Journal of Biomolecular NMR **30:** 11–23, 2004.
© 2004 *Kluwer Academic Publishers. Printed in the Netherlands.*

Mars – robust automatic backbone assignment of proteins

Young-Sang Jung & Markus Zweckstetter*
Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, D-37077 Göttingen, Germany

Received 8 January 2004; Accepted 12 May 2004

Key words: automated assignment, NMR, software, structural genomics, triple resonance, unfolded protein

Abstract

MARS a program for robust automatic backbone assignment of $^{13}C/^{15}N$ labeled proteins is presented. MARS does not require tight thresholds for establishing sequential connectivity or detailed adjustment of these thresholds and it can work with a wide variety of NMR experiments. Using only $^{13}C^{\alpha}/^{13}C^{\beta}$ connectivity information, MARS allows automatic, error-free assignment of 96% of the 370-residue maltose-binding protein. MARS can successfully be used when data are missing for a substantial portion of residues or for proteins with very high chemical shift degeneracy such as partially or fully unfolded proteins. Other sources of information, such as residue specific information or known assignments from a homologues protein, can be included into the assignment process. MARS exports its result in SPARKY format. This allows visual validation and integration of automated and manual assignment.

Introduction

Backbone resonance assignment is a prerequisite for structure determination of proteins by NMR (Wüthrich, 2003). Especially useful for backbone assignment are triple-resonance experiments on $^{13}C/^{15}N$-labeled protein, such as HNCA, HN(CO)CA, HNCACB and CBCA(CO)NH or HN(CO)CACB. These experiments are the most sensitive triple-resonance experiments and they are also applicable to large deuterated proteins (Bax and Grzesiek, 1993; Riek et al., 1999). They provide information on $^{1}H^{N}_{i}$, $^{15}N_{i}$, $^{13}C^{\alpha}_{i}$, $^{13}C^{\beta}_{i}$ chemical shifts of residue (i) and $^{13}C^{\alpha}_{i-1}$, $^{13}C^{\beta}_{i-1}$ chemical shifts of residue (i − 1). The chemical shifts are assembled into arrays called pseudoresidues, each of them associated with a single $^{1}H^{N}$, ^{15}N root (a single resonance in a ^{15}N-^{1}H HSQC spectrum). Additional connectivity information, as obtained from experiments such as HNCO and HN(CA)CO, is also often included. In the assignment process these pseudoresidues are sequentially linked. The connected segments are then mapped onto the known protein sequence based on the very sensitive relationship between amino acid type and $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts (Moseley and Montelione, 1999).

The assignment process is conceptually very simple and several algorithms have been developed in recent years to automate it. The different approaches can be grouped into two classes. The first group comprises numerical optimization algorithms that try to minimize a global pseudoenergy function or maximize a global 'goodness of fit'. These include simulated annealing (Bartels et al., 1997; Bernstein et al., 1993; Buchler et al., 1997; Lukin et al., 1997), threshold accepting (Leutner et al., 1998), and neuronal networks (Hare and Prestegard, 1994). The second class is based on best-first search strategies (Friedrichs et al., 1994; Meadows et al., 1994; Olson and Markley, 1994). The Montelione group expanded this strategy in their program AUTOASSIGN by propagating constraints from initial confident assignments towards later stages of the assignment process (Zimmerman and Montelione, 1995). A similar approach is used by the program TATAPRO (Atreya et al., 2000). The program MAPPER by Güntert et al. performs an exhaustive search to place connected segments onto the primary sequence and PACES performs an exhaustive search both for es-

*To whom correspondence should be addressed. E-mail: mzwecks@gwdg.de tablishing sequential connectivity and for assignment (Coggins and Zhou, 2003; Guntert et al., 2000).

Both strategies have their advantages and disadvantages. The problem of global optimization algorithms is that they can be trapped in local minima and assess only alternative complete assignments. Best-first strategies, on the other hand, are prone to propagation of errors made in the initial phases of the assignment process. Overall, good progress has been made in automation of backbone assignment for small to medium-sized proteins up to ~20 kDa (Moseley and Montelione, 1999). Especially for larger or partially unfolded proteins, however, automation of resonance assignment is still difficult. Spectral overlap, chemical exchange or incomplete back-exchange of amide protons in deuterated proteins result in an incomplete set of resonances. These missing resonances severely deteriorate commonly used assignment algorithms. Therefore, for proteins above ~20 kDa a significant fraction of manual assignment is still required.

Here we present MARS a program for robust automatic backbone assignment of ^{13}C/^{15}N labeled proteins. MARS simultaneously optimizes the local and global quality of assignment to minimize propagation of initial assignment errors and to extract reliable assignments. Using only ^{13}C$^{\alpha}$/^{13}C$^{\beta}$ connectivity information, MARS allows automatic, error-free assignment of unfolded and large proteins. We demonstrate that MARS is highly robust against missing chemical shifts and reliably distinguishes correct from incorrect assignments. MARS results can be directly read into the program SPARKY, where reliable assignments together with not assigned spin systems can be viewed as sequentially aligned strips. MARS has been tested on 14 proteins ranging in size from the 71-residue Z domain of Staphylococcal protein A to 723-residue malate synthase G, including experimental data from a natively unfolded protein.

Methods

Resonance assignment of ^{13}C/^{15}N-labeled proteins is commonly performed using a five step analysis scheme: (1) pick and filter peaks, and reference resonances across different spectra; (2) group resonances into pseudoresidues (PRs); (3) identify the amino acid type of pseudoresidues; (4) find and link sequential pseudoresidues into segments; (5) map pseudoresidue segments onto the primary sequence (Moseley and Montelione, 1999). Steps (1) and (2) are essential for manual assignment as well as for automatic approaches. Therefore, most NMR analysis software, like FELIX (Hare Research, Bothwell, WA), AURELIA (Neidig et al., 1995), XEASY (Bartels et al., 1995), SPARKY (Kneller and Kuntz, 1993) and NMRView (Johnson and Blevins, 1994) provide tools for peak picking and referencing of multiple NMR spectra (Bartels et al., 1995). For assignment using MARS pseudoresidues should be generated using one of these programs. In principle, steps (1) and (2) could also be performed automatic, however, the key to any successful assignment is reliable distinction between protein resonances and spectral noise. Therefore, in practice, 3D spectra, picked peaks and pseudoresidues are always inspected manually before starting the assignment process, as this can rapidly be done and the quality of picked peaks and pseudoresidues (or assignment strips) is crucial for successful assignment. The approach is further motivated by the fact that in most cases (especially for large proteins) assignment will be done semiautomatically, i.e., assignment results obtained by MARS will be refined visually on the screen.

Key features of MARS are: (1) simultaneous optimization of the local and global quality of assignment, (2) exhaustive search for fragment lengths comprising up to five PRs during linking and mapping, (3) best-first elements for both linking and mapping, (4) combination of the secondary structure prediction program PSIPRED (McGuffin et al., 2000) with statistical chemical shift distributions, which were corrected for neighboring residue effects (Wang and Jardetzky, 2002), to improve identification of likely positions in the primary sequence and (5) assessment of the reliability of fragment mapping by performing multiple assignment runs with 'noise-disturbed' chemical shifts. The overall MARS strategy is outlined in Figure 1 and detailed below.

