 2001; Blackledge, 2005; Donald & Martin, 2009). The main concern of RDC-based algorithms is then to gather constraints in order to define a single reference frame. The traditional approach is to measure a specific set of RDCs – the more the merrier – in two alignment media. Despite being undesirable; because of the investment required in terms of sample preparation and experimental time, this is the only way to overcome the high degeneracy of the orientations of the peptide planes. Indeed, even multiple RDCs, but from a single

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

Table 3. *Principle components of RDC-based algorithms for automatic protein structure determination*

	program						
	3P	Fast NMR	ITAS	MECCANO	MFR	RDC-PANDA	RDC-Rosetta
chemical shift assignment	required	performs	performs	required	required	required	required
labeling	^{15}N	^{15}N & ^{13}C[a]	^{15}N & ^{13}C[a]	^{15}N & ^{13}C	^{15}N	^{15}N	^{15}N & ^{13}C[a]
RDCs per residue[b]	5	3-4	3-4	6	1-6[c]	2	2-6
number of alignment media	1	1	1	2	1-2	1-2	1-2
database	no	yes	yes	no	yes	yes[d]	yes
NOEs	no	full set	no	no	sparse[a]	full set	sparse to full set[a,e]
TALOS	yes	no	no	no	yes	yes	no

[a]Optional.
[b]Per alignment medium.
[c]'A dozen or more' (Kontaxis *et al.* 2005) per fragment of 7–10 amino acids.
[d]Only for side-chain atoms.
[e]See Rosetta section.

alignment medium, leave many ambiguities unresolved and the additional information may only improve the precision of the structure bundle, rather than the accuracy of the atomic coordinates (Hus *et al.* 2008). An alternative is to use a database of known structures, NOEs and/or dihedral angle restraints to wipe out the degeneracies (Annila *et al.* 1999; Delaglio *et al.* 2000; Andrec *et al.* 2002; Rohl & Baker, 2002). Note that for most methods (see Table 3 for a list), ^{15}N enrichment is sufficient (optimal performance is nonetheless achieved with fully labelled samples); therefore, the extra complication of producing aligned samples is offset to an extent by a reduction in labelling costs. To reiterate though, no labelling strategy can circumvent the requirement of measurements in several alignment media.

It is clear, however, that an optimal set of restraints has not yet been defined. Indeed, depending on the availability of the different experimental constraints, any one of the algorithms described in the following may be the most favourable for a particular target. Another issue concerning software developed so far is that the measurement of RDCs is not as popular and widespread as say, NOEs; therefore, most programs have only been tested so far on simulated data (using e.g. PALES (Zweckstetter & Bax, 2000; Zweckstetter *et al.* 2004; Zweckstetter, 2008)) or experimental values recorded on model proteins.

3.2.1 Exploiting previously solved structures

3.2.1.1 Molecular fragment replacement (MFR)

Unlike NOE-based structure calculation algorithms, Monte-Carlo or simulated annealing-type methods do not regularly obtain native folds from RDC constraints, mainly because of local minima caused by degenerate solutions to global RDC restraints. This limitation does not, however, apply to elements of the structure; hence, the idea behind MFR (Annila *et al.* 1999) to reconstruct protein structures using the most likely fragments selected from a database.

In MFR, fragments from the query protein 7–10 amino acids in length are compared to those in a database of high-resolution X-ray structures (a subset of the PDB). The best 20 fragments

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

are selected according to a χ^2 fit to the experimental RDCs and to a lesser extent, the chemical shifts and primary structure. A dozen or so RDCs are required for each fragment, optimally measured in two alignment media. A similar approach has been proposed by Andrec *et al.* but using a single dataset (Andrec *et al.* 2001; Andrec *et al.* 2002). And indeed for favourable cases, MFR can also make do with fewer constraints (Kontaxis *et al.* 2005).

Small structural changes can lead to large variations in RDCs. Conversely, in order to whittle away these false positives, the top fragments are energy minimised with respect to the experimental RDCs by low-temperature simulated annealing. The selection is further refined by only accepting well-clustered fragments within the top 20. Any ambiguities are removed during the next stage of fragment assembly whereby sequential units are appended to the existing structure while minimising the RMSD to the latter. Subsequent appendages have many residues in common therefore in general overlapping fragments will have a single shared orientation. If however, this procedure does not converge to a single fold – axially symmetric alignment is a typical problem, the remaining ambiguities are easily resolved using a few specific NOE constraints.

In this arrangement, MFR solved both domains of the murine gamma-S crystalline (Wu *et al.* 2005) (PDB IDs: 1ZWM & 1ZWO) that were recombined with the help of small-angle X-ray scattering (SAXS) data (Grishaev *et al.* 2005) (2A5M). In addition, the latest version of the program, MFR+, has been tested on both experimental and simulated data sets for six model proteins (sequence length 56–153 and ~4–5 RDCs per residue in two alignment media) (Kontaxis *et al.* 2005). The structures converged in all cases with backbone RMSDs to the reference structure in the range 0·44–1·95 Å.

3.2.1.2 RDC-Rosetta

We have already seen how experimental constraints – in this case chemical shifts, described in CS-ROSETTA section – can be used to increase the reliability of Rosetta structure predictions. Similarly, RDCs can be used to prune a database of the most relevant fragments and guide Rosetta backbone reconstruction towards low-energy regions of fold space (Rohl & Baker, 2002; Das & Baker, 2008).

Ten calculations were performed on simulated data from seven proteins and a converged structure, with a backbone RMSD below 3·0 Å was obtained in five cases. In addition, experimental RDCs were used for four model proteins, for which three converged to the native fold (backbone RMSD to the reference structure from 1·1 to 6·0 Å). In each case, highly homologous proteins were excluded from the database. Unsurprisingly, the more RDC restraints included, the more accurate was the Rosetta model.

3.2.2 Non-database methods

Since the theoretical link between NMR observable and structure is well-established in the case of RDCs, the tantalising prospect exists of accessing novel folds by calculating structures directly from the experimental values, thereby waiving the database homology requirement.

3.2.2.1 MECCANO

Planar arrangements of RDCs are eightfold degenerate with regard to a single alignment medium whereas chiral structures are fourfold degenerate. In two alignment media, the ambiguities are

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

reduced by a factor of four meaning that as long as suitable chiral elements can be constructed, the orientation of the alignment tensor can be determined *de novo* (Blackledge, 2005).

In MECCANO (Hus *et al.* 2001), the six RDCs measured per residue, in two different alignment media define the peptide plane (N^{i}–H_i^{N}, C'_{i-1}–N_i, C'_{i-1}–H_iN, $C^{\alpha}{}_{i-1}$–C'_{i-1}), and the tetrahedral trans-peptide bond ($C^{\alpha}{}_{i-1}$–$H^{\alpha}{}_i$, $C^{\alpha}{}_{i-1}$–$C^{\beta}{}_{i-1}$). Together, these define the required sub-structure. For the calculation of ubiquitin, experimental values from two media were available for all couplings other than C^{α}–C^{β} and totalled 648. The backbone RMSD between MECCANO and reference structures was 1·0 Å.

When data are missing for a particular plane, these can be simulated but the results are far less precise, hereby the prerequisite of extensive RDC constraints is the main limit for this approach. Note however, that of all the RDC-based programs described here, MECCANO is the only one to rely exclusively on such constraints. Since assigned chemical shifts are required anyway, dihedral angle constraints (Cornilescu *et al.* 1999; Berjanskii *et al.* 2006; Shen *et al.* 2009a; Cheung *et al.* 2010) could easily be incorporated into the protocol and thereby extend its applicability.

3.2.2.2 3P

Like MECCANO, the 3P algorithm (Wang *et al.* 2007) constructs chiral elements from internuclear vectors. However now, instead of relying on experiments in a second alignment medium, the degeneracy of the RDCs is lifted in regular secondary structure elements by exploiting periodic relationships between the coupling strengths and the orientation of the peptide planes (Mesleh *et al.* 2002; Lee *et al.* 2003; Mascioni & Veglia, 2003; Mesleh *et al.* 2003; Mesleh & Opella, 2003; Walsh *et al.* 2004; Walsh & Wang, 2005). In addition, dihedral angle restraints are included to overcome problems if distinct solutions are not well defined – primarily ambiguities caused by missing or noisy data – and to constrain regions of non-regular secondary structure.

A Levenberg–Marquardt algorithm (Nocedal & Wright, 2006) is used to fit peptide-plane orientations to the RDC data. The three best results are then considered during the fragment assembly stage. In a process similar to MFR, starting with the residues for which the most restraints are available, the backbone structure is reconstructed iteratively by appending each time the best-matching candidates in terms of RDCs.

The influence of imprecise or missing constraints has been investigated using data simulated for the protein cutinase (PDB ID: 1CEX). 3P remained robust with up to 30% noise added to the simulated RDCs (RMSD bias $<3{\cdot}0$ Å) or with up to 30% of these missing (RMSD bias $<2{\cdot}5$ Å). However, in the absence of (highly precise) dihedral angle restraints, the algorithm was not able to identify the correct fold (19 Å RMSD bias). For the three model proteins ubiquitin (1UBQ), GB3 (1IGD) and DinI (1GHH), also used to test the MFR + program, the performance of 3P is similar with backbone RMSDs to the reference structures of less than 1·0 Å.

3.2.2.3 RDC-PANDA

RDC-PANDA (Zeng *et al.* 2009) builds on recent work by Wang *et al.* (Wang & Donald, 2004; Wang *et al.* 2006) to efficiently convert RDC measurements into structural constraints. The aim then is to connect RDC-defined secondary structure elements using unambiguous NOEs, and then to use this preliminary structure to re-evaluate the NOE spectra in the usual way. In this sense, RDC-PANDA has a lot more in common with early NOE-assignment

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

algorithms (Güntert *et al.* 1991; Güntert *et al.* 1993; Mumenthaler *et al.* 1997) than with the other programs discussed in this section. It also shares their weakness in terms of assignment ambiguity and applicability for larger proteins. The extra experimental effort is therefore not justified and faced with the choice, it is clear that one would forgo the measurement of RDCs and opt for one of the more robust and proven NOE-assignment algorithms described in the previous section.

3.2.3 Combined chemical shift assignment and structure determination

On the other hand, the recording of RDCs as *an alternative* to chemical shift assignment is an alluring prospect. Note that the methods described so far in this section all require chemical shift assignment. As mentioned above, this process is notoriously labour-intensive and uncongenial to traditional automatic approaches.

3.2.3.1 ITAS and FastNMR

In ITAS (Jung *et al.* 2004), chemical shift assignment is performed automatically using MARS, first without and then with a structural model as the two rounds are interspersed with Rosetta structure prediction. Rosetta fragment assembly is however, notoriously slow and is prone to errors for larger proteins ($> \sim 100$ residues) or novel folds (Bradley *et al.* 2005; Raman *et al.* 2010b). In FastNMR (Korukottu *et al.* 2007), the second Rosetta step is replaced by a CYANA-MARS-CYANA sequence to complete the chemical shift assignment and refine the structure. FastNMR is therefore more robust and computationally efficient than ITAS, but incurs an extra experimental cost in the form of NOE spectra.

ITAS has been tested on a combination of experimental and simulated data from eight proteins (56–153 residues) (Jung *et al.* 2004). Out of ten calculations, seven were successful with backbone RMSD biases less than 3·0 Å. MARS was the culprit for the two Z domain (PDB ID: 2SPZ) runs. For DinI (1GHH), ITAS (Rosetta) was unable to obtain the correct fold even starting from previously published assignments. Starting from experimental data, the FastNMR approach on the other hand yielded high-resolution models of the proteins ubiqutin (1D37), Conk-S1 (2CA7) and Conk-S2 (Korukottu *et al.* 2007). In each case, the RMSD to the mean of the CYANA bundle was less than 1 Å as was the deviation from the reference structure (not available for Conk-S2).

