J Biomol NMR (2008) 41:29–41
DOI 10.1007/s10858-008-9238-2

ARTICLE

Sequence specific resonance assignment via Multicanonical Monte Carlo search using an ABACUS approach

Alexander Lemak · Carlos A. Steren · Cheryl H. Arrowsmith · Miguel Llinás

Received: 10 May 2007 / Accepted: 8 April 2008 / Published online: 6 May 2008

Abstract ABACUS [Grishaev et al. (2005) *Proteins* 61:36–43] is a novel protocol for automated protein structure determination via NMR. ABACUS starts from molecular fragments defined by unassigned *J*-coupled spin-systems and involves a Monte Carlo stochastic search in assignment space, probabilistic sequence selection, and assembly of fragments into structures that are used to guide the stochastic search. Here, we report further development of the two main algorithms that increase the flexibility and robustness of the method. Performance of the BACUS [Grishaev and Llinás (2004) *J Biomol NMR* 28:1–101] algorithm was significantly improved through use of sequential connectivities available from through-bond correlated 3D-NMR experiments, and a new set of likelihood probabilities derived from a database of 56 ultra high resolution X-ray structures. A Multicanonical Monte Carlo procedure, Fragment Monte Carlo (FMC), was developed for sequence-specific assignment of spin-systems. It relies on an enhanced assignment sampling and provides the uncertainty of assignments in a quantitative manner. The efficiency of the protocol was validated on data from four proteins of between 68–116 residues, yielding 100% accuracy in sequence specific assignment of backbone and side chain resonances.

Keywords BACUS · NOE identification · Fragment Monte Carlo · Resonance assignment

A. Lemak (✉) · C. H. Arrowsmith
The Ontario Cancer Institute and Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada M5G 2M9
e-mail: alemak@uhnres.utoronto.ca

C. A. Steren · M. Llinás
Department of Chemistry, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Introduction

The assignment of resonances to their original nuclei is usually one of the most time-consuming stages of protein structure determination from NMR data. Most of the automated and semi-automated assignment protocols available today rely on sequential information derived from a large suite of triple resonance NMR spectra (see Moseley and Montelione (1999); Zimmerman and Montelione (1995) and Malmodin and Billeter (2005) for review). A number of approaches to establish inter-residue sequential connectivities mainly from NOE data have been described previously as well; these include main-chain directed method (MCD; Wand and Nelson 1991), JIGSAW algorithm (Bailey-Kellog et al. 2000), and more recently ABACUS (Grishaev et al. 2005), a direct extension of the CLOUDS protocol (Grishaev and Llinás 2002a, b). MCD and JIGSAW assignment strategies are based on secondary structure pattern identification in NOESY spectra by probabilistic reasoning techniques or applying a graph theory type of analysis. The ABACUS protocol relies on NOE identities established via BACUS (Grishaev and Llinás 2004), an automated Bayesian analysis procedure for NOESY cross-peaks identification prior to sequence-specific resonance assignment.

ABACUS was originally developed and blind-tested using spin-systems obtained from manually sorted data. As published, the protocol calls for knowledge of the amino acid spin systems (AA-fragments) as a prerequisite. However, in order to use ABACUS directly on data derived from experimental spectra, the algorithm requires improvements. In particular, robust spin-system identification is needed. Nowadays, many commonly acquired and more sensitive NMR spectra for $^{13}C/^{15}N$ double-labeled proteins allow one to perform spin-system identification in terms of peptide bonded (PB) fragments, rather than

amino-acid residues, via correlations through the peptide bond (see Fig. 1). Here, we report on further developments of the ABACUS protocol that increase the flexibility and robustness of the protocol. The two main modules of the protocol were improved: BACUS, the program that interprets NOESY data, and LINKMAP, the program that finds the sequential placement of the fragments.

BACUS, was originally created to analyze NMR data from unlabeled and ^{15}N-labelled proteins via COSY and TOCSY J-correlations. We therefore extended the classes of connectivities used by BACUS to encompass 3D triple resonance NMR experiments on ^{13}C/^{15}N double-labeled proteins that provide information on other types of connectivities.

A new procedure that replaces LINKMAP, named Fragment Monte Carlo (FMC), was developed aimed at the sequence-specific assignment of PB- or AA-fragments. It employs the Multi-Canonical (MUCA) method (Berg and Neuhaus 1991) which provides enhanced assignment sampling as compared to the simulated annealing Monte Carlo method used by LINKMAP. Essentially, MUCA sampling generates a random walk in one-dimensional assignment space, with the energy sampled with equal probability. The effective sampling of both optimal and non-optimal assignments provides more statistical–mechanical information about the system that, in turn, allows for exploiting the thermodynamic analogy more extensively. In particular, an estimate of the certainty of the optimal assignment can be made based on the statistical properties of the entire ensemble of assignments.

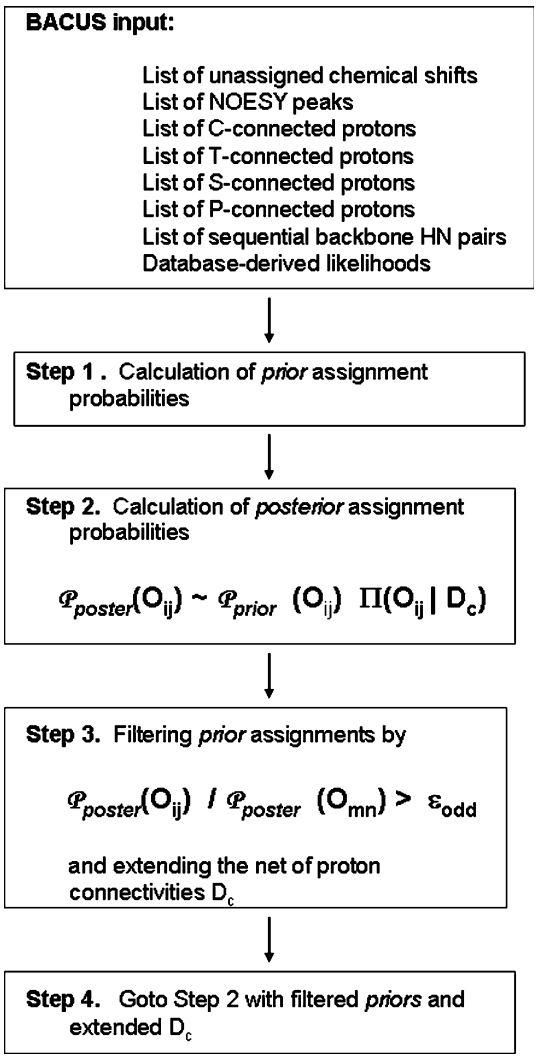

Fig. 2 Flowchart describing the BACUS procedure

Materials and methods

BACUS procedure

A flowchart of the BACUS algorithm (Grishaev and Llinás 2004) is depicted in Fig. 2. The input data for the program consist of lists of unassigned proton chemical shifts, NOESY cross-peak coordinates, and information on the connectivity between protons, as obtained from NOESY and through bond NMR experiments. BACUS starts by calculating for each NOE peak O^{α}, *prior* assignment probabilities $\mathcal{P}^{\alpha}_{prior}(\Omega_{ij})$, where Ω_{ij} indicates a pair of protons i and j that is a candidate for the cross-peak assignment. *Prior* probabilities are evaluated by matching the proton chemical shifts to the coordinates of the NOESY cross-peaks and assuming a Gaussian probability distribution function to score the matching probabilities (see Grishaev and Llinás 2004). Additional scoring of a prior assignment is performed whenever the corresponding mirror NOESY cross-peak is observed. In this case, the prior probability is multiplied by a user defined weight W_s, followed by renormalization of prior probabilities.

