Published in final edited form as:

Prog Nucl Magn Reson Spectrosc. 2010 May ; 56(4): 311–328. doi:10.1016/j.pnmrs.2010.02.001.

Structure-oriented methods for protein NMR data analysis

Guillermo A. Bermejo^a and Miguel Llinás^{*}

Department of Chemistry, Carnegie Mellon University, Pittsburgh PA 15213, USA.

Keywords

Automated protein NMR structure determination; Direct NMR methods; Clouds; Structure-oriented; Top-down protein NMR structure

1. Introduction

The standard approach to nuclear magnetic resonance (NMR) protein structure determination is based on the near complete assignment of the resonances' chemical shifts to atoms in the molecule, *prior* to the structure calculation [1]. Computational methods that adhere to this approach are intrinsically assignment-oriented in that first priority is given to solving the assignment problem from which the rest of the process readily follows [2]. More recently, however, an increasing number of researchers have started to question this strategy [3]. Indeed, if the goal is to obtain a protein fold, why should the derivation of assignments be given so much weight upfront? Considerations of this kind have led to the formulation of alternative structure-oriented methods that do not hinge on the availability of complete, accurate assignments; instead, the focus is placed on the structure from the onset by generating initial, often approximate structural models that subsequently can be improved in iterative automated or semi-automated fashion. Thus, by deemphasizing the arduous and time-consuming assignment stage, the aim is to expedite structure determination, an aspect that is particularly important when high throughput is the end.

1.1. Assignment-oriented methods

Conventional automated structure determination methods typically tackle the assignment stage as follows [2]: (1) grouping of chemical shifts into spin systems, (2) identification of amino acid type for each spin system, (3) linking sequential spin systems into segments, (4) mapping spin-system segments onto the known protein sequence. To this end, an experimental NMR dataset is recorded that, in the preferred case of isotopically ¹⁵N/¹³C double-labeled samples, provides information on through-bond interactions (via scalar *J*-couplings) and chemical environments (via chemical shifts). Although such information is relatively limited with regards to the overall molecular topology (for advances in structure calculation from chemical shifts see Refs. [4-6]), its rather exhaustive analysis provides chemical shift assignments required, in turn, for the interpretation of additional, more structurally relevant experiments

© 2009 Elsevier B.V. All rights reserved.

^{*}Corresponding author. Tel: +1 412 268 3140; fax: +1 412 268 1061. llinas@andrew.cmu.edu..

^aCurrent address: Division of Computational Bioscience, Center for Information Technology, National Institutes of Health, Bethesda MD 20892, USA.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

(e.g., nuclear Overhauser effect spectroscopy; NOESY) in terms of three-dimensional (3D) models.

1.2. Structure-oriented methods

Structure-oriented methods propose a trial 3D model early in the process of fold generation. This initial structure may consist of a high-resolution model of the molecule determined via another technique, like X-ray crystallography, or, by contrast, a random spatial distribution of covalently unconnected atoms (i.e., a “gas” of atoms). Between these two extremes lie methods whose starting structural assumptions incorporate varying degrees of covalent information. Structure-oriented protocols are based on the concept that the initial structure is iteratively improved by exploiting the topological information conveyed by NMR spectra such as NOESY, producing chemical shift assignments during the cycling process. Hence, a shift in paradigm vis-à-vis the assignment-oriented approach is implied, that places more weight on the structurally rich experiments, often allowing for reduced datasets, as spectra traditionally used for assignment purposes may not be needed. While expediting the analysis of data with high structural information content, however, such often-called “assignment-free” strategies [7] may, in the end, yield by-product assignments that are only partially accurate. Gronwald and Kalbitzer [3] refer to structure-oriented methods as “top-down”, contrasting them to the “bottom-up” approach of assignment-oriented protocols.

1.3. Scope

This review focuses on structure-oriented strategies for solution NMR structure determination of proteins. Protocols that might be considered structure-oriented, but whose main objective is that of facilitating the assignment of chemical shifts by the determination of secondary structure elements (e.g., Refs. [8,9]) are ignored. However, methods that rely on a structure-oriented stage to compensate for sparseness of conventionally generated assignments are included. Such is the case of ITAS [10] and FastNMR [11], which rely on partial chemical shift assignment information to calculate the “initial structure” of a subsequent structure-oriented step. Although proxy residue [12] and floating chirality [13] methods are fully structure-oriented, in that they start with structure calculation based on minimal NMR data, both methods hinge on previous assignments obtained via additional experiments and, as a result, are associated with a conventional experimental dataset. Table 1 provides a loose outline of the following discussion. It lists representative structure-oriented protocols—sorted according to the extent of covalent information of their initial trial structures—and indicates the associated input datasets.

2. Direct methods

These methods aim at obtaining protein structures *directly* from NMR data, without prior assignments, by translating the spectral information into distances between pairs of atoms. Since such atoms are (at least initially) unassigned and only labeled by their chemical shifts, their covalent connectivity cannot be established upfront. Historically, exclusion of covalent constraints to account for a lack of assignment information preceded the introduction of direct methods per se. The floating chirality approach [13-16] is the case in point where, in absence of stereospecific assignments for methylene and isopropyl groups, each chemical-shift-labeled proton or methyl within a stereo-related pair is allowed to “float” between pro-R and pro-S configurations during structure calculations based on nuclear Overhauser effects (NOEs), by relaxing the constraints that enforce chirality. In the case of direct methods—the subject of this section—atoms float even more freely due to, in certain instances, the complete absence of covalent interactions.

When NOESY spectra are used as sole experimental input to drive a direct protocol, NOE-derived interproton distances, complemented by non-bonded repulsive interactions, are the only restraints enforced during structure calculations that start from a random spatial distribution of ^1H atoms—i.e., the initial trial 3D model of structure-oriented methods. Such calculations, which may be based on molecular dynamics (MD), distance geometry (DG), or inferential structure determination (ISD) [17,18], condense the initial proton distribution into a spatially organized configuration, henceforth referred to as a “cloud” due to its lack of covalent support. Subsequently, the atoms in the cloud can be mapped to the chemical framework of the protein molecule. This yields chemical shift and NOE assignments, which, along with the introduction of covalent constraints, allows for the calculation of improved structures. In the remainder of this section we review NOE-based direct methods. A protocol that follows the above-described general strategy, but that relies on residual dipolar couplings as the main source of structural information is discussed in Section 4.3.

2.1. Nuages

Conceptually, the direct approach was first formulated in 1992 by Malliavin et al. [19], who tested it vis-à-vis the X-ray crystallographic structure of lysozyme (129 residues). The protein was assumed ^{15}N -labeled, and only $\text{H}^{\text{N}}\text{--}\text{H}^{\text{N}}$ NOEs were considered. Simulated NOE intensities yielded 302 distances (0.5% intensity threshold), considered determined with 5% precision. Unobserved NOEs were assumed to reflect distances $> 4.5 \text{ \AA}$. Out of a total of 100 H^{N} -only clouds (*nuages* in the French original) generated by distance geometry, two were rejected owing to distance violations. Although relative to the crystallographic coordinates the overall accuracy of the clouds was poor (11.47 $\text{\AA}$ root-mean-square deviation; RMSD), regions corresponding to elements of secondary structure were more accurately determined. Subsequently, an H^{N} chain was threaded in each cloud by iteratively merging fragments (initially consisting of single H^{N} atoms) assuming likely sequential $\text{H}^{\text{N}}\text{--}\text{H}^{\text{N}}$ distances derived from structure database statistics. After establishing a consensus from the ensemble of clouds, 91.2% correct assignments were obtained.

The obtained sequential assignments assumed knowledge of the primary structure and the $\text{N} \rightarrow \text{C}$ chain directionality. The latter is a crucial issue not clearly addressed by the paper [19]. Notwithstanding this caveat, and the overall simplicity of the cloud interpretation routine, this study stands as a seminal attempt at direct protein structure determination from NOE data only.

2.2. ANSRS

The ANSRS (assignment of NOESY spectra in real space) protocol was reported in 1994 by Kraulis [20]. Input data are peak intensities and chemical shifts from ^1H atoms and their directly bound ^{15}N or ^{13}C , as extracted from 3D or 4D ^{15}N - or ^{13}C -edited NOESY spectra. The algorithm proceeds in three stages. First, NOE distance restraints are used to generate proton clouds via a molecular dynamics simulated annealing (MD/SA) protocol; the average cloud is used in subsequent steps. Second, for each $^1\text{H}\text{--}\text{X}$ ($\text{X} = ^{15}\text{N}, ^{13}\text{C}$) chemical shift-labeled proton in the cloud, a chemical shift score is calculated according to proton type (H^{α} , H^{β} , etc.) and residue type (Ala, Leu, etc.) probabilities, as assessed from a chemical shift database. Groups of proximal protons in the cloud that represent possible amino acid residues are identified via a recursive combinatorial search. Each group is scored according to the probability of belonging to a particular residue type by combining the above-described chemical shift scores and a score on the overall spatial configuration of the group within the cloud (as compared to conformations in known structures). Third, assignments are found via a Monte Carlo simulated annealing (MC/SA) protocol that minimizes the cost function $E = E_{\text{score}} + E_{\text{overlap}} + E_{\text{dist}}$, where E_{score} is the sum of group probability scores described above, E_{overlap} penalizes assignments of multiple groups to the same residue, and E_{dist} represents the sum of sequential distances within the amino acid residue chain.

ANSRS was tested on a segment of the DNA binding domain of GAL4 (residues 9–41) and the bovine pancreatic trypsin inhibitor (BPTI; 58 residues), on the basis of experimental chemical shifts and simulated NOE restraints corresponding to interproton distances $< 4.0 \text{ \AA}$ in reported structures. The calculated NOEs were used to generate sparser datasets by random removal of up to 30% of the total. In all cases, average clouds exhibited $< 2 \text{ \AA}$ RMSD from the known structures. The backbone trace and side chain positions could be determined from datasets with up to 20% NOE removal; better than 95% correct assignments were produced for the most complete dataset.

2.3. Antidistance constraints and chemical shift degeneracy

Often, direct methods supplement the NOE-derived distances with antidistance constraints (ADCs, or non-NOEs). ADCs assume that when an NOE connectivity between a pair of protons is not observed, the protons are likely to be removed from each other; thus, their proximity in the course of molecular dynamics (MD) calculations is discouraged by the application of a repulsive potential [21–23]. A similar strategy can be implemented in distance geometry (DG) schemes by use of a threshold distance, as exemplified in the original formulation of the direct approach [19] (Section 2.1).

In 1996, Atkinson and Saudek [24] studied the effect of ADCs and chemical shift degeneracy on the accuracy of proton cloud calculations. BPTI (58 residues) was chosen as a protein model. All ^1H – ^1H separations $< 4.0 \text{ \AA}$ in the crystallographic structure were incorporated as distance restraints (i.e., synthetic NOEs), and supplemented by ADCs in a combined DG/MD protocol. Omission of the ADCs resulted in collapsed clouds with decreased accuracy relative to the reference coordinates. Chemical shift degeneracy was taken into consideration for the first time in a direct protocol by merging protons assumed degenerate into a single “observable” ^1H that accounts for the NOE links of the merged atoms. Without chemical shift degeneracy, clouds as close as $1.91 \text{ \AA}$ RMSD from the reference were obtained. However, inclusion of 36 previously known cases of degeneracy increased the RMSD to $8.46 \text{ \AA}$.

In 1997, the authors tackled the degeneracy problem via a Monte Carlo simulated annealing (MC/SA) protocol that simultaneously fits ^1H cloud coordinates and chemical shifts directly to the NOESY spectrum [25]. In order to account for degeneracy during the MC/SA search, two or more atoms were allowed to have the same chemical shift, and the intensity at a given point in the spectrum could potentially arise from any number of spin pairs. The algorithm was tested on a six-proton fragment extracted from the X-ray structure of BPTI. NOESY spectra for both crystal reference and configurations evolving during the fit were synthetic, modeled according to the isolated spin pair approximation (ISPA). Correct assignments and accurate cloud coordinates were obtained both for a fully resolved spectrum and for a case involving a two-proton degeneracy. Despite the encouraging results, the proposed MC/SA algorithm has, to our knowledge, not yet been tested on intact protein sequences, conceivably because the search space becomes unmanageably large. In order to mitigate this problem the authors suggested enforcing covalent connectivity, predicting chemical shifts from the evolving structure, and supplementing the NOESY spectrum with heteronuclear J -correlated data. Interestingly, such search restrictions—including other more significant ones—have been implemented more recently in a similar, although considerably more sophisticated approach [26] (discussed in Section 5.2).