3.3 RosettaNMR

It is clear that both RDC- and chemical shift-based protein structure determination is still in its infancy and the methods described over the last few pages cannot match the range or the robustness of the more mature NOE-based algorithms. But this is to be expected since the new algorithms have to make do with sparser (RDCs) or more approximative (chemical shifts) restraints.

To overcome these limitations and with the aim of establishing a rival protocol for the most challenging protein structures, a couple of recent studies have focused on different variations of RosettaNMR (Rohl, 2005; Raman *et al.* 2010a, b). NMR constraints can either be used in advance – as it were – to filter the Rosetta database fragments (as described above for CS- and RDC-Rosetta), during the assembly process, to guide the sampling towards low energy regions

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

of conformation space, or subsequently, to rank the resulting structures according to how well they reproduce the experimental data (Rohl, 2005).

3.3.1 CS-DP-Rosetta

We have seen that NMR peak picking is hard to perform automatically and that NOE-assignment algorithms are extremely sensitive to artefacts. In the course of most NMR protein structure calculations therefore, a lot of time is spent manually refining peak lists. One option to expedite this process is simply to avoid it. However, methods such as CS-Rosetta are only effective for small proteins. As a compromise therefore between efficiency and robustness, Raman *et al.* propose to filter CS-Rosetta models against NOE peak lists, thereby achieving the range of traditional NOE-assignment algorithms but with raw (or 'minimally edited') peak lists (Raman *et al.* 2010a).

The thousand or so lowest-energy CS-Rosetta models out of an initial set of 50 000 are filtered against the NOE peak lists using the RPF software (Huang *et al.* 2005). This evaluates how well a structure explains the experimental peak pattern. The top 20 conformers (energy + DP score) are then extensively refined according to the Rosetta 'rebuild and refine' protocol (Qian *et al.* 2007; Ramelot *et al.* 2009). When refined peak lists are available, an alternative strategy, 'AssignNOE-Rosetta' is to use a traditional NOE-assignment algorithm to produce a preliminary structure subsequently refined using Rosetta.

The CS-DP-Rosetta protocol was tested on six proteins, ranging in length from 97 to 145 amino acids. Three of these calculations were performed blind (i.e., without knowledge of the reference structure). Only in one case, surprisingly for the smallest protein, was the RMSD bias to the reference structure greater than 3·0 Å (Raman *et al.* 2010a). Moreover, AssignNOE-Rosetta performed systematically better than CYANA or AutoStructure with refined peak lists: for a set of five proteins (100–160 amino acids) the RMSD biases for the structured regions were in the range 1·0–1·4 Å as compared to 1·3–2·5 Å for the reference calculations (Raman *et al.* 2010a).

One would need to weigh up the extra computational cost incurred through the use of Rosetta (each CS-Rosetta model takes ∼5–10 CPU minutes to build (Shen *et al.* 2008b)) and the time-cost of manual peak-list refinement. Or perhaps a more relevant comparison could have been made with the ATNOS/CANDID protocol for (unsupervised) automatic peak picking and NOE assignment. Nonetheless, these results suggest that CS-DP-Rosetta is a viable alternative to the traditional NOE algorithms for medium-sized proteins.

3.3.2 CS-RDC-(NOE)-Rosetta

A different strategy has been proposed for larger proteins. Fig. 2 shows that for proteins longer than 200 residues solution NMR in general, and automated algorithms in particular are not a popular choice. Raman *et al.* seek to address this problem by supplying Rosetta with backbone chemical shifts and sparse NOEs (Raman *et al.* 2010b). The authors argue that by exclusively using experimental constraints from the protein backbone, this approach should be more conducive to solving larger targets for which deuteration significantly reduces the density of the NOE network.

CS-Rosetta is robust up to approximately 100 residues. Adding RDC restraints to guide the conformational search towards low-energy regions (Rohl & Baker, 2002) helps CS-RDC-Rosetta

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

to consistently solve structures up to a length of 120 amino acids (Raman *et al.* 2010b). To push the size limit even further, backbone amide NOEs (HN–HN) were included at the 'rebuild and refine' stage (Raman *et al.* 2010b). This CS-RDC-NOE-Rosetta protocol was tested for 15 proteins (122–266 residues) such that 13 structures were obtained that had converged for more than half of the sequence. Of these, ten had an RMSD bias to the reference structure of less than 3·5 Å (Raman *et al.* 2010b).

This approach is advantageous since no experimental information is required for the side-chain atoms as these are added automatically during the refinement stage. Rosetta is not in the business of proving free lunches however: according to this study, the calculation of a 200 amino-acid protein takes in the order of 82 000 CPU hours. In terms of range and robustness moreover, it is not clear that CS-RDC-NOE-Rosetta actually constitutes an improvement on pre-existing methods.

3.4 Structure calculation with unassigned chemical shifts

Although efforts have concentrated so far on overhauling the traditional constraint-gathering paradigm, it can be argued that the most problematic and time-consuming step in NMR protein structure determination is chemical shift assignment and that efforts should on the contrary concentrate on eliminating this stage. Indeed, no standard approach has yet emerged – either in terms of experiments or algorithms – for automatic chemical shift assignment. An attractive proposition is to rely solely on the distance information contained in NOE spectra.

3.4.1 CLOUDS

Indeed, these always encode the spatial arrangement of protons in terms of correlated frequencies, whether the latter have been attributed to specific atoms or not. The pioneering work on this subject dates from the early 90 s (Malliavin *et al.* 1992) but so far, these ideas have only been pursued by a couple of research groups (Atkinson & Saudek, 1997, 2002; Grishaev & Llinás, 2002a, b). Of the algorithms proposed, CLOUDS is the only one to have been successful starting from experimental data. For two small proteins (60 & 83 residues), the CLOUDS approach yielded backbone structures with less than 1·5 Å RMSD from the reference (Grishaev & Llinás, 2002a).

Parallels may be drawn between CLOUDS and X-ray methods for protein structure determination, in that NOESY spectra are used to create a spatial distribution (of protons here, rather than electrons) into which a model structure is then fit. Upper- and lower-distance constraints are extracted from the NOE spectra by relaxation matrix analysis. An MD simulation is then performed with electrostatic repulsions to assemble 1000 proton clouds. The candidates that deviate most from the mean are then discarded and the resulting family of clouds defines a sleeve into which the covalent structure is folded. Sequential connectivities are established through the clouds based on the likelihood of H^N–H^N, H^α–H^N and H^α–H^α contacts obtained from a Bayesian interpretation of database values. A similar approach is used to attach the side chains. Finally, the fully assembled protein is squeezed into shape using MD constrained by the proton density.

A variation on the CLOUDS protocol has recently been proposed (Bermejo & Llinás, 2008). SC-CLOUDS is designed to work with selectively labelled perdeuterated proteins and uses

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

Rosetta to assemble the final structure. Rosetta is required because the distance constraints are too sparse to collapse the proton clouds down to a single structure. SC-CLOUDS borrows this approach – roughly speaking: to use Rosetta instead of assigned chemical shifts – from the work of Jens Meiler and David Baker (Meiler & Baker, 2003), and we have seen that ITAS employs a similar strategy. We refer the interested reader to the comparison of these three methods in (Bermejo & Llinás, 2010).

4. Outlook

Perhaps one of the main conclusions to be drawn from this whole discussion is that when attempting to hasten the process of protein structure determination it is very hard to avoid gathering long-range constraints in general, and NOE distance limits in particular: one will have noted the use of NOE restraints to improve the robustness of many of the non-classical methods.

None of the protocols proposed as an alternative to classical NMR structure calculation have (yet?) imposed themselves in the field. This is maybe a question of maturity more than anything: the more established protocols have been gradually refined over more than ten years and extensively tested. Of all the algorithms available to the community for the automatic analysis of NMR data, only a limited number of NOE-assignment programs are routinely used.

When predicting the future state of a research field, a tempting manoeuvre consists in comparing the limitations of a rival approach, five-years-old, with the potential of one's own method, five years hence. Suffice to say here that since all the procedures come with specific advantages and drawbacks, and since most of the algorithms described so far are still in active development, the field seems destined to remain as vibrant as ever. In this context, we shall now briefly cover some emerging or related research topics that could influence automated NMR protein structure determination in the years to come.

4.1 Towards more realistic structure models

A structure should always be interpreted in the context of the assumptions and limitations of the modelling process. In NMR for example, the experimental data are invariably indirect, incomplete, insufficient and imprecise. Wherefrom stem four questions, namely; how accurate is the intensity–distance mapping? How ambiguous are the experimental constraints? How should they be weighted relative to physical parameters and how can the errors be quantified? By formulating protein structure determination as a problem of Bayesian inference, inferential structure determination (ISD) seeks objective rather than *ad hoc* answers (Rieping *et al.* 2005; Habeck *et al.* 2005a; Rieping *et al.* 2008).

It is illogical, the authors argue, to interpret NMR data on the basis of a single structure (as is done at present) if the former are uncertain. A Bayesian interpretation is preferable because it naturally accounts for all possible solutions and leads to an objective measure of the precision of a structural ensemble – the width of the probability distribution – based on the inferred accuracy of the experimental constraints (Rieping *et al.* 2005; Habeck *et al.* 2005a).

ISD is based on a Markov chain Monte Carlo algorithm (Habeck *et al.* 2005b) wherein, rather than keeping intensities and their calibration fixed, these parameters are sampled

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

randomly and iteratively with conformation space to infer the errors, both experimental (for the intensities) and theoretical (for the distance calibration). Potential structures are represented by a posterior probability distribution of conformers given the experimental data and prior information (e.g. the covalent geometry). A replica exchange (Swendsen & Wang, 1986; Earl & Deem, 2005) approach ensures that the calculation does not become trapped in local minima.

So far, ISD has been used to calculate meaningful errors on atomic coordinates (Rieping *et al.* 2005) and to optimise the balance between experimental and theoretical constraints (Habeck *et al.* 2006). The same authors have also applied a similar procedure to calculate torsion angles from *J* couplings (Habeck *et al.* 2005c) and for the analysis of RDCs (Habeck *et al.* 2008). Even though this approach needs to be tested in more challenging situations, the prospect of more objective structures is alluring, if only to obviate lengthy cross validation and provide a true measure of NMR resolution.

Still, even ISD assumes that the experimental data can be explained by a single structure. By interfering with the NOEs, internal dynamics do contribute to the width of the ISD distribution, but this reflects uncertainties in the fold caused by a lack of information rather than true protein motions.

4.2 New frontiers for automated protein structure determination

To this date, a key influence on the development of algorithms for automatic protein structure determination has been the demand for high throughput in the ultimate aim of mapping out a sizeable fraction of protein fold space (Levitt, 2009; Terwilliger *et al.* 2009). It now seems that structural genomics is moving into a new age where the emphasis will shift from quantity to quality. The latest PSI initiative for example, PSI:Biology, seeks to foster tighter interplay between the structural and functional aspects of biology, as already done by other centres such as the SGC (Structural Genomics Consortium) and the Targeted Proteins Research Program (TPRP) (Ledford, 2010). This means that rather than choosing targets based mainly on their amenability to automatic calculation, selections will be made according to the current interests of the biological community, notably in terms of drug development. The implications for automated structure determination are numerous.

We have already mentioned two approaches geared specifically towards large proteins, combining selective labelling and extensive automation (Takeda *et al.* 2007a; Raman *et al.* 2010b). More challenging targets mean that a more generalised implementation of innovations – technical, experimental and in terms of sample preparation and data processing – will probably be required. These in turn will surely require new tools for automatic data analysis and interpretation.

Of course, NMR is not a panacea. In situations where high-quality crystals are readily obtained, NMR will continue to play second fiddle to faster methods based on X-ray diffraction. And in terms of protein sizes, the range of crystallography is unrivalled. However, structures in the PDB, solved both by NMR and crystallography, are intrinsically more reliable than those determined by one or other method (see validation section). This is particularly important as these databases are widely used for homology modelling and in general, for all fragment assembly methods. It is crucial therefore that the confidence in the structures should be maximised. In addition, small differences are interesting in terms of the behaviour of proteins in the two environments.