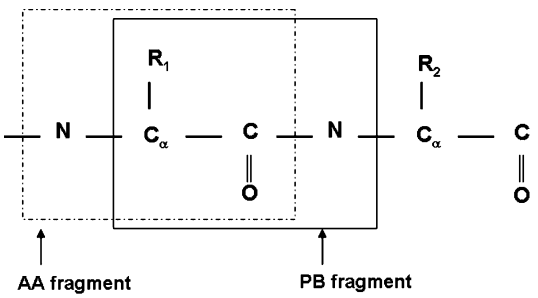

Fig. 1 Schematic description of two type of molecular fragments: AA-fragment include all the atoms belonging to the same residue; PB-fragment include all the atoms from one residue except the backbone amide group, plus the amide group from the next residue in the protein sequence

$\mathcal{P}^{\alpha}_{prior}(\Omega_{ij})$ are then refined in an iterative fashion as described below. In each iteration, *posterior* assignment probabilities $\mathcal{P}^{\alpha}_{post}(\Omega_{ij})$ are evaluated using Bayes' formula

$$\mathcal{P}^{\alpha}_{post}(\Omega_{ij}|D_c) \propto \mathcal{P}^{\alpha}_{prior}(\Omega_{ij}) \cdot \Pi(\mathrm{O}_{ij}|D_c) \quad (1)$$

where D_c denotes information extracted from the protons connectivity network data, and $\Pi(\mathrm{O}_{ij}|D_c)$ is the likelihood of observing an NOE between protons i an $j(\mathrm{O}_{ij})$ conditional on D_c. Next, pairs of protons with unambiguous *posterior* assignment are considered NOE connected and the network of known proton connectivities, D_c, is extended by adding the new NOEs as fake TOCSY cross-peaks. The subsequent iteration starts with trimmed *prior* assignments and an extended network of proton connectivities. The iteration process stops when the entropy criterion is satisfied (Grishaev and Llinás 2004).

Originally, two classes of connectivites, COSY and TOCSY, were defined (Grishaev and Llinás 2004). In this work, four types of connectivities are specified. C- and T-connectivities correspond to COSY and TOCSY correlations between protons of the same residue but excluding the backbone H^N. S-connectivity specifies a TOCSY correlation between the backbone H^N and another proton of the same residue. P-connectivities correspond to those that go through the peptide bond (Fig. 1). All four types of connectivities correspond to correlations between resonances that are directly observed in through-bond J-correlated 3 and 4D NMR experiments commonly employed for resonance assignments of double-labeled proteins. We denote by Λ_{ij} the type of connectivity that exists between protons i and j. For completeness of description, two more classes were added; $\Lambda_{ij} = \mathrm{U}$ when there is no connection between protons i and j, and $\Lambda_{ij} = \Delta$ if $i = j$. Thus, Λ_{ij} can be C, T, S, P, U, and Δ.

The likelihood $\Pi(\mathrm{O}_{ij}|D_c)$ in Eq. 1 depends on the class of connectivity between protons i and j. The likelihood of observing an NOE between a pair of connected protons, i.e. $\Lambda_{ij} \neq \mathrm{U}$, is $\Pi(\mathrm{O}_{ij}|D_c) = \Pi(\mathrm{O}_{ij}|\Lambda_{ij})$, and can be calculated directly from the database of known protein structures. On the other hand, when protons i and j are not connected, i.e. $\Lambda_{ij} = \mathrm{U}$, $\Pi(\mathrm{O}_{ij}|D_c)$ is calculated using the following relationship

$$\Pi(\mathrm{O}_{ij}|D_c) = \sum_n \sum_m \Pi(\mathrm{O}_{ij}|\mathrm{O}_{mn}, \Lambda_{im}, \Lambda_{jn}) \cdot \mathcal{P}(\mathrm{O}_{mn}) \quad (2)$$

where n and m correspond to other protons, called reporters, known to be connected to i and j respectively, $\mathcal{P}(\mathrm{O}_{mn}) = \max_{\alpha}\left\{\mathcal{P}^{\alpha}_{prior}(\mathrm{O}_{mn})\right\}$ stands for a probability of observing an NOE between the pair (m, n), and $\Pi(\mathrm{O}_{ij}|\mathrm{O}_{mn}, \Lambda_{im}, \Lambda_{jn})$ is a database-derived likelihood of observing an NOE between a pair of protons (i, j) under the condition that a NOESY cross-peak between protons m and n is observed.

Database-derived likelihoods

The likelihoods $\Pi(\mathrm{O}_{ij}|\Lambda_{ij})$ and $\Pi(\mathrm{O}_{ij}|\mathrm{O}_{mn}, \Lambda_{im}, \Lambda_{jn})$ for the four classes of connections, C, T, P and S, were calculated from distance distributions obtained from a set of known protein structures. The database entries where selected from the RCSB Protein Data Bank (Berman et al. 2000) following the criteria that: (a) structures were solved crystallographically by X-ray diffraction (XRD) with a resolution better than 1.0 Å, (b) pair-wise primary sequence similarity was less than 25%. A total of 56 XRD protein structures meeting these conditions were selected (see Table 1). The positions of missing hydrogen atoms were calculated with the program REDUCE (Word et al. 1999).

$\Pi(\mathrm{O}_{ij}|\Lambda_{ij})$ can be expressed in terms of two probabilities,

$$\Pi(\mathrm{O}_{ij}|\Lambda_{ij}) = \int \mathrm{H}(r_{ij}|\Lambda_{ij}) \cdot \mathcal{P}(\mathrm{O}|r_{ij}, r_0) \cdot dr_{ij} \quad (3)$$

where $\mathrm{H}(r_{ij}|\Lambda_{ij})$ is the probability that two protons i and j, with connectivity Λ_{ij}, are separated by a given distance r_{ij}, and $\mathcal{P}(\mathrm{O}|r, r_0)$ is the probability to observe a NOESY cross-peak originated from a pair of protons which are separated by a given distance r. Parameter r_0 denotes the average inter-proton distance at which the probability to observe an NOE is equal to 0.5 (discussed below). Thus, the calculated values of the likelihoods depend on the value specified for r_0.

$\mathrm{H}(r_{ij}|\Lambda_{ij})$ is extracted from the statistical distribution of distances between relevant protons observed in the database of known protein structures. The probability to observe an NOE between two protons separated by a distance r can be estimated from the expression

$$\mathcal{P}(\mathrm{O}|r, r_0) = \int \mathcal{P}(\mathrm{O}|V/V_0) \cdot \mathcal{P}(V|r) \cdot dV \quad (4)$$

where $\mathcal{P}(V|r)$ is the probability of a given NOESY cross-peak volume conditional on the distance between two protons, $\mathcal{P}(\mathrm{O}|V/V_0)$ encodes for the sensitivity of the NOE detection,

$$\mathcal{P}(\mathrm{O}|V/V_0) = \begin{cases} 0.5 + 0.5 * erf\left\{\frac{3}{\sqrt{2}}(V/V_0 - 1)\right\}, & V \geq V_0 \\ 0.5 - 0.5 * erf\left\{-\frac{3}{\sqrt{2}}(V/V_0 - 1)\right\}, & V < V_0 \end{cases} \quad (5)$$

and V_0 is a "characteristic" volume which can be specified from measured spectral noise and known linewidths for two resonances, assuming that the noise has a Gaussian distribution (see Eq. 6 in Grishaev and Llinás 2004). The probability $\mathcal{P}(V|r)$ was estimated from a relaxation matrix back-calculations of volumes V using distances from the 56 selected protein structures. It should be noted that the volume V appears in $\mathcal{P}(V|r)$ in arbitrary units while according to

Table 1 List of proteins in the database used for statistical analysis of distance between protons

PDB Code	Resolution (Å)	PDB Code	Resolution (Å)	PDB Code	Resolution (Å)	PDB Code	Resolution (Å)
1A6M	1.0	1G66	0.9	1K5C	0.96	1NQJ	1.0
1AHO	0.96	1G6X	0.86	1KWF	0.94	1NWZ	0.82
1BYI	0.97	1GA6	0.86	1L9L	0.92	1O7J	1.0
1C75	0.97	1GCI	0.78	1LKK	1.0	1OAI	1.0
1C7K	1.0	1GKM	1.0	1LNI	1.0	1OD3	1.0
1CEX	1.0	1GVK	0.94	1M1Q	0.97	1OEW	0.9
1DY5	0.87	1GWE	0.88	1M40	0.85	1RB9	0.92
1EB6	1.0	1I1W	0.89	1MC2	0.85	1UCS	0.62
1EJG	0.54	1IQZ	0.92	1MN8	1.0	2ERL	1.0
1ET1	0.9	1IUA	0.83	1MSB	1.0	2FDN	0.94
1F94	0.97	1IX9	0.90	1MUW	0.86	2PVB	0.91
1F9Y	1.0	1IXH	0.98	1MXT	0.95	3LZT	0.92
1FN8	0.81	1JFB	1.0	1N55	0.83	7A3H	0.95
1G2Y	1.0	1K4I	0.98	1NLS	0.94	1MSO	1.0

Eq. 5 the expression for $\mathcal{P}(\mathrm{O}|V/V_0)$ involves volume in reduced form V/V_0. Thus, the result of the integration in the right-hand side of Eq. 4 depends on how the volume V is scaled. Noting that $\mathcal{P}(\mathrm{O}|V/V_0) = 0.5$ when $V = V_0$ (see Fig. 3), we choose to scale back-calculated volumes V in such a way that $V(r_0) = V_0$. Figure 4 shows the resulting probabilities $\mathcal{P}(\mathrm{O}|r, r_0)$ as a function of the inter-proton distance r, calculated for a few values of parameter r_0.