2.4. J-correlated information

Similar to NOESY connectivities, crosspeaks from other types of multi-dimensional experiments can be interpreted as distance restraints as well. Thus, an ^{15}N – ^1H heteronuclear single quantum coherence (HSQC) connectivity reveals an H–N bond distance, and an HN (CO)CA peak yields distances between $\text{H}^{\text{N}}\text{--C}^{\alpha}_{i-1}$ and $\text{N--C}^{\alpha}_{i-1}$, fixed by the peptide bond

geometry. In 2002, Atkinson and Saudek [27] extended the direct approach beyond NOESY data, to include heteronuclear *J*-correlated spectral datasets generally used in assignment-oriented approaches. The principle was tested via a MD/SA protocol on data simulated for ^{15}N -, ^{13}C -labeled ubiquitin (76 residues). A total of 1,647 distance restraints were derived from a set of ten simulated heteronuclear experiments commonly exploited for chemical shift assignments. 2,060 NOE restraints were obtained from the X-ray reference structure (involving ^1H – ^1H distances within 4.0 Å); 92,570 ADCs (^1H – ^1H distances > 4.0 Å) completed the dataset. Computed clouds had average C^α RMSDs ≤ 2.32 Å when superimposing segments of secondary structure. Despite the addition of heavy atoms, however, the clouds do not afford bona fide structural models mainly because: (1) chirality was not enforced in the calculations so that topological mirror images were generated; (2) prolines introduced gaps in the backbone chain; (3) side chain atoms were not linked to the backbone, as experiments based on total correlation spectroscopy (TOCSY) were not included. On the other hand, in contrast to ^1H -only clouds (derived exclusively from NOESY experiments), (^1H , ^{15}N , ^{13}C)-clouds expectedly resemble standard structural models more closely.

2.5. CLOUDS

The feasibility of the direct approach using experimental NOE intensities was reported in 2002 by Grishaev and Llinás [28,29] (reviewed in Ref. [30]). In this study, a routing for the entire process, from the recorded spectral data to the protein structure calculation, was achieved. The method, named CLOUDS (computed location of unassigned spins), was tested on the col 2 domain of human matrix metalloproteinase-2 (60 residues) and the kringle 2 domain of human plasminogen (83 residues), starting from a list of unassigned, unambiguous experimental NOESY data available from the previously reported assignment-oriented structural determinations [31,32]. The protocol starts by generating ~1,000 ^1H -only clouds via a MD/SA protocol that incorporates accurate distance restraints resulting from a relaxation matrix analysis of homonuclear NOESY spectra [33], antidistance constraints, and van der Waals repulsions. After filtering by reference to the cloud closest to the mean, a minimal dispersion proton distribution is obtained [28]. Such distribution, named FOC (family of clouds), effectively represents a proton density of the protein to which the covalent framework is to be fit [29]. To this end, the polypeptide backbone is traced through the H^{N} and H^α protons in the clouds via a Bayesian approach where the probabilities of sequential connectivity hypotheses are inferred from likelihoods of H^{N} – H^{N} , H^{N} – H^α , and H^α – H^α distances, as well as chemical shifts, derived from public databases. Once the polypeptide sequence of (H^{N} , H^α) protons becomes identified, a similar procedure is followed to link the side chain protons to the main chain. The thus assigned proton density is transformed into a potential energy to which the molecular structure is embedded via a MD protocol, after selection of the correct mirror image configuration by an embedding energy criterion.

Overall, clouds exhibited < 0.8 Å H^{N} RMSD from the known NMR structures used to test the approach. MD embedding of the covalent template into the FOC yielded models within 1.0–1.5 Å (N , C^α , C') RMSDs from reported X-ray and NMR structures. CLOUDS has been integrated to the CcpNmr Analysis software for spectral analysis, under the Collaborative Computing Project for NMR (CCPN) data model framework [34].

2.6. ISD-generated clouds

Starting from assigned restraints, inferential structure determination (ISD) has been shown to produce structures with accuracies similar to or better than those generated with conventional MD-based structure calculation packages [17,18]. On the basis of Bayes' theorem,

$$\mathcal{P}(X, \xi | D, I) \propto \mathcal{P}(D | X, \xi, I) \mathcal{P}(X | I) \mathcal{P}(\xi | I), \quad (1)$$

ISD infers atomic coordinates by maximizing the joint posterior probability density $\mathcal{P}(X, \xi | D, I)$ on the coordinate set X and nuisance parameters ξ , given the data D and prior information I . The likelihood $\mathcal{P}(D | X, \xi, I)$ is expressed in terms of a model to estimate D from X , assuming a log-normal distribution for the associated error. The conformational prior $\mathcal{P}(X | I)$ is the probability density on the coordinates given the bonded and non-bonded interaction energy between atoms, assumed as a Boltzmann distribution. $\mathcal{P}(\xi | I)$ is expressed as Jeffreys' prior [35]. Posterior maximization is achieved via a Markov chain Monte Carlo algorithm based on the replica exchange method, specifically designed for the structure determination problem [36].

In 2004, Habeck et al. [37] adapted the method to deal with the problem of generating clouds from unassigned NOE intensities, where the lack of assignment information is reflected in a conformational prior that takes only van der Waals-like interactions into account. The approach was tested on simulated NOE intensities for the H^N protons of the Fyn SH3 domain (59 residues, assumed perdeuterated). A relaxation matrix model was used for computing the likelihood so that coordinates and relaxation rates are simultaneously optimized. In addition, a single optimization process can account for NOEs arising from experiments recorded with different mixing times. A ten-mixing-time NOESY series (100 ms–5 s) was used for testing the approach, assuming 10% error and a 0.001 intensity threshold. When inputting each NOESY on separate optimizations, the resulting proton clouds exhibited average RMSDs to the known coordinates between ~ 7.5 and ~ 0.8 Å, with improved accuracy for long mixing times. However, simultaneous optimization of all ten NOESY experiments yielded an average RMSD < 0.5 Å. These results are encouraging in view of the relatively high accuracy of the computed clouds.

2.7. Sparse-constraint CLOUDS

In 2008, Bermejo and Llinás [38] proposed the sparse-constraint CLOUDS (SC-CLOUDS) protocol as an application of the direct approach to the analysis of selectively methyl-protonated, otherwise perdeuterated, proteins. As a result of the isotopic labeling [39], the generated clouds contain only CH_3 protons from Val, Leu, and Ile ($\delta 1$) residues, as well as labile backbone and side chain NH protons. Clouds are calculated via a MD/SA scheme that incorporates NOE restraints, ADCs, and van der Waals repulsions. In earlier instances of the direct approach, the ADCs consisted of restraining potentials similar to those typically used for NOEs, excluding close distances with lower bounds at 4–5 Å [24,27,28]. In SC-CLOUDS, however, ADCs are based on NOE intensities simulated from a database of known protein structures. The simulation replicates the conditions of the experimental NOESY (mixing time, rotational correlation time, etc.) available for the protein under study (of unknown structure). Simulated NOEs are deemed “observable” if their intensities exceed a threshold determined from the experimental NOESY spectra, and the fractional probabilities $\mathcal{P}_{\text{noe}}$ of observing an NOE as a function of the 1H – 1H distance are calculated for the different interproton pairs (H^N – H^N , H^N – CH_3 , etc.). In the course of cloud calculations, the $\mathcal{P}_{\text{noe}}$ profiles are directly implemented as “ADC potentials”. As in standard practice, ADCs are enforced only between those protons that fail to yield an experimentally observable NOE, thus biasing them to assume distances corresponding to low $\mathcal{P}_{\text{noe}}$ values (i.e., low probabilities of producing a detectable NOE). ADC curves are shown in Fig. 1.

SC-CLOUDS was tested with experimental 3D ^{15}N - and ^{13}C -edited NOESY data on the Z domain of *Staphylococcal* protein A (58 residues; Fig. 2a), from which a total of 234 NOEs and 4,483 ADCs (Fig. 1; implemented as energy terms in Xplor-NIH [40,41]) were derived for cloud calculation. ADCs had a significant effect in preventing the collapse of the cloud under the attractive NOE forces (compare Fig. 2c and Fig. 2d). Z domain clouds (Fig. 2c and Fig. 3) had an H^N RMSD of 6.1 Å to the previously reported structure generated from a

conventional NMR study of a fully protonated sample [42] (Fig. 2a and 2b). This RMSD-based accuracy is lower than that of clouds generated via ANSRS [20] or CLOUDS [28] (Sections 2.2 and 2.5, respectively), a reflection of the NOE sparseness resulting from deuteration. Thus, the cloud interpretation routines developed for the above-mentioned protocols are inapplicable to the sparse-constraint case, as they implicitly assume cloud coordinates to be of a quality comparable to those in high-resolution protein structures. To circumvent this problem, a graph-theoretic strategy based on a sum-of-distances minimization criterion tolerant to coordinate errors was developed.

The strategy relies on the construction of graphs from atomic coordinates, where each vertex represents an atom (or group of atoms), and each edge linking vertices i and j is assigned a cost $C(i, j)$ equal to the i - j interatomic distance. Specifically, given a cloud, a graph is constructed from the H^N s, including edges whose associated H^N - H^N distance is shorter than a specified cutoff (i.e., $C(i, j) < \text{cutoff}$). The depth-first search algorithm is used to find in this graph the chain that minimizes $\sum_{i,j} C(i, j)$, where the sum extends over all pairs of adjacent vertices in the chain. The latter represents the sequential H^N connectivity in the polypeptide backbone. Identification of the H^N s—or, what is equivalent, the determination of the $N \rightarrow C$ chain direction—and the side chain groups in the cloud is subsequently achieved via bipartite graph matching (Fig. 4), a technique that has been applied to the chemical shift assignment problem by other authors [9,43-47]. Assuming one of the two possible chain directions, a putative identity for each H^N is established given the known amino acid sequence. A bipartite graph of the type shown in Fig. 4a is subsequently built, where one vertex set represents the H^N s from residues with protonated side chains (Leu, Val, etc.; according to the assumed chain direction), and the other vertex set represents the side chain groups in the cloud. A matching M in such a graph is a set of its edges so that no two edges share the same vertex (Fig. 4b). $C(M)$, the cost of M , is the sum of M 's edge costs (i.e., sum of H^N -side chain group distances) and, when all vertices in the bipartite graph are involved in the matching (as in Fig. 4b), it can be cast in terms of a permutation π of vertices $\{1, 2, \dots, n\}$ in one vertex set:

$$C(M) = \sum_{i=1}^n C(i, \pi(i)). \quad (2)$$

The lowest-cost matching, namely the permutation that minimizes Eq. 2, is found subject to the constraint that only H^N atoms and side chain groups of the appropriate type can be matched (e.g., Leu H^N atoms can be matched only to isopropyl groups). Since this matching is more likely to have lower cost when generated from the bipartite graph that assumes the correct chain direction, the identity of the H^N s is thus established and, simultaneously, that of the side chain groups via their matching to the identified H^N s. In practice, however, a side chain group is considered reliably identified only if it appears frequently matched to a given H^N in a set of low-cost matchings. Consequently, side chain groups not satisfying this frequency criterion remain unidentified at this point. The identification is completed once preliminary structures are calculated, by following an approach similar to that proposed by AB et al. [12] (see Section 3.1 for details).

Finally, since the above identification of cloud protons is equivalent to assigning the NOEs, the latter can be used to calculate standard structural models. Structure calculation is achieved via RosettaNMR [48,49] (reviewed in Ref. [50]), a knowledge-based method that assembles structures with native-like global properties from fragments of known structures, while enforcing NMR restraints. The structural bias of RosettaNMR is needed to overcome the fact that SC-CLOUDS relies on a list of sparse NOEs and no other information.

Application of the above-described strategy to the interpretation of the lowest-energy Z domain cloud yielded 79% correctly identified cloud groups. The remaining misidentified 21% resulted from local cloud inaccuracies that caused, for example, the swapping of sequential H^N identities. The resulting misassigned NOEs had a negligible effect on the quality of the final structures as (1) NOEs are excluded from RosettaNMR's selection of protein fragments (to be subsequently used for structure calculation), and (2) the subsequent NOE-driven assembly of the fragments into compact structures is coarse-grained. The final SC-CLOUDS structures had a backbone (N, C^α , C') RMSD of 2.8 Å relative to the previously reported NMR model of the Z domain, obtained from a fully protonated sample [42]. This is comparable to the accuracy of reported structural models generated via an assignment-oriented method that used the same NOESY data as SC-CLOUDS, in addition to *J*-correlated experiments [51]. That correct structures can be obtained with a sizable fraction of incorrect assignments has been reported for other structure-oriented methods [10,13,26] (discussed below). Additional SC-CLOUDS tests with restraints simulated from structures spanning a range of sizes and topological complexities suggest a wider applicability of the protocol. SC-CLOUDS is summarized in Fig. 5, along with other structure-oriented methods that exploit RosettaNMR in a similar fashion. Table 2 compares SC-CLOUDS and CLOUDS.

2.8. Distance restraints extracted from highly NOE-embedded protons

An alternative approach to dealing with the NOE sparseness associated with highly deuterated protein samples has been proposed recently by Schedlbauer et al. [52]. As earlier demonstrated by the SC-CLOUDS method [38] (Section 2.7), a viable possibility is the implementation of error-tolerant techniques able to identify atoms in sparsely constrained, inherently distorted clouds, along with a structure calculation protocol that is robust against misassigned data. In contrast, Kontaxis and coworkers [52] focus on improving the accuracy of the clouds to the point where they can be input, in combination with a minimal set of triple-resonance experiments, to preexisting structure-driven chemical shift assignment programs (Section 5.6), in place of standard protein models. Briefly, clouds are generated via a MD/SA protocol that enforces NOE-derived distance restraints and van der Waals repulsions. The distance between two protons *i* and *j*, not directly linked by an NOE, is converted into a new, "artificial" restraint provided that (1) the distance is consistently conserved across the family of clouds, and (2) *i* and *j* are separated by a specified number of links in the NOE network. Thus, a new set of distance restraints is generated which, combined with the one directly derived from the NOEs, helps compute improved clouds in subsequent MD/SA runs. This approach is implemented iteratively until no new conserved distances can be identified. After the first iteration, more stringent conditions are used for the selection of conserved distances, involving protons that are more embedded in the NOE network.