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

Alternatively, two sets of experimental data may be integrated directly in the structure calculation process. Automatic treatment of data from a number of sources can be modelled on programs for the prediction of protein interactions. For instance, a number of recent studies have demonstrated the value of combining SAXS and NMR restraints (Grishaev *et al.* 2005, 2008a, b; Gabel *et al.* 2008; Wang *et al.* 2009; Jehle *et al.* 2010; Zuo *et al.* 2010). Similarly, cryo-electron microscopy (cryo-EM) is an emerging field that seems destined for fertile collaborations with NMR. Cryo-EM can provide large-scale density maps into which atomic-resolution NMR structures are folded (Ritchie, 2008; Byeon *et al.* 2009; Fujii *et al.* 2010). For large multi-domain proteins, the association may prove particularly effective with solid-state NMR, as this technique does not intrinsically suffer from the same size limit as its solution counterpart.

Another avenue to follow involving solid-state NMR may be its association with constraints from paramagnetic effects, the potential whereof has recently been demonstrated for the protein MMP12 (Balayssac *et al.* 2008; Bertini *et al.* 2010), the largest protein solved so far using solid state NMR. Since the turn of the century, several small proteins have been calculated using restraints gathered manually from solid-state NMR spectra (Bockmann, 2008). An automatic approach has recently been demonstrated with ATNOS/CANDID, thanks to which the structure of ubiquitin was obtained directly from experimental spectra (Manolikas *et al.* 2008). Furthermore, the ARIA protocol was used to determine the structure of the Crh dimer, starting from spectra filtered against the reference structure (Loquet *et al.* 2008, 2010).

Similarly, the potential of paramagnetic effects to provide (very) long-distance constraints has long been recognised, but its application has been restricted to naturally occurring metalloproteins (Bertini *et al.* 2001; Banci *et al.* 2010). The recent development of techniques to tag proteins with paramagnetic ions (Keizers & Ubbink, 2010) should incur a greater interest in this approach from the structural biology community in general.

For more run-of-the-mill targets, full automation offers the prospect of fast and fully objective data interpretation. In this context, the critical assessment of structure determination (CASD) initiative seeks to evaluate and disseminate the potential of algorithms for automated NMR protein structure determination (Rosato *et al.* 2009). During the first round, the programs ARIA, AutoStructure, CANDID and CYANA all performed well in determining the nine blind targets. The objective in the future is to push the investigation further and start directly from raw NMR data, as was done already for two targets with the ATNOS/CANDID algorithm. It is important that computational developments, in particular those made through large-scale initiatives, are advertised and made available to the community in this way. Indeed, a criticism made of NMR in terms of structural genomics projects is that too few of the technical advances have trickled down to individual research groups (Billeter *et al.* 2008; Williamson & Craven, 2009). In this context, CCPN (Fogh *et al.* 2002; Vranken *et al.* 2005) seeks to facilitate the use of popular NMR programs through a single underlying data model. Furthermore, the eNMR grid and web server (Bonvin *et al.* 2010) through which commonly used software and vast computer resources may be accessed is particularly propitious.

Much hard work is therefore still required in the field of automated protein-structure determination, unless of course, the field is completely overrun by computer-game playing monkeys.

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

Appendix

Table A1. *List of references for programs performing automated chemical shift assignment published in the last 15 years*

1	ABACUS (Lemak *et al.* 2008)
2	ASCAN (Fiorito *et al.* 2008)
3	ASSTOOL (Reed *et al.* 2003)
4	AUTOASSIGN (Zimmerman *et al.* 1997; Moseley *et al.* 2001)
5	Bhavesh *et al.* (Bhavesh *et al.* 2001)
6	'Bipartite Matching' (Xu *et al.* 2002)
7	CISA (Wan & Lin, 2007a)
8	'Contact Based' (Kamisetty *et al.* 2006)
9	'Contact Replacement' (Xiong *et al.* 2008)
10	CONTRAST (Olson & Markley, 1994)
11	CRAACK (Benod *et al.* 2006)
12	DYNASSIGN (Schmucki *et al.* 2009)
13	GANA (Lin *et al.* 2005)
14	GARANT with structure (Bartels *et al.* 1996)
15	GARANT without structure (Bartels *et al.* 1997)
16	GASA (Wan & Lin, 2007b)
17	Hus *et al.* (Hus *et al.* 2002)
18	Inferential (Vitek *et al.* 2006)
19	IPASS (Alipanahi *et al.* 2009a)
20	JIGSAW (Bailey-Kellogg *et al.* 2000)
21	MAPPER (Güntert *et al.* 2000)
22	MARS with structure (Jung & Zweckstetter, 2004a)
23	MARS without structure (Jung & Zweckstetter, 2004b)
24	MATCH (Volk *et al.* 2008)
25	MC_ASSIGN1 (Tycko & Hu)
26	MONTE (Hitchens *et al.* 2003)
27	NOEnet (Stratmann *et al.*)
28	NVR 1 (Langmead & Donald, 2004)
29	NVR 2 (Langmead *et al.* 2004)
30	NVR 3 (Apaydin *et al.* 2008)
31	PACES (Coggins & Zhou, 2003)
32	PASA (Xu *et al.* 2006)
33	PASTA (Leutner *et al.* 1998)
34	PINE (Bahrami *et al.* 2009)
35	PISTACHIO (Eghbalnia *et al.* 2005b)
36	PRODECOMP/SHABBA (Staykova *et al.* 2008)
37	RIBRA (Wu *et al.* 2006)
38	SAGA (Crippen *et al.* 2010)
39	Sanctuary 1 (Xu & Sanctuary, 1993)
40	Sanctuary 2 (Xu *et al.* 1994)
41	Sanctuary 3 (Li & Sanctuary, 1997b)
42	Sanctuary 4 (Li & Sanctuary, 1997a)
43	TATAPRO (Atreya *et al.* 2000)
44	Wan *et al.* (Wan *et al.* 2003)

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

Table A2. *PDB search strings for chemical shift assignment programs*

		Query[a]	
		Internal	External[b]
All programs …	'program name'	'program name' *and not* 'associated structural genomics centre'	
	…except		
	MONTE	*MONTE and not MONTE CARLO*	as above
	ASCAN	*ASCAN and not UNIO* + *UNIO and not ASCAN*	*ASCAN and not UNIO and not JCSG* + *UNIO and not ASCAN and not JCSG*
	MATCH	*APSY*	*APSY and not JCSG*

[a]A PDB 'Advanced Search' was performed with the listed queries as a 'Text Search' within all solution NMR depositions.
[b]Depositions for which the author list included one of the program authors were discarded.

Table A3. *PDB search strings for NOE assignment programs*

		Query[a] internal[b]
All programs …		'Program name'
	…except	
	CANDID[c]	CYANA and (KUJIRA or CANDID or automated or automatic or atnoscandid or UNIO) + KUJIRA and not (CYANA or CANDID or radar or atnoscandid or UNIO) + CANDID and not (CYANA or radar or atnoscandid or KUJIRA or UNIO) + atnoscandid and not (CYANA or CANDID or radar or KUJIRA or UNIO) + radar and not (CYANA or CANDID or KUJIRA or atnoscandid or UNIO) + UNIO and not (CYANA or CANDID or KUJIRA or atnoscandid or RADAR)

[a]A PDB 'Advanced Search' was performed with the listed queries as a 'Text Search' within all solution NMR depositions.
[b]For external hits, the internal query was supplemented with '*and not*' 'associated structural genomics centre'; i.e. for CANDID this was '*and not JCSG and not RIKEN*'. Also, depositions were discarded for which one of the structure's authors was/is also involved in program development.
[c]CYANA is also used for simulated-annealing structure calculation, but not necessarily with CANDID. The number of CYANA hits (2460) overestimates the use of CANDID, but excluding all those that do not also reference other associated software (as was done here) is surely an underestimate.

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

Table A4. *PDB search strings for peak-picking algorithms*

		Query[a] internal[b]
All programs …		'Program name'
	…except	
	ATNOS	ATNOS and not (atnoscandid or radar or athnos) + atnoscandid and not (ATNOS or radar or ATHNOS) + ATHNOS and not (atnoscandid or radar or ATNOS) + radar and not (atnoscandid or ATHNOS or ATNOS)

[a]A PDB 'Advanced Search' was performed with the listed queries as a 'Text Search' within all solution NMR depositions.

[b]For external hits, the internal query was supplemented with '*and not*' 'associated structural genomics centre'. Also, depositions were discarded for which one of the structure's authors was/is also involved in program development.

Table A5. *Search string for PDB entries for which some automation was used*

	Query[a]
Some automation	(CYANA and (KUJIRA or CANDID or automated or automatic or atnoscandid)) or KUJIRA or CANDID or atnoscandid or radar or UNIO or ARIA or autostructure or AUREMOL or pasd or noa or garant or autoassign or abacus or pine

[a]A PDB 'Advanced Search' was performed with the listed queries as a 'Text Search' within all solution NMR depositions.

Table A6. *PDB search strings for torsion-angle prediction programs*

	Query[a]
All programs …	'Program name'

[a]A PDB 'Advanced Search' was performed with the listed queries as a 'Text Search' within all solution NMR depositions.

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

5. References

Alipanahi, B., Gao, X., Karakoc, E., Balbach, F., Donaldson, L. & Li, M. (2009a). IPASS: error tolerant NMR backbone resonance assignment by linear programming. In *University of Waterloo Technical Reports*. Waterloo: University of Waterloo.

Alipanahi, B., Gao, X., Karakoc, E., Donaldson, L. & Li, M. (2009b). PICKY: a novel SVD-based NMR spectra peak picking method. *Bioinformatics* **25**, i268–i275.

Andrec, M., Du, P. & Levy, R. M. (2001). Protein backbone structure determination using only residual dipolar couplings from one ordering medium. *Journal of Biomolecular NMR* **21**, 335–347.

Andrec, M., Harano, Y., Jacobson, M. P., Friesner, R. A. & Levy, R. M. (2002). Complete protein structure determination using backbone residual dipolar couplings and sidechain rotamer prediction. *Journal of Structural and Functional Genomics* **2**, 103–111.

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

Apaydin, M. S., Conitzer, V. & Donald, B. R. (2008). Structure-based protein NMR assignments using native structural ensembles. *Journal of Biomolecular NMR* **40**, 263–276.

Atkinson, A. R. & Saudek, V. (1997). Direct fitting of structure and chemical shift to NMR spectra. *Journal of the Chemical Society Faraday Transactions* **93**, 3319–3323.

Atkinson, R. A. & Saudek, V. (2002). The direct determination of protein structure by NMR without assignment. *FEBS Letters* **510**, 1–4.

Atreya, H. S., Sahu, S. C., Chary, K. V. & Govil, G. (2000). A tracked approach for automated NMR assignments in proteins (TATAPRO). *Journal of Biomolecular NMR* **17**, 125–136.

Atreya, H. S. & Szyperski, T. (2005). Rapid NMR data collection. *Methods in Enzymology* **394**, 78–108.

Bahrami, A., Assadi, A. H., Markley, J. L. & Eghbalnia, H. R. (2009). Probabilistic interaction network of evidence algorithm and its application to complete labeling of peak lists from protein NMR spectroscopy. *PLoS Computational Biology* **5**, e1000307.

Bailey-Kellogg, C., Widge, A., Kelley, J. J., Berardi, M. J., Bushweller, J. H. & Donald, B. R. (2000). The NOESY jigsaw: automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. *Journal of Computational Biology* **7**, 537–558.

Balayssac, S., Bertini, I., Bhaumik, A., Lelli, M. & Luchinat, C. (2008). Paramagnetic shifts in solid-state NMR of proteins to elicit structural information. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 17284–17289.

Baldwin, A. J. & Kay, L. E. (2009). NMR spectroscopy brings invisible protein states into focus. *Nature Chemical Biology* **5**, 808–814.