Input data

The input data for MARS consist of: 1) the primary sequence of the protein, (2) secondary structure prediction data (for example obtained from PSIPRED), (3) an ASCII file that defines assignment parameters, such as the type of available information and chemical shift tolerances for establishing sequential connectivity, and (4) observed intra- and inter-residual chemical shifts grouped into pseudoresidues. A pseudoresidue (PR) comprises experimental chemical shifts that can

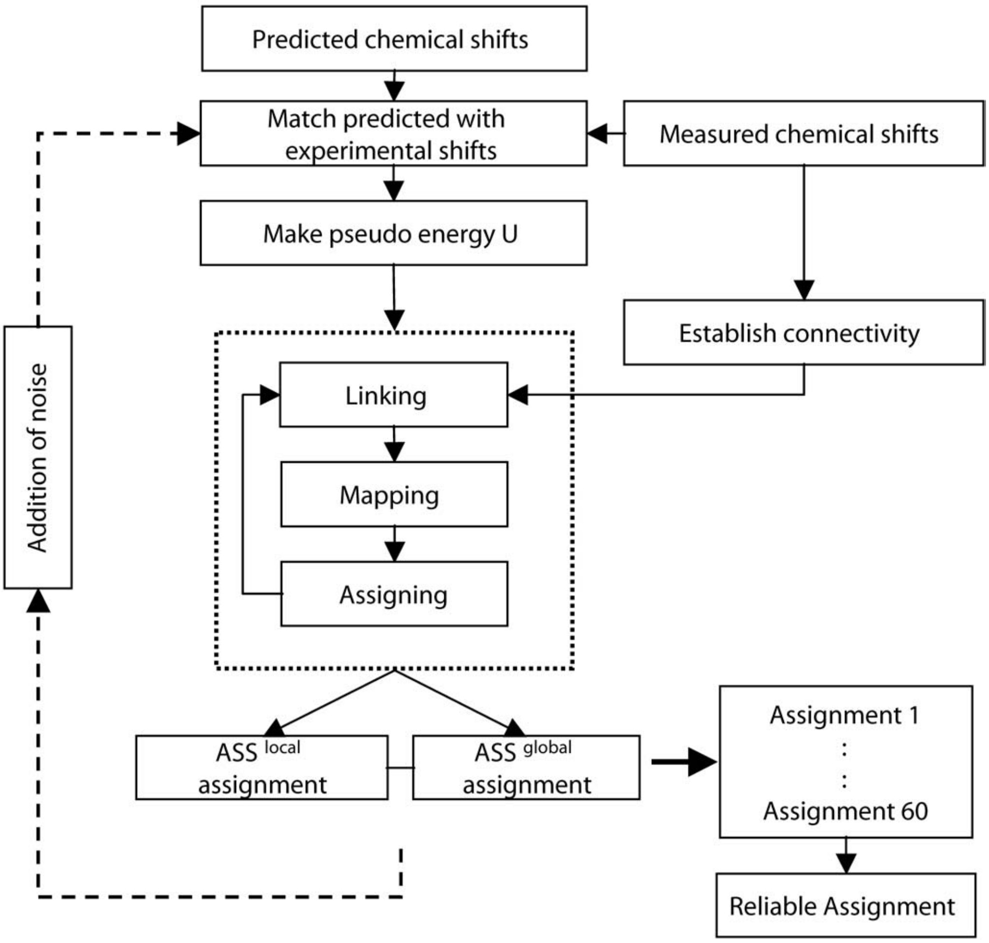

Figure 1. Overview of the MARS assignment procedure. See text for a definition of the two assignment solutions ASS$^{\mathrm{local}}$ and ASS$^{\mathrm{global}}$.

be related to a single amino acid such as $\delta(\mathrm{H}_i^N)$, $\delta(\mathrm{N_i})$, $\delta(\mathrm{C}'_{i-1})$, $\delta(\mathrm{C}_i^\alpha)$, $\delta(\mathrm{C}_{i-1}^\alpha)$, $\delta(\mathrm{C}_i^\beta)$, $\delta(\mathrm{C}_{i-1}^\beta)$ depending on the type of spectra available. All results presented here were obtained with pseudoresidues that contained at least $^1\mathrm{H}^N$ and ^{15}N of residue *i* and ^{13}C′ of residue $i-1$.

MARS does not perform peak picking, referencing of spectra or grouping of peaks into pseudoresidues. In our lab we use SPARKY (Kneller and Kuntz, 1993) to perform these tasks. This allows visual control and refinement of pseudoresidues. When manually inspecting PRs, amide degeneracy can often be resolved, as peak shapes and the higher resolution in a 2D HSQC spectrum can be taken into account. If H^N/N overlap remains, multiple spin systems should be provided to MARS comprising the full set of possible combinations of peaks. In order to avoid an unreasonable high number of PRs in these cases, ambiguous peaks can also be partially discarded, as MARS does not favor pseudoresidues with more complete chemical shift information during the assignment process. The suspicious peaks can be reinserted when running MARS a second or third time, after an initial MARS run was performed, the assignment results were visually validated using SPARKY and verified assignments were fixed.

Besides C$^\alpha$/C$^\beta$ connectivity information, MARS can use sequential information from HNCO/ HN(CA)CO and H^N-H^N NOESY spectra. Moreover, information about the amino acid type of a pseudoresidue can be included into MARS assignment. This information can come from a variety of sources, such as amino acid specific labeling (Lemaster and Richards, 1985; Ou et al., 2001), backbone resonance experiments that select only signals from specific amino acids (Dotsch et al., 1996; Schubert et al., 1999) or amide peaks in a (H)C(CO)NH-TOCSY spectrum

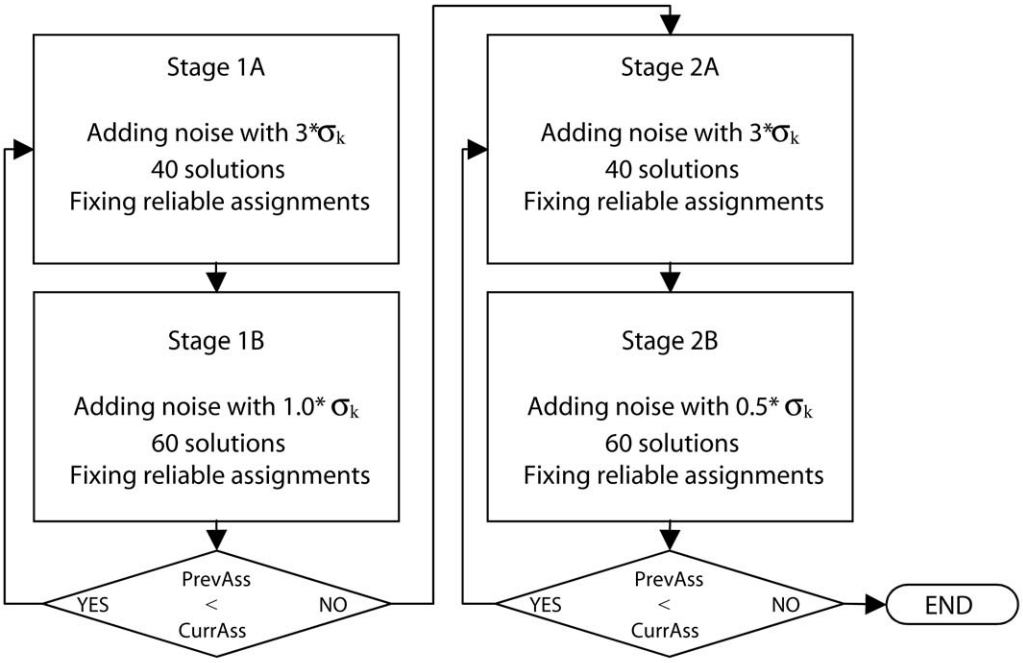

Figure 2. Empirically optimized scheme for avoiding errors due to inaccuracies in predicted chemical shifts when mapping pseudoresidue segments to the protein sequence. Stages 1A and 2A are identical except that the solution space is decreased when going from 1A to 2A due to assignments fixed in previous assignment stages. Stages 1B and 2B are also identical except that the amount of noise that is added to chemical shifts (which are calculated from the protein sequence) is decreased. σ_k is the standard deviation of the statistical chemical shift distribution that is used for calculating chemical shifts from the protein sequence. *PrevAss* and *CurrAss* is the number of assignments after stages A and B, respectively. Arrows indicate the program flow, i.e., if the number of assignments obtained from stage 1B (*CurrAss*) is larger than that from stage 1A (*PrevAss*) the program returns to stage 1A and reruns stage 1A but now with the reduced space of assignment solutions.