The method was tested on several proteins, including Cyclophilin D (165 residues; β barrel), for which high-resolution NMR structures, assigned chemical shifts and NOE restraint lists were available from previous conventional assignment-oriented studies on fully protonated samples. In the case of Cyclophilin D, 462 NOE restraints were selected, assuming a deuteration scheme where only methyl groups from Val, Leu, and Ile (δ 1) residues, and labile backbone amides remain protonated [39] (i.e., same deuteration scheme used by SC-CLOUDS; Section 2.7). After two rounds of MD/SA, 1,532 additional, artificial distance restraints were identified from the clouds, which combined with the original 462, yielded final (H^N , CH_3)-clouds with H^N RMSD of 1.92 Å relative the high-resolution reference (considering the structural core only).

While HNCO, HNCA and HN(CO)CA experiments can be acquired with high sensitivity, even in large molecular weight systems, often the sequential $^{13}C^\alpha$ connectivity they provide is insufficient to link the spin systems in an assignment-oriented approach (step (3) in Section

1.1), owing to $^{13}\text{C}^\alpha$ chemical shift degeneracy. However, combination of this minimal set of J -correlated data with NOE information and a structural model facilitates the assignment. Using the MONTE algorithm [53], only 15% of the chemical shifts in the assumed highly deuterated Cyclophilin D could be reliably assigned via HNC0/HNCA/HN(CO)CA-type connectivities only. However, the percentage increased to 97% when including H^N – H^N NOE data, and the above-described clouds as input structural models. Once this cloud-driven assignment is complete, a standard covalent model of the protein can be obtained by conventional MD-based protocols. Final Cyclophilin D models had a full backbone RMSD of 2.13 Å to the high-resolution reference. Similar results were found for other test cases involving proteins of different size and fold topology. Improvements were observed when using a residue-type selective isotope labeling strategy [52].

2.9. Conclusions

As noted at the beginning of this section, the relaxation of the covalent information on the system was anticipated by the floating chirality approach [13–16], before direct methods were introduced. Indeed, the relative similarity between the two approaches is apparent from inspection of Fig. 3, which shows the methyl groups of a cloud superposition for the Z domain of *Staphylococcal* protein A, generated via SC-CLOUDS. Provided the NOEs for each methyl within an isopropyl group differ, the group adopts a preferential orientation within the cloud (Fig. 3b and Fig. 3d), similar to what is the case with the non-stereospecifically assigned moieties within floating chirality structures. Another characteristic common to floating chirality [13] and SC-CLOUDS, in particular, is that despite improved or correct structures being obtained, respectively, the structure-to-assignment mapping is not necessarily failure-proof.

Even after the NOEs have been assigned within direct methods that rely only on NOE data, the calculation of standard covalent structural models may not be straightforward, particularly when dealing with sparse NOE information. CLOUDS and SC-CLOUDS have been shown to complete the structure elucidation from unassigned experimental NOE intensities. In both CLOUDS and SC-CLOUDS the structural interpretation of atomic clouds resembles model-building from crystallographic electron densities, which hinges on tracing the polypeptide backbone to which the side chains are subsequently connected. In particular, the H^N chain tracing protocol in SC-CLOUDS is akin to ARP/wARP strategy for building a C^α chain from X-ray diffraction data [54]. Conversely, the X-ray field may benefit from the developments presented above, as suggested by Pozharski et al. [55] who tested proton clouds for molecular fragment replacement in the absence of close structural homologues. Furthermore, a recently developed protocol for the structure determination of solid nanostructured materials—which present only short-range order and thus cannot be solved via crystallographic techniques—is akin to the above-discussed direct NMR methods: unassigned distances obtained from powder diffraction measurements are used as restraints in a distance geometry approach [56].

With the already noted exceptions [27,52] (Sections 2.4 and 2.8) the focus of direct methods is on NOESY experiments exclusively. While NOE assignment is not required, all protocols—with the exception of those presented in Section 2.3 [24,25]—assume NOEs are unambiguously *identified*. In other words, it is known with certainty that a particular NOESY crosspeak results from protons i – j and not i – k , independent of whatever protons i , j , and k happen to be, i.e., independent of their *assignments*. Peak lists of this kind, however, may not be straightforward to obtain owing to chemical shift degeneracy. Indeed, as pointed out by Atkinson and Saudek [27], “two resonances from nuclei that are far apart in the structure with identical chemical shifts but distinct sets of neighbors would be represented by a single atom with one set of neighbors, leading to gross distortion of the calculated [cloud]”. In this context, the results obtained with SC-CLOUDS and by Schedlbauer et al. [52] (Section 2.8) are

encouraging as the combination of highly deuterated protein samples with heteronuclear-edited experiments reduces the likelihood of degeneracy. Furthermore, Grishaev and Llinás [57,58] have developed a suite of algorithms for removing NOE ambiguities which, as it relies on neither chemical shift assignments nor an initial structural model, produce a suitable input for CLOUDS [30]. However, such algorithms require *J*-correlated experiments and, as discussed in Section 4 below, the through-bond information they provide can be used to group the otherwise unconnected chemical shifts, giving rise to covalently linked molecular fragments that facilitate structure determination.

3. Proxy residues

The proxy residue approach developed by AB et al. [12] aims at assisting in the protein structure calculation and chemical shift assignment stages, when partial assignments are already available. Unassigned chemical shifts lead to additional atoms or groups of atoms, i.e., the proxy residues, that interact with each other and with atoms in the protein—represented by a standard all-atom model—exclusively through distance restraints during MD calculations. Thus, proxy residues are treated in a similar fashion as atoms within clouds in direct methods (Section 2), namely, stripped of covalent connections to the rest of the system. The distance restraints include NOEs and so-called identity restraints: ambiguous distance restraints that force the proxy atoms to be proximal to at least one of the unassigned atoms in the protein sharing the same type (methylene, aromatic, etc., as determined, for example, from characteristic chemical shifts). In addition, spatial overlap of proxy residues is avoided by enforcing a repulsive potential between them. Non-bonded interactions between proxy and protein atoms are switched off in order to allow for the former to gain access to the protein core.

The proxy residue approach was incorporated to ARIA/CNS [59-61] and CANDID/CYANA [7,62] software packages, and tested with experimental data on the proteins DWNN (86 residues) and IIB^{Chb} (106 residues), which, for comparison, were also analyzed in a conventional assignment-oriented manner. The two main proposed applications of the method are discussed next and summarized in Fig. 6.

3.1. Chemical shift assignment

Final structures—each consisting of protein + proxies—suggest possible assignments for proxy residues based on their proximity to the protein atoms they represent (Fig. 6). This approach was tested on DWNN, for which several ¹³C-edited NOESY resonances were unassigned. 38 proxy residues were defined from which 10 assignments were proposed and later confirmed using *J*-correlated spectra. The approach was additionally tested on a cluster of three aromatic rings from IIB^{Chb}, a particularly problematic situation for conventional assignment strategies. The rings were assumed unassigned, producing the corresponding proxies that, after convergence of CANDID runs, located themselves in their correct loci within the protein.

Based on the above results, a similar strategy was implemented within SC-CLOUDS (Section 2.7), in order to improve on the cloud interpretation routines which usually leave unidentified a number of side chain groups (the “proxy residues”). SC-CLOUDS scores the different assignment possibilities via bipartite graph matching (Fig. 4). A complete bipartite graph is built (Fig. 4a), where one vertex set represents the proxies, the other set the unassigned side chains in the protein, and each edge is given a cost equal to the corresponding distance in the final (protein + proxies) configuration. The lowest-cost matchings yielded by several final configurations lead to consensus assignments [38].

3.2. Improved convergence of ARIA and CANDID

Even when a proxy residue is unassigned, structural improvements in the protein conformation may result since regions distant in the primary sequence may interact indirectly via a “bridging” proxy (Fig. 6). Thus, NOE restraints that would otherwise be discarded can still be exploited through proxy residues. This is particularly valuable for the performance of ARIA (ambiguous restraints for iterative assignment) and CANDID calculations, which are sensitive to input data completeness. Indeed, while CANDID with proxy residues produced relatively accurate DWNN and IIB^{Chb} structures when randomly removing up to 25–30% of the known assignments, without proxy residues this could be achieved with < 10–15% missing assignments. In addition, removal of all side chain assignments, except for those at β positions, from the DWNN dataset (35% missing assignments), resulted in final proxy-based structures with a backbone RMSD of 1.37 Å relative to the reference structure generated with all the assignments. The RMSD increased to ~9 Å when proxy residues were omitted.

4. Fragment-based methods

4.1. DGPA

It should be apparent that “directly” solving the structure (Section 2) is basically equivalent to “indirectly” assigning the chemical shifts: one implies the other. In 1993, Oshiro and Kuntz [63] proposed the DGPA (distance geometry proton assignment) method for deriving chemical shift assignments from the interpretation of clouds of amino acid residue fragments. Although in essence the approach is very similar to the direct methods described above in Section 2, in that it relies on the calculation of clouds from unassigned, unambiguous NOEs, extra information is assumed from *J*-correlated experiments that allows for the grouping of H^N, H ^{α} , and H ^{β} protons into (partial) spin systems. As a result, such protons result linked together as covalent fragments that contain all heavy atoms within an amino acid residue, up to the C ^{β} position.

DGPA starts by calculating an ensemble of clouds of such fragments via a distance geometry protocol that enforces NOE restraints. A graph is made where the vertices represent the fragments and the edges the NOE connectivity. Removal of articulation points, found via a depth-first search algorithm, yields different sets of cloud residue fragments or “domains”, for which their associated average coordinates are calculated from the cloud ensemble. Within each domain a trial sequence of fragments is chosen and the associated average coordinates compared to a standard α -helix or β -sheet of the same length. The fragment sequence that produces the minimum C ^{α} RMSD is assumed correct.

The method was tested on the protein BPTI (58 residues) with distance restraint sets both simulated from the known X-ray model and experimentally determined from NOESY spectra. Cloud interpretation was focused only on domains associated with the reported secondary structure of the protein. The interpretation protocols are based on exhaustive sorting of residues within the cloud domains according to their putative order in the sequence. For example, all 3,628,800 sequence ordering possibilities (permutations) of the 10 fragments within the cloud domain associated with BPTI's single α -helix were assumed and RMSD-compared to standard α -helical coordinates. Although the protocol performed well with simulated data, it yielded inconclusive results with experimental NOEs. On a pessimistic note, the authors concluded that the “approach is only useful with excellent quality stereo-resolved data” [63].

4.2. ABACUS

The requirement of unambiguously identified, albeit unassigned, NOEs by DGPA and the direct methods discussed in Section 2 (see Section 2.9 for the conceptual differentiation between NOE identification and assignment), motivated Grishaev and Llinás [57] to develop

BACUS (Bayesian analysis of coupled unassigned spins; reviewed in Ref. [30]). Based on Bayesian inference, the algorithm automatically establishes probabilistic identities for NOESY crosspeaks. The input is the grid of chemical shifts, grouped into spin systems, as generated from *J*-correlated experiments on fully protonated proteins. Since BACUS relies on neither prior chemical shift assignments nor a structural model, the produced unambiguous NOEs can be used as input for direct methods such as CLOUDS [30]. However, a BACUS/CLOUDS approach would not take full advantage of the through-bond information since CLOUDS, as originally designed [28,29], purposefully disregards it. This dichotomy led Grishaev et al. [64] to the development of ABACUS (applied BACUS).

In ABACUS, each spin system defines a covalent fragment during MD/SA calculations that enforce BACUS-identified NOEs. The resulting clouds of fragments are subsequently interpreted to yield the chemical shift and NOE assignments. Similar to the matching problem discussed in Section 2.7 (Fig. 4), the assignment of each fragment to a position in the protein sequence can be cast in terms of a permutation. The LINKMAP routine performs a Monte Carlo (MC) search for the permutation π that minimizes the pseudo-energy

$$E = \sum_{i=1}^n E_{cs}(i, \pi(i)) + \sum_{i=1}^{n-1} E_{dist}(\pi(i), \pi(i+1)), \quad (3)$$

where n is the number of residues. E_{cs} measures the contribution from the assignment of fragment $\pi(i)$ to residue i via the fit between the fragment's chemical shifts and those database-estimated for residue i . This chemical shift-based fragment classification into amino acid types is performed by the separate TYPESYST routine. E_{dist} represents the probability for the sequential linkage between fragments $\pi(i)$ and $\pi(i+1)$, derived from their separation distance within the ensemble of fragment clouds. The above-described approach, outlined in Fig. 7, results in assigned chemical shifts and NOEs which are subsequently used for the calculation of conventional structural models via a standard MD-based protocol. NOE identities resulting in restraint violations $> 1 \text{ \AA}$ in the generated structure ensemble are removed (STRNOE routine), followed by a BACUS search for new identities and new MD structure calculations. This MD/STRNOE/BACUS cycle is performed until convergence. In the final stages, hydrogen bond and TALOS-derived [65] backbone dihedral angle constraints are incorporated into the structure calculations.