Banci, L., Bertini, I., Luchinat, C. & Mori, M. (2010). NMR in structural proteomics and beyond. *Progress in Nuclear Magnetic Resonance Spectroscopy* **56**, 247–266.

Baran, M. C., Huang, Y. J., Moseley, H. N. & Montelione, G. T. (2004). Automated analysis of protein NMR assignments and structures. *Chemical Reviews* **104**, 3541–3556.

Bardiaux, B., Bernard, A., Rieping, W., Habeck, M., Malliavin, T. E. & Nilges, M. (2009). Influence of different assignment conditions on the determination of symmetric homodimeric structures with ARIA. *Proteins: Structure Function and Bioinformatics* **75**, 569–585.

Bartels, C., Billeter, M., Güntert, P. & Wüthrich, K. (1996). Automated sequence-specific NMR assignment of homologous proteins using the program GARANT. *Journal of Biomolecular NMR* **7**, 207–213.

Bartels, C., Güntert, P., Billeter, M. & Wüthrich, K. (1997). GARANT – a general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. *Journal of Computational Chemistry* **18**, 139–149.

Bartels, C., Xia, T. H., Billeter, M., Güntert, P. & Wüthrich, K. (1995). The program Xeasy for computer-supported NMR spectral-analysis of biological macromolecules. *Journal of Biomolecular NMR* **6**, 1–10.

Bax, A., Kontaxis, G. & Tjandra, N. (2001). Dipolar couplings in macromolecular structure determination. *Nuclear Magnetic Resonance of Biological Macromolecules, Pt B* **339**, 127–174.

Benod, C., Delsuc, M. A. & Pons, J. L. (2006). CRAACK: consensus program for NMR amino acid type assignment. *Journal of Chemical Information and Model* **46**, 1517–1522.

Berg, B. A. & Celik, T. (1992). New approach to spin-glass simulations. *Physical Review Letters* **69**, 2292–2295.

Berg, B. A. & Neuhaus, T. (1991). Multicanonical algorithms for 1st order phase-transitions. *Physics Letters B* **267**, 249–253.

Berjanskii, M. V., Neal, S. & Wishart, D. S. (2006). PREDITOR: a web server for predicting protein torsion angle restraints. *Nucleic Acids Research* **34**, W63–W69.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research* **28**, 235–242.

Bermejo, G. A. & Llinás, M. (2008). Deuterated protein folds obtained directly from unassigned nuclear Overhauser effect data. *Journal of the American Chemical Society* **130**, 3797–3805.

Bermejo, G. A. & Llinás, M. (2010). Structure-oriented methods for protein NMR data analysis. *Progress in Nuclear Magnetic Resonance Spectroscopy* **56**, 311–328.

Bertini, I., Bhaumik, A., De Paepe, G., Griffin, R. G., Lelli, M., Lewandowski, J. R. & Luchinat, C. (2010).

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

High-resolution solid-state NMR structure of a 17.6 kDa protein. *Journal of the American Chemical Society* **132**, 1032–1040.

Bertini, I., Luchinat, C. & Parigi, G. (2001). *Solution NMR of Paramagnetic Molecules: Applications to Metallobiomolecules and Models*. Amsterdam; New York: Elsevier Science Ltd.

Bhattacharya, A., Tejero, R. & Montelione, G. T. (2007). Evaluating protein structures determined by structural genomics consortia. *Proteins: Structure Function and Bioinformatics* **66**, 778–795.

Bhavesh, N. S., Panchal, S. C. & Hosur, R. V. (2001). An efficient high-throughput resonance assignment procedure for structural genomics and protein folding research by NMR. *Biochemistry* **40**, 14727–14735.

Billeter, M., Wagner, G. & Wüthrich, K. (2008). Solution NMR structure determination of proteins revisited. *Journal of Biomolecular NMR* **42**, 155–158.

Blackledge, M. (2005). Recent progress in the study of biomolecular structure and dynamics in solution from residual dipolar couplings. *Progress in Nuclear Magnetic Resonance Spectroscopy* **46**, 23–61.

Bockmann, A. (2008). 3D protein structures by solid-state NMR spectroscopy: ready for high resolution. *Angewandte Chemie* (*International Edition in English*) **47**, 6110–6113.

Bonvin, A. M., Rosato, A. & Wassenaar, T. A. (2010). The eNMR platform for structural biology. *Journal of Structural and Functional Genomics* **11**, 1–8.

Bouvignies, G., Markwick, P. R. & Blackledge, M. (2007). Simultaneous definition of high resolution protein structure and backbone conformational dynamics using NMR residual dipolar couplings. *ChemPhysChem* **8**, 1901–1909.

Bowers, P. M., Strauss, C. E. & Baker, D. (2000). De novo protein structure determination using sparse NMR data. *Journal of Biomolecular NMR* **18**, 311–318.

Bradley, P., Misura, K. M. & Baker, D. (2005). Toward high-resolution de novo structure prediction for small proteins. *Science* **309**, 1868–1871.

Brunger, A. (1992). *X-PLOR version 3.1: a System for X-ray Crystallography and NMR*. Yale: Yale University Press.

Brunger, A. T., Adams, P. D., Clore, G. M., Delano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). Crystallography and NMR system: a new software suite for macromolecular structure determination. *Acta crystallographica. Section D, Biological crystallography* **54**(Pt 5), 905–921.

Byeon, I. J., Meng, X., Jung, J., Zhao, G., Yang, R., Ahn, J., Shi, J., Concel, J., Aiken, C., Zhang, P. & Gronenborn, A. M. (2009). Structural convergence between Cryo-EM and NMR reveals intersubunit interactions critical for HIV-1 capsid function. *Cell* **139**, 780–790.

Case, D. A. (2002). Molecular dynamics and NMR spin relaxation in proteins. *Accounts of Chemical Research* **35**, 325–331.

Case, D. A., Cheatham, 3RD, T. E., Darden, T., Gohlke, H., Luo, R., Merz, JR., K. M., Onufriev, A., Simmerling, C., Wang, B. & Woods, R. J. (2005). The AMBER biomolecular simulation programs. *Journal of Computational Chemistry* **26**, 1668–1688.

Cavalli, A., Salvatella, X., Dobson, C. M. & Vendruscolo, M. (2007). Protein structure determination from NMR chemical shifts. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 9615–9620.

Cavanagh, J. (2007). *Protein NMR Spectroscopy: Principles and Practice*, 2nd edn. Amsterdam, Boston: Academic Press.

Celda, B., Biamonti, C., Arnau, M. J., Tejero, R. & Montelione, G. T. (1995). Combined use of 13C chemical shift and 1H alpha-13C alpha heteronuclear NOE data in monitoring a protein NMR structure refinement. *Journal of Biomolecular NMR* **5**, 161–172.

Cheung, M. S., Maguire, M. L., Stevens, T. J. & Broadhurst, R. W. (2010). DANGLE: A Bayesian inferential method for predicting protein backbone dihedral angles and secondary structure. *Journal of Magnetic Resonance* **202**, 223–233.

Chiang, T. H., Hsu, D. & Latombe, J. C. (2010). Markov dynamic models for long-timescale protein motion. *Bioinformatics* **26**, i269–i277.

Christen, M., Hunenberger, P. H., Bakowies, D., Baron, R., Burgi, R., Geerke, D. P., Heinz, T. N., Kastenholz, M. A., Krautler, V., Oostenbrink, C., Peter, C., Trzesniak, D. & Van Gunsteren, W. F. (2005). The GROMOS software for biomolecular simulation: GROMOS05. *Journal of Computational Chemistry* **26**, 1719–1751.

Clore, G. M. & Gronenborn, A. M. (1998). New methods of structure refinement for macromolecular structure determination by NMR. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 5891–5898.

Coggins, B. E. & Zhou, P. (2003). PACES: Protein sequential assignment by computer-assisted exhaustive search. *Journal of Biomolecular NMR* **26**, 93–111.

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., Popovic, Z. & Players, F. (2010). Predicting protein structures with a multiplayer online game. *Nature* **466**, 756–760.

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

Crippen, G. M., Rousaki, A., Revington, M., Zhang, Y. & Zuiderweg, E. R. (2010). SAGA: rapid automatic main chain NMR assignment for large proteins. *Journal of Biomolecular NMR* **46**, 281–298.

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

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

Dancea, F. & Gunther, U. (2005). Automated protein NMR structure determination using wavelet de-noised NOESY spectra. *Journal of Biomolecular NMR* **33**, 139–152.

Das, R. & Baker, D. (2008). Macromolecular modeling with Rosetta. *Annual Review of Biochemistry* **77**, 363–382.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

De Oliveira, J. F., Sforca, M. L., Blumenschein, T. M., Goldfeder, M. B., Guimaraes, B. G., Oliveira, C. C., Zanchin, N. I. & Zeri, A. C. (2010). Structure, dynamics, and RNA interaction analysis of the human SBDS protein. *Journal of Molecular Biology* **396**, 1053–1069.

Delaglio, F., Kontaxis, G. & Bax, A. (2000). Protein structure determination using molecular fragment replacement and NMR dipolar couplings. *Journal of the American Chemical Society* **122**, 2142–2143.

Donald, B. R. & Martin, J. (2009). Automated NMR assignment and protein structure determination using sparse dipolar coupling constraints. *Progress in Nuclear Magnetic Resonance Spectroscopy* **55**, 101–127.

Donoho, D. L. & Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–455.

Donoho, D. L. & Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* **90**, 1200–1224.

Doreleijers, J. F., Nederveen, A. J., Vranken, W., Lin, J., Bonvin, A. M., Kaptein, R., Markley, J. L. & Ulrich, E. L. (2005). BioMagResBank databases DOCR and FRED containing converted and filtered sets of experimental NMR restraints and coordinates from over 500 protein PDB structures. *Journal of Biomolecular NMR* **32**, 1–12.

Doreleijers, J. F., Rullmann, J. A. & Kaptein, R. (1998). Quality assessment of NMR structures: a statistical survey. *Journal of Molecular Biology* **281**, 149–164.

Doreleijers, J. F., Vriend, G., Raves, M. L. & Kaptein, R. (1999). Validation of nuclear magnetic resonance structures of proteins and nucleic acids: hydrogen geometry and nomenclature. *Proteins: Structure Function and Bioinformatics* **37**, 404–416.

Duggan, B. M., Legge, G. B., Dyson, H. J. & Wright, P. E. (2001). SANE (structure assisted NOE evaluation): an automated model-based approach for NOE assignment. *Journal of Biomolecular NMR* **19**, 321–329.

Earl, D. J. & Deem, M. W. (2005). Parallel tempering: theory, applications, and new perspectives. *Physical Chemistry Chemical Physics* **7**, 3910–3916.

Eghbalnia, H. R., Bahrami, A., Tonelli, M., Hallenga, K. & Markley, J. L. (2005a). High-resolution iterative frequency identification for NMR as a general strategy for multidimensional data collection. *Journal of the American Chemical Society* **127**, 12528–12536.

Eghbalnia, H. R., Bahrami, A., Wang, L., Assadi, A. & Markley, J. L. (2005b). Probabilistic identification of spin systems and their assignments including coil-helix inference as oOutput (PISTACHIO). *Journal of Biomolecular NMR* **32**, 219–233.

Elber, R. (2005). Long-timescale simulation methods. *Current Opinion in Structural Biology* **15**, 151–156.

Ernst, R. R., Bodenhausen, G. & Wokaun, A. (1987). Principles of Nuclear Magnetic Resonance in One and Two Dimensions. Oxford [Oxfordshire], New York: Clarendon Press; Oxford University Press.

Fan, H. & Mark, A. E. (2003). Relative stability of protein structures determined by X-ray crystallography or NMR spectroscopy: a molecular dynamics simulation study. *Proteins: Structure Function and Bioinformatics* **53**, 111–120.

Fiorito, F., Herrmann, T., Damberger, F. F. & 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. *Journal of Biomolecular NMR* **42**, 23–33.

