ABACUS, a Direct Method for Protein NMR Structure Computation via Assembly of