The likelihoods $\Pi(\mathrm{O}_{ij}|\mathrm{O}_{mn}, \Lambda_{im}, \Lambda_{jn})$ were calculated from the expression

$$\Pi(\mathrm{O}_{ij}|\mathrm{O}_{mn}, \Lambda_{im}, \Lambda_{jn}) = \int\int \mathcal{P}(\mathrm{O}_{ij}|r_{ij}, r_0) \cdot \mathrm{H}(r_{ij}|r_{mn}, \Lambda_{im}, \Lambda_{jn}) \cdot \mathcal{P}(r_{mn}|\mathrm{O}_{mn}) dr_{mn} dr_{ij} \quad (6)$$

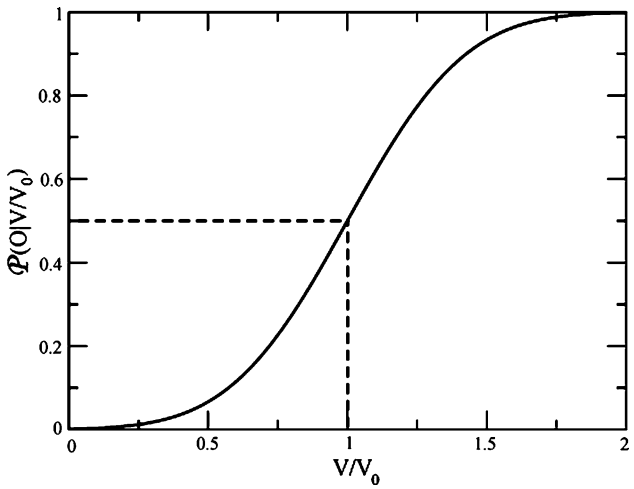

Fig. 3 Probability of observation of 2D NOESY cross-peak as a function of reduced volume V/V_0. Here $V_0 = 6\pi\sigma_1\sigma_2\sigma_{noise}$, σ_{noise} is spectral noise, and σ_1 and σ_2 are the linewidths for two dimensions, respectively

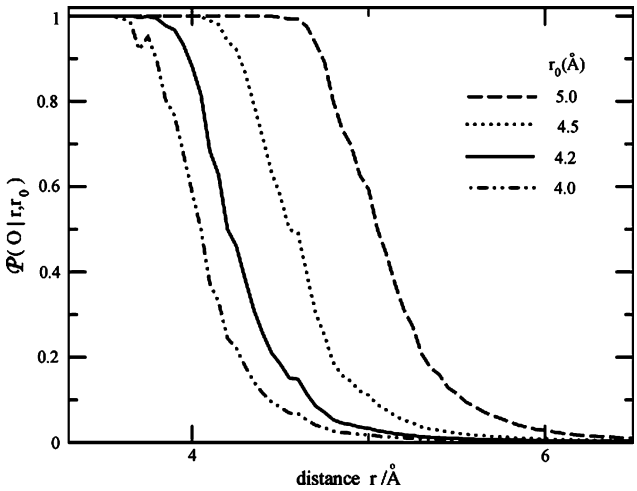

Fig. 4 Probability $\mathcal{P}(\mathrm{O}|r, r_0)$ to observe 2D NOESY cross-peak between two protons as a function of inter-proton distance r. Different curves corresponds to different values of parameter r_0, an average inter-proton distance at which the probability to observe NOESY cross-peak is equal to 0.5

where $\mathrm{H}(r_{ij}|r_{mn}, \Lambda_{im}, \Lambda_{jn})$ is a database-derived statistical distribution of distances between pair of protons i and j that are not connected ($\Lambda_{ij} = \mathrm{U}$) conditional on the distance between reporters m and n. Since the database used by Grishaev and Llinás (2004) to calculate the likelihoods included only five structures, it was not enough to derive the likelihoods $\Pi(\mathrm{O}_{ij}|\mathrm{O}_{mn}, \Lambda_{im}, \Lambda_{jn})$. Thus, a Monte Carlo procedure was implemented to generate the distance distributions, in contrast with calculating them from the database, as done in this work.

With these new classes of connectivity added to BACUS, the program can now exploit information on correlations of backbone amide protons with protons belonging to the preceding residue. Hence, two sequential

backbone H^N protons share a common subset of atoms connected to them although they are in the U class of protons not connected within a spin system. For this reason, class U was divided into two subgroups, U_1 and $\bar{U}_1$. U_1 corresponds to pairs of backbone H^N that are sequential in the protein and $\bar{U}_1$ corresponds to all other proton pairs that belong to class U. Independent probabilities were calculated for both cases. In summary, we have expanded the number of categories of database-derived likelihoods to 26, in contrast to the 7 categories considered by the original version of BACUS.

Tables 2 and 3 summarize the calculated likelihoods. Table 2 shows the dependency of $\Pi(O_{ij}|\Lambda_{ij})$ on the value of r_0. When r_0 changes from 4.0 Å to 5.0 Å, an increase of 20–27% in the value of the likelihoods occurs. Table 3 shows that $\Pi(O_{ij}|O_{mn}, \Lambda^t_{im}, \Lambda^t_{jn})$ for sequential H^N proton pairs is not symmetrical with respect to the swapping of protons *i*and *j*. This reflects the fact that directionality of the position of backbone H^N along the protein sequence has to be taken into account in the calculations.

Fragment Monte Carlo procedure

Two programs, FINDSEQ and LINKMAP, were originally created, as part of the ABACUS protocol, to predict the sequential placements of the AA-fragments, (Grishaev et al. 2005). FINDSEQ performs a probabilistic fragment assignment using Bayesian inference and relies on assembling fragments in structures ("clouds of fragments"). LINKMAP, on the other hand, makes use of a thermodynamic analogy to find a minimum of pseudo-energy that corresponds to the optimal mapping of fragments onto the primary sequence. The construction of the pseudo-energy that scores a particular fragment mapping is not necessarily restricted to the results of the "clouds of fragments". It can include information on fragment connectivity from other sources, such as NOESY data interpreted by BACUS. This avoids the need to run time-consuming MD simulations to generate the "clouds of fragments". However, LINKMAP has the following restrictions: (1) the number of fragments, N_{frag}, should be the same as the number of positions in the sequence, N_{pos}, and (2) fragments are allowed to occupy only those positions which are consistent with the fragment amino acid type prediction. The first restriction puts out of consideration cases where the experimental data does not allow for discerning all residues of the protein. The second restriction can partition the assignment space in such a way that the stochastic search can be trapped in an assignment corresponding to a local minimum and the only way to attain the optimal assignment is through sampling of unlikely, high-energy assignments.