ABACUS was originally run as a blind test on the genomic protein mth1743 (70 residues) from *M. thermoautotrophicum*, the assignments and structure being independently and simultaneously determined via a conventional assignment-oriented approach based on the same experimental NMR data [64]. LINKMAP yielded 100% correct sequence specific assignments for the mth1743 fragments. The conventional and the ABACUS approaches converged to similar structures (1.2 $\text{\AA}$ backbone RMSD) and NOE assignments (97% similarity).

The two main modules of ABACUS, namely BACUS and LINKMAP, have been recently improved by Lemak et al. [66]. BACUS, originally designed to analyze NMR data from unlabeled and ^{15}N -labeled proteins via COSY (correlation spectroscopy) and TOCSY *J*-correlations, has been extended to encompass 3D triple resonance NMR experiments on $^{13}\text{C}/^{15}\text{N}$ double-labeled proteins. A new algorithm, named Fragment Monte Carlo (FMC), has been developed to overcome the following LINKMAP limitations: (1) the number of fragments and positions in the sequence should be equal (implicit in the use of the permutation π in Eq. (3)), and (2) fragments are allowed to occupy only those positions consistent with the fragment amino acid type prediction by TYPESYST. The first restriction does not allow for cases where a fraction of residues in the protein cannot be discerned from the experimental

data. The second restriction can partition the assignment space in such a way that the stochastic search may become trapped in a local minimum. FMC allows for an arbitrary user-defined number of fragments, and penalizes (but allows for) the assignment of fragments to residues of different type. In addition, the E_{dist} function in Eq. (3) is modified so that the probability of sequential linkage between two fragments depends on both their relative positions in the cloud of fragments and their NOE connectivity. FMC uses the multicanonical (MUCA) MC method [67,68] for enhanced sampling of assignment space, instead of the standard canonical MC method [69] employed by LINKMAP.

The improved ABACUS version recently has been applied to the structure determination of the CPH domain of human Cul7 (101 residues) [70], a domain common to other proteins but of previously unknown fold. The input spectral dataset consisted in HNCO, CBCA(CO)NH, HBHA(CO)NH, CC(CO)NH-TOCSY, HC(CO)NH-TOCSY, HC(C)H-TOCSY, (H)CCH-TOCSY, ^{15}N - and ^{13}C -edited NOESY. Assignments for 94.8% of Cul7-CPH domain's chemical shifts (^1H , ^{15}N , and ^{13}C) and 2,144 NOEs were obtained, leading to the structure determination. In addition to automation, a salient feature of the method is the avoidance of triple resonance spectra with low signal/noise, such as HNCACB and NHHAHB, conferring ABACUS an active role in on-going structural genomics projects [71] (for example applications of the method see Refs. [70,72-75]). Structures of genomic target proteins determined with the help of ABACUS are shown in Fig. 8; the set comprises proteins with up to 137 residues per polypeptide chain, and includes two homodimers: ATU0232 (2×62 residues; protein data bank (PDB) code 2k7i) and ATC0852 (2×122 residues; PDB code 2kjj), both from *A. tumefaciens*.

4.3. Direct RDC implementation

A valuable source of restraints for NMR molecular structure determination is provided by residual dipolar couplings (RDCs) (reviewed in Refs. [76-80]). In contrast to NOEs, which provide short-range ^1H - ^1H distance restraints, RDCs yield long-range orientational information of interatomic vectors within the molecule, weakly aligned in the magnetic field. This orientation dependence can be expressed by

$$D_{ij} = \frac{D_{ij}^{\max}}{r_{ij}^3} \sum_{k,l} S_{kl} \cos \alpha_k \cos \alpha_l, \quad (4)$$

where D_{ij} , the dipolar coupling between atoms i and j , is written in terms of a collection of physical constants $D_{ij}^{\max}$, the interatomic distance r_{ij} , and direction cosines $\cos \alpha_{kl}$ that describe the orientation of the interatomic vector relative to an arbitrary molecule (or molecular fragment) frame whose overall order and orientation is encoded by the elements S_{kl} of the Saupe order matrix. Despite providing a wealth of information, RDCs have been largely relegated to supplement NOE-derived distance restraints in de novo structure determination. Within few exceptions where RDCs play a central role [49,81-86], the method originally proposed in 2001 by Tian et al. [83]—the subject of this section—additionally represents a bona fide structure-oriented approach that establishes the assignments simultaneously with the structure. Indeed, the method can be classified as an *RDC-based direct* approach. An important difference with the direct NOE-based methods (Section 2) is that the experiments that yield RDCs readily provide additional through-bond connectivity information, which leads to initial structures that consist of covalent fragments instead of unconnected atoms.

A recent implementation of the direct RDC approach [87,88] relies on only five experiments applied to protein samples which are highly ^{15}N -labeled but only 15–20% ^{13}C -labeled. Three core experiments yield RDCs and J -couplings: a coupled ^{15}N -HSQC ($^1D_{\text{N-HN}}$), an E.COSY-

HNCA ($^1D_{C\alpha-H\alpha}$, $^4D_{HN-H\alpha(i-1)}$, $^3D_{HN-H\alpha}$ and $^3J_{HN-H\alpha}$) and a coupled HNC0 ($^2DC'-HN$ and $^1DC'-N$); the measured couplings indicated in parenthesis. Two supplementary experiments, an ^{15}N -edited NOESY and an ^{15}N -edited TOCSY, provide distance and residue type information, respectively. Acquisition of this dataset takes about a third of the time required for conventional assignment-oriented, NOE-based approaches.

The structure calculation strategy starts by generating backbone fragments consistent with the measured RDCs, using the program REDCRAFT (residual dipolar coupling based residue assembly and filter tool) [87,88]. The latter begins by considering fragments consisting of two peptide planes connected by a C^α carbon with variable ϕ and ψ angles (Fig. 9). All possible ϕ/ψ combinations are generated using 10° increments. For each geometry, a system of equations of the type of Eq. (4) (i.e., one equation for each of the potentially nine observed RDCs in the fragment (Fig. 9)) is solved by singular value decomposition [89] to obtain the order parameters S_{kl} that, in turn, help, calculate the RDCs via Eq. (4). The fragment geometry is scored based on the comparison of its calculated and experimental RDCs, as well as the extent of agreement with Ramachandran map and Karplus curve for the $^3J_{HN-H\alpha}$ coupling. A list of fragment geometries is produced where each ϕ/ψ combination is ranked according to the above criteria. Subsequently, longer backbone fragments are generated by merging dipeptide planes chosen primarily on the basis of $^{13}C^\alpha$ chemical shift matching in HNCA-type patterns. Only the top-ranked geometries for each dipeptide fragment are used for merging, thus restricting the possible conformations of the longer fragment. Such conformations go through the scoring process described above for the dipeptide planes, followed by further fragment extension. The extension eventually halts owing to $^{13}C^\alpha$ chemical shift ambiguities or lack of experimental data, particularly at proline sites. Fragments of five or more C^α carbons are usually sufficient to proceed to the next stage.

Based on their geometry and $^{13}C^\alpha$ chemical shifts, the multiple fragments generated by REDCRAFT are placed in the sequence by the program SEASCAPE [90]. Validation and further assignments are obtained from the ^{15}N -edited TOCSY data. Coordinates for the structurally defined fragments are transformed to a common principal alignment frame with the program REDCAT [91], where the problem of assembling a complete backbone structure becomes one of translation. (Inherent degeneracies in fragment orientations are solved by considering RDCs collected in a second alignment medium [92].) Fragments are translated manually to satisfy a small set of NOEs and covalent end-to-end connections. The final structure is obtained after refinement with Xplor-NIH [40,41], on the basis of all the available experimental data.

The method has been applied to three proteins from *P. furiosus*. In the initial proof-of-principle application [83], the structure of Zn-substituted rubredoxin (54 residues) was obtained after assembly of five final fragments, yielding a backbone RMSD to the known X-ray structure of 1.8 Å for the 2–50 residue stretch. Subsequent applications have focused on genomic targets lacking significant sequence identity with proteins of known structure. Target PF1061 (67 residues) yielded five final fragments ranging from 5 to 18 residues, calculated from an average of 5.5 RDCs per residue [87]. Only 12 interfragment NOEs were required for assembling the structure, which had similar topology and a backbone RMSD of 3.4 Å (residues 2–63 only) to the structure of the closest homologue (33% sequence identity). In addition, partial results on protein target PF0255 have been published [88], where the structure for the 16-residue C-terminal fragment was determined from a total of 124 RDCs, yielding a backbone RMSD of 1.1 Å (residues 45–57 only) relative to the simultaneously solved X-ray structure.

5. Standard molecular representations

Conceptually, the simplest approach towards a structure-oriented implementation is the calculation of standard structural models under each possible combination of assignments for the observed chemical shifts, followed by an evaluation of each structure via its compatibility with the experimental data and covalent information. Indeed, such an approach was attempted in 1982 by Brown et al. [93] (reviewed in Ref. [94]) to overcome the lack of assignments for six residues on the detergent micelle-bound peptide melittin (26 residues). NOE-driven distance geometry calculations assuming different assignment possibilities were performed, suggesting the assignment for two residues. Exhaustive approaches of this kind, however, are not suitable for larger systems. In the remainder of this section we cover efficient structure-oriented methods whose initial trial structures, in contrast to the protocols discussed thus far, already make chemical sense in terms of prior knowledge on covalent connectivity, bond lengths, bond angles, etc.

5.1. AUREMOL

AUREMOL [3,95] is a program package that aims at the full automation of NMR protein structure determination, following a structure-oriented (or “molecule-centered”) approach. The proposed strategy starts from an arbitrary initial structure, for example, an extended conformation of the protein. Chemical shifts are tentatively assigned based on databases containing information on random coil chemical shifts, *J*-couplings, amino acid sequence, available structures from homologous proteins, etc. NMR spectra are simulated and compared with the experimental counterparts. The process proceeds iteratively, by reassigning chemical shifts and recalculating structures until satisfactory convergence.

While AUREMOL remains under development, the basic computational framework and several subroutines are available. Comprehensive computational data structures have been developed to deal with the databases required to store general (chemical structure of amino acids, random coil chemical shifts, etc.) and specific (protein sequence, sample composition, structures from homologous proteins, experimental spectral datasets, etc.) information on the system under study. AUREMOL incorporates existing tools for the preprocessing of NMR spectra, which include those within the program AURELIA [96], an automatic pick picking routine, a Bayesian algorithm for the separation of true signals from noise and other spectral artifacts [97,98], and peak volume integration via iterative segmentation combined with a region-growing algorithm [99]. Since the structure calculation protocol relies on fitting experimental and simulated spectra, RELAX [100] is used for the simulation of 2D and 3D NOESY spectra via a relaxation matrix protocol. For validation of final structures AUREMOL implements the recently developed program AUREMOL-RFAC-3D [101] (a generalization of the older RFAC method [102]), which automatically calculates NMR R-factors from 2D homonuclear and 3D ^{13}C - or ^{15}N -edited NOESY spectra.

5.2. Selection and refinement of de novo predicted structures

The method introduced by Meiler and Baker [26] (here referred to as Rosetta+MCassign for didactical reasons) is based on a Monte Carlo search in coordinate and chemical shift space for a 3D structure and assignments that best satisfy the input NMR data stemming from unassigned NOE, residual dipolar coupling and heteronuclear *J*-correlated experiments. While in spirit the approach is similar to that originally proposed by Atkinson and Saudek [25] (Section 2.3), Meiler and Baker [26] significantly restrict the search for coordinates to those from structures predicted de novo by Rosetta [103] (reviewed in Ref. [104]). Additional guidance is provided by chemical shift prediction from the proposed structures using PROSHIFT [105]. Specifically, representative structures are selected from a large ensemble of models generated from sequence information via Rosetta. An optimal assignment of all

chemical shifts is obtained for the selected structures via a Monte Carlo (MC) search procedure that scores individual assignments based on the consistency of the chemical shift, NOE and RDC data with the 3D model. Subsequently, the structures are scored according to their overall fit to the NMR data, assuming the identified assignments. When tested against eight proteins of known structure ranging in size between 56 and 128 residues, the protocol produced best-scoring models within 3.43–6.67 Å backbone RMSD of the reference, all displaying correct overall folds. It is noteworthy that in the above cases, the NMR data are not enforced as restraints during structure calculation, but instead used as a filter to select de novo predicted models.

The number of correct assignments determined by the MC procedure was found correlated with the quality of the structural model under consideration. Thus, for models 3–6 Å RMSD from the known structure, only 5–40% of the assignments were found correct. For a given structure, however, the frequency of observation of a particular chemical shift assignment in independent MC optimization runs can be exploited to gauge assignment confidence. Highly frequent assignments are assumed correct, leading to assigned NOEs and RDCs, which are enforced as restraints in subsequent Rosetta calculations (RosettaNMR) [48–50]. This leads to improved structures from which additional assignments can be obtained. The iteration of these steps can be successfully implemented provided that the fraction of correct assignments stemming from the initial MC search is not too low. Among the eight test proteins mentioned above, four afforded structural refinement to within 0.60–1.97 Å backbone RMSD, increasing the fraction of correctly assigned chemical shifts up to 70%. Fig. 5 outlines the combined initial model selection stage and subsequent iterative refinement. Similar to other structure-oriented methods [10,13,38], correct folds are obtained without completely correct assignments. This contrasts with the paradigm of conventional high-resolution structure elucidation approaches.