indicating methyl containing residues (Gardner et al., 1996). Information about the amino acid type of a pseudoresidue is most useful, when C^α and C^β chemical shift information is incomplete and for proteins above 40 kDa.

MARS not only allows restriction of possible amino acid types, the user can also fix connectivity between two pseudoresidues. This is useful in an iterative approach, where a MARS assignment is refined manually on the screen, manually validated sequential connectivites are fixed and MARS is rerun with the reduced space of possible assignment solutions. Moreover, when assignment of a PR is known, i.e., the residue in the primary sequence of the protein that corresponds to the pseudoresidue has been identified, this assignment can be fixed.

Establishing sequential connectivity

In a first step, all possible sequential connectivities are detected. The approach taken in MARS is that initially each PR is assumed to be sequentially connected to every other PR and only connectivities not in agreement with experimental intra- and inter-residual chemical shifts are removed. Within the tolerance set for the individual nuclei, all matching shifts are equally accepted: there is no preference for the ‘best match’ to avoid a bias from insignificant chemical shift differences. In addition, missing chemical shifts are not given a penalty, i.e., only when an atom type has chemical shift values for both pseudoresidues (in one case the intra-residual and in the other case the inter-residual chemical shift) and the difference between these two values is larger than the user-specified threshold the connectivity is deleted. This is especially important for assignment of proteins that miss chemical shifts for a substantial portion of residues. Another important feature of MARS is that all pseudoresidues are used in all phases of the assignment procedure. PRs are not classified according to the number of chemical shifts they contain or the intensity of their corresponding NMR resonances. Therefore, PRs strongly affected by chemical exchange or by the presence of a paramagnetic ion can be fully utilized.

Matching of experimental chemical shifts to the protein sequence

The second key step in assignment is to map segments that comprise sequentially linked pseudoresidues onto the primary sequence. Particularly useful in this respect is comparison of experimental C^{α} and C^{β} chemical shifts with values that were obtained for each residue from a statistical analysis of chemical shifts deposited in the BMRB (Doreleijers et al., 2003). In MARS this process is further improved by using chemical shift distributions that are corrected for neighbor residue effects (Wang and Jardetzky, 2002). Besides the type of amino acid (and the type of neighbors in the primary sequence), however, chemical shifts very much depend on the type of secondary structure an amino acid is involved in. This is addressed in MARS by using the secondary structure prediction program PSIPRED (McGuffin et al., 2000) to identify regions in the protein sequence that are likely to be involved in regular secondary structure elements. For each residue a theoretical chemical shift is calculated as the normalized sum of the random coil value and the value expected when this residue is involved in an α helix or a β strand. The probability of being in this secondary structure element, as identified by PSIPRED, is used as a weighting factor. Chemical shifts calculated in this way are of comparable quality as values predicted for proteins with known structure using the program SHIFTS (Xu and Case, 2002) (data not shown). If the protein under study is predeuterated, MARS can be directed to adjust the calculated chemical shifts accordingly (Venters et al., 1996).

In order to map PR fragments onto the protein sequence, MARS calculates for all experimentally observed pseudoresidues the deviation of their experimental chemical shifts from predicted values according to

$$D(i,j) = \sum_{k=1}^{N_{CS}} \left\{ \frac{\delta(i)_k^{\text{exp}} - \delta(j)_k}{\sigma_k} \right\}^2, \qquad (1)$$

where $\delta(i)_k^{\text{exp}}$ is the measured chemical shift of type k (e.g., $^{13}C^{\alpha}$ or $^{13}C^{\beta}$) of pseudoresidue i, $\delta(j)_k$ is the predicted chemical shift of type k of residue j, N_{CS} is the number of chemical shift types and σ_k^2 is the variance of the statistical chemical shift distribution that is used for calculating $\delta(j)_k$. For $^1H^N$, ^{15}N, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$ and $^1H^{\alpha}$ σ_k values of 0.82, 4.3, 1.2, 1.1, 1.7 and 0.82 ppm were used, respectively. In case a chemical shift of type k is missing, $[\delta(i)_k^{\text{exp}} - \delta(j)_k]$ is set to zero.

If calculation of chemical shifts from the protein sequence would be perfect, comparison with experimental values would be sufficient to complete assignment (Gronwald et al., 1998). This, however, is not achievable with current prediction methods and additional connectivity information is required. In order to further increase the reliability of the mapping process, MARS does not rely directly on chemical shift deviations. Instead these values are converted into a pseudoenergy $U(i,j)$ by ranking all residues j according to their chemical shift deviation (as calculated in Equation 1) with respect to pseudoresidue i. This makes MARS even more robust against unusual chemical shifts as not the exact fit of calculated to experimental chemical shifts is important, but the overall quality of the chemical shift fit.

Exhaustive search for establishing sequential connectivity and mapping

At the start of a MARS assignment process all pseudoresidues are assigned randomly to the protein sequence. This information is stored as ASS^{local}. In order to refine ASS^{local}, MARS randomly selects a pseudoresidue. Starting from this PR it searches in the direction of the primary sequence ('forward direction') for all pseudoresidue segments of length five that can be assembled based on the available connectivity information. In the next step, all these N_{seg} segments are mapped onto all possible positions of the protein sequence. The probability that a fragment belongs to a specific position in the protein sequence is evaluated by calculating a summed pseudoenergy according to

$$U_i^m(j) = \sum_{k=i}^{i+n} U(k, j_i), \qquad (2)$$

where i is the number of the pseudoresidue that was randomly selected as the start of the segment, n is the length of the fragment (in this case $n = 5$), m is the fragment number ($m \in [1, N_{seg}]$) and j_i are the residue numbers to which pseudoresidues i to $i + n$ are tentatively assigned to (j is the starting position). Next, all $U_i^m(j)$ are ranked. The minimum $U_i^m(j)$ identifies the best-fitting pseudoresidue segment, which starts with pseudoresidue i, and its corresponding position in the primary sequence. The information about this segment and the corresponding amino acid sequence is stored in SEG_{for} and ASS_{for}, respectively. In order to validate this assignment, the same procedure is repeated but now starting from the last pseudoresidue of SEG_{for} providing an additional assignment possibility (SEG_{back}/ASS_{back}). If $SEG_{for} = SEG_{back}$, the assignment of the segment to the protein sequence is regarded as reliable and following approach is adopted to refine ASS^{local}. When $SEG_{for} = SEG_{back}$ but $ASS_{for} \neq ASS^{local}$ the overall assignment is updated, i.e. $ASS_{for} \rightarrow ASS^{local}$. In case of $SEG_{for} = SEG_{back}$ and $ASS_{for} = ASS^{local}$, this would have no effect. In order, however, to favor an assignment that is retained from previous assignment phases a penalty is given to all other assignments, which are possible for the PRs and residues that comprise SEG_{for} and ASS_{for}. Thus, the total energy of the system is changed in such a way that the correct assignment is favored. When, on the other hand, $SEG_{for} \neq SEG_{back}$, the suggested assignment solution is regarded as unreliable and ASS^{local} is kept unchanged. The whole optimization phase is repeated until all pseudoresidues have been used once as segment starting point.