Table 2 Likelihood $\Pi(O_{ij}|\Lambda_{ij})$ to observe NOESY cross-peak between two connected protons *i* and *j* as estimated from database of 56 ultra high resolution XRD protein structures using different values of r_0^a (see text)

r_0 (Å)	Connectivity between *i* an $j\Lambda_{ij}$			
	C	T	S	P
4.0	0.999648	0.778514	0.827644	0.687568
4.2	0.999932	0.823273	0.852674	0.734630
4.5	0.999987	0.880704	0.892865	0.805322
5.0	0.999993	0.941538	0.948085	0.879277

[a] Parameter r_0 is the average inter-proton distance at which the probability to observe NOESY cross-peak is equal to 0.5

Table 3 Likelihood $\Pi(O_{mn}|O_{ij}, \Lambda_{im}, \Lambda_{jn})$ to observe NOESY cross-peak between two disconnected protons *m* and *n* as estimated from the database of 56 protein structures using $r_0 = 5.0$ Å (see text)

Λ_{im}	Λ_{jn}				
	Δ	C	T	S	P
	For protons *m* and *n* from class U_1:				
Δ	1.0	–	–	0.2939	0.9483
S	0.8681	–	–	0.3414	0.9764
P	0.4244	–	–	0.1646	0.3265
	For protons *m* and *n* from class $\bar{U}_1$:				
Δ	1.0	0.7241	0.5533	0.6054	0.581
C	0.7241	0.5896	0.4429	0.4345	0.4173
T	0.5533	0.4429	0.3677	0.3264	0.3347
S	0.6054	0.4345	0.3264	0.4472	0.5503
P	0.5810	0.4173	0.3347	0.5503	0.512

The new procedure developed in this work, Fragment Monte Carlo (FMC), allows for fragments to occupy any position in the assignment space which consists of the primary protein sequence and a pool of unassigned fragments. The number of positions in the pool is specified by the user and should be large enough to accommodate the expected spin-systems from the protein primary sequence as well as any additional systems such as those from an expression-tag, possible minor conformations of the protein, or contaminants in the sample. However, the scoring algorithm for mapping of each fragment to the sequence only counts those fragments mapped to the primary sequence position. For stochastic sampling of assignment space the multicanonical (MUCA) MC method (Berg and Neuhaus 1991; Berg and Celic 1992) was used instead of the standard canonical MC method (Metropolis et al. 1953) employed in LINKMAP.

The essential ideas of MUCA are summarized below. A simulation with this method is performed in an iterative fashion. In each iteration, MC is run with a weight factor $e^{-S(E)}$. Each MC step consists of a randomly chosen swap of the position of two fragments. In the first iteration $S_0(E) = E/T_0$, which corresponds to the conventional

canonical MC simulation. At the first run T_0 is set sufficiently high. During the k-th simulation, the energy histogram $H_k(E)$ is constructed and the weight factor is updated by

$$S_{k+1}(E) = S_k(E) + \ln H_k(E) \quad (7)$$

The iterative process stops when the energy histogram $H(E)$ is flat within the sampling interval ΔE.

To address the minimization problem, a variant of the MUCA method was introduced where an upper bound for the energy was set, rejecting all attempts beyond this bound. A new iteration starts with the bound moved in the low-energy direction, in such a way that the sampling region includes the lowest energy sampled during the previous iterations, while the sampling interval ΔE is kept fixed. This strategy forces the sampling into the low-energy region of the assignment space.

The pseudo-energy E, which scores a particular fragment's assignment, is the sum of two terms (Grishaev et al. 2005). The first term, $E_1(i,p)$, evaluates the compatibility of the amino acid type in the position p of the protein sequence with the type identification of fragment i. The second, non-local term, $E_2(i,j)$, evaluates the possibility that two fragments i and j occupy adjacent positions in the sequence. E_2 is calculated only for pairs of fragments mapped sequentially on the sequence. It is defined in the form $E_2(i,j) \propto -\ln\Pi(i,j)$, where $\Pi(i,j)$ is a likelihood of the upstream sequential connectivity from fragment i to fragment j. Two main modifications were introduced in the way the pseudo-energy is constructed.

(1) A penalty is introduced in E_1 energy for each fragment such that the possible fragment's typing is not in agreement with the amino acid type at the sequence position to which the fragment is assigned. In the result, E_1 energy is defined by the following expression

$$E_1(i,p) = \begin{cases} -\ln\mathcal{P}(i,T_p) - E_1^{\max} - 1, & \text{if } \mathcal{P}(i,T_p) > 0 \\ 1, & \text{otherwise} \end{cases} \quad (8)$$

Here T_p is the amino acid type of residue p in the sequence, $\mathcal{P}(i,T)$ is the probability of fragment i having amino acid type T, and $E_1^{\max} = \max\{-\ln\mathcal{P}(i,T)\}$. Also, in such a manner, a "repulsion" energy is added to the E_2 term for each pair of fragments that are assigned to adjacent sequence positions in disagreement with their possible typing.

$$E_2(i,j) = \begin{cases} -\ln\Pi(i,j) - E_2^{\max} - 1, & \Pi(i,j) > 0 \\ 1, & \Pi(i,j) = 0 \end{cases} \quad (9)$$

Here $E_2^{\max} = \max\{-\ln\Pi(i,j)\}$. These changes are required in order to properly score *any* possible assignment.

(2) The second modification concerns the calculation of likelihoods $\Pi(i,j)$. In brief, the two alternative strategies to estimate $\Pi(i,j)$ as proposed by Grishaev et al. (2005) were combined in the following way

$$\Pi(i,j) = (1 - W_{cloud}) \cdot \Pi_{NOE}(i,j) + W_{cloud} \cdot \Pi_{CLOUD}(i,j) \quad (10)$$

Here, W_{CLOUD} is a user defined parameter, $\Pi_{NOE}(i,j)$ are likelihoods derived from the BACUS-processed NOESY peak list, and $\Pi_{CLOUD}(i,j)$ are likelihoods derived from the spatial proximity of fragments observed in structures of assembled fragments, "clouds of fragments", (see Grishaev et al. 2005). $\Pi_{NOE}(i,j)$ are calculated by considering each possible proton pair consisting of the H^N atom of fragment i and H^α or H^β atom of fragment j. Namely, $\Pi_{NOE}(i,j)$ is increased by the amount of W_{prior} if there is NOESY peak with a *prior* assignment to the proton pair in question and by the additional amount of W_{poster} if there is corresponding *posterior* assignment as well, were W_{prior} and W_{poster} are user defined parameters. The combination given by Eq. 10 can be justified by the fact that $\Pi_{CLOUD}(i,j)$ can provide information on fragment connectivity complimentary to that encoded in $\Pi_{NOE}(i,j)$. Indeed, $\Pi_{NOE}(i,j)$ is estimated based on the existence of NOESY cross-peaks only between main-chain protons H^N, H^α, and H^β. In the case, when no main-chain connectivities for two sequential fragments are observed in the NOE data, the missing information can be readily obtained from a cloud of fragments if the fragments in question are within globular parts of the protein and have significant density of inter-side-chain NOEs. Another advantage of constructing a pseudo-energy in two steps is that knowledge of a partial fragment assignment which is identified as highly reliable by the FMC procedure (based on only $\Pi_{NOE}(i,j)$ likelihoods) can help to generate more accurate cloud of fragments, which in turn, allows for constructing a better quality pseudo-energy.

Experimental NMR data sets

For the evaluation of the performance of BACUS and FMC procedures the experimental data sets of four proteins of known structure *mth1743* (70a.a.), *mth0256* (68a.a.), *pa0128* (116a.a.), and the *CPH* domain of *Cul7* (105a.a.) were used. Solution NMR structures of three of these proteins *mth1743, mth0256*, and *pa0128* (PDB entries: 1RYJ, 1NE3, and 2AKL, respectively) were determined by a conventional approach while the structure of *Cul7-CPH* (PDB entry 2JNG) was determined using the ABACUS approach. Details on sample preparation, NMR data collection, resonance assignment, and structure calculation of these proteins have been reported previously (Yee et al. 2002; Grishaev et al. 2005; Wu et al. 2003; Srisailam et al. 2006; Wang et al. 2007; Kaustov et al. 2007) . Briefly, the assignments of ^1H, ^{13}C, and ^{15}N resonances of *mth1743* and *mth0256* proteins were based primarily on the

following triple resonance experiments: CBCA(CO)NH, HNCACB, HNHA, HNCO, CC(CO)NH, HC(CO)NH, HCCH-TOCSY, and HCCH-COSY. A simultaneous ^{13}C- and ^{15}N-NOESY spectrum of *mth1743* was collected with a mixing time of 150 ms. The ^{13}C-edited NOESY and ^{15}N-edited NOESY spectra of *mth0256* were collected with 120 ms and 150 ms mixing times, respectively. All spectra for these two proteins were recorded on VARIAN INOVA 600 MHz and 750 MHz spectrometers. Spectra were processed using the NMR pipe (Delaglio et al. 1995) and manually peak-picked and analyzed using XEASY (Bartels et al. 1995) and SPARKY (Goddard and Kneller 2003).