Fogh, R., Ionides, J., Ulrich, E., Boucher, W., Vranken, W., Linge, J. P., Habeck, M., Rieping, W., Bhat, T. N., Westbrook, J., Henrick, K., Gilliland, G., Berman, H., Thornton, J., Nilges, M., Markley, J. & Laue, E. (2002). The CCPN project: an interim report on a data model for the NMR community. *Nature Structural Biology* **9**, 416–418.

Foley, J. D. (1996). *Computer Graphics: Principles and Practice*, 2nd ed. in C. edn. Reading, MA; Wokingham: Addison-Wesley.

Fujii, T., Iwane, A. H., Yanagida, T. & Namba, K. (2010). Direct visualization of secondary structures of F-actin by electron cryomicroscopy. *Nature* **467**, 724–728.

Gabel, F., Simon, B., Nilges, M., Petoukhov, M., Svergun, D. & Sattler, M. (2008). A structure refinement protocol combining NMR residual dipolar couplings and small angle scattering restraints. *Journal of Biomolecular NMR* **41**, 199–208.

Ginzinger, S. W., Weichenberger, C. X. & Sippl, M. J. (2010). Detection of unrealistic molecular environments in protein structures based on expected electron densities. *Journal of Biomolecular NMR* **47**, 33–40.

Goddard, T. D. & Kneller, D. G. (2008). Sparky 3. San Francisco, CA: University of California.

Gong, H. P., Shen, Y. & Rose, G. D. (2007). Building native protein conformation from NMR backbone chemical shifts using Monte Carlo fragment assembly. *Protein Science* **16**, 1515–1521.

Grishaev, A. & Llinás, M. (2002a). CLOUDS, a protocol for deriving a molecular proton density via NMR. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 6707–6712.

Grishaev, A. & Llinás, M. (2002b). Protein structure elucidation from NMR proton densities. *Proceedings of the*

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

National Academy of Sciences of the United States of America **99**, 6713–6718.

Grishaev, A. & Llinás, M. (2004). BACUS: A Bayesian protocol for the identification of protein NOESY spectra via unassigned spin systems. *Journal of Biomolecular NMR* **28**, 1–10.

Grishaev, A., Tugarinov, V., Kay, L. E., Trewhella, J. & Bax, A. (2008a). Refined solution structure of the 82-kDa enzyme malate synthase G from joint NMR and synchrotron SAXS restraints. *Journal of Biomolecular NMR* **40**, 95–106.

Grishaev, A., Wu, J., Trewhella, J. & Bax, A. (2005). Refinement of multidomain protein structures by combination of solution small-angle X-ray scattering and NMR data. *Journal of the American Chemical Society* **127**, 16621–16628.

Grishaev, A., Ying, J., Canny, M. D., Pardi, A. & Bax, A. (2008b). Solution structure of tRNAVal from refinement of homology model against residual dipolar coupling and SAXS data. *Journal of Biomolecular NMR* **42**, 99–109.

Gronwald, W. & Kalbitzer, H. R. (2004). Automated structure determination of proteins by NMR spectroscopy. *Progress in Nuclear Magnetic Resonance Spectroscopy* **44**, 33–96.

Gronwald, W., Kirchhofer, R., Gorler, A., Kremer, W., Ganslmeier, B., Neidig, K. P. & Kalbitzer, H. R. (2000). RFAC, a program for automated NMR R-factor estimation. *Journal of Biomolecular NMR* **17**, 137–151.

Gronwald, W., Moussa, S., Elsner, R., Jung, A., Ganslmeier, B., Trenner, J., Kremer, W., Neidig, K. P. & Kalbitzer, H. R. (2002). Automated assignment of NOESY NMR spectra using a knowledge based method (KNOWNOE). *Journal of Biomolecular NMR* **23**, 271–287.

Gryk, M. R., Vyas, J. & Maciejewski, M. W. (2010). Biomolecular NMR data analysis. *Progress in Nuclear Magnetic Resonance Spectroscopy* **56**, 329–345.

Güntert, P. (1998). Structure calculation of biological macromolecules from NMR data. *Quarterly Reviews of Biophysics* **31**, 145–237.

Güntert, P. (2003). Automated NMR protein structure calculation. *Progress in Nuclear Magnetic Resonance Spectroscopy* **43**, 105–125.

Güntert, P. (2004). Automated NMR structure calculation with CYANA. *Methods in Molecular Biology* **278**, 353–378.

Güntert, P., Berndt, K. D. & Wüthrich, K. (1993). The program Asno for computer-supported collection of NOE upper distance constraints as input for protein-structure determination. *Journal of Biomolecular NMR* **3**, 601–606.

Güntert, P., Braun, W. & Wüthrich, K. (1991). Efficient computation of 3-dimensional protein structures in solution from nuclear magnetic resonance data using the program DIANA and the supporting programs CALIBA, HABAS and GLOMSA. *Journal of Molecular Biology* **217**, 517–530.

Güntert, P., Mumenthaler, C. & Wüthrich, K. (1997). Torsion angle dynamics for NMR structure calculation with the new program DYANA. *Journal of Molecular Biology* **273**, 283–298.

Güntert, P., Salzmann, M., Braun, D. & Wüthrich, K. (2000). Sequence-specific NMR assignment of proteins by global fragment mapping with the program MAPPER. *Journal of Biomolecular NMR* **18**, 129–137.

Habeck, M., Nilges, M. & Rieping, W. (2005a). Bayesian inference applied to macromolecular structure determination. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* **72**(3 Pt 1), 031912.

Habeck, M., Nilges, M. & Rieping, W. (2005b). Replica-exchange Monte Carlo scheme for Bayesian data analysis. *Physical Review Letters* **94**, 018105.

Habeck, M., Nilges, M. & Rieping, W. (2008). A unifying probabilistic framework for analyzing residual dipolar couplings. *Journal of Biomolecular NMR* **40**, 135–144.

Habeck, M., Rieping, W. & Nilges, M. (2005c). Bayesian estimation of Karplus parameters and torsion angles from three-bond scalar couplings constants. *Journal of Magnetic Resonance* **177**, 160–165.

Habeck, M., Rieping, W. & Nilges, M. (2006). Weighting of experimental evidence in macromolecular structure determination. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 1756–1761.

Hamelberg, D., Mongan, J. & Mccammon, J. A. (2004). Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *The Journal of Chemical Physics* **120**, 11919–11929.

Herrmann, T., Güntert, P. & Wüthrich, K. (2002a). Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. *Journal of Biomolecular NMR* **24**, 171–189.

Herrmann, T., Güntert, P. & Wüthrich, K. (2002b). Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. *Journal of Biomolecular NMR* **24**, 171–189.

Hiller, S., Fiorito, F., Wüthrich, K. & Wider, G. (2005). Automated projection spectroscopy (APSY). *Proceedings of the National Academy of Sciences of the United States of America* **102**, 10876–10881.

Hiller, S., Wider, G. & Wüthrich, K. (2008). APSY-NMR with proteins: practical aspects and backbone assignment. *Journal of Biomolecular NMR* **42**, 179–195.

Hitchens, T. K., Lukin, J. A., Zhan, Y., Mccallum, S. A. & Rule, G. S. (2003). MONTE: An automated Monte Carlo based approach to nuclear magnetic resonance assignment of proteins. *Journal of Biomolecular NMR* **25**, 1–9.

Hooft, R. W., Vriend, G., Sander, C. & Abola, E. E. (1996). Errors in protein structures. *Nature* **381**, 272.

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A. & Simmerling, C. (2006). Comparison of multiple AMBER force fields and development of improved protein backbone parameters. *Proteins: Structure Function and Bioinformatics* **65**, 712–725.

Huang, Y. J., Powers, R. & Montelione, G. T. (2005). Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics. *Journal of the American Chemical Society* **127**, 1665–1674.

Huang, Y. J., Tejero, R., Powers, R. & Montelione, G. T. (2006). A topology-constrained distance network algorithm for protein structure determination from NOESY data. *Proteins: Structure Function and Bioinformatics* **62**, 587–603.

Hung, L. H. & Samudrala, R. (2006). An automated assignment-free Bayesian approach for accurately identifying proton contacts from NOESY data. *Journal of Biomolecular NMR* **36**, 189–198.

Hus, J. C., Marion, D. & Blackledge, M. (2001). Determination of protein backbone structure using only residual dipolar couplings. *Journal of the American Chemical Society* **123**, 1541–1542.

Hus, J. C., Prompers, J. J. & Bruschweiler, R. (2002). Assignment strategy for proteins with known structure. *Journal of Magnetic Resonance* **157**, 119–123.

Hus, J. C., Salmon, L., Bouvignies, G., Lotze, J., Blackledge, M. & Bruschweiler, R. (2008). 16-fold degeneracy of peptide plane orientations from residual dipolar couplings: analytical treatment and implications for protein structure determination. *Journal of the American Chemical Society* **130**, 15927–15937.

Ikeya, T., Takeda, M., Yoshida, H., Terauchi, T., Jee, J. G., Kainosho, M. & Güntert, P. (2009). Automated NMR structure determination of stereo-array isotope labeled ubiquitin from minimal sets of spectra using the SAIL-FLYA system. *Journal of Biomolecular NMR* **44**, 261–272.

Ikeya, T., Terauchi, T., Güntert, P. & Kainosho, M. (2006). Evaluation of stereo-array isotope labeling (SAIL) patterns for automated structural analysis of proteins with CYANA. *Magnetic Resonance in Chemistry* **44** (Spec No), S152–S157.

Iwadate, M., Asakura, T. & Williamson, M. P. (1999). C-alpha and C-beta carbon-13 chemical shifts in proteins from an empirical database. *Journal of Biomolecular NMR* **13**, 199–211.

Jee, J. & Güntert, P. (2003). Influence of the completeness of chemical shift assignments on NMR structures obtained with automated NOE assignment. *Journal of Structural and Functional Genomics* **4**, 179–189.

Jehle, S., Rajagopal, P., Bardiaux, B., Markovic, S., Kuhne, R., Stout, J. R., Higman, V. A., Klevit, R. E., Van Rossum, B. J. & Oschkinat, H. (2010). Solid-state NMR and SAXS studies provide a structural basis for the activation of alphaB-crystallin oligomers. *Nature Structural & Molecular Biology* **17**, 1037–1042.

Johnson, B. A. & Blevins, R. A. (1994). NMR View – a computer-program for the visualization and analysis of NMR data. *Journal of Biomolecular NMR* **4**, 603–614.

Jung, Y. S., Sharma, M. & Zweckstetter, M. (2004). Simultaneous assignment and structure determination of protein backbones by using NMR dipolar couplings. *Angewandte Chemie* (*International Edition in English*) **43**, 3479–3481.

Jung, Y. S. & Zweckstetter, M. (2004a). Backbone assignment of proteins with known structure using residual dipolar couplings. *Journal of Biomolecular NMR* **30**, 25–35.

Jung, Y. S. & Zweckstetter, M. (2004b). Mars – robust automatic backbone assignment of proteins. *Journal of Biomolecular NMR* **30**, 11–23.

Kainosho, M. & Güntert, P. (2009). SAIL–stereo-array isotope labeling. *Quarterly Reviews of Biophysics* **42**, 247–300.

Kainosho, M., Torizawa, T., Iwashita, Y., Terauchi, T., Ono, A. M. & Güntert, P. (2006). Optimal isotope labelling for NMR protein structure determinations. *Nature* **440**, 52–57.

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

Keizers, P. H. J. & Ubbink, M. (2010). Paramagnetic tagging for protein structure and dynamics analysis. *Progress in Nuclear Magnetic Resonance Spectroscopy* **58**, 88–96.

Kim, Y. & Prestegard, J. H. (1989). A dynamic model for the structure of acyl carrier protein in solution. *Biochemistry* **28**, 8792–8797.

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. *Journal of Biomolecular NMR* **39**, 31–52.