The above-described implementation of the method [26] addresses a worst case scenario since it: (1) assumes no homologous structures are available (hence the reliance on de novo predicted models), (2) takes no advantage of the covalent connectivity information provided by the *J*-correlated experiments (except for pairs of chemical shifts associated to observed RDCs, assumed separated by one or two bonds, respectively), and (3) considers no a priori established assignments. The effect of such extra information on further restricting conformational and assignment space was studied for the fumarate sensor DcuS (140 residues) [106]. The initial structures were generated by de novo prediction and comparative modeling, and used as input to the MC assignment procedure, along with 1,100 chemical shifts (^1H , ^{13}C , and ^{15}N), ~3,000 NOEs, 209 RDCs ($^1D_{\text{N-HN}}$, $^1D_{\text{N-C}\alpha}$, $^1D_{\text{C}\alpha\text{-H}\alpha}$, and $^1D_{\text{C}\alpha\text{-C'}}$), and 315 covalent connectivity restraints derived from the RDCs and one-bond HSQC-type patterns. The NMR data were obtained from ^{13}C - and ^{15}N -edited NOESY, TROSY-HNCO, and CBCACONH experiments, and led to the selection of a model with a backbone RMSD of 6.03 Å (“6-Å model”) relative to the known structure. Addition of 25% of all possible connectivity restraints involving sequential backbone heavy-atom chemical shifts (111 restraints, assumed known from analysis of triple resonance spectra) did not significantly change the model selection outcome, but increased the initial fraction of correctly established assignments, allowing for the determination of confident assignments and refinement of the 6-Å model to 3.67 Å RMSD. Alternatively, the effect of partial assignments obtained in early stages of a traditional manual analysis of the NMR data [107] was tested by refining the initially obtained 6-Å model with 92 assigned N–H $^{\text{N}}$ RDCs and 130 backbone–backbone NOE restraints, which lowered the backbone RMSD to 2.83 Å.

5.3. ITAS

Proposed by Jung et al. [10] for simultaneous assignment and backbone structure determination via residual dipolar couplings, ITAS (iterative assignment and structure) relies on *J*-correlated

experiments to perform partial backbone chemical shift and RDC assignments, that assist in the computation of initial structural models subsequently used in a structure-oriented stage. Specifically, assignments are obtained via the MARS protocol [108], using the general strategy outlined in Section 1.1. The assigned backbone chemical shifts and RDCs are used as input restraints for RosettaNMR [48,49] (reviewed in Ref. [50]), yielding structures which, in turn, help establish additional assignments via the RDC-enhanced version of MARS [109] (Section 5.6.2.3). Only assignments consistent across the 20 lowest-energy RosettaNMR models (out of a total of 1,000) are assumed correct. The structure calculation and RDC-based assignment steps are applied iteratively until the number of assignments converges. Final structures are subjected to energy minimization. The protocol is outlined in Fig. 5, and a potential experimental dataset is listed in Table 1.

ITAS was tested with eight proteins (56–153 residue range) from which 2–82% of the chemical shifts could be initially assigned by MARS. The initial assignments were used to generate 20-structure bundles via RosettaNMR, with average backbone RMSDs as large as 14.5 Å from the a priori known coordinates. RDCs consisted in $^1D_{N-HN}$, $^1D_{C\alpha-C'}$, $^1D_{C\alpha-H\alpha}$ and $^1D_{N-C'}$. All four types of RDC were available for two proteins, while the remaining six proteins involved only three types. After convergence of the ITAS cycle (4–8 iterations), assigned RDCs amounted to 1.7–3.6 per residue, and backbone chemical shifts were, on average, 96% correctly assigned. Despite the small number of misassignments, final structures displayed correct folds within 0.7–5.1 Å backbone RMSD, relative to reference structures. This robustness of RosettaNMR against misassigned data has also been observed in the context of SC-CLOUDS [38] (Section 2.7).

5.4. FastNMR

Building on their ITAS protocol [10] (Section 5.3), Korukottu et al. [11] developed FastNMR. The method improves ITAS folds by incorporating side chain information obviated by the backbone-centered ITAS approach. Specifically, side chain 1H and ^{13}C chemical shifts are assigned by matching the observed chemical shifts to those predicted from ITAS structures with the programs PROSHIFT [105] and SHIFTS [110,111]. The side chain assignments, jointly with the structure and backbone assignments produced by ITAS, are used as input for CANDID/CYANA which automatically assigns NOESY peaks and calculates structures in an iterative fashion [7,62]. The improved NOE-based structures allow for a better prediction of side chain chemical shifts, which, in turn, are used for a second round of chemical shift assignment followed by a new CANDID/CYANA run. Finally, all the available data, namely CANDID-assigned NOEs, ITAS-assigned RDCs, and all assigned chemical shifts, are combined to obtain a high-resolution structure via standard protocols of NMR structural refinement. The method is outlined in Fig. 5.

FastNMR was tested on the proteins konkunitzin-S1 (Conk-S1; 60 residues), konkunitzin-S2 (Conk-S2; 65 residues), and ubiquitin (76 residues). While the structures of both Conk-S1 and ubiquitin had previously been solved by standard NMR methods, the structure of Conk-S2 was determined de novo via FastNMR. A representative set of the NMR experiments used is shown in Table 1. FastNMR structures were determined with an average of 7.7–8.8 NOEs, and 2.1–3.2 RDCs per residue, respectively. RDCs consisted of a combination of either three or all of $^1D_{N-HN}$, $^1D_{C\alpha-C'}$, $^1D_{C\alpha-H\alpha}$, and $^1D_{N-C'}$ types. Conk-S1 and ubiquitin structures deviated by 0.4 and 0.6 Å RMSD from previous NMR models, respectively and, along with the de novo Conk-S2 structure, had backbone precisions < 0.7 Å RMSD within the bundle of the 20 lowest-energy structures.

The separation of the structure calculation process into stages where either RDCs (ITAS stage) or NOEs (CANDID/CYANA stage) are used as the main source of restraints, strengthens the confidence on the accuracy of the produced structures. An incorrect initial RDC-based fold

may be detected by either failure of subsequent CANDID/CYANA iterations to converge, or convergence to a fold significantly different from the initial structure, which, consequently, does not satisfy the RDCs. In addition, the method is fully automated and the complete calculation for each of the three protein test cases was achieved in < 24 hours.

5.5. Structural homology via unassigned residual dipolar couplings

Methods that involve the calculation of residual dipolar couplings from a structural model need to assume certain values for the elements of the 3×3 order matrix S , (Eq. 4; Section 4.3). Indeed, estimating S is the main task of the protocols presented in this section. S contains five independent elements, which, after matrix diagonalization, can be expressed in terms of the eigenvalues that yield the magnitude D_a and rhombicity R of the alignment tensor, as well as the three Euler angles that define the rotation of the initial molecular frame to the one where S is diagonal (the principal axis system). Assuming a structural model, there are several possible strategies for computing the order matrix parameters, two of which have been exploited for rapid database detection of structural homologues, as they do not require prior RDC assignment. They are: (1) prediction from the protein's shape [112], and (2) rotation of the structural model (i.e., variation of the Euler angles) to best fit model-calculated (using Eq. 4) and experimentally measured RDC distributions. The latter strategy requires prior knowledge of D_a and R , which can be obtained via a powder pattern analysis [113]. While (1) is more straightforward since it requires no rotation search, (2) is more general as it makes no assumption on the mechanism of molecular alignment.

The general approach for structural homology search proposed by Valafar and Prestegard [114], and Langemead et al. [47] consists of estimating S from each structure in a database, from which backbone $^1\text{H}^{\text{N}}\text{-}^{15}\text{N}$ RDCs can be calculated (Eq. 4) and their distribution compared to that experimentally obtained from an ^{15}N -labeled protein of unknown structure. Similar $^1D_{\text{N-HN}}$ distributions are interpreted to indicate structural similarity (Fig. 10). Valafar and Prestegard tested both options (1) and (2) for estimating the order matrix, while Langemead et al. considered (2) only. Using the more general option (2), Valafar and Prestegard succeeded in finding the correct fold family of a 79-residue helical protein, regardless of sequence similarity, from a database of 20 structures comprising nine different folds. Langemead et al. compiled a database of ~2,500 structures, used to search for structural homologues of ubiquitin (76 residues), the B1 domain of *Streptococcal* protein G (56 residues) and lysozyme (129 residues). In each case, homologous structures were found with as little as 10% sequence similarity. Langemead et al. incorporated their algorithm within the program NVR (nuclear vector replacement) [47] (Section 5.6.2.2). Thus, these methods may help identify structural genomics targets that have no structural homology in existing databases.

5.6. Assignments from known structures

In those cases where the structure of the protein is already known, the main task of structure-oriented methods is the assignment of chemical shifts to be used in further studies. For example, the X-ray model of the protein under consideration can be used to assist in the assignment process, in turn enabling the calculation of the solution structure, which might exhibit differences relative to its crystallized counterpart. Furthermore, availability of assigned chemical shifts is a prerequisite for ligand binding and protein dynamics studies, beyond the scope of this review. The relevance of structure-driven assignment methods is also underscored by the success of protocols presented in previous sections where the structure is "known" thanks to a powerful de novo prediction algorithm [26], availability of structural homologues [106], or the combination of partial chemical shift assignments with a structure calculation protocol that incorporates significant prior structural knowledge [10,11,38].

The problem of assigning chemical shifts given the 3D structure has been of interest since early in the development of peptide and protein NMR (e.g., Refs. [115,116]). As a result, several automated or semi-automated methods—mostly based on either NOEs or RDCs—have emerged for this purpose. The common thread behind these methods is the search for assignments that minimize the difference between the above-mentioned NMR parameters calculated from the structure and those experimentally observed. In what follows we describe several methods for structure-assisted chemical shift assignment, paying particular attention to stand-alone programs not discussed in previous sections.

5.6.1. NOE-based assignment—NOE-based methods search for chemical shift assignments that, given the input 3D model, reproduce the experimental NOESY connectivities. The fit between calculated and observed NOEs can be cast in terms of a pseudo-energy to be minimized. The resulting energy landscape is usually rough, and several strategies are implemented to avoid entrapment in local minima. The program ALFA (algorithm for fast assignment) [117] uses a combinatorial optimization strategy, focusing on the assignment of small parts of the amino acid sequence at any given time. ST2NMR [118], on the other hand, minimizes its energy function via a Monte Carlo procedure, restricting search space by only allowing assignments of spin systems to residues of the appropriate type. Similarly, MONTE [53] performs a Monte Carlo search combined with simulated annealing, and has recently been implemented as part of a direct method, using an unconnected ^1H configuration (a cloud) as input 3D model [52] (Section 2.8). The program GARANT (general algorithm for resonance assignment) [119,120] overcomes the roughness of the energy surface by resorting to a general evolutionary algorithm in conjunction with a specific local optimization routine that finds suitable combinations of solutions in the parent generation that will yield improved solutions in the following generation.

GARANT has been integrated with spectral peak picking routines as a complete automated process that leads from raw spectra to assignments [121]. In this approach the peaks originating from different spectra are automatically identified by the program AUTOPSY [122]. After calibration and filtering of the resulting peak lists with the program PICS [121], GARANT assigns the chemical shifts by searching for an optimal mapping of peaks experimentally observed to peaks expected from the input 3D structure. The latter, however, is optional, so that when omitted the method becomes purely assignment-oriented (this feature is shared by ALFA and MONTE, discussed above). In addition, chemical shifts from NMR studies on homologous proteins may be incorporated to improve the expected peak positions.

Xiong et al. [123] have cast the assignment problem in terms of graph theory, adopting a minimalist approach in the requirements for isotopic labeling (^{15}N only) and experimental dataset (^{15}N -HSQC, HNHA, ^{15}N -edited TOCSY, and ^{15}N -edited NOESY). The input structure is represented by a “contact” graph, where protein residues become vertices, and edges link nearby residue positions (e.g., $< 5 \text{ \AA}$ proton-proton separation). Similarly, the spectral information takes the form of an “NMR interaction” graph, whose vertices represent pseudo-residues, and the edges indicate interacting protons (e.g. via NOE). Essentially, the interaction graph is a corrupted, ambiguous version of the contact graph, the goal being to uncover the correspondence in a process referred to as contact replacement. The randomized algorithm used for this purpose exploits the property that both the contact and interaction graphs contain a “chain” of vertices (technically, a Hamilton path) that represents the polypeptide chain. The authors show via synthetic and experimental datasets that the algorithm is efficient and robust, achieving assignment accuracies of 80% in α -helices, 70% in β -sheets, and 60% in loop regions.

The general strategy of finding an assignment hypothesis that minimizes the difference between experimental and structure-calculated NOE data has been applied to the special case of

obtaining stereospecific assignments for methylene protons and isopropyl groups [124,125]. An additional approach to this problem [14-16] has been discussed in Section 2.