For *pa0128* and *Cul7-CPH*, the assignments of ^{1}H, ^{15}N, and ^{13}C nuclear resonances were obtained by both ABACUS and conventional approaches using the following experiments: CBCA(CO)NH, HBHA(CO)NH, HNCA, HNHA, HNCACB, CC(CO)NH, HC(CO)NH, HC(C)H-TOCSY and (H)CCH-TOCSY. The ^{13}C-edited NOESY and ^{15}N-edited NOESY spectra were collected with 120 ms and 150 ms mixing times, respectively. All NMR spectra were recorded on BRUKER AVANCE 600MHz spectrometer. Spectra were processed using the NMR pipe and manually peak-picked and analyzed using SPARKY.

NMR data used as input for the validation of FMC

The FMC procedure requires the protein sequence, a list of unassigned ^{1}H, ^{15}N, and ^{13}C resonances grouped in spin-systems (PB or AA fragments), and ^{13}C- and ^{15}N-edited 3D NOESY peak lists as an input.

For *mth1743* and *mth0256* proteins, the list of unassigned spin-systems was generated from known sequence-specific resonance assignments available from BioMagResBank (accession codes: *mth1743*, 5106; *mth0256*, 5620).

In the case of *pa0128* and *Cul7-CPH*, the grouping of ^{1}H, ^{15}N, and ^{13}C nuclear resonances in spin-systems (PB fragments) was performed by analyzing the raw data from HNCO, CBCA(CO)NH, HBHA(CO)NH, CC(CO)NH, HC(CO)NH, HC(C)H-TOCSY and (H)CCH-TOCSY experiments. The resulting spin-systems were practically identical with those obtained based on conventional assignment procedure. For both proteins $\sim$98% of all chemical shifts were identical for both methods with the remaining discrepancies attributed to minor differences in CH_2 groups. A conventional manual resonance assignment was performed independently (Srisailam S. and Kaustov L., personal communication) for the purpose of validation and involve assignment of backbone resonances using the data from additional experiments HNCA, HNHA, and HNCACB that contain information on sequential connectivity.

For all four proteins we have utilized the same NOESY peak lists that were used in the original structure determination. These lists, except for the case of *mth1743*, are 'raw' lists which contain both the peaks that were assigned and used to generate distance constraints in the final stage of structure calculation and the unassigned peaks that were not compatible with the final structure. The unassigned NOESY peaks can be attributed to noise peaks, peaks that originated from unassigned resonances, and other possible artifacts often seen in NOESY data. The percentages of the unassigned (spurious) peaks for the proteins in question are between 0% and 10% .

NMR data used for the validation of the automated NOE interpretation with BACUS

Not all five connectivity lists shown in Fig. 2 as an input are necessary for BACUS to perform an automated NOE interpretation. Lists of S- and P-connected protons, as well as the list of sequential backbone H^N pairs are optional and could be provided to BACUS when available. BACUS is used in different stages of the ABACUS protocol (Grishaev et al. 2005) when different information on proton connectivity is available. For example, in the resonance assignment stage, in the case of unassigned resonances grouped in PB spin-systems, both the list of S-connected protons and the list of sequential backbone H^N pairs are not available. However, when BACUS is used in the structure calculation stage, sequence-specific resonance assignment are known and all five connectivity lists are available. Therefore, we have tested BACUS performance with different sets of connectivity lists as input. The data for two proteins *mth1743* and *mth0256* were used in the test. The list of chemical shifts and connectivity lists included in the input of BACUS were generated from known sequence-specific resonance assignment.

It should be noted here that an implicit assumption underlying the NOE interpretation with BACUS is that all input NOE peaks are "real" and the prime objective of the BACUS procedure is to resolve the assignment ambiguities for each input peak. No validation of the NOESY peaks is incorporated in BACUS. Therefore, the NOESY peak lists used in the test of BACUS performance do not include the original peaks identified as spurious in the final stage of the original structure calculation with CYANA/CANDID. The original NOESY peak list of *mth0256* contained $\sim$8% peaks that were recognized as spurious, while the NOESY peak list of *mth1743* was clean.

Details of the calculations

For all calculations with BACUS reported in this work tolerances of 0.03, 0.05, and 0.5 were used for chemical shift matching in the direct ^{1}H, indirect ^{1}H, and ^{13}C or ^{15}N dimensions, respectively. Parameters W_{CLOUD}, W_{prior} and W_{poster} used in the calculation of pseudo-energy were set to

0.5, 1, and 10, respectively. Each MUCA simulation performed in this work consists of 25 iterations with 5×10^5 MC steps per iteration. Temperature T_0 for the first iteration was set to 10.0.

Results and discussion

Quality of NOE cross-peak assignment

NOE assignments produced by our modified BACUS with respect to reference assignments for two proteins, *mth1743* and *mth0256* (PDB code 1RYZ and 1NE3, respectively) were analyzed. Data for protein *mth0256* is an example of less than ideal NMR data: the spin system data for three residues was missing, and the sensitivity of the NOESY spectra were of poor quality. For both proteins, the reference NOE assignments were obtained by semi automated assignment methods (Herrmann et al. 2002; Güntert 2004) performed in the context of a conventional structure determination with manual assessment of the NOE assignments. The assignment of NOESY cross-peaks by BACUS was performed with the same NOESY cross-peak lists used to obtain the corresponding reference assignments. For every NOESY cross-peak a direct comparison of the BACUS assignment with the corresponding reference assignment was made. The results of such comparison for both proteins are shown in Table 4.

The following conclusions can be made from the results presented in Table 4.

(1) A change of the set of likelihoods calculated with $r_0 = 5.0$ Å to those with $r_0 = 4.0$ Å results in fewer unambiguous, but more accurately assigned, peaks. The percentage of 'correctly' assigned inter-fragment peaks increased by ~2–6% and ~6–10% for 3D and 2D NOESY spectra, respectively.

(2) Adding the information on the position of the residues in the protein sequence increases the overall quality of NOE assignments by ~3% and ~10% for 3D and 2D NOESY data, respectively. Most noticeable is an increase of ~30–50% of the correctly assigned inter-residue cross-peaks for 2D NOESY data, in which case the network of proton connectivities remains the only basis to resolve assignment ambiguities.

(3) Differences in NOE assignments depending on the type of fragments used, PB-fragments (C, T, P classes) and AA-fragments (C, T, S classes), are also shown in Table 4. For PB-fragments, there are more inter-fragment assignments with a higher accuracy than for AA-fragments. The number of correctly assigned inter-fragment peaks increases by ~6%, ~15%, and ~27% for 3D data of *mth0256*, and 3D and 2D data of *mth1743*, respectively.

(4) The weighting of symmetry-related cross-peak assignments improves the quality of NOE assignments and could increase the number of unambiguously resolved peaks. For example, in the case of PB-fragments of *mth1743*, the number of inter-fragment assignments obtained at Ws = 3 is increased by 3.5% and their accuracy is increased by 2.3% comparing to the case when no symmetry-related weighting was applied (Ws = 1). Overall on the test cases, the optimal performance of BACUS was achieved for Ws within the 3–5 range .

The additional information on sequential connectivities, as implemented in our modified BACUS, affects mostly the assignments of sequential and long-range NOEs (between residues separated by more than 5 positions in the sequence). For example, in the case of 3D NOESY data of protein *mth1743*, ~20% of the sequential NOE correlations were assigned incorrectly as long-range NOEs when the original connectivity lists were used. The validation of the long-range assignments by performing "goodness of fit" to the known NMR structure ensemble indicates that the additional information on sequential connectivities reduces the number of incorrect long-range assignments by approximately one half: there are 57 (~17% of all long-range assignments) incorrect constraints produced by the original version of BACUS and only 25 (~9%) incorrect constraints produced by the new version.

Assigning spin-systems to the primary sequence

Effective sampling of the low-energy region of assignment space using the FMC procedure yields a set of assignments which are optimal or close to optimal according to the scoring function $E(A)$. We have tested the performance of FMC, with both AA- and PB-fragments, using data for four proteins of known structure: *mth1743, mth0256, pa0128*, and *CPH* domain of *Cul7* (PDB code 1RYJ, 1NE3, 2AKL, and 2JNG, respectively). NOESY peak lists used as input for FMC contained spurious peaks (see Table 5) in order to demonstrate the performance of the algorithm under condition of noisy data. BACUS, with set of $r_0 = 5.0$ Å likelihoods and Ws = 5, was used to assign the NOESY cross-peaks. The results of the FMC calculations are summarized in Table 5.