Kohlhoff, K. J., Robustelli, P., Cavalli, A., Salvatella, X. & Vendruscolo, M. (2009). Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. *Journal of the American Chemical Society* **131**, 13894.

Kontaxis, G., Delaglio, F. & Bax, A. (2005). Molecular fragment replacement approach to protein structure determination by chemical shift and dipolar homology database mining. *Nuclear Magnetic Resonance of Biological Macromolecules, Part C* **394**, 42.

Koradi, R., Billeter, M., Engeli, M., Güntert, P. & Wüthrich, K. (1998). Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. *Journal of Magnetic Resonance* **135**, 288–297.

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

Koradi, R., Billeter, M. & Güntert, P. (2000). Point-centered domain decomposition for parallel molecular dynamics simulation. *Computer Physics Communications* **124**, 139–147.

Korukottu, J., Bayrhuber, M., Montaville, P., Vijayan, V., Jung, Y. S., Becker, S. & Zweckstetter, M. (2007). Fast high-resolution protein structure determination by using unassigned NMR data. *Angewandte Chemie* (*International Edition in English*) **46**, 1176–1179.

Korzhnev, D. M. & Kay, L. E. (2008). Probing invisible, low-populated States of protein molecules by relaxation dispersion NMR spectroscopy: an application to protein folding. *Accounts of Chemical Research* **41**, 442–451.

Korzhnev, D. M., Religa, T. L., Banachewicz, W., Fersht, A. R. & Kay, L. E. (2010). A transient and low-populated protein-folding intermediate at atomic resolution. *Science* **329**, 1312–1316.

Korzhnev, D. M., Salvatella, X., Vendruscolo, M., Di Nardo, A. A., Davidson, A. R., Dobson, C. M. & Kay, L. E. (2004). Low-populated folding intermediates of Fyn SH3 characterized by relaxation dispersion NMR. *Nature* **430**, 586–590.

Korzhneva, D. M., Ibraghimov, I. V., Billeter, M. & Orekhov, V. Y. (2001). MUNIN: application of three-way decomposition to the analysis of heteronuclear NMR relaxation data. *Journal of Biomolecular NMR* **21**, 263–268.

Kuszewski, J., Gronenborn, A. M. & Clore, G. M. (1995a). The impact of direct refinement against proton chemical shifts on protein structure determination by NMR. *Journal of Magnetic Resonance Series B* **107**, 293–297.

Kuszewski, J., Gronenborn, A. M. & Clore, G. M. (1995b). The impact of direct refinement against proton chemical-shifts on protein-structure determination by NMR. *Journal of Magnetic Resonance Series B* **107**, 293–297.

Kuszewski, J., Qin, J., Gronenborn, A. M. & Clore, G. M. (1995c). The impact of direct refinement against 13C alpha and 13C beta chemical shifts on protein structure determination by NMR. *Journal of Magnetic Resonance Series B* **106**, 92–96.

Kuszewski, J., Schwieters, C. D., Garrett, D. S., Byrd, R. A., Tjandra, N. & Clore, G. M. (2004). Completely automated, highly error-tolerant macromolecular structure determination from multidimensional nuclear Overhauser enhancement spectra and chemical shift assignments. *Journal of the American Chemical Society* **126**, 6258–6273.

Langmead, C. J. & Donald, B. R. (2004). An expectation/maximization nuclear vector replacement algorithm for automated NMR resonance assignments. *Journal of Biomolecular NMR* **29**, 111–138.

Langmead, C. J., Yan, A., Lilien, R., Wang, L. & Donald, B. R. (2004). A polynomial-time nuclear vector replacement algorithm for automated NMR resonance assignments. *Journal of Computational Biology* **11**, 277–298.

Laskowski, R. A. (2003). Structural quality assurance. *Methods of Biochemical Analysis* **44**, 273–303.

Laskowski, R. A., Macarthur, M. W., Moss, D. S. & Thornton, J. M. (1993). Procheck – a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography* **26**, 283–291.

Laskowski, R. A., Rullmannn, J. A., Macarthur, M. W., Kaptein, R. & Thornton, J. M. (1996). AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *Journal of Biomolecular NMR* **8**, 477–486.

Ledford, H. (2010). Protein mapping gains a human focus. *Nature* **466**, 544.

Lee, S., Mesleh, M. F. & Opella, S. J. (2003). Structure and dynamics of a membrane protein in micelles from three solution NMR experiments. *Journal of Biomolecular NMR* **26**, 327–334.

Lehtivarjo, J., Hassinen, T., Korhonen, S. P., Perakyla, M. & Laatikainen, R. (2009). 4D prediction of protein (1)H chemical shifts. *Journal of Biomolecular NMR* **45**, 413–426.

Lemak, A., Steren, C. A., Arrowsmith, C. H. & Llinás, M. (2008). Sequence specific resonance assignment via multicanonical Monte Carlo search using an ABACUS approach. *Journal of Biomolecular NMR* **41**, 29–41.

Leutner, M., Gschwind, R. M., Liermann, J., Schwarz, C., Gemmecker, G. & Kessler, H. (1998). Automated backbone assignment of labeled proteins using the threshold accepting algorithm. *Journal of Biomolecular NMR* **11**, 31–43.

Levitt, M. (2009). Nature of the protein universe. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 11079–11084.

Li, K. B. & Sanctuary, B. C. (1997a). Automated resonance assignment of proteins using heteronuclear 3D NMR. 2. Side chain and sequence-specific assignment. *Journal of Chemical Information and Computer Sciences* **37**, 467–477.

Li, K. B. & Sanctuary, B. C. (1997b). Automated resonance assignment of proteins using heteronuclear 3D NMR. Backbone spin systems extraction and creation of polypeptides. *Journal of Chemical Information and Computer Sciences* **37**, 359–366.

Lin, H. N., Wu, K. P., Chang, J. M., Sung, T. Y. & Hsu, W. L. (2005). GANA – a genetic algorithm for NMR backbone resonance assignment. *Nucleic Acids Research* **33**, 4593–4601.

Lindorff-Larsen, K., Best, R. B., Depristo, M. A., Dobson, C. M. & Vendruscolo, M. (2005). Simultaneous determination of protein structure and dynamics. *Nature* **433**, 128–132.

Linge, J. P., Habeck, M., Rieping, W. & Nilges, M. (2004). Correction of spin diffusion during iterative automated NOE assignment. *Journal of Magnetic Resonance* **167**, 334–342.

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

Linge, J. P. & Nilges, M. (1999). Influence of non-bonded parameters on the quality of NMR structures: a new force field for NMR structure calculation. *Journal of Biomolecular NMR* **13**, 51–59.

Linge, J. P., O'Donoghue, S. I. & Nilges, M. (2001). Automated assignment of ambiguous nuclear Overhauser effects with ARIA. *Methods in Enzymology* **339**, 71–90.

Linge, J. P., Williams, M. A., Spronk, C. A., Bonvin, A. M. & Nilges, M. (2003). Refinement of protein structures in explicit solvent. *Proteins-Structure Function and Bioinformatics* **50**, 496–506.

Lopez-Mendez, B. & Güntert, P. (2006). Automated protein structure determination from NMR spectra. *Journal of the American Chemical Society* **128**, 13112–13122.

Loquet, A., Bardiaux, B., Gardiennet, C., Blanchet, C., Baldus, M., Nilges, M., Malliavin, T. & Bockmann, A. (2008). 3D structure determination of the Crh protein from highly ambiguous solid-state NMR restraints. *Journal of the American Chemical Society* **130**, 3579–3589.

Loquet, A., Gardiennet, C. & Bockmann, A. (2010). Protein 3D structure determination by high-resolution solid-state NMR. *Comptes Rendus Chimie* **13**, 423–430.

Lovell, S. C., Davis, I. W., Arendall, 3RD, W. B., De Bakker, P. I., Word, J. M., Prisant, M. G., Richardson, J. S. & Richardson, D. C. (2003). Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins: Structure Function and Bioinformatics* **50**, 437–450.

Mackerell, Jr., A. D. (2004). Empirical force fields for biological macromolecules: overview and issues. *Journal of Computational Chemistry* **25**, 1584–1604.

Mallat, S. G. (1999). *A wavelet Tour of Signal Processing*, 2nd edn. San Diego, CA: Academic Press.

Malliavin, T. E., Rouh, A., Delsuc, M. A. & Lallemand, J. Y. (1992). A Direct approach of molecular-structures determination from nuclear Overhauser effect. *Comptes Rendus De L Academie Des Sciences Serie Ii* **315**, 653–659.

Malmodin, D. & Billeter, M. (2005). High-throughput analysis of protein NMR spectra. *Progress in Nuclear Magnetic Resonance Spectroscopy* **46**, 109–129.

Malmodin, D., Papavoine, C. H. M. & Billeter, M. (2003). Fully automated sequence-specific resonance assignments of heteronuclear protein spectra. *Journal of Biomolecular NMR* **27**, 69–79.

Manolikas, T., Herrmann, T. & Meier, B. H. (2008). Protein structure determination from C-13 spin-diffusion solid-state NMR spectroscopy. *Journal of the American Chemical Society* **130**, 3959–3966.

Markwick, P. R., Bouvignies, G., Salmon, L., Mccammon, J. A., Nilges, M. & Blackledge, M. (2009). Toward a unified representation of protein structural dynamics in solution. *Journal of the American Chemical Society* **131**, 16968–16975.

Markwick, P. R., Malliavin, T. & Nilges, M. (2008). Structural biology by NMR: structure, dynamics, and interactions. *PLoS Computational Biology* **4**, e1000168.

Markwick, P. R., Sprangers, R. & Sattler, M. (2005). Local structure and anisotropic backbone dynamics from cross-correlated NMR relaxation in proteins. *Angewandte Chemie* (*International Edition in English*) **44**, 3232–3237.

Mascioni, A. & Veglia, G. (2003). Theoretical analysis of residual dipolar coupling patterns in regular secondary structures of proteins. *Journal of the American Chemical Society* **125**, 12520–12526.

Mcguffin, L. J., Bryson, K. & Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404–405.

Meiler, J. (2003). PROSHIFT: protein chemical shift prediction using artificial neural networks. *Journal of Biomolecular NMR* **26**, 25–37.

Meiler, J. & Baker, D. (2003). Rapid protein fold determination using unassigned NMR data. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 15404–15409.

Mesleh, M. F., Lee, S., Veglia, G., Thiriot, D. S., Marassi, F. M. & Opella, S. J. (2003). Dipolar waves map the structure and topology of helices in membrane proteins. *Journal of the American Chemical Society* **125**, 8928–8935.

Mesleh, M. F. & Opella, S. J. (2003). Dipolar Waves as NMR maps of helices in proteins. *Journal of Magnetic Resonance* **163**, 288–299.

Mesleh, M. F., Veglia, G., Desilva, T. M., Marassi, F. M. & Opella, S. J. (2002). Dipolar waves as NMR maps of protein structure. *Journal of the American Chemical Society* **124**, 4206–4207.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092.

Mielke, S. P. & Krishnan, V. V. (2009). Characterization of protein secondary structure from NMR chemical shifts. *Progress in Nuclear Magnetic Resonance Spectroscopy* **54**, 141–165.

Moseley, H. N., Monleon, D. & Montelione, G. T. (2001). Automatic determination of protein backbone resonance assignments from triple resonance nuclear magnetic resonance data. *Methods in Enzymology* **339**, 91–108.

Moseley, H. N. & Montelione, G. T. (1999). Automated analysis of NMR assignments and structures for proteins. *Current Opinion in Structural Biology* **9**, 635–642.

Mulder, F. A. A. & Filatov, M. (2010). NMR chemical shift data and *ab initio* shielding calculations: emerging tools for protein structure determination. *Chemical Society Reviews* **39**, 578–590.

Mumenthaler, C. & Braun, W. (1995). Automated assignment of simulated and experimental NOESY

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

spectra of proteins by feedback filtering and self-correcting distance geometry. *Journal of Molecular Biology* **254**, 465–480.