5.6.2. RDC-based assignment—In contrast to methods described in the previous section, and protocols for chemical shift assignment without prior knowledge of the structure (see Section 1.1), the global or long-range nature of RDCs allows in certain cases [45,109] for obviating through-bond (*J*-correlated) and through-space (NOE) sequential connectivity information between (pseudo-)residues. Methods for structure-assisted chemical shift assignment based on RDCs search for the best fit between experimental RDCs and those calculated from the structure via Eq. 4. As discussed in Section 5.5 such calculation involves determining the order matrix.

5.6.2.1. RDC matching: In 2002 Hus et al. [45] explored the possibility of assigning backbone and $^{13}\text{C}^\beta$ chemical shifts by exploiting RDC ($^1D_{\text{N-HN}}$, $^1D_{\text{C}^\beta\text{-N}}$ and $^1D_{\text{C}^\beta\text{-C}\alpha}$ in a single alignment medium), as well as $^{13}\text{C}^\alpha$ and $^{13}\text{C}^\beta$ chemical shift data. Observed chemical shifts are grouped into C^β_{i-1} , C^α_{i-1} , C^β_{i-1} , N_i , H^{N}_i pseudo-residues via 3D CBCA(CO)NH-type experiments which supplement those that yield the RDCs. Bipartite graph matching (Fig. 4) is used to assign each experimentally determined pseudo-residue *i* to each counterpart *j* in the input 3D structure, via a cost matrix $C(i, j) = C_{cs}(i, j) + C_{rdc}(i, j)$, where

$$C_{cs}(i, j) = w \sum_{k=1}^{N_{cs}} \left[\frac{\delta(i)_k^{\text{exp}} - \delta(j)_k^{\text{calc}}}{\sigma_k^{cs}} \right]^2 \quad (5)$$

and

$$C_{rdc}(i, j) = \sum_{l=1}^{N_{rdc}} \left[\frac{D(i)_l^{\text{exp}} - D(j)_l^{\text{calc}}}{\sigma_l^{rdc}} \right]^2. \quad (6)$$

Eq. 5 measures the fit between experimental $\delta(i)_k^{\text{exp}}$ and structure-derived $\delta(j)_k^{\text{calc}}$ chemical shifts of type *k* ($^{13}\text{C}^\alpha$ or $^{13}\text{C}^\beta$; $N_{cs} = 2$). Similarly, Eq. 6 measures the fit to the experimental RDC data, which, as mentioned above, involves three different types of internuclear vectors ($N_{rdc} = 3$). σ_k^{cs} and σ_l^{rdc} are the standard deviation of database chemical shift distributions (used for estimating $\delta(j)_k^{\text{calc}}$ and RDC experimental errors, respectively; *w* (Eq. 5) weights C_{cs} relative to C_{rdc} in the overall cost. Indeterminacy of the order matrix *S* is solved via powder pattern analysis [113] and a grid search over the Euler angles (see Section 5.5). At each step in this search, the bipartite graph matching routine is run to define a set of pseudo-residues that show low C_{rdc} (Eq. 6), used to refine *S*. The method was tested on experimental data on ubiquitin (76 residues), using an X-ray model as the input structure; it produced correct assignments for > 90% of the residues.

5.6.2.2. Nuclear vector replacement: The nuclear vector replacement (NVR) approach, originally proposed by Langmead et al. [46,47] (reviewed in Ref. [126]), assigns backbone $^1\text{H}^{\text{N}}$ and ^{15}N chemical shifts, as well as a limited number of NOEs. NVR relies on $^1D_{\text{N-HN}}$ RDCs in two alignment media, H-D exchange, and sparse $\text{H}^{\text{N}}\text{-H}^{\text{N}}$ NOE data extracted from an ^{15}N -edited NOESY, thus requiring only ^{15}N isotope enrichment and short experimental acquisition time compared to that of traditional datasets used for assignment. The estimation of the order matrix, *S*, and the chemical shift assignment are performed in two stages. *S* is first determined in each alignment medium by rotating the input structural model to find the best fit between the histogram of calculated and experimental RDCs (as detailed in Section

5.5, this step can be used independently for structural homology detection [47]). Next, the probability of assignment for each experimentally observed $N-H^N$ vector i to each counterpart j in the input structure is calculated for both alignment media, taking into account the RDC fit and consistency with $H-D$ exchange and NOE data. A best-first strategy is adopted where the most probable assignment consistent across both media is selected. An iterative procedure is implemented in which the confidently assigned RDCs are used to refine S via singular value decomposition [89], and recalculate the assignment probabilities. In the final iterations bipartite graph matching (Fig. 4) is used to establish remaining assignments that could not be unambiguously determined via the above approach. The protocol was tested with the proteins ubiquitin (76 residues), the B1 domain of *Streptococcal* protein G (56 residues) and lysozyme (129 residues), using their structures and those of close homologues, available from X-ray crystallography and NMR studies. The accuracy of the resulting assignments was within the 92–100% range.

The NVR algorithm performs satisfactorily when the structure of either the protein under study or a close homologue ($< 2 \text{ \AA}$ backbone RMSD from the former) is available. Recent improvements have increased NVR's robustness against structural noise, and extended its circle of convergence so that distant models can be used [43]. Considering that NVR and the contact replacement method [123] (Section 5.6.1) rely on ^{15}N -labeled samples but different structural probes (RDCs vs. NOEs, respectively), Xiong et al. [123] have proposed the two as complementary protocols.

5.6.2.3. MARS: Reported by Jung and Zweckstetter [108,109] in 2004, MARS is a program for backbone chemical shift assignment that applies to a wide variety of NMR experiments. In addition to deriving assignments in the absence of structural information (i.e., in strict assignment-oriented mode) [108], the method's effectiveness is significantly enhanced by the incorporation of RDCs and high-resolution crystallographic models [109]. The basic strategy consists in mapping experimentally derived pseudo-residues (e.g., of the type described in Section 5.6.2.1) or pseudo-residue segments onto the protein sequence. Goodness of fit is assessed via a cost function $C(i, j) = C_{cs}(i, j) + C_{rdc}(i, j)$, as specified by Eqs. 5 and 6, respectively. In order to calculate C_{rdc} (Eq. 6), MARS provides several options for the estimation of the order matrix S , the most general of which consists of a grid search over different orientations. All sampled orientations are ranked according to their corresponding $C(i, j)$ values, the lowest one indicating the best estimate for the experimental order matrix. Similar to the NVR method [47] (Section 5.6.2.2), once reliable assignments are obtained in a first run of the algorithm, singular value decomposition generates more accurate order matrix parameters [89], to be used in a second assignment run. Given an order matrix estimate, the MARS assignment strategy is similar to the one it implements when no structural model is available (i.e., $C(i, j) = C_{cs}(i, j)$; Eq. 5) [108], which consists of an exhaustive search that simultaneously establishes pseudo-residue segments (step (3) in Section 1.1) and segment mapping onto the protein sequence (step (4) in Section 1.1). The incorporation of RDC and structural data allows, in favorable cases, for the mapping of individual pseudo-residues (instead of segments), obviating sequential connectivity information.

The MARS RDC-enhanced assignment strategy was tested on ubiquitin (76 residues) and the two-domain maltose-binding protein (MBP; 370 residues), by resorting to crystallographic models as input structures [109]. In the case of ubiquitin, 91.7% and 83.3% correct assignments (using 1.8 and 2.3 $\text{\AA}$ resolution models, respectively) could be obtained with $^1D_{N-HN}$, $^1D_{C'-N}$ and $^1D_{C'-C\alpha}$ RDCs, without sequential connectivity information. Addition of $^{13}\text{C}^\alpha$ -based sequential connectivity increased the percentage of correct assignments to 100% with both structural models. For the analysis of MBP, the RDC set was similar to that of ubiquitin, supplemented by ($^{13}\text{C}^\alpha$, $^{13}\text{C}^\beta$)-based sequential connectivity. Despite the larger size of the protein, the algorithm yielded 99.1% correct assignments, slightly

dropping to 94.3% when randomly removing 20% of the experimentally determined pseudo-residues. MARS has also been tested successfully using RosettaNMR structural models calculated with partial assignments, as part of the ITAS [10] and FastNMR [11] methods discussed above (Sections 5.3 and 5.4, respectively; Fig. 5).

6. Concluding remarks

Structure-oriented approaches have been documented early in the development of NMR applications to the protein structure elucidation field. Expectedly, because of their strong structural focus, structure-oriented methods have benefited from recent progress in the area of protein structure prediction. Fig. 5 summarizes methods that rely on a salient exponent of that field. Despite the obvious underlying similarities, such protocols exemplify another characteristic of the structure-oriented approaches, namely, their heterogeneity. While SC-CLOUDS [38] starts from random spatial distributions of unconnected protons/proton groups, Rosetta+MCassign [26], ITAS [10], and FastNMR [11] resort to initial all-covalent structures representing compact folds. Furthermore, while both SC-CLOUDS and Rosetta+MCassign start by proposing structural models that immediately launch the structure determination pipeline, ITAS and FastNMR postpone their formulation until after partial assignments are obtained in a conventional fashion.

The structure-oriented approach has been successfully applied to structural genomics protein targets (e.g., proteins in Fig. 8), and has been extended to the determination of the binding pose of small ligands within protein–ligand complexes [44,127]. It is our conviction that structure-oriented methods shall continue to expand, if not as exclusive tools, by offering a different paradigm to tackle the variety of specific practical challenges that arise when solving protein structure and function via NMR.

Acknowledgments

Alexander Lemak provided the list of protein structures used to generate Fig. 8. This work was supported by the U. S. Public Health Service, NIH Grant GM-067964.

Abbreviations

2D	two-dimensional
3D	three-dimensional
4D	four-dimensional
ABACUS	applied BACUS
ADC	antidistance constraint
ALFA	algorithm for fast assignment
ANSRS	assignment of NOESY spectra in real space
ARIA	ambiguous restraints for iterative assignment
AURELIA	automated resonance line assignment
AUTOPSY	automated peak picking for NMR spectroscopy
BACUS	Bayesian analysis of coupled unassigned spins
BPTI	bovine pancreatic trypsin inhibitor
CCPN	collaborative computing project for NMR
CLOUDS	computed location of unassigned spins

CNS	crystallographic/NMR refinement software
Conk-S1	conkunitzin-S1
Conk-S2	conkunitzin-S2
COSY	correlation spectroscopy
D_a	magnitude of the alignment tensor
DG	distance geometry
DGPA	distance geometry proton assignment
FMC	fragment Monte Carlo
FOC	family of clouds
GARANT	general algorithm for resonance assignment
HSQC	heteronuclear single quantum coherence
ISD	inferential structure determination
ISPA	isolated spin pair approximation
ITAS	iterative assignment and structure
MBP	maltose-binding protein
MC	Monte Carlo
MD	molecular dynamics
MUCA	multicanonical
NMR	nuclear magnetic resonance
NOE	nuclear Overhauser effect
NOESY	nuclear Overhauser effect spectroscopy
NVR	nuclear vector replacement
PDB	protein data bank
PDP	probability density profile
PICS	peak improvement by calibration and selection
R	rhombicity of the alignment tensor
RDC	residual dipolar coupling
REDCAT	residual dipolar coupling analysis tool
REDCRAFT	residual dipolar coupling based residue assembly and filter tool
RMSD	root-mean-square deviation
S	Saupe order matrix
SA	simulated annealing
SC-CLOUDS	sparse-constraint CLOUDS
SEASCAPE	sequential assignment by structure and chemical shift assisted probability estimation
TALOS	torsion angle likelihood obtained from shifts and sequence similarity