For PB-fragments the resulting assignments were 100% correct for all four proteins. When AA-fragments were used, for proteins *mth0256* and *Cul7-CPH* the assignments were not fully correct. The better performance of FMC with PB-fragments than with AA-fragments can be rationalized by the fact that for some residues no *inter*-residue NOESY cross-peaks are observed while *intra*-residue NOEs between backbone H^N and other protons from

Table 4 Comparison of NOESY cross-peak assignments produced by BACUS with the reference assignment[a]

BACUS input[b]			Number of NOESY cross-peaks[c]			$N^{d,e}_{sequential}$
Lists	r_0 (Å)	Ws	N^{c}_{total}	$N_{unambiguous}$	$N^{d}_{inter-fragment}$	
Protein mth1743 *3D NOESY spectra:*						
Original	–	–	2465	2027 (91.6%)	800 (87.7%)	51
C, T, S	5.0	5	2465	2091 (91.3%)	851 (87.9%)	57
C, T, S	4.0	5	2465	2053 (92.5%)	825 (89.8%)	58
C, T, P	5.0	5	2465	2115 (89.5%)	955 (89.2%)	61
C, T, P	4.0	5	2465	2038 (91.7%)	903 (92.1%)	60
All	5.0	5	2465	2039 (93.5%)	858 (90.9%)	67
All	4.0	5	2465	1996 (94.7%)	824 (93.1%)	67
C, T, P	5.0	1	2465	2100 (88.4%)	933 (87.4%)	61
C, T, P	5.0	2	2465	2125 (88.9%)	961 (88.4%)	60
C, T, P	5.0	3	2465	2127 (89.7%)	966 (89.7%)	62
C, T, P	5.0	7	2465	2119 (89.4%)	960 (88.9%)	61
C, T, P	5.0	10	2465	2124 (89.6%)	965 (89.4%)	61
C, T, P	5.0	20	2465	2124 (89.4%)	967 (88.9%)	61
Protein mth1743 2D *NOESY spectra:*						
Original	–	–	2374	1617 (72.1%)	484 (56.6%)	26
C, T, S	5.0	5	2374	1724 (70.0%)	569 (54.5%)	27
C, T, S	4.0	5	2374	1597 (72.8%)	464 (59.0%)	24
C, T, P	5.0	5	2374	1761 (67.9%)	613 (63.9%)	35
C, T, P	4.0	5	2374	1641 (70.0%)	532 (66.0%)	34
All	5.0	5	2374	1568 (78.2%)	586 (69.1%)	60
All	4.0	5	2374	1515 (82.6%)	543 (79.0%)	62
Protein mth0256 *3D NOESY spectra:*						
Original	–	–	2255	1848 (89.6%)	792 (85.0%)	41
C, T, S	5.0	5	2255	1865 (89.4%)	815 (85.2%)	43
C, T, S	4.0	5	2255	1818 (91.2%)	772 (87.6%)	44
C, T, P	5.0	5	2255	1873 (89.1%)	872 (85.8%)	47
C, T, P	4.0	5	2255	1810 (90.4%)	804 (87.8%)	45
All	5.0	5	2255	1809 (92.0%)	824 (87.5%)	48
All	4.0	5	2255	1800 (92.5%)	799 (88.5%)	48
C, T, P	5.0	1	2255	1874 (87.6%)	865 (83.4%)	43
C, T, P	5.0	2	2255	1881 (88.2%)	875 (84.4%)	44
C, T, P	5.0	3	2255	1878 (88.9%)	872 (86.0%)	46
C, T, P	5.0	7	2255	1868 (88.0%)	867 (83.4%)	44
C, T, P	5.0	10	2255	1858 (88.8%)	857 (85.2%)	45
C, T, P	5.0	20	2255	1866 (89.1%)	865 (85.8%)	46

[a] The reference assignment is the final assignment obtained after 7 cycles of structure-calculation/NOE-assignment using CAYANA /CANDID (Güntert 2004; Güntert et al. 1997; Herrmann et al. 2002)

[b] A number of different assignments with BACUS were performed using different kind of information available on proton connectivities and different sets of database likelihoods

C, T, P, and S specifies the connectivity lists which were used as an input for BACUS; "*All*" denotes the case when all available information on proton connectivity (including the list of sequential H^N-H^N pairs) was used while "*original*" refers to the calculations with the original version of BACUS

Parameter r_0 is the average inter-proton distance at which the probability to observe NOESY cross-peak is equal to 0.5. Different values of r_0 specify different sets of database likelihoods used by BACUS

[c] Only those NOESY peaks that have the reference assignment were used as input for BACUS. N_{total} specifies the total number of reference assignments. The numbers in brackets show percentage of unambiguous BACUS assignments which are identical to the reference assignments

[d] Different types of fragments were used to classify restraints on *inter-* and *intra*-fragment depending on what connectivity lists were used as an input for BACUS. PB-fragments are used when list on P-connectivity is available but list on S-connectivity do not. In all other cases AA-fragments are used

[e] $N_{sequential}$ is the number of contacts between fragments adjacent in sequence according to the NOE assignment. Two fragments are considered to be in contact if there is at least one unambiguously assigned NOE between H^N proton of one fragment and H^{α}/H^{β} protons of another fragment

the same residue are present in the data. In the case of PB-fragments these *intra*-residue NOEs are *inter*-fragment, so that more information on sequential connectivity is available than for the case of AA-fragments. For example, on Table 4 one can see that for *mth0256* in the case of PB-fragments there are 4 more sequential inter-fragment contacts than in the case of AA-fragments (Two fragments are considered to be in contact if there is at least one

Table 5 Comparison of the results of mapping of two types of spin-systems onto the primary sequence using the FMC procedure

Protein (PDB id)	*mth1743* (1RYJ)	*mth0256* (1NE3)	*pa0128* (2AKL)	*Cul7-CPH* (2JNG)
Number of residues in primary sequence	70	68	116	105
Number of spin-systems (fragments)	70	65	130[a]	104
Conventional resonance assignment completeness	98%	95%	96%	94.8%
Number of NOESY peaks	2792	2746	5024	4450
Percentage of spurious[b] NOESY peaks	0%	8%	2.2%	9.3%
Number of assigned[c] AA-fragments	70 (100%)	65 (85%)	116 (100%)	104 (96%)
Number of assigned[c] PB-fragments	70 (100%)	65 (100%)	116 (100%)	104 (100%)

[a] 14 spin-systems originate from His-tag and minor protein conformations

[b] Peaks that were not included in the final cycle of the original structure calculations

[c] Fragments that were placed on the protein sequence positions. Percentage of the correctly assigned fragments is given in brackets

unambiguously assigned NOE between H^N proton of one fragment and H^{α}/H^{β} protons of another fragment). The distribution of the difference between the numbers of intra-residue and sequential NOEs expected for a backbone amide proton, extracted from an ensemble of distances between backbone H^N proton and H^{α}/H^{β} protons obtained from 56 proteins of known structure, is shown on Fig. 5. The results suggest that in globular proteins backbone H^N's exhibit, on average, 1.4 more intra-residue than sequential NOEs. On this basis, one can expect that sequence-specific fragment assignments be more effective with PB-fragments than with AA-fragments.