So far, assignment has been optimized only with segments in which five PRs could be sequentially linked. The assignment is further refined in a second round, where the exhaustive search is restricted to segments in which four PRs are linked, then in a third and fourth round with tri- and dipeptide fragments. The procedure is conducted with decreasing fragment sizes based on the assumption that the longest matching segments have the greatest certainty of leading to correct assignments. Finally, the whole phase comprising refinement of ASS^{local} by five, four, three and two PR segments is repeated four times. As each phase is based on pseudoenergies U(i,j) that were refined in the previous phase, the assignment procedure finally converges. All assignment results reported here comprised a total of five phases.

The maximum segment length of five linked pseudoresidues is a compromise between the desired total execution time of a MARS assignment run and the ability to reliably place PR segments onto the protein sequence. When connectivity information from C^{α} and C^{β} chemical shifts is available with an accuracy better than 0.5 ppm, MARS execution times for proteins as big as 370-residue maltose-binding protein are below 90 minutes on a single 1.7 GHz PC. At the same time, PR fragments with length five can in most cases be placed uniquely into the protein sequence when intra- and inter-residual C^{α} and C^{β} chemical shifts are available.

Identification of reliable assignments

The algorithm described above results in a final optimized assignment ASS^{local}. This assignment is mainly driven by the local fit of fragments, comprising up to five pseudoresidues, to the protein sequence. In addition, however, pseudoenergy values U(i,j), which qualitatively describe the mapping of a single residue j to pseudoresidue i, have been changed during the process: This approach is similar to assignment algorithms where an energy function is optimized globally. Thus, a second assignment ASS^{global} can be extracted from U(i,j) at the end of the MARS assignment process. Each pseudoresidue i is assigned to that residue j for which U(i,j) is the minimum among all U(i,1), U(i,2), ..., U(i,N_{res}) values. The two alternative assignment solutions, ASS^{global} and ASS^{local}, are compared and only consistent assignments are retained.

A major factor influencing the final assignment is the quality of chemical shifts predicted from the primary sequence as these values guide the mapping of PR segments to the protein sequence. To overcome this problem, MARS repeats the complete assignment process described above many times (Figure 2). For each assignment run predicted chemical shifts $\delta(j)_k$ are modulated by addition of noise according to a Gaussian distribution. For the first 20 assignment runs, which generate a total of 40 assignment solutions (20 ASS^{global} and 20 ASS^{local} assignments), the width of this Gaussian is set to three times the standard deviation σ_k of the statistical chemical shift distributions. By selecting assignments that are consistent across all 40 solutions, only the most reliable assignments are retained. These highly reliable assignments are fixed and the corresponding PRs and residues are excluded from future assignment runs. In subsequent assignment runs the amount of added noise is reduced according to an empirically optimized scheme (Figure 2). This gradually increases the number of consistent assignments. Thus, MARS uses best-first features both for establishing sequential connectivity (assignment is started with long connectivity segments) and for mapping PR segments onto the primary sequence (PR segments that are less affected by changes in calculated chemical shifts are mapped first).

Output data

The output of MARS consists of different ASCII files: (1) '*assignment_AA.out*', a file listing pseudoresidues assigned reliably to residues, i.e., the final assignment result, (2) ‘*assignment_AAs.out*’, an extended assignment including alternative assignment possibilities that show up with a 10% probability, (3) ‘*assignment_PR.out*’, the most likely assignment for each pseudoresidue (this is useful in order to find out what is the most likely assignment for PRs that have not been assigned reliably to any residue), (4) ‘*connectivity.out*’, a summary of all possible sequential connectivities and (5) ‘*mars.log*’, which contains detailed information about predicted chemical shifts, number of reliable assignments, number of constraints for each pseudoresidue, matrices matching experimental and back-calculated chemical shifts and pseudoenergy matrices at each iteration step. In addition, chemical shift tables with updated assignments are stored (‘*sparky_all.out*’, ‘*sparky_CA.out*’, ‘*sparky_CA-1.out*’, ‘*sparky_CB.out*’, ...) that can directly be read into the analysis program SPARKY using the ‘*Read peak list*’ feature of SPARKY (Kneller and Kuntz, 1993) and allow visual inspection of the assignment result. Assigned pseudoresidues can be viewed as sequentially linked strips together with PRs that have not been assigned so far, alternative assignments can be evaluated on the screen using the information provided in the files ‘*assignment_AAs.out*’ and ‘*connectivity.out*‘, and assignment suggestions for pseudoresidues that have not been assigned so far are provided in ‘*assignment_PR.out*’. After validation on the screen safe assignments and sequential connectivities can be fixed and MARS can be rerun with the reduced space of possible assignment solutions.

Implementation

The core of MARS was written using the C programming language. This core is embedded into a shell script that uses the UNIX utility awk for formatting of input and output files. This integrated approach has the advantage that improved programs for chemical shift prediction, chemical shifts from homologues proteins or chemical shifts from a previous assignment can easily be used.

Testing of MARS

MARS has been tested on 14 proteins ranging in size from the 71-residue Z domain of Staphylococcal protein A to 723-residue malate synthase G (Alattia et al., 2000; Gardner et al., 1998; Garrett et al., 1997; Ikura et al., 1991; Liu et al., 2000; Schwaiger et al., 1998; Tashiro et al., 1997; Tugarinov et al., 2002; Vathyam et al., 1999; Wang et al., 1995). Special focus was put on proteins that are challenging with respect to assignment either by their size or because chemical shifts are missing for a substantial portion of residues (Table 1). MARS was tested primarily using only C^{α} and C^{β} connectivity information as intra-residual carbonyl chemical shifts are most difficult to obtain experimentally due to the lower sensitivity of HN(CA)CO spectra. For selected proteins the effect of including C' connectivity information was evaluated and for ubiquitin the performance was tested using only C^{α} sequential connectivity. In addition, two threshold conditions for establishing connectivity were tested, namely 0.5, 0.5 and 0.25 ppm (condition I) and 0.2, 0.4 and 0.15 ppm (condition II) for C^{α}, C^{β} and C', respectively.