Mumenthaler, C., Güntert, P., Braun, W. & Wüthrich, K. (1997). Automated combined assignment of NOESY spectra and three-dimensional protein structure determination. *Journal of Biomolecular NMR* **10**, 351–362.

Nabuurs, S. B., Nederveen, A. J., Vranken, W., Doreleijers, J. F., Bonvin, A. M., Vuister, G. W., Vriend, G. & Spronk, C. A. (2004a). DRESS: a database of refined solution NMR structures. *Proteins-Structure Function and Bioinformatics* **55**, 483–486.

Nabuurs, S. B., Spronk, C. A., Krieger, E., Maassen, H., Vriend, G. & Vuister, G. W. (2003). Quantitative evaluation of experimental NMR restraints. *Journal of the American Chemical Society* **125**, 12026–12034.

Nabuurs, S. B., Spronk, C. A., Vuister, G. W. & Vriend, G. (2006). Traditional biomolecular structure determination by NMR spectroscopy allows for major errors. *PLoS Computational Biology* **2**, e9.

Nabuurs, S. B., Spronk, C. A. E. M., Vriend, G. & Vuister, G. W. (2004b). Concepts and tools for NMR restraint analysis and validation. *Concepts in Magnetic Resonance Part A* **22A**, 90–105.

Neal, S., Nip, A. M., Zhang, H. Y. & Wishart, D. S. (2003). Rapid and accurate calculation of protein H-1, C-13 and N-15 chemical shifts. *Journal of Biomolecular NMR* **26**, 215–240.

Nederveen, A. J., Doreleijers, J. F., Vranken, W., Miller, Z., Spronk, C. A., Nabuurs, S. B., Güntert, P., Livny, M., Markley, J. L., Nilges, M., Ulrich, E. L., Kaptein, R. & Bonvin, A. M. (2005). RECOORD: a recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank. *Proteins: Structure Function and Bioinformatics* **59**, 662–672.

Neidig, K. P., Geyer, M., Gorler, A., Antz, C., Saffrich, R., Beneicke, W. & Kalbitzer, H. R. (1995). Aurelia, a program for computer-aided analysis of multi-dimensional NMR-spectra. *Journal of Biomolecular NMR* **6**, 255–270.

Neuhaus, D. & Williamson, M. P. (2000). *The Nuclear Overhauser Effect in Structural and Conformational Analysis*, 2nd edn. New York: Wiley.

Nilges, M. (1993). A calculation strategy for the structure determination of symmetric dimers by 1H NMR. *Proteins: Structure Function and Bioinformatics* **17**, 297–309.

Nilges, M. (1995). Calculation of protein structures with ambiguous distance restraints. Automated assignment of ambiguous NOE crosspeaks and disulphide connectivities. *Journal of Molecular Biology* **245**, 645–660.

Nilges, M. & O'Donoghue, S. I. (1998). Ambiguous NOEs and automated NOE assignment. *Progress in Nuclear Magnetic Resonance Spectroscopy* **32**, 107–139.

Nocedal, J. & Wright, S. J. (2006). *Numerical Optimization*, 2nd ed. New York, London: Springer.

Ohki, S. Y. & Kainosho, M. (2008). Stable isotope labeling methods for protein NMR spectroscopy. *Progress in Nuclear Magnetic Resonance Spectroscopy* **53**, 208–226.

Olson, J. B. & Markley, J. L. (1994). Evaluation of an algorithm for the automated sequential assignment of protein backbone resonances – a demonstration of the connectivity tracing assignment tools (Contrast) software package. *Journal of Biomolecular NMR* **4**, 385–410.

Orekhov, V. Y., Ibraghimov, I. V. & Billeter, M. (2001). MUNIN: a new approach to multi-dimensional NMR spectra interpretation. *Journal of Biomolecular NMR* **20**, 49–60.

Ösapay, K. & Case, D. A. (1991). A new analysis of proton chemical-shifts in proteins. *Journal of the American Chemical Society* **113**, 9436–9444.

Ösapay, K., Theriault, Y., Wright, P. E. & Case, D. A. (1994). Solution structure of carbonmonoxy myoglobin determined from nuclear magnetic resonance distance and chemical shift constraints. *Journal of Molecular Biology* **244**, 183–197.

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kale, L. & Schulten, K. (2005). Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry* **26**, 1781–1802.

Prestegard, J. H., Bougault, C. M. & Kishore, A. I. (2004). Residual dipolar couplings in structure determination of biomolecules. *Chem Rev* **104**, 3519–3540.

Pugalenthi, G., Shameer, K., Srinivasan, N. & Sowdhamini, R. (2006). HARMONY: a server for the assessment of protein structures. *Nucleic Acids Research* **34**, W231–W234.

Qian, B., Raman, S., Das, R., Bradley, P., Mccoy, A. J., Read, R. J. & Baker, D. (2007). High-resolution structure prediction and the crystallographic phase problem. *Nature* **450**, 259–264.

Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology* **7**(1), 95–&.

Raman, S., Huang, Y. J. P., Mao, B. C., Rossi, P., Aramini, J. M., Liu, G. H., Montelione, G. T. & Baker, D. (2010a). Accurate automated protein NMR structure determination using unassigned NOESY data. *Journal of the American Chemical Society* **132**, 202–207.

Raman, S., Lange, O. F., Rossi, P., Tyka, M., Wang, X., Aramini, J., Liu, G. H., Ramelot, T. A., Eletsky, A., Szyperski, T., Kennedy, M. A., Prestegard, J., Montelione, G. T. & Baker, D. (2010b). NMR structure determination for larger proteins using backbone-only data. *Science* **327**, 1014–1018.

Ramelot, T. A., Raman, S., Kuzin, A. P., Xiao, R., Ma, L. C., Acton, T. B., Hunt, J. F., Montelione, G. T., Baker, D. & Kennedy, M. A. (2009). Improving NMR protein structure quality by Rosetta refinement: a

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

molecular replacement study. *Proteins: Structure Function and Bioinformatics* **75**, 147–167.

Ramirez, B. E. & Bax, A. (1998). Modulation of the alignment tensor of macromolecules dissolved in a dilute liquid crystalline medium. *Journal of the American Chemical Society* **120**, 9106–9107.

Ratnaparkhi, G. S., Ramachandran, S., Udgaonkar, J. B. & Varadarajan, R. (1998). Discrepancies between the NMR and X-ray structures of uncomplexed barstar: analysis suggests that packing densities of protein structures determined by NMR are unreliable. *Biochemistry* **37**, 6958–6966.

Reed, M. A., Hounslow, A. M., Sze, K. H., Barsukov, I. G., Hosszu, L. L., Clarke, A. R., Craven, C. J. & Waltho, J. P. (2003). Effects of domain dissection on the folding and stability of the 43 kDa protein PGK probed by NMR. *Journal of Molecular Biology* **330**, 1189–1201.

Rieping, W., Habeck, M. & Nilges, M. (2005). Inferential structure determination. *Science* **309**, 303–306.

Rieping, W., Nilges, M. & Habeck, M. (2008). ISD: a software package for Bayesian NMR structure calculation. *Bioinformatics* **24**, 1104–1105.

Ritchie, D. W. (2008). Recent progress and future directions in protein–protein docking. *Current Protein and Peptide Science* **9**, 1–15.

Robustelli, P., Cavalli, A., Dobson, C. M., Vendruscolo, M. & Salvatella, X. (2009). Folding of small proteins by Monte Carlo simulations with chemical shift restraints without the use of molecular fragment replacement or structural homology. *Journal of Physical Chemistry B* **113**, 7890–7896.

Robustelli, P., Cavalli, A. & Vendruscolo, M. (2008). Determination of protein structures in the solid state from NMR chemical shifts. *Structure* **16**, 1764–1769.

Robustelli, P., Kohlhoff, K., Cavalli, A. & Vendruscolo, M. (2010). Using NMR chemical shifts as structural restraints in molecular dynamics simulations of proteins. *Structure* **18**, 923–933.

Rohl, C. A. (2005). Protein structure estimation from minimal restraints using Rosetta. *Methods in Enzymology* **394**, 244–260.

Rohl, C. A. & Baker, D. (2002). De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. *Journal of the American Chemical Society* **124**, 2723–2729.

Rosato, A., Bagaria, A., Baker, D., Bardiaux, B., Cavalli, A., Doreleijers, J. F., Giachetti, A., Guerry, P., Güntert, P., Herrmann, T., Huang, Y. J., Jonker, H. R., Mao, B., Malliavin, T. E., Montelione, G. T., Nilges, M., Raman, S., Van Der Schot, G., Vranken, W. F., Vuister, G. W. & Bonvin, A. M. (2009). CASD-NMR: critical assessment of automated structure determination by NMR. *Nature Methods* **6**, 625–626.

Saccenti, E. & Rosato, A. (2008). The war of tools: how can NMR spectroscopists detect errors in their structures? *Journal of Biomolecular NMR* **40**, 251–261.

Schmucki, R., Yokoyama, S. & Güntert, P. (2009). Automated assignment of NMR chemical shifts using peak-particle dynamics simulation with the DYNASSIGN algorithm. *Journal of Biomolecular NMR* **43**, 97–109.

Schwieters, C. D., Kuszewski, J. J. & Clore, G. M. (2006). Using Xplor-NIH for NMR molecular structure determination. *Progress in Nuclear Magnetic Resonance Spectroscopy* **48**, 47–62.

Schwieters, C. D., Kuszewski, J. J., Tjandra, N. & Clore, G. M. (2003). The Xplor-NIH NMR molecular structure determination package. *Journal of Magnetic Resonance* **160**, 65–73.

Serban, N. (2007). MICE: multiple-peak identification, characterization, and estimation. *Biometrics* **63**, 531–539.

Shen, Y. & Bax, A. (2007). Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *Journal of Biomolecular NMR* **38**, 289–302.

Shen, Y. & Bax, A. (2010). Prediction of Xaa-Pro peptide bond conformation from sequence and chemical shifts. *Journal of Biomolecular NMR* **46**, 199–204.

Shen, Y., Delaglio, F., Cornilescu, G. & Bax, A. (2009a). TALOS plus: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *Journal of Biomolecular NMR* **44**, 213–223.

Shen, Y., Lange, O., Delaglio, F., Rossi, P., Aramini, J. M., Liu, G., Eletsky, A., Wu, Y., Singarapu, K. K., Lemak, A., Ignatchenko, A., Arrowsmith, C. H., Szyperski, T., Montelione, G. T., Baker, D. & Bax, A. (2008a). Consistent blind protein structure generation from NMR chemical shift data. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 4685–4690.

Shen, Y., Lange, O., Delaglio, F., Rossi, P., Aramini, J. M., Liu, G. H., Eletsky, A., Wu, Y. B., Singarapu, K. K., Lemak, A., Ignatchenko, A., Arrowsmith, C. H., Szyperski, T., Montelione, G. T., Baker, D. & Bax, A. (2008b). Consistent blind protein structure generation from NMR chemical shift data. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 4685–4690.

Shen, Y., Vernon, R., Baker, D. & Bax, A. (2009b). De novo protein structure generation from incomplete chemical shift assignments. *Journal of Biomolecular NMR* **43**, 63–78.

Shin, J. & Lee, W. (2008). Structural proteomics by NMR spectroscopy. *Expert Rev Proteomics* **5**, 589–601.

Sims, G. E. & Kim, S. H. (2006). A method for evaluating the structural quality of protein models by using higher-order phi-psi pairs scoring. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 4428–4432.

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

Sippl, M. J. (1993). Recognition of errors in 3-dimensional structures of proteins. *Proteins-Structure Function and Genetics* **17**, 355–362.

Snyder, D. A., Bhattacharya, A., Huang, Y. J. & Montelione, G. T. (2005). Assessing precision and accuracy of protein structures derived from NMR data. *Proteins: Structure Function and Bioinformatics* **59**, 655–661.