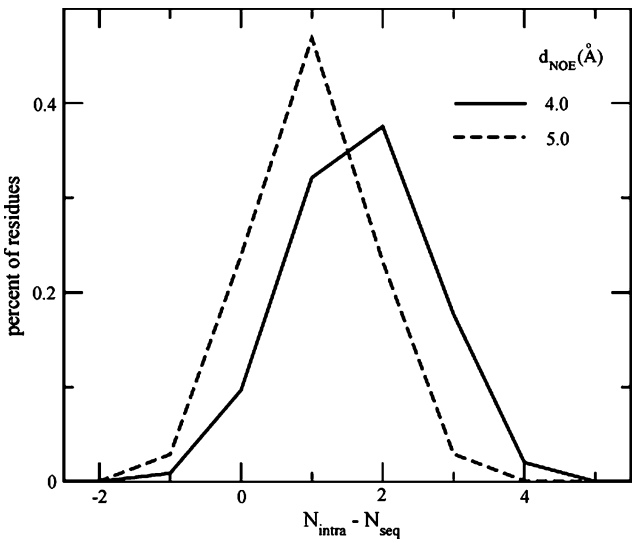

Fig. 5 The distribution of the difference $N_{intra} - N_{seq}$ extracted from database of 56 high resolution XRD structures. N_{intra} and N_{seq} are the number of expected NOESY cross-peaks between backbone amide H^N proton and H^{α}/H^{β} protons from the same and preceding residue, respectively. NOESY cross-peak between two protons is expected if the corresponding proton distance $< d_{NOE}$ in a known protein structure, where $d_{NOE} = 4$ Å (solid line) and 5 Å (broken line). The statistics was made over 8301 residues

The poorer performance of FMC observed for *mth0256* can be explained by the lower quality of its NMR data. For this protein, residues 17, 19, and 20 failed to yield signals, so that only 65 fragments could be mapped to 68 positions. Upon typing, 62 fragments were unique, and 3 fragments (GLN8, ASP11, GLU9) were ambiguous. The properties of the pseudo-energy $E(A)$ constructed from these data depends on the type of fragments (PB- or AA-) used. The pseudo-energy surface profile along variable Q(A), that measures the "quality" of an assignment A, was monitored. Q(A) for a particular assignment A is defined as the number of positions in the sequence with correct fragment assignments. The residues with "missing" data (17, 19, and 20) were also counted if no fragment was assigned to them, so that $Q = 68$ for the correct assignment and $Q < 68$ for all other cases. Figure 6 shows the pseudo-energy profile along Q as sampled in a MUCA run consisting of 25 cycles of 10^5 MC steps each. It is noteworthy that in the case of PB-fragments the pseudo-energy exhibits the desired property, i.e. the global minimum of $E(A)$ corresponds to the correct assignment (see Fig. 6B). The different assignments for a given $Q < 68$ can have different energies taken from some energy interval $\delta E(Q)$. As Q decreases, the size of the interval $\delta E(Q)$ increases and its position shifts to higher energies.

In the case of AA-fragments the global minimum of $E(A)$ is degenerate (Fig. 6A): there are 5 assignments of different quality ($Q = 58$, 60, 62, 64, and 66 respectively) corresponding to the global minimum; moreover, the pseudo-energy of the correct assignment is higher than at the global minimum. Thus, in this case we are not able to choose one assignment as optimal on the basis of the defined scoring function.

Assignment confidence estimation

In order to determine a unique solution using the FMC procedure it is required that the global minimum of a

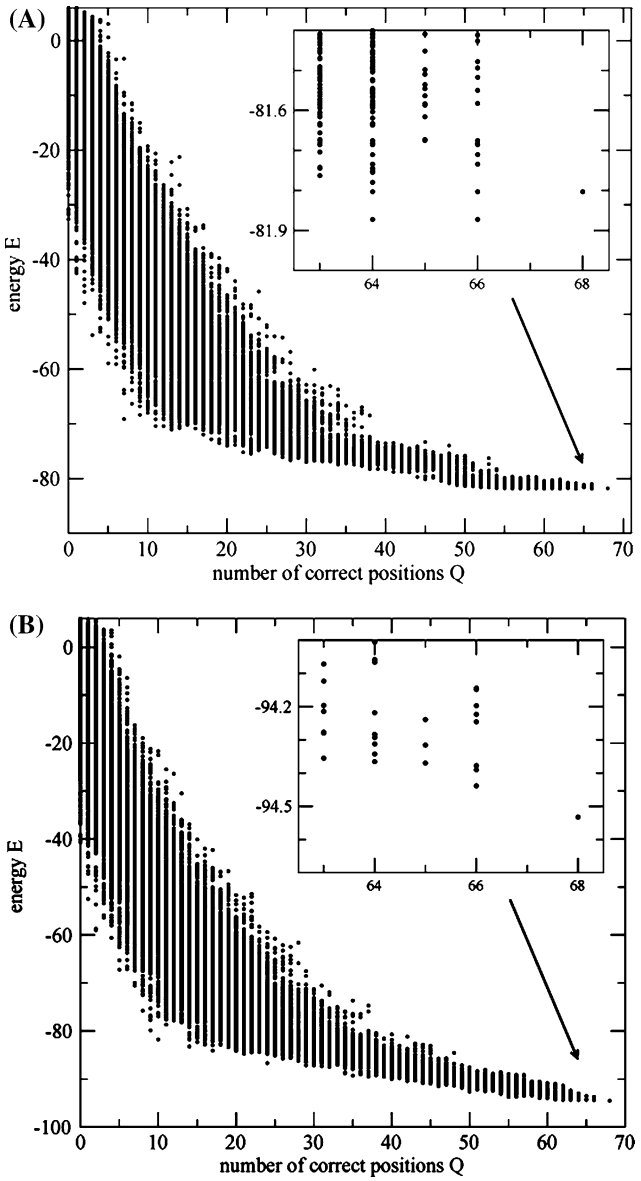

Fig. 6 Pseudo-energy E versus quality parameter Q for sequence assignments of AA-(**A**) and PB-(**B**) fragments of protein *mth0256* obtained during MUCA runs. The value of Q = 68 corresponds to the correct assignment. The insets show the low pseudo-energy regions in detail

pseudo-energy function corresponds to the correct fragment assignment. However, data obtained under conditions such as low signal-to-noise, poor dispersion of chemical shifts, chemical exchange and conformational exchange may not be of sufficient quality to meet this requirement. Poorly defined and extra spin-systems could result in some regions of the protein sequence exhibiting ambiguous or incorrect assignments. The example of protein *mth0256* described above demonstrates a pseudo-energy constructed from sub-optimal NMR data that had a global minimum that did not correspond to the correct fragment assignment.

It is of great practical importance to know the reliability of the results obtained. The MUCA simulation provides a means to quantitatively estimate this reliability within the theoretical framework without additional experimental information. From the MUCA simulation run, one can not only locate the global-energy minimum, but also calculate various properties that depend on the ensemble of all possible assignments (i.e. 'thermodynamic properties'). In particular, the probability of fragment k to occupy sequence position s, P_k (s, T), is defined as a canonical expectation value of the corresponding fragment occupancy by

$$P_k(s,T) = \sum_A \delta^k_{i(A,s)} \cdot e^{-E(A)/T} \Big/ \sum_A e^{-E(A)/T} \qquad (11)$$

Here, the summation is over all possible fragment assignments, A denotes a particular assignment for the complete set of fragments, $i(A, s)$ denotes a fragment that occupies a sequence position s in assignment A, δ^m_n is the Kronecker's delta, and T stands for temperature. The canonical ensemble average value given by Eq. 11 can be calculated for *any* temperature from *just one* MUCA run by using umbrella sampling reweighting techniques (Ferrenberg and Swendsen 1989; Hansmann and Okamoto 1993).

The occupancy probabilities $P_k(s,T)$ provide a probabilistic assignment and serve as a gauge of the certainty of the obtained optimal assignment. Figure 7 shows $P_k(s,T)$

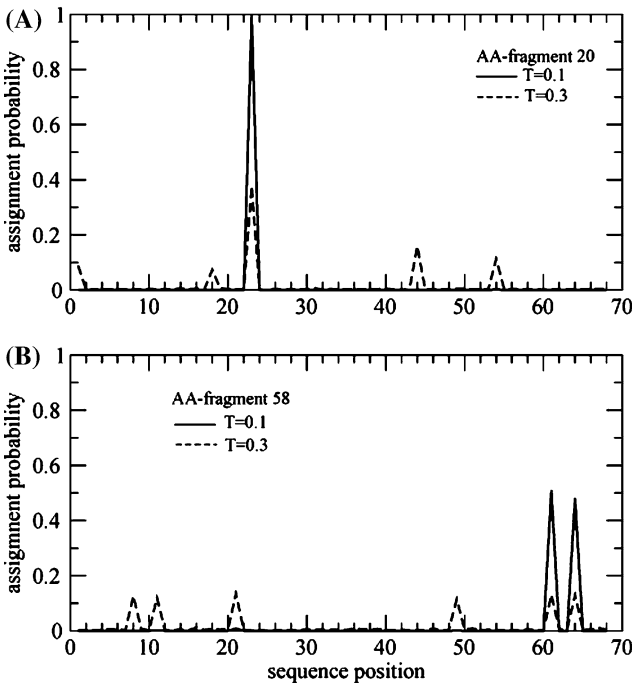

Fig. 7 Fragment assignment probabilities (see Eq. 11) versus sequence position calculated for two AA-fragments of protein *mth0256* at two temperatures: T = 0.3 (dashed line) and 0.1 (solid line), respectively