TOCSY total correlation spectroscopy

References

- [1]. Wüthrich, K. NMR of Proteins and Nucleic Acids. Wiley; New York: 1986.
- [2]. Moseley HN, Montelione GT. Curr. Opin. Struct. Biol 1999;9:635–642. [PubMed: 10508776]
- [3]. Gronwald W, Kalbitzer HR. Prog. Nucl. Magn. Reson. Spectrosc 2004;44:33–96.
- [4]. Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D, Bax A. Proc. Natl. Acad. Sci. USA 2008;105:4685–4690. [PubMed: 18326625]
- [5]. Cavalli A, Salvatella X, Dobson CM, Vendruscolo M. Proc. Natl. Acad. Sci. USA 2007;104:9615–9620. [PubMed: 17535901]
- [6]. Robustelli P, Cavalli A, Dobson CM, Vendruscolo M, Salvatella X. J. Phys. Chem. B 2009;113:7890–7896. [PubMed: 19425536]
- [7]. Güntert P. Prog. Nucl. Magn. Reson. Spectrosc 2003;43:105–125.
- [8]. Englander SW, Wand AJ. Biochemistry 1987;26:5953–5958. [PubMed: 3689754]
- [9]. Xu Y, Xu D, Kim D, Olman V, Razumovskaya J, Jiang T. Comput. Sci. Eng 2002;4:50–52.
- [10]. Jung YS, Sharma M, Zweckstetter M. Angew. Chem. Int. Edit 2004;43:3479–3481.
- [11]. Korukottu J, Bayrhuber M, Montaville P, Vijayan V, Jung YS, Becker S, Zweckstetter M. Angew. Chem. Int. Edit 2007;46:1176–1179.
- [12]. AB E, Pugh DJR, Kaptein R, Boelens R, Bonvin AMJJ. J. Am. Chem. Soc 2006;128:7566–7571. [PubMed: 16756312]
- [13]. Folmer RHA, Hilbers CW, Konings RNH, Nilges M. J. Biomol. NMR 1997;9:245–258. [PubMed: 9229503]
- [14]. Holak TA. FEBS Lett 1989;242:218–224. [PubMed: 2914604]
- [15]. Pardi A, Hare DR, Selsted ME, Morrison RD, Bassolino DA, Bach AC. J. Mol. Biol 1988;201:625–636. [PubMed: 2843652]
- [16]. Weber PL, Morrison R, Hare D. J. Mol. Biol 1988;204:483–487. [PubMed: 3221395]
- [17]. Rieping W, Habeck M, Nilges M. Science 2005;309:303–306. [PubMed: 16002620]
- [18]. Habeck M, Nilges M, Rieping W. Phys. Rev. E 2005;72:031912. doi: 10.1103/PhysRevE.72.031912.
- [19]. Malliavin TE, Rouh A, Delsuc MA, Lallemand JY. C. R. Acad. Sci. Paris Serie II 1992;315:653–659.
- [20]. Kraulis PJ. J. Mol. Biol 1994;243:696–718. [PubMed: 7525970]
- [21]. Brüschweiler R, Blackledge M, Ernst RR. J. Biomol. NMR 1991;1:3–11. [PubMed: 1841688]
- [22]. de Vlieg J, Boelens R, Scheek RM, Kaptein R, van Gunsteren WF, Israel J. Chem 1986;27:181–188.
- [23]. Rejante MR, Llinás M. Eur. J. Biochem 1994;221:939–949. [PubMed: 8181476]
- [24]. Atkinson, RA.; Saudek, V. Dynamics and the Problem of Recognition in Biological Macromolecules. Jardetzky, O.; Lefèvre, JF., editors. Plenum Press; New York: 1996. p. 49-55.
- [25]. Atkinson RA, Saudek V. J. Chem. Soc. Faraday T 1997;93:3319–3323.
- [26]. Meiler J, Baker D. Proc. Natl. Acad. Sci. USA 2003;100:15404–15409. [PubMed: 14668443]
- [27]. Atkinson RA, Saudek V. FEBS Lett 2002;510:1–4. [PubMed: 11755519]
- [28]. Grishaev A, Llinás M. Proc. Natl. Acad. Sci. USA 2002;99:6707–6712. [PubMed: 12011433]
- [29]. Grishaev A, Llinás M. Proc. Natl. Acad. Sci. USA 2002;99:6713–6718. [PubMed: 12011434]
- [30]. Grishaev A, Llinás M. Method. Enzymol 2005;394:261–295.
- [31]. Briknarová K, Grishaev A, Bányai L, Tordai H, Patthy L, Llinás M. Structure Fold. Des 1999;7:1235–1245. [PubMed: 10545322]
- [32]. Marti DN, Schaller J, Llinás M. Biochemistry 1999;38:15741–15755. [PubMed: 10625440]
- [33]. Madrid M, Llinás E, Llinás M. J. Magn. Reson 1991;93:329–346.

- [34]. Vranken WF, Boucher W, Stevens TJ, Fogh RH, Pajon A, Llinás M, Ulrich EL, Markley JL, Ionides J, Laue ED. *Proteins* 2005;59:687–696. [PubMed: 15815974]
- [35]. Jeffreys H. *Proc. Roy. Soc. A* 1946;186:453–461.
- [36]. Habeck M, Nilges M, Rieping W. *Phys. Rev. Lett* 2005;94:018105. doi: 10.1103/PhysRevLett.94.018105. [PubMed: 15698139]
- [37]. Habeck, M.; Rieping, W.; Nilges, M. In: Rainer, F.; Roland, P.; von Toussaint, U., editors. *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, AIP; 2004; p. 119–126.
- [38]. Bermejo GA, Llinás M. *J. Am. Chem. Soc* 2008;130:3797–3805. [PubMed: 18318535]
- [39]. Goto NK, Gardner KH, Mueller GA, Willis RC, Kay LE. *J. Biomol. NMR* 1999;13:369–374. [PubMed: 10383198]
- [40]. Schwieters CD, Kuszewski JJ, Clore GM. *Prog. Nucl. Magn. Reson. Spectrosc* 2006;48:47–62.
- [41]. Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM. *J. Magn. Reson* 2003;160:65–73. [PubMed: 12565051]
- [42]. Tashiro M, Tejero R, Zimmerman DE, Celda B, Nilsson B, Montelione GT. *J. Mol. Biol* 1997;272:573–590. [PubMed: 9325113]
- [43]. Apaydin MS, Conitzer V, Donald BR. *J. Biomol. NMR* 2008;40:263–276. [PubMed: 18365752]
- [44]. Constantine KL, Davis ME, Metzler WJ, Mueller L, Claus BL. *J. Am. Chem. Soc* 2006;128:7252–7263. [PubMed: 16734479]
- [45]. Hus JC, Prompers JJ, Brüschweiler R. *J. Magn. Reson* 2002;157:119–123. [PubMed: 12202140]
- [46]. Langmead CJ, Donald BR. *J. Biomol. NMR* 2004;29:111–138. [PubMed: 15014227]
- [47]. Langmead CJ, Yan A, Lilien R, Wang LC, Donald BR. *J. Comput. Biol* 2004;11:277–298. [PubMed: 15285893]
- [48]. Bowers PM, Strauss CEM, Baker D. *J. Biomol. NMR* 2000;18:311–318. [PubMed: 11200525]
- [49]. Rohl CA, Baker D. *J. Am. Chem. Soc* 2002;124:2723–2729. [PubMed: 11890823]
- [50]. Rohl CA. *Method. Enzymol* 2005;394:244–260.
- [51]. Zheng DY, Huang YPJ, Moseley HNB, Xiao R, Aramini J, Swapna GVT, Montelione GT. *Protein Sci* 2003;12:1232–1246. [PubMed: 12761394]
- [52]. Schedlbauer A, Auer R, Ledolter K, Tollinger M, Klobner K, Lichtenecker R, Ruedisser S, Hommel U, Schmid W, Konrat R, Kontaxis G. *J. Biomol. NMR* 2008;42:111–127. [PubMed: 18762865]
- [53]. Hitchens TK, Lukin JA, Zhan YP, McCallum SA, Rule GS. *J. Biomol. NMR* 2003;25:1–9. [PubMed: 12566995]
- [54]. Morris RJ, Perrakis A, Lamzin VS. *Acta Crystallogr. D* 2002;58:968–975. [PubMed: 12037299]
- [55]. Pozharski E, Grishaev A, Greenspan D, Tulinsky A, Llinás M, Petsko G, Ringe D. *Biophys. J* 2002;82:470A.
- [56]. Juhás P, Cherba DM, Duxbury PM, Punch WF, Billinge SJL. *Nature* 2006;440:655–658. [PubMed: 16572167]
- [57]. Grishaev A, Llinás M. *J. Biomol. NMR* 2004;28:1–10. [PubMed: 14739635]
- [58]. Grishaev A, Llinás M. *J. Biomol. NMR* 2002;24:203–213. [PubMed: 12522308]
- [59]. Brünger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL. *Acta Crystallogr. D* 1998;54:905–921. [PubMed: 9757107]
- [60]. Linge JP, Habeck M, Rieping W, Nilges M. *Bioinformatics* 2003;19:315–316. [PubMed: 12538267]
- [61]. Linge JP, O'Donoghue SI, Nilges M. *Method. Enzymol* 2001;339:71–90.
- [62]. Herrmann T, Güntert P, Wüthrich K. *J. Mol. Biol* 2002;319:209–227. [PubMed: 12051947]
- [63]. Oshiro CM, Kuntz ID. *Biopolymers* 1993;33:107–115. [PubMed: 8381306]
- [64]. Grishaev A, Steren CA, Wu B, Pineda-Lucena A, Arrowsmith C, Llinás M. *Proteins* 2005;61:36–43. [PubMed: 16080153]
- [65]. Cornilescu G, Delaglio F, Bax A. *J. Biomol. NMR* 1999;13:289–302. [PubMed: 10212987]
- [66]. Lemak A, Steren CA, Arrowsmith CH, Llinás M. *J. Biomol. NMR* 2008;41:29–41. [PubMed: 18458824]

- [67]. Berg BA, Celik T. Phys. Rev. Lett 1992;69:2292–2295. [PubMed: 10046447]
- [68]. Berg BA, Neuhaus T. Phys. Lett. B 1991;267:249–253.
- [69]. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. J. Chem. Phys 1953;21:1087–1092.
- [70]. Kaustov L, Lukin J, Lemak A, Duan SL, Ho M, Doherty R, Penn LZ, Arrowsmith CH. J. Biol. Chem 2007;282:11300–11307. [PubMed: 17298945]
- [71]. Yee A, Gutmanas A, Arrowsmith CH. Curr. Opin. Struct. Biol 2006;16:611–617. [PubMed: 16942869]
- [72]. Sheng Y, Laister RC, Lemak A, Wu B, Tai E, Duan SL, Lukin J, Sunnerhagen M, Srisailam S, Karra M, Benchimol S, Arrowsmith CH. Nat. Struct. Mol. Biol 2008;15:1334–1342. [PubMed: 19043414]
- [73]. Srisailam S, Lukin JA, Lemak A, Yee A, Arrowsmith CH. J. Biomol. NMR 2006;36(Suppl. 1):27. [PubMed: 16703420]
- [74]. Wu B, Yee A, Pineda-Lucena A, Semesi A, Ramelot TA, Cort JR, Jung JW, Edwards A, Lee W, Kennedy M, Arrowsmith CH. Protein Sci 2003;12:2831–2837. [PubMed: 14627743]
- [75]. Yee A, Chang XQ, Pineda-Lucena A, Wu B, Semesi A, Le B, Ramelot T, Lee GM, Bhattacharyya S, Gutierrez P, Denisov A, Lee CH, Cort JR, Kozlov G, Liao J, Finak G, Chen L, Wishart D, Lee W, McIntosh LP, Gehring K, Kennedy MA, Edwards AM, Arrowsmith CH. Proc. Natl. Acad. Sci. USA 2002;99:1825–1830. [PubMed: 11854485]
- [76]. Blackledge M. Prog. Nucl. Magn. Reson. Spectrosc 2005;46:23–61.
- [77]. Lipsitz RS, Tjandra N. Annu. Rev. Biophys. Biomol. Struct 2004;33:387–413. [PubMed: 15139819]
- [78]. de Alba E, Tjandra N. Prog. Nucl. Magn. Reson. Spectrosc 2002;40:175–197.
- [79]. Bax A, Kontaxis G, Tjandra N. Method. Enzymol 2001;339:127–174.
- [80]. Prestegard JH, Al-Hashimi HM, Tolman JR. Q. Rev. Biophys 2000;33:371–424. [PubMed: 11233409]
- [81]. Delaglio F, Kontaxis G, Bax A. J. Am. Chem. Soc 2000;122:2142–2143.
- [82]. Hus JC, Marion D, Blackledge M. J. Am. Chem. Soc 2001;123:1541–1542. [PubMed: 11456746]
- [83]. Tian F, Valafar H, Prestegard JH. J. Am. Chem. Soc 2001;123:11791–11796. [PubMed: 11716736]
- [84]. Andrec M, Du P, Levy RM. J. Biomol. NMR 2001;21:335–347. [PubMed: 11824753]
- [85]. Wedemeyer WJ, Rohl CA, Scheraga HA. J. Biomol. NMR 2002;22:137–151. [PubMed: 11883775]
- [86]. Haliloglu T, Kolinski A, Skolnick J. Biopolymers 2003;70:548–562. [PubMed: 14648765]
- [87]. Valafar H, Mayer KL, Bougault CM, LeBlond PD, Jenney FE Jr, Brereton PS, Adams MW, Prestegard JH. J. Struct. Funct. Genomics 2004;5:241–254. [PubMed: 15704012]
- [88]. Prestegard JH, Mayer KL, Valafar H, Benison GC. Method. Enzymol 2005;394:175–209.
- [89]. Losonczi JA, Andrec M, Fischer MW, Prestegard JH. J. Magn. Reson 1999;138:334–342. [PubMed: 10341140]
- [90]. Morris LC, Valafar H, Prestegard JH. J. Biomol. NMR 2004;29:1–9. [PubMed: 15017135]
- [91]. Valafar H, Prestegard JH. J. Magn. Reson 2004;167:228–241. [PubMed: 15040978]
- [92]. Al-Hashimi HM, Valafar H, Terrell M, Zartler ER, Eidsness MK, Prestegard JH. J. Magn. Reson 2000;143:402–406. [PubMed: 10729267]
- [93]. Brown LR, Braun W, Kumar A, Wüthrich K. Biophys. J 1982;37:319–328. [PubMed: 6275926]
- [94]. Braun W. Q. Rev. Biophys 1987;19:115–157. [PubMed: 3310077]
- [95]. Gronwald W, Brunner K, Kirchhöfer R, Nasser A, Trenner J, Ganslmeier B, Riepl H, Ried A, Scheiber J, Elsner R, Neidig KP, Kalbitzer HR. Bruker Reports 2004;154-155:11–14.
- [96]. Neidig KP, Geyer M, Görler A, Antz C, Saffrich R, Beneicke W, Kalbitzer HR. J. Biomol. NMR 1995;6:255–270.
- [97]. Antz C, Neidig KP, Kalbitzer HR. J. Biomol. NMR 1995;5:287–296.
- [98]. Schulte AC, Görler A, Antz C, Neidig KP, Kalbitzer HR. J. Magn. Reson 1997;129:165–172. [PubMed: 9441881]
- [99]. Geyer M, Neidig KP, Kalbitzer HR. J. Magn. Reson. Series B 1995;109:31–38.