Chemical shifts were taken from the BMRB data base (Doreleijers et al., 2003), with all H^{N} and N chemical shifts entered as spin-systems and with the carbon chemical shifts of the preceding residue entered as inter-residue chemical shifts. To put MARS to a more rigorous test, we also started from raw peak lists obtained from automatic peak picking of NMR spectra recorded on Z domain of Staphylococcal protein A. These raw peak lists were taken from the distribution package of the AUTOASSIGN software (Zimmerman and Montelione, 1995). Pseudoresidues for testing of MARS were generated from these peak lists by reading them into AUTOASSIGN and using the ‘*Create Ladders*’ feature. This produces the generic spin system objects (GS) that are equivalent to pseudoresidues in MARS. Overlapping GSs/PRs are thereby automatically separated (Zimmerman and Montelione, 1995). In addition, MARS was applied to the assignment of the fully unfolded, soluble N-terminal 110-residues of intimin receptor Tir (Tir110). 3D HNCA, CBCA(CO)NH, HNCACB, HNCO and HNCACO experiments were collected on a Bruker DRX800 spectrometer and processed using NMRPipe (Delaglio et al., 1995). Calibration of spectra, peak picking and grouping of peaks into pseudoresidues was done using SPARKY (Kneller and Kuntz, 1993). Pseudoresidues were saved to an ASCII file using the ‘*Save Assignment table*’ feature of SPARKY and read into MARS without further modification.

For proteins that lacked experimental data the robustness of MARS against missing chemical shifts was tested by random removal of entire pseudoresidues as well as deletion of certain chemical shifts within the pseudoresidues. In addition, it was evaluated how chemical shifts that are outside the connectivity threshold δ due to peak overlap or distortion (although in reality they are sequentially connected)

Table 1. Proteins and data quality used for testing MARS

Protein	BMRB code	# of residues	# of PRO/GLY	$C^{\alpha}_{i}/C^{\alpha}_{i-1}$ (%)[a]	$C^{\beta}_{i}/C^{\beta}_{i-1}$ (%)[a]	C'_{i}/C'_{i-1} (%)[a]	$H^{\alpha}_{i}/H^{\alpha}_{i-1}$ (%)[a]
Malate synthase G	5471	723	31/51	95/95	94/94	94/95	–
Maltose binding protein	4354	370	21/29	96/96	95/96	–	–
Rous Sarcoma Virus capsid	4384	262	23/20	92/92	89/91	92/93	–
Human carbonic anhydrase I	4022	260	17/16	100/100	100/100	95/96	–
N-terminal domain of enzyme I (EIN)	4106	259	4/15	96/97	96/97	–	–
E-cadherin domains II and III	4457	227	14/12	78/63	78/63	–	–
Human prion protein	4402	210	15/43	98/97	98/97	–	–
Superoxide dismutase	4341	192	8/14	64/64	62/63	48/61	–
Calmodulin/M13 complex	547	148	2/11	99/99	–	99/99	–
Profilin	4082	139	4/16	99/99	100/98	–	–
E. coli EmrE	4136	110	5/12	86/84	57/60	73/77	–
Human ubiquitin	–	76	3/6	100/100	100/100	–	–
Z domain	–	71	3/0	90/96	51/82	–	89/100
Tir110		110	12/15	100/100	100/100	100/100	–

[a] Percentage of available chemical shifts of a given type.

affect automatic assignment by MARS. For this, random noise $d = N(0, \delta/2.5)$ was added to each inter-residual chemical shift, where $N(\mu, \sigma)$ represents a random variable of normal density with mean μ and standard deviation σ. In this way, about 2–3% of connectivities were affected (condition III). For the N-terminal domain of enzyme I of the phosphoenolpyruvate the percentage of wrong inter-residual chemical shifts was further increased up to 50%. This corresponds to $d = N(0, \delta/1.1)$.

In all tests assignment was performed by MARS without manual intervention and the results are reported in Table 2. Running times (not CPU times) on a 1.7 GHz Linux PC varied from about 30 s for ubiquitin to about 90 min in case of maltose-binding protein (only C^{α}, C^{β} connectivity with a common threshold of 0.5 ppm). For malate synthase G running times vary from 2 h (C^{α}, C^{β} and C' connectivity with thresholds of 0.2, 0.4 and 0.15 ppm, respectively) to 13 h (only C^{α} and C^{β} connectivity with thresholds of 0.2 and 0.4 ppm, respectively) and up to 150 h when only C^{α} and C^{β} connectivity information is available with a resolution of 0.5 and 0.5 ppm, respectively.

Results and discussion

Small proteins

76-residue ubiquitin serves as a first basic test case. Using C^{α}/C^{β} connectivity information all 72 non-proline residues (excluding the N-terminus) could be assigned correctly and reliably for both threshold conditions. When only C^{α} chemical shift information was used, the total number of correct assignments dropped to 32 and 9 were identified as reliable. This rather strong decrease in reliable assignment is expected due to the higher degeneracy and the less precise determination of amino acid types in the absence of C^{β} chemical shifts. Only if fragments are sufficiently long or if they contain residues with very characteristic C^{α} chemical shifts, such as glycines, a mapping to the sequence is identified as reliable by MARS. However, none of the nine reliable assignments was wrong.

MARS was further tested on the 67 pseudoresidues of Z domain of Staphylococcal protein A as obtained from raw peak lists (Zimmerman and Montelione, 1995). The number of pseudoresidues agrees with the expected number taking into account the three prolines and the N-terminal amino acid, i.e., no additional, spurious PRs are present. For 19% of Z domain's PRs the H^{N}/N root frequencies partially overlap and 90% of all expected intra-residual C^{α} chemical shifts are present. However, C^{β} connecticity information is far from complete with only 51% of all expected intra-

Table 2. Mars assignment results for proteins of varying size and data completeness

Protein	# of residues with data[a]	Used chemical shifts	Condition I[b] Assignment #		Condition II[c] Assignment #		Condition III[d] Assignment #	
			All[e]	Reliable/ Errors[g]	All[e]	Reliable/ Errors[g]	All[e]	Reliable/ Errors[g]
Malate synthase G	654	C′, Cα, Cβ	652	639/0	652	639/0	651	623/0
		Cα, Cβ[f]	500	207/0	639	584/2	622	511/0
Maltose binding protein	335	Cα, Cβ	323	303/0	333	324/0	330	313/1
Rous Sarcoma Virus capsid	221	C′, Cα, Cβ	214	205/0	218	207/0	218	199/0
Human carbonic anhydrase I	243	C′, Cα, Cβ	242	235/0	242	237/0	242	225/0
N-terminal domain of enzyme I (EIN)	248	Cα, Cβ	246	232/0	246	246/0	248	245/0
E-cadherin domains II and III	167	Cα, Cβ	116	77/1	134	102/0	136	70/1
Human prion protein	190	Cα, Cβ	138	103/0	155	127/0	154	118/0
Superoxide dismutase	117	C′, Cα, Cβ	112	101/0	112	104/0	111	100/0
		Cα, Cβ	111	101/0	112	104/0	112	103/0
Calmodulin/M13	144	Cα, C′	97	37/0	144	142/0	136	119/0
Profilin	132	Cα, Cβ	130	132/2	132	132/0	132	123/0
E. coli EmrE	74	C′, Cα, Cβ	61	35/0	70	58/0	64	50/0
Human ubiquitin	72	Cα, Cβ	72	72/0	72	72/0	72	70/0
		Cα	32	9/0	58	18/0	58	9/0
Z domain	67	Cα,Cβ, Hα[h]	65	65/0	–	–	–	–
		Cα, Cβ[i]	57	34/0	–	–	–	–
Tir110[j]	97	C′, Cα, Cβ	91	80/0	–	–	–	–

[a] Includes only those residues for which HN and N chemical shifts were reported.

[b] Condition I: 0.5, 0.5 and 0.25 ppm are used for establishing connectivity for Cα, Cβ and C′, respectively.

[c] Condition II: 0.2, 0.4 and 0.15 ppm are used for establishing connectivity for Cα, Cβ and C′, respectively.