for two AA-fragments of protein *mth0256* ($k = 58$ and 20, respectively) as a function of the sequence position s calculated at two different temperature values. At high temperatures, $T \geq 0.3$, the assignment probability is distributed rather uniformly over many different sequence positions. However, at low temperatures, $T \approx 0.1$, the assignment probability is concentrated only on one or a few sequence positions. For example, AA-fragment 58 has almost equal probability (≈ 0.5) to occupy positions 61 and 64 at $T = 0.1$ (see Fig. 7B), which indicates an ambiguity of the fragment assignment. To measure quantitatively the confidence of the assignment of fragment k we apply the following ratio

$$\rho(k,T) = P_k^{\max 1}(T)/P_k^{\max 2}(T) \quad (12)$$

where $P_k^{\max 1}(T)$ and $P_k^{\max 2}(T)$ are the first and the second top probabilities $P_k(s,T)$, $s = 1,..,N_{pos}$. From our experience, the assignment of a fragment can be considered to be reliable if its confidence $\rho \geq 4$, which corresponds to $P_k^{\max 1} > 0.8$. Table 6 summarizes the assignment confidence calculated for proteins *mth1743* and *mth0256* at low temperature $T = 0.1$. The results reflect differences in the quality of the NMR data of these proteins: all PB-fragments of *mth1743* are assigned with high confidence while five PB-fragments of *mth0256* have low assignment confidence ($\rho < 4$). The assignment probabilities $P_k(s,T)$ identify the portion of the optimal assignments that are accurate and those that are not. Thus, in the case of AA-fragments of *mth0256*, the assignments corresponding to the global minimum of the pseudo-energy are not 100% correct. For example, fragment 17 is assigned incorrectly to sequence position 1, the same in all optimal assignments. At the same time, the certainty ρ calculated for fragment 17 is 2.1 which indicates that the fragment has significant occupancy on the other sequence positions, so that the assignment of fragment 17 cannot be considered reliable. It should be noted here that performing simulated annealing runs does not provide this information since they supposedly yield the optimal assignment.

Table 6 Statistic of the fragments assignment confidence ρ[a]

Confidence $\rho_{\max}$	Number of fragments[b] with $\rho > \rho_{\max}$		
	mth1743 PB-fragments	*mth0256* PB-fragments	*mth0256* AA-fragments
2	70 (100%)	65 (100%)	57 (87.6%)
3	70 (100%)	60 (92.3%)	55 (84.6%)
5	70 (100%)	60 (92.3%)	49 (75.3%)
10	70 (100%)	57 (87.7%)	46 (70.8%)
20	70 (100%)	56 (86.2%)	39 (60.0%)

[a] The confidence is calculated using Eq. 12 with $T = 0.1$

[b] The number of all fragments used for assignment are 70 and 65 for proteins *mth1743* and *mth0256*, respectively. The number in brackets indicate what percentage of all fragments were assigned with specified (or better) confidence

Conclusions

We have further developed the ABACUS protocol. The classes of proton connectivities that BACUS exploits for automated interpretation of NOE data were extended. This allows for the BACUS algorithm to incorporate information on sequential, as well as intra-residue, connectivities. The likelihoods that BACUS requires were estimated from a database of 56 high resolution X-ray crystal structures. Performance of the program when tested on different proteins show that the quality of the NOE assignments increases as the information on proton connectivities is extended. Up to 94% of the unambiguously assigned NOEs match the reference assignments obtained by conventional methods. This underscores the positive influence of the changes introduced in the new version of BACUS.

The mapping of AA- and PB-fragments onto the primary sequence made by FMC was tested for four different proteins. The results indicate that it is more robust and reliable to map PB-fragments than AA-fragments. It is revealing that the analysis of occupancy probabilities allows one to obtain a partial, yet highly reliable assignment, even when NMR data are sub-optimal. This is particularly desirable for many experimental situations such as proteins with poor solubility or when spectra are complicated by conformational averaging.

The developments reported here enable an ABACUS resonance assignment strategy that is based on PB-spin-system identification and has the advantage that it does not rely on sequential connectivities from less sensitive experiments such as HNCACB which most traditional sequential assignment procedures rely on. Software implementing the developed algorithms is available from authors upon request.

Acknowledgements We wish to thank Dr. Lilia Kaustov and Dr. Sampas Srisailam for providing us with resonance assignments of *CPH* domain of *Cul7* and protein *pa0128*, respectively, which were performed by conventional manual procedure. We also thank Dr. Bin Wu for providing us with complete resonance assignment and NOESY peak lists of proteins *mth1743* and *mth0256*. This work was supported by the Ontario Research and Development Challenge Fund, the US National Institute of Health Protein Structure Initiative (P50-GM62413-01 and GM67965) through the Northeast Structural Genomics Consortium, Genome Canada through the Ontario Genomics Institute, and The Canadian Institutes of Health Research.

References

Bailey-Kellogg C, Widge A, Kelley JJ, Berardi MJ, Bushweller JH, Donald BR (2000) The NOESY jigsaw: automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. J Comput Biol 7:537–558

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

Berg BA, Neuhaus T (1991) Multicanonical algorithms for first order phase transitions. Phys Lett B 267:249–253

Berg BA, Celik T (1992) New approach to spin-glass simulation. Phys Rev Lett 69:2292–2295

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242

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

Ferrenberg AM, Swendsen RH (1989) New Monte Carlo technique for studing phase transitions. Phys Rev Lett 63:1658

Goddard TD, Kneller DG (2003) Sparky - NMR assignment and integration software. University of California, San Francisco

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

Grishaev A, Llinás M (2002b) Protein structure elucidation from NMR proton density. Proc Natl Acad Sci USA 99:6713–6718

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

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

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

Güntert P (2004) Automated NMR structure calculation with CYANA. Methods Mol Biol 278:353–378

Hansmann UH, Okamoto Y (1993) Prediction of peptide conformation by multicanonical algorithm: new approach to the multiple-minima problem. J Comp Chem 14:1333–1338

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

Kaustov L, Lukin J, Lemak A, Duan S, Ho M, Doherty R, Penn LZ, Arrowsmith CH (2007) The conserved CPH domains of Cul7 and PARC are protein-protein interaction modules that bind the tetramerization domain of p53. J Biol Chem 282:11300–11307

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

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. J Chem Phys 21:1087–1092

Moseley HN, Montelione GT (1999) Automated analysis of NMR assignments and structures for proteins. Curr Opin Struct Biol 9:635–642

Srisailam S, Lukin JA, Lemak A, Yee A, Arrowsmith CH (2006) Sequence Specific resonance assignment of a hypothetical protein PA0128 from pseudomonas aeruginosa. J Biomol NMR 36:27

Wand AJ, Nelson SJ (1991) Refinement of the main chain directed assignment strategy for the analysis of 1H NMR spectra of proteins. Biophys J 59:1101–1112

Wang X, Srisailam S, Yee A, Lemak A, Arrowsmith CH, Prestergard JH, Tian F (2007) Domain-domain motions in proteins from time-modulated pseudocontact shifts. J Biomol NMR 39:53–61

Word JM, Lovell SC, Richardson JS, Richardson DC (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol 285: 1735–1747

Wu B, Yee A, Pineda-Lucena A, Semesi A, Ramelot TA, Cort JR, Jung JW, Edwards A, Lee W, Kennedy M, Arrowsmith CH (2003) Solution structure of ribosomal protein S28E from methanobacterium thermoautotrophicum. Protein Sci 12: 2831–2837

Yee A, Chang X, Pineda-Lucena A, Wu B, Semesi A, Le B, Ramelot T, Lee GM, Bhattacharyya S, Gutierrez A, Denisov A, Lee C, Cort JR, Kozlov G, Liao J, Finak G, Chen L, Wishart D, Lee W, McIntosh LP, Gehring K, Kennedy MA, Edwards AM, Arrowsmith CH (2002) An NMR approach to structural proteomics. Proc Nat Acad Sci USA 99:1825–1830

Zimmerman DE, Montelione GT (1995) Automated analysis of nuclear magnetic resonance assignment for proteins. Curr Opin Struct Biol 5:664–673