Spronk, C. A., Linge, J. P., Hilbers, C. W. & Vuister, G. W. (2002). Improving the quality of protein structures derived by NMR spectroscopy. *Journal of Biomolecular NMR* **22**, 281–289.

Spronk, C. A. E. M., Nabuurs, S. B., Krieger, E., Vriend, G. & Vuister, G. W. (2004). Validation of protein structures derived by NMR spectroscopy. *Progress in Nuclear Magnetic Resonance Spectroscopy* **45**, 315–337.

Staykova, D. K., Fredriksson, J., Bermel, W. & Billeter, M. (2008). Assignment of protein NMR spectra based on projections, multi-way decomposition and a fast correlation approach. *Journal of Biomolecular NMR* **42**, 87–97.

Steen, H. & Mann, M. (2004). The ABCs (and XYZs) of peptide sequencing. *Nature Reviews. Molecular Cell Biology* **5**, 699–711.

Stratmann, D., Guittet, E. & Van Heijenoort, C. (2010). Robust structure-based resonance assignment for functional protein studies by NMR. *Journal of Biomolecular NMR* **46**, 157–173.

Sundström, M., Norin, M. & Edwards, A. (2006). *Structural Genomics and high Throughput Structural Biology*. Boca Raton, FL: Taylor&Francis.

Swendsen, R. H. & Wang, J. S. (1986). Replica Monte-Carlo simulation of spin-glasses. *Physical Review Letters* **57**, 2607–2609.

Szilagyi, L. (1995). Chemical-shifts in proteins come of age. *Progress in Nuclear Magnetic Resonance Spectroscopy* **27**, 325–443.

Takeda, M., Chang, C. K., Ikeya, T., Güntert, P., Chang, Y. H., Hsu, Y. L., Huang, T. H. & Kainosho, M. (2008). Solution structure of the c-terminal dimerization domain of SARS coronavirus nucleocapsid protein solved by the SAIL-NMR method. *Journal of Molecular Biology* **380**, 608–622.

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

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

Terwilliger, T. C., Stuart, D. & Yokoyama, S. (2009). Lessons from structural genomics. *Annual Review of Biophysics* **38**, 371–383.

Torda, A. E., Scheek, R. M. & Van Gunsteren, W. F. (1990). Time-averaged nuclear Overhauser effect distance restraints applied to tendamistat. *Journal of Molecular Biology* **214**, 223–235.

Tosatto, S. C. & Battistutta, R. (2007). TAP score: torsion angle propensity normalization applied to local protein structure evaluation. *BMC Bioinformatics* **8**, 155.

Tosatto, S. C. E. (2005). The Victor/FRST function for model quality estimation. *Journal of Computational Biology* **12**, 1316–1327.

Tycko, R. & Hu, K. N. (2010). A Monte Carlo/simulated annealing algorithm for sequential resonance assignment in solid state NMR of uniformly labeled proteins with magic-angle spinning. *Journal of Magnetic Resonance* **205**, 304–314.

Van Der Kamp, M. W., Shaw, K. E., Woods, C. J. & Mulholland, A. J. (2008). Biomolecular simulation and modelling: status, progress and prospects. *Journal of the Royal Society Interface* **5** (Suppl 3), S173–190.

Vitek, O., Bailey-Kellogg, C., Craig, B. & Vitek, J. (2006). Inferential backbone assignment for sparse data. *Journal of Biomolecular NMR* **35**, 187–208.

Volk, J., Herrmann, T. & Wüthrich, K. (2008). Automated sequence-specific protein NMR assignment using the memetic algorithm MATCH. *Journal of Biomolecular NMR* **41**, 127–138.

Vranken, W. F., Boucher, W., Stevens, T. J., Fogh, R. H., Pajon, A., Llinás, M., Ulrich, E. L., Markley, J. L., Ionides, J. & Laue, E. D. (2005). The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins: Structure Function and Bioinformatics* **59**, 687–696.

Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. *Journal of Molecular Graphics* **8**, 52–56, 29.

Walsh, J. D., Cabello-Villegas, J. & Wang, Y. X. (2004). Periodicity in residual dipolar couplings and nucleic acid structures. *Journal of the American Chemical Society* **126**, 1938–1939.

Walsh, J. D. & Wang, Y. X. (2005). Periodicity, planarity, residual dipolar coupling, and structures. *Journal of Magnetic Resonance* **174**, 152–162.

Wan, X. & Lin, G. (2007a). CISA: combined NMR resonance connectivity information determination and sequential assignment. *IEEE/ACM Transactions of Computational Biology & Bioinformatics* **4**, 336–348.

Wan, X. & Lin, G. (2007b). GASA: a graph-based automated NMR backbone resonance sequential assignment program. *Journal of Bioinformatics and Computational Biology* **5**, 313–333.

Wan, X., Xu, D., Slupsky, C. M. & Lin, G. (2003). Automated protein NMR resonance assignments. *Proceedings of IEEE Computers Society and Bioinformatics Conference* **2**, 197–208.

Wang, J., Walsh, J. D., Kuszewski, J. & Wang, Y. X. (2007). Periodicity, planarity, and pixel (3P): a program using the intrinsic residual dipolar coupling periodicity-to-peptide plane correlation and phi/psi angles to derive

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

protein backbone structures. *Journal of Magnetic Resonance* **189**, 90–103.

Wang, J. B., Zuo, X. B., Yu, P., Byeon, I. J. L., Jung, J. W., Wang, X. X., Dyba, M., Seifert, S., Schwieters, C. D., Qin, J., Gronenborn, A. M. & Wang, Y. X. (2009). Determination of multicomponent protein structures in solution using global orientation and shape restraints. *Journal of the American Chemical Society* **131**, 10507–10515.

Wang, L. & Donald, B. R. (2004). Exact solutions for internuclear vectors and backbone dihedral angles from NH residual dipolar couplings in two media, and their application in a systematic search algorithm for determining protein backbone structure. *Journal of Biomolecular NMR* **29**, 223–242.

Wang, L., Mettu, R. R. & Donald, B. R. (2006). A polynomial-time algorithm for de novo protein backbone structure determination from nuclear magnetic resonance data. *Journal of Computational Biology* **13**, 1267–1288.

Wang, Y. J. (2004). Secondary structural effects on protein NMR chemical shifts'. *Journal of Biomolecular NMR* **30**, 233–244.

Wiederstein, M. & Sippl, M. J. (2007). ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Research* **35**(Web Server issue), W407–W410.

Williamson, M. P. & Craven, C. J. (2009). Automated protein structure calculation from NMR data. *Journal of Biomolecular NMR* **43**, 131–143.

Wilm, M., Shevchenko, A., Houthaeve, T., Breit, S., Schweigerer, L., Fotsis, T. & Mann, M. (1996). Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature* **379**, 466–469.

Wishart, D. S. (2011). Interpreting protein chemical shift data. *Progress in Nuclear Magnetic Resonance Spectroscopy* **58**, 62–87.

Wishart, D. S., Arndt, D., Berjanskii, M., Tang, P., Zhou, J. & Lin, G. (2008). CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data. *Nucleic Acids Research* **36**, W496–W502.

Wishart, D. S. & Case, D. A. (2001). Use of chemical shifts in macromolecular structure determination. *Nuclear Magnetic Resonance of Biologica Macromolecules Part A,* **338**, 3–34.

Wu, K. P., Chang, J. M., Chen, J. B., Chang, C. F., Wu, W. J., Huang, T. H., Sung, T. Y. & Hsu, W. L. (2006). RIBRA – an error-tolerant algorithm for the NMR backbone assignment problem. *Journal of Computational Biology* **13**, 229–244.

Wu, Z., Delaglio, F., Wyatt, K., Wistow, G. & Bax, A. (2005). Solution structure of (gamma)S-crystallin by molecular fragment replacement NMR. *Protein Science* **14**, 3101–3114.

Wüthrich, K. (1986). *NMR of Proteins and Nucleic Acids*. New York: John Wiley & Sons.

Xiong, F., Pandurangan, G. & Bailey-Kellogg, C. (2008). Contact replacement for NMR resonance assignment. *Bioinformatics* **24**, i205–i213.

Xu, J. & Sanctuary, B. C. (1993). Cpa – constrained partitioning algorithm for initial assignment of protein H-1 resonances from Mqf-Cosy. *Journal of Chemical Information and Computer Sciences* **33**, 490–500.

Xu, J., Straus, S. K., Sanctuary, B. C. & Trimble, L. (1994). Use of fuzzy Mathematics for complete automated assignment of peptide H-1 2d NMR-spectra. *Journal of Magnetic Resonance Series B* **103**, 53–58.

Xu, X. P. & Case, D. A. (2001). Automated prediction of N-15, C-13(alpha), C-13(beta) and C-13′ chemical shifts in proteins using a density functional database. *Journal of Biomolecular NMR* **21**, 321–333.

Xu, Y., Wang, X., Yang, J., Vaynberg, J. & Qin, J. (2006). PASA – a program for automated protein NMR backbone signal assignment by pattern-filtering approach. *Journal of Biomolecular NMR* **34**, 41–56.

Xu, Y., Xu, D., Kim, D., Olman, V., Razumovskaya, J. & Jiang, T. (2002). Automated assignment of backbone NMR peaks using constrained bipartite matching. *Computing in Science and Engineering* **4**, 50–62.

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. (2002). An NMR approach to structural proteomics. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 1825–1830.

Yee, A., Gutmanas, A. & Arrowsmith, C. H. (2006). Solution NMR in structural genomics. *Current Opinion in Structural Biology* **16**, 611–617.

Yokoyama, S., Hirota, H., Kigawa, T., Yabuki, T., Shirouzu, M., Terada, T., Ito, Y., Matsuo, Y., Kuroda, Y., Nishimura, Y., Kyogoku, Y., Miki, K., Masui, R. & Kuramitsu, S. (2000a). Structural genomics projects in Japan. *Nature Structural Biology* **7** (Suppl.), 943–945.

Yokoyama, S., Matsuo, Y., Hirota, H., Kigawa, T., Shirouzu, M., Kuroda, Y., Kurumizaka, H., Kawaguchi, S., Ito, Y., Shibata, T., Kainosho, M., Nishimura, Y., Inoue, Y. & Kuramitsu, S. (2000b). Structural genomics projects in Japan. *Progress in Biophysics and Molecular Biology* **73**, 363–376.

Yokoyama, S., Terwilliger, T. C., Kuramitsu, S., Moras, D. & Sussman, J. L. (2007). RIKEN aids international structural genomics efforts. *Nature* **445**, 21.

Zeng, J., Boyles, J., Tripathy, C., Wang, L., Yan, A., Zhou, P. & Donald, B. R. (2009). High-resolution protein structure determination starting with a global

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press

fold calculated from exact solutions to the RDC equations. *Journal of Biomolecular NMR* **45**, 265–281.

Zimmerman, D. E., Kulikowski, C. A., Huang, Y., Feng, W., Tashiro, M., Shimotakahara, S., Chien, C., Powers, R. & Montelione, G. T. (1997). Automated analysis of protein NMR assignments using methods from artificial intelligence. *Journal of Molecular Biology* **269**, 592–610.

Zuo, X., Wang, J., Yu, P., Eyler, D., Xu, H., Starich, M. R., Tiede, D. M., Simon, A. E., Kasprzak, W., Schwieters, C. D., Shapiro, B. A. & Wang, Y. X. (2010). Solution structure of the cap-independent translational enhancer and ribosome-binding element in the 3′ UTR of turnip crinkle virus. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 1385–1390.

Zweckstetter, M. (2008). NMR: prediction of molecular alignment from structure using the PALES software. *Nature Protocols* **3**, 679–690.

Zweckstetter, M. & Bax, A. (2000). Prediction of sterically induced alignment in a dilute liquid crystalline phase: aid to protein structure determination by NMR. *Journal of the American Chemical Society* **122**, 3791–3792.

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

https://doi.org/10.1017/S0033583510000326 Published online by Cambridge University Press