[d] Condition III: Same as condition II but with simulated error.

[e] # of correct assignments in Ass^{global} ; Ass^{global} was obtained from a MARS run without addition of noise.

[f] The maximum length of pseudoresidue segments, which were searched exhaustively, was four (instead of five).

[g] Assignments that were identified as reliable but are incorrect, i.e., the number of errors.

[h] Experimental data. Connectivity thresholds of 0.5, 0.7 and 0.05 ppm were used for Cα, Cβ and H^{α}, respectively.

[i] Experimental data. Connectivity thresholds of 0.3 and 0.5 ppm were used for Cα and Cβ, respectively.

[j] Experimental data. Connectivity thresholds of 0.2, 0.5 and 0.25 ppm were used for Cα, Cβ and C′, respectively.

residual C^{β} chemical shifts available. Employing a common connectivity threshold of 0.5 ppm for both C^{α} and C^{β} MARS assigned 34 PRs reliably. In addition, the correct assignment was indicated for another 23 pseudoresidues, providing valuable starting points for manual assignment. Upon inclusion of H^{α} connectivity information the number of assignments was raised to 65 with no errors present.

Partially and completely disordered proteins

In case of the 210-residue full-length human prion protein, the N-terminal half (residues 1-125) is completely disordered. This results in a very narrow chemical shift dispersion, severe degeneracy and poses a significant challenge to sequential assignment. Using only C^{α}/C^{β} chemical shifts for establishing connectivity (with a common threshold of 0.5 ppm) MARS assigned 138 out of 190 available pseudoresidues correctly and 103 of these were identified as reliable. All assignments identified as reliable were correct, i.e., 103 residues were assigned by MARS without false positives. When the threshold was reduced to 0.2 and 0.4 ppm for C^{α} and C^{β}, respectively, the number of reliable and correct assignments increased to 127, i.e., an assignment score of 67%.

Similar, high quality results were obtained using experimental chemical shift lists that were prepared from triple-resonance spectra recorded on the completely unfolded, soluble N-terminal 110-residues of intimin receptor Tir. Using C^{α}, C^{β} and C' chemical shifts MARS assigned 80 out of 97 experimental pseudoresidues and indicated the correct assignment for a total of 91 PRs (Table 2). Based on the 80 reliable assignments and the assignment suggestions provided by MARS for the remaining PRs, the assignment could be quickly completed by visual inspection of assignment strips (pseudoresidues) using SPARKY.

Big proteins

N-terminal domain of enzyme I of the phosphoenolpyruvate (EIN), human carbonic anhydrase I, rous sarcoma virus capsid, maltose-binding protein (MBP) and malate synthase G (MSG) are challenging for assignment due to their size of 259, 260, 262, 370 and 723-residues. With C^{α} and C^{β} chemical shifts at an accuracy of better than 0.2 and 0.4 ppm, respectively, 99% of EIN and 97% of MBP could be assigned reliably and for almost 100% the correct assignment was indicated. For 723-residue MSG 89% of pseudoresidues could be assigned, however, two of these were wrong. Inclusion of C' connectivity removed the two errors and increased the reliable assignment score to 98%. Whereas in case of C^{α}, C^{β} and C' connectivity information the number of possible connectivities for each pseudoresidue is 1.02 on average (note that this is just an average value), it is raised to 4.48 when C' connectivity is not available. Therefore, it was necessary to reduce the maximum fragment length, which is searched exhaustively during the linking process, to four PRs and it still took several days to complete the assignment process on MSG. In case of condition III, 207 pseudoresidues of MSG were assigned reliably, out of a total of 500 correct ones, and not a single reliable assignment was wrong. The long duration of the assignment process for such difficult cases can significantly be shortened if MARS is run on several PCs in parallel on a Linux cluster.

Proteins with incomplete chemical shift data

EmrE, superoxide dismutase and E-cadherin are missing H^N/N chemical shifts for a substantial portion of their residues. For superoxide dismutase only 61% of expected pseudoresidues were observed in triple-resonance NMR spectra as a result of paramagnetic relaxation of residues in the vicinity of an Fe^{3+} ion. In addition, about half of the available PRs are scattered throughout the length of the protein, separated by numerous small gaps. MARS was able to efficiently handle these difficult cases and assigned 101 out of 117 pseudoresidues reliably using only C^{α} and C^{β} connectivity information (threshold of 0.5 ppm for both). Including C' data or reducing the thresholds to 0.2 and 0.4 ppm for C^{α} and C^{β}, respectively, did not significantly affect the assignment score.

Required chemical shift data and thresholds for establishing connectivity

MARS is highly flexible and does not require a specific set or resolution of NMR spectra. Whether only C^{α} or C^{α}, C^{β}, C' and H^{α} connectivity information is available, assignments identified by MARS as reliable will have a very low to zero error rate. When only C^{α} chemical shifts are available, reliable assignment is restricted to very small proteins with very complete data. With C^{α} and C^{β} information available for more than 80% of residues and with an accuracy better than 0.2 and 0.4 ppm, respectively, an assignment score of more than 95% is possible without errors. For proteins above 40 kDa or less complete or more degenerate data it is highly useful to have access to additional C' connectivity information. Assignment is less susceptible to errors (see results on malate synthase G) and thresholds for establishing sequential connectivity have to be less tight. For example, for superoxide dismutase and malate synthase G similar results are obtained with thresholds of 0.5, 0.5, 0.25 ppm and 0.2, 0.4, 0.15 ppm for C^{α}, C^{β} and C', respectively. This is especially important, as overlap and weak resonances often require higher connectivity thresholds as anticipated on the basis of the digital resolution of the NMR spectra. In addition, the reduced degeneracy for establishing sequential connectivity significantly shortens execution times of MARS.

Robustness against missing data

When chemical shift information is close to complete and NMR spectra were recorded with a resolution better than 0.2, 0.4 and 0.15 ppm for C^{α}, C^{β} and C', respectively – as for ubiquitin, calmodulin or EIN – MARS allows automatic assignment of 99 to 100% of observed pseudoresidues. Such favorable situations, however, are rarely encountered in real applications. More important is, therefore, the reliability of the assignment procedure in case of incomplete chemical shift data. Table 2 shows that only for some selected test cases one or two reliable assignments were wrong. In all other situations assignments labeled as reliable by MARS were correct (i.e., zero error rate).

The robustness of MARS was further tested by randomly deleting a fraction of the observed pseudoresidues. Random deletion of pseudoresidues is particularly challenging as it introduces many gaps

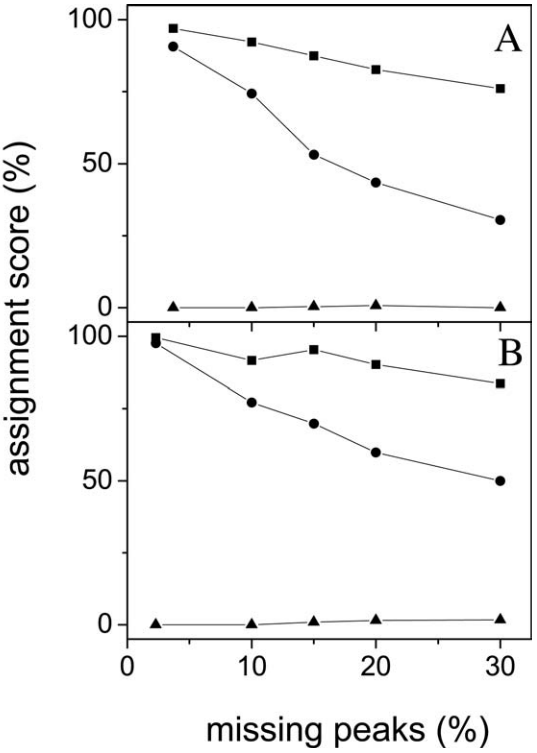

Figure 3. Dependence of MARS assignment on the percentage of missing pseudoresidues. Pseudoresidues were deleted randomly. ■ indicate the percentage of all assignments that were correct (not tested for reliability). ● show the percentage of residues that could be assigned reliably (relative to the total number of assignable residues) and ▲ indicate assignments that were identified as reliable but are wrong, i.e., the error rate of MARS. C^α and C^β chemical shifts with a common threshold of 0.5 ppm for establishing sequential connectivity were used. (A) Results for the 370-residue maltose-binding protein. (B) Results for the 259-residue N-terminal domain of enzyme I. Note the very small to zero error rate.