- [100]. Görler A, Gronwald W, Neidig KP, Kalbitzer HR. J. Magn. Reson 1999;137:39–45. [PubMed: 10053131]
- [101]. Gronwald W, Brunner K, Kirchofer R, Trenner J, Neidig KP, Kalbitzer HR. J. Biomol. NMR 2007;37:15–30. [PubMed: 17136423]
- [102]. Gronwald W, Kirchhofer R, Görler A, Kremer W, Ganslmeier B, Neidig KP, Kalbitzer HR. J. Biomol. NMR 2000;17:137–151. [PubMed: 10921778]
- [103]. Simons KT, Kooperberg C, Huang E, Baker D. J. Mol. Biol 1997;268:209–225. [PubMed: 9149153]
- [104]. Rohl CA, Strauss CEM, Misura KMS, Baker D. Method. Enzymol 2004;383:66–93.
- [105]. Meiler J. J. Biomol. NMR 2003;26:25–37. [PubMed: 12766400]
- [106]. Meiler J, Baker D. J. Magn. Reson 2005;173:310–316. [PubMed: 15780923]
- [107]. Pappalardo L, Janausch IG, Vijayan V, Zientz E, Junker J, Peti W, Zweckstetter M, Unden G, Griesinger C. J. Biol. Chem 2003;278:39185–39188. [PubMed: 12907689]
- [108]. Jung YS, Zweckstetter M. J. Biomol. NMR 2004;30:11–23. [PubMed: 15452431]
- [109]. Jung YS, Zweckstetter M. J. Biomol. NMR 2004;30:25–35. [PubMed: 15452432]
- [110]. Ösapay K, Case DA. J. Biomol. NMR 1994;4:215–230. [PubMed: 8019135]
- [111]. Xu XP, Case DA. J. Biomol. NMR 2001;21:321–333. [PubMed: 11824752]
- [112]. Zweckstetter M, Bax A. J. Am. Chem. Soc 2000;122:3791–3792.
- [113]. Clore GM, Gronenborn AM, Bax A. J. Magn. Reson 1998;133:216–221. [PubMed: 9654491]
- [114]. Valafar H, Prestegard JH. Bioinformatics 2003;19:1549–1555. [PubMed: 12912836]
- [115]. Llinás M. Struct. Bond 1973;17:135–220.
- [116]. Dubs A, Wagner G, Wüthrich K. Biochim. Biophys. Acta 1979;577:177–194. [PubMed: 311660]
- [117]. Bernstein R, Cieslar C, Ross A, Oschkinat H, Freund J, Holak TA. J. Biomol. NMR 1993;3:245–251.
- [118]. Pristovšek P, Rüterjans H, Jerala R. J. Comput. Chem 2002;23:335–340. [PubMed: 11908496]
- [119]. Bartels C, Billeter M, Güntert P, Wüthrich K. J. Biomol. NMR 1996;7:207–213.
- [120]. Bartels C, Güntert P, Billeter M, Wüthrich K. J. Comput. Chem 1997;18:139–149.
- [121]. Malmodin D, Papavoine CHM, Billeter M. J. Biomol. NMR 2003;27:69–79. [PubMed: 12878842]
- [122]. Koradi R, Billeter M, Engeli M, Güntert P, Wüthrich K. J. Magn. Reson 1998;135:288–297. [PubMed: 9878459]
- [123]. Xiong F, Pandurangan G, Bailey-Kellogg C. Bioinformatics 2008;24:i205–i213. [PubMed: 18586716]
- [124]. Güntert P, Braun W, Wüthrich K. J. Mol. Biol 1991;217:517–530. [PubMed: 1847217]
- [125]. Pristovšek P, Franzoni L. J. Comput. Chem 2006;27:791–797. [PubMed: 16526035]
- [126]. Donald BR, Martin J. Prog. Nucl. Magn. Reson. Spectrosc 2009;55:101–127. [PubMed: 20160991]
- [127]. Metzler, WJ.; Claus, BL.; McDonnell, PA.; Johnson, SR.; Goldfarb, V.; Davis, ME.; Mueller, L.; Constantine, KL. Fragment-Based Drug Discovery: A Practical Approach. Zartler, E.; Shapiro, M., editors. John Wiley & Sons; Hoboken, N.J.: 2008. p. 99–134.
- [128]. Koradi R, Billeter M, Wüthrich K. J. Mol. Graphics 1996;14:51–55.

Fig. 1.

Probability of NOE observation (ADC potential). Curves represent different types of proton (s)–proton(s) pairs. From shorter to longer distances at which curves intersect $P_{\text{noe}}=0.5$ (dashed line), pairs are: $\text{NH}_2\text{--NH}_2$, $\text{H}^{\text{N}}\text{--NH}_2$, $\text{H}^{\text{N}}\text{--H}^{\text{N}}$, $\text{CH}_3\text{--NH}_2$, $\text{H}^{\text{N}}\text{--CH}_3$, and $\text{CH}_3\text{--CH}_3$, respectively. Curves were calculated by simulating ^{15}N - and ^{13}C -edited NOESY data on the Z domain of *Staphylococcal* protein A, assuming a 350-ms mixing time and a rotational correlation time of 3.2 ns. Figure adapted from Ref. [38].

Fig. 2. SC-CLOUDS on the Z domain of *Staphylococcal* protein A. (a) Backbone trace of reference NMR structure generated via an assignment-oriented method using a fully protonated sample (PDB code: 2spz). (b) Backbone H^N atoms from 2spz. (c) SC-CLOUDS-derived backbone H^N atoms using ADCs (lowest-energy cloud). (d) SC-CLOUDS-derived backbone H^N atoms without using ADCs (lowest-energy cloud). The coloring represents a blue (N-terminus) to red (C-terminus) gradient. Figure adapted from Ref. [38], generated with the help of MOLMOL [128].

Fig. 3.

Methyl distribution in clouds of the Z domain of *Staphylococcal* protein A. Methyl pseudo-atoms from the 50 lowest-energy clouds generated via SC-CLOUDS are superimposed to the minimum-energy cloud. (a) All methyls are shown with the exception of those of the N-terminal Val residue, which are poorly defined. Methyls within individual Val/Leu isopropyl groups are equally shaded. (b, c, and d) Selected isopropyl groups are expanded, and the two methyls within each isopropyl are identified by different shades. Figure generated with the help of MOLMOL [128].

Fig. 4. Bipartite graph matching. (a) Fully connected (i.e., *complete*) bipartite graph with two sets of n vertices each. (b) Matching involving all vertices in the graph (a *perfect* matching).

Fig. 5. SC-CLOUDS [38], Rosetta+MCassign [26], ITAS [10], and FastNMR [11] strategies. The three methods share a RosettaNMR-based structure calculation/assignment cycle. The structure-aided assignment stage consists of a distance-based bipartite graph matching protocol (SC-CLOUDS), a Monte Carlo assignment optimization (Rosetta+MCassign), or the RDC-enhanced MARS implementation [109] (ITAS, FastNMR). FastNMR further refines the “final” structures (see text).

Fig. 6.

Proxy atoms (labeled with a question mark) interact with other atoms via unassigned NOEs (dotted lines). A proxy atom may place itself close to an unassigned atom (dark grey), suggesting the assignment, and/or indirectly connect assigned atoms (light grey). As a result, the fold calculation, which also includes all available assigned NOEs (dashed lines), is improved. Figure adapted from Ref. [12].

Fig. 7.
Basic ABACUS assignment strategy.

Fig. 8.

Genomic protein structures determined with the help of ABACUS as of December 2009. From left to right, top to bottom, the PDB codes are (number of residues indicated in parenthesis): 2k2c (137), 2jn4 (66), 2akl (116), 2k1b (55), 2k2d (79), 2kp6 (82), 2k4v (125), 2kgo (108), 2jq5 (128), 2gpf (72), 2kk4 (95), 2ida (102), 2jxx (97), 2jq4 (83), 2k54 (123), 2keo (89), 2jtv (65), 2jya (103), 2joq (91), 1ne3 (68), 2jng (101), 2kfv (92), 2knr (121), 2kzj (2×122, homodimer), 1ryj (70), 2k7i (2×62, homodimer), 2k2p (85), 2kco (133), 2klc (101), 2k8e (129), and 2k28 (60). The coloring (shown in Web version only) represents a blue (N-terminus) to red (C-terminus) gradient. Figure generated with the help of MOLMOL [128].

Fig. 9. Basic backbone fragment in REDCRAFT. Peptide planes are indicated by continuous lines. Interatomic vectors associated to potentially measurable RDCs are indicated by dashed lines. Figure generated with the help of MOLMOL [128].

Figure 10.

Probability density profile (PDP) estimated from backbone N-H^N RDC distributions calculated for three similar-sized proteins. 1IIR and 1F0K are structurally homologous glycosyltransferases (2.2-Å backbone RMSD) with low sequence identity (< 15%). ARF has a dissimilar structure to the glycosyltransferases. Order matrices were predicted from protein shape (see text). Figure adapted from Ref. [114].

Table 1

Automated structure-oriented methods for protein NMR data analysis

Method	Section ^a	Experimental input	Initial structure
CLOUDS / SC-CLOUDS ^b	2.5 / 2.7	¹ H- ¹ H NOESY ^c / ¹⁵ N-NOESY, ¹³ C-NOESY ^c	Random cloud of protons ± small covalent fragments
DGPA	4.1	<i>J</i> -correlated (to establish H ^N , H ^α , H ^β spin systems), NOESY ^c	Random cloud of amino acid fragments
ABACUS	4.2	HNCO, CBCA(CO)NH, HBHA(CO)NH, CC(CO)NH-TOCSY, HC(CO)NH-TOCSY HC(C)H-TOCSY, (H)CCH-TOCSY, ¹⁵ N-NOESY, ¹³ C-NOESY ^d	Random cloud of amino acid fragments
Direct RDC	4.3	¹⁵ N-HSQC, E.COSY-HNCA, HNCO, ¹⁵ N-NOESY, ¹⁵ N-TOCSY	Backbone fragments defined in torsion angle space with no translational information
Proxy residues	3	Conventional ^e	Random protein conformation + unconnected fragments (proxy residues)
Floating chirality	2, 2.9	Conventional ^e	Random protein conformation
Rosetta+MCassign	5.2	¹⁵ N-NOESY, ¹³ C-NOESY, TROSY-HNCO, CBCACONH ^d	Protein model predicted by Rosetta de novo
ITAS ^f	5.3	HNCACB, TROSY-HNCO, CBCA(CO)NH ^g	RosettaNMR model from partial backbone chem. shift and RDC assignments
FastNMR ^f	5.4	HNCACB, H(CCO)NH-TOCSY, C(CO)NH-TOCSY, 2D ¹⁵ N- ¹³ C ^γ and ¹³ C- ¹³ C ^γ diff. experiments, TROSY-HNCO, CBCA(CO)NH, ¹⁵ N-NOESY, ¹³ C-NOESY	RosettaNMR model from partial backbone chem. shift and RDC assignments
NVR and related (structural homology search)	5.5	Experiments for ¹⁵ N- ¹ H ^N RDCs	Protein structure from database
Contact replacement	5.6.1	¹⁵ N-HSQC, HNHA, ¹⁵ N-TOCSY, ¹⁵ N-NOESY	Known structure of the protein or close homologue
RDC matching	5.6.2.1	Experiments for backbone RDCs, 3D CBCA(CO)NH-type	Known structure of the protein or close homologue
NVR	5.6.2.2	Experiments for ¹⁵ N- ¹ H ^N RDCs, H-D exchange, sparse H ^N -H ^N NOEs	Known structure of the protein or close homologue

^aSection in the text.^bDirect NOE-based methods that demonstrated the complete routing from experimental data to structure calculation.^cUnambiguous crosspeaks.^dExperiments used in the latest application of the method (ABACUS: Ref. [66], Rosetta+MCassign: Ref. [106]).^ePartial assignments obtained via conventional NMR analysis are required prior to the implementation of the method.^fExploit the MARS method [109] (Section 5.6.2.3), thus not explicitly included in the table.^gITAS paper [10] does not explicitly specify the experiments required; those used in the ITAS cycle within FastNMR [11] are listed instead.

Table 2

Comparison of CLOUDS and Sparse-Constraint CLOUDS methods

	CLOUDS	SC-CLOUDS
Protein labeling	Fully protonated	Highly deuterated
Experiments used	Homonuclear NOESY	¹⁵ N-NOESY, ¹³ C-NOESY
NOE-distance estimation	Relaxation matrix	Isolated spin pair approximation
Accuracy of clouds (RMSD/Å) ^a	0.5–0.8	6.1
Basis of cloud interpretation	Database of high-resolution structures	Sum-of-distances minimization
Structure calculation	Fit to proton density from a family of clouds (FOC)	RosettaNMR

^aH¹N RMSD to structure obtained via traditional NMR approach on a fully protonated sample.