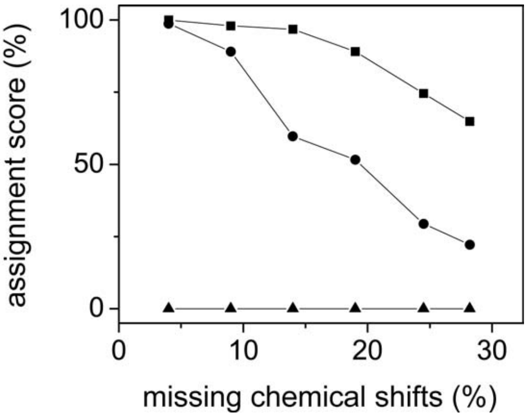

Figure 4. Dependence of MARS assignment on the percentage of missing chemical shifts within pseudoresidues for the 259-residue N-terminal domain of enzyme I. Chemical shifts were deleted randomly. ■, ● and ▲ indicate correct, reliable and wrong reliable assignments, respectively. C^α and C^β chemical shifts with thresholds of 0.2 and 0.4 ppm for establishing sequential connectivity were used.

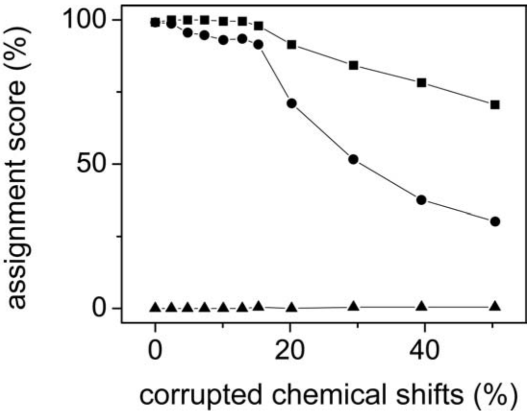

Figure 5. Dependence of MARS assignment on the percentage of chemical shifts falling outside the connectivity thresholds for the 259-residue N-terminal domain of enzyme I. Connectivity thresholds were 0.2 and 0.4 ppm for C^α and C^β, respectively. ■, ● and ▲ indicate correct, reliable and wrong reliable assignments, respectively.

into the sequential connectivity path. Removing 10% of EIN's pseudoresidues decreased the reliable assignment from 95% to 78% (Figure 3). However, the assignment remains without error. When 20 or 30% of pseudoresidues are removed the number of reliable assignments is further reduced to 122 and 89 (out of a total of 204 and 178 remaining pseudoresidues of EIN, respectively). For MBP, on the other hand, the percentage of reliable assignments dropped to 30% when 30% of pseudoresidues were randomly deleted. This strong decrease is expected due to the large size of maltose-binding protein. However, even is such a challenging situation the number of assignment errors is kept at a minimum. Both for EIN and MBP the number of errors is always less than three (zero for MBP, three for EIN at 30% randomly deleted pseudoresidues). In addition, for many proteins missing data are concentrated into a specific region of the protein sequence, such as for EmrE where NMR data for residues 32 to 76 are missing. This is less problematic than random deletion, as reliable assignment can be obtained efficiently for the remainder of the sequence.

The robustness of MARS against missing data was also tested by randomly deleting chemical shifts within pseudoresidues of EIN. Similar to the case where complete PRs are deleted, the number of overall correct assignments remained almost unchanged up to 15% missing chemical shifts (Figure 4). For even more incomplete data the assignment score started dropping and ended up at 65% when 28% of chemical shifts were removed. At the same time the number of reliable assignments reduced more quickly with an assignment score of 52% for 19% missing chemical shifts. In agreement with the tests where complete pseudoresidues were removed, assignments termed reliable by MARS are indeed very reliable with zero errors even at 30% missing chemical shifts.

The low error rate of MARS is sometimes a tradeoff with the completeness of assignment. For example, for ubiquitin (using only C^{α} chemical shifts with a threshold of 0.2 ppm) 58 assignments were correct, but only 18 were identified as reliable (Table 2). MARS, however, should be used together with analysis software that allows visual inspection, such as SPARKY, and the 58 correct assignments of ubiquitin provide a very valuable starting point to manually complete assignment. In addition, they can give hints on what additional information, such as selective labeling, is required.

Robustness against chemical shifts outside the connectivity threshold

The connectivity information provided by inter- and intra-residual chemical shifts is an essential component of the assignment process. At the same time, however, peaks are often distorted or overlapped and corresponding chemical shifts fall outside the connectivity thresholds. The effect of chemical shift errors was tested by addition of noise to each inter-residual chemical shift, such that about 2–3% of connectivities were affected. For all tested proteins the overall assignment scores were virtually unchanged upon introduction of the distorted chemical shifts (Table 2). In addition, the reliable assignments were only slightly affected. The strongest decreases in the number of reliable assignments were seen for E-cadherin and the calmodulin/M13 complex. For E-cadherin this can be attributed to the high number of missing chemical shifts and the fact that only C^{α} and C^{β} chemical shift information was available (Table 1). For superoxide dismutase, on the other hand, where even more pseudoresidues and C^{α} and C^{β} chemical shifts are missing, the assignment is almost unchanged due to the availability of C' chemical shifts (Table 2). This demonstrates that using slightly too tight connectivity thresholds is not problematic for MARS. For EIN we further took these tests to the extreme by strongly increasing the amount of added noise such that up to 45% of sequential connectivities were lost (Figure 5). As long as less than 15% of inter-residual chemical shifts were outside the connectivity thresholds both the overall and the reliable assignment scores remained high. Only when even more chemical shifts were corrupted the number of assignments started to rapidly decrease. However, even when 45% of connectivities were lost (corresponding to 50% of chemical shifts outside the connectivity thresholds) only a single reliable assignment was wrong.

Concluding remarks

We have introduced a software for backbone assignment of proteins that can be applied independent of the assignment complexity, that does not require tight thresholds for establishing sequential connectivity or detailed adjustment of these thresholds, that uses always all available data during the assignment process and that does not require a specific set of NMR experiments. The key for any automatic assignment is that one can trust the answer the program returns. When the amount and quality of available information is poor, this will always result in a decrease in the number of assignments that will be regarded as reliable, independent of whether the assignment is performed manually or automatically. In these difficult cases MARS retains a good assignment score and, at the same time, assignments that are identified as reliable are almost always correct.

Compared to other currently available programs MARS is applicable to proteins above 15 kDa using only C^{α} and C^{β} chemical shift information with connectivity thresholds as high as 0.5 ppm and it is applicable to proteins with very high degeneracy such as partially or fully unfolded proteins. It offers improved assignment scores for proteins where data are missing for a substantial portion of residues and it has a good tolerance against erroneous chemical shifts. MARS assignment results can be directly read into the program SPARKY (Kneller and Kuntz, 1993). This allows visual validation of the assignment results. Thus, several cycles of automatic assignment using MARS and manual validation on the screen can be performed, in order to complete assignment even in difficult cases. We therefore believe that MARS can proof highly useful for the protein NMR community.

MARS is available for SGI, Linux and OSX machines via the Internet at http://www.mpibpc.mpg.de/abteilungen/030/zweckstetter.

Acknowledgements

We thank Sigrun Rumpel for the experimental data of Tir110. M.Z. is funded by the DFG Emmy Noether-research programme (ZW 71/1-3).

References

Alattia, J.R., Tong, F.K., Tong, K.I. and Ikura, M. (2000) *J. Biomol. NMR*, **16**, 181–182.

Atreya, H.S., Sahu, S.C., Chary, K.V.R. and Govil, G. (2000) *J. Biomol. NMR*, **17**, 125–136.

Bartels, C., Güntert, P., Billeter, M. and Wüthrich, K. (1997) *J. Comput. Chem.*, **18**, 139–149.

Bartels, C., Xia, T.H., Billeter, M., Güntert, P. and Wüthrich, K. (1995) *J. Biomol. NMR*, **6**, 1–10.

Bax, A. and Grzesiek, S. (1993) *Accounts Chem. Res.*, **26**, 131–138.

Bernstein, R., Cieslar, C., Ross, A., Oschkinat, H., Freund, J. and Holak, T.A. (1993) *J. Biomol. NMR*, **3**, 245–251.

Buchler, N.E.G., Zuiderweg, E.R.P., Wang, H. and Goldstein, R.A. (1997) *J. Magn. Reson.*, **125**, 34–42.

Coggins, B.E. and Zhou, P. (2003) *J. Biomol. NMR*, **26**, 93–111.

Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J. and Bax, A. (1995) *J. Biomol. NMR*, **6**, 277–293.

Doreleijers, J.F., Mading, S., Maziuk, D., Sojourner, K., Yin, L., Zhu, J., Markley, J.L. and Ulrich, E.L. (2003) *J. Biomol. NMR*, **26**, 139–146.

Dotsch, V., Oswald, R.E. and Wagner, G. (1996) *J. Magn. Reson. Ser. B*, **110**, 107–111.

Friedrichs, M.S., Mueller, L. and Wittekind, M. (1994) *J. Biomol. NMR*, **4**, 703–726.

Gardner, K.H., Konrat, R., Rosen, M.K. and Kay, L.E. (1996) *J. Biomol. NMR*, **8**, 351–356.

Gardner, K.H., Zhang, X.C., Gehring, K. and Kay, L.E. (1998) *J. Am. Chem. Soc.*, **120**, 11738–11748.

Garrett, D.S., Seok, Y.J., Liao, D.I., Peterkofsky, A., Gronenborn, A.M. and Clore, G.M. (1997) *Biochemistry*, **36**, 2517–2530.

Gronwald, W., Willard, L., Jellard, T., Boyko, R.E., Rajarathnam, K., Wishart, D.S., Sonnichsen, F.D. and Sykes, B.D. (1998) *J. Biomol. NMR*, **12**, 395–405.

Güntert, P., Salzmann, M., Braun, D. and Wüthrich, K. (2000) *J. Biomol. NMR*, **18**, 129–137.

Hare, B.J. and Prestegard, J.H. (1994) *J. Biomol. NMR*, **4**, 35–46.

Ikura, M., Kay, L.E., Krinks, M. and Bax, A. (1991) *Biochemistry*, **30**, 5498–5504.

Johnson, B.A. and Blevins, R.A. (1994) *J. Biomol. NMR*, **4**, 603–614.

Kneller, D.G. and Kuntz, I.D. (1993) *J. Cell. Biochem.*, 254–254.

Lemaster, D.M. and Richards, F.M. (1985) *Biochemistry*, **24**, 7263–7268.

Leutner, M., Gschwind, R.M., Liermann, J., Schwarz, C., Gemmecker, G. and Kessler, H. (1998) *J. Biomol. NMR*, **11**, 31–43.

Liu, A.Z., Riek, R., Wider, G., von Schroetter, C., Zahn, R. and Wüthrich, K. (2000) *J. Biomol. NMR*, **16**, 127–138.

Lukin, J.A., Gove, A.P., Talukdar, S.N. and Ho, C. (1997) *J. Biomol. NMR*, **9**, 151–166.

McGuffin, L.J., Bryson, K. and Jones, D.T. (2000) *Bioinformatics*, **16**, 404–405.

Meadows, R.P., Olejniczak, E.T. and Fesik, S.W. (1994) *J. Biomol. NMR*, **4**, 79–96.

Moseley, H.N.B. and Montelione, G.T. (1999) *Curr. Opin. Struct. Biol.*, **9**, 635–642.

Neidig, K.P., Geyer, M., Gorler, A., Antz, C., Saffrich, R., Beneicke, W. and Kalbitzer, H.R. (1995) *J. Biomol. NMR*, **6**, 255–270.

Olson, J.B. and Markley, J.L. (1994) *J. Biomol. NMR*, **4**, 385–410.

Ou, H.D., Lai, H.C., Serber, Z. and Dotsch, V. (2001) *J. Biomol. NMR*, **21**, 269–273.

Riek, R., Wider, G., Pervushin, K. and Wüthrich, K. (1999) *Proc. Natl. Acad. Sci. USA*, **96**, 4918–4923.

Schubert, M., Smalla, M., Schmieder, P. and Oschkinat, H. (1999) *J. Magn. Reson.*, **141**, 34–43.

Schwaiger, M., Lebendiker, M., Yerushalmi, H., Coles, M., Groger, A., Schwarz, C., Schuldiner, S. and Kessler, H. (1998) *Eur. J. Biochem.*, **254**, 610–619.

Tashiro, M., Tejero, R., Zimmerman, D.E., Celda, B., Nilsson, B. and Montelione, G.T. (1997) *J. Mol. Biol.*, **272**, 573–590.

Tugarinov, V., Muhandiram, R., Ayed, A. and Kay, L.E. (2002) *J. Am. Chem. Soc.*, **124**, 10025–10035.

Vathyam, S., Byrd, R.A. and Miller, A.F. (1999) *J. Biomol. NMR*, **14**, 293–294.

Venters, R.A., Farmer, B.T., Fierke, C.A. and Spicer, L.D. (1996) *J. Mol. Biol.*, **264**, 1101–1116.

Wang, A.C., Grzesiek, S., Tschudin, R., Lodi, P.J. and Bax, A. (1995) *J. Biomol. NMR*, **5**, 376–382.

Wang, Y.J. and Jardetzky, O. (2002) *J. Am. Chem. Soc.*, **124**, 14075–14084.

Wüthrich, K. (2003) *Angew. Chem.-Int. Edit.*, **42**, 3340–3363.

Xu, X.P. and Case, D.A. (2002) *Biopolymers*, **65**, 408–423.

Zimmerman, D.E. and Montelione, G.T. (1995) *Curr. Opin. Struct. Biol.*, **5**, 664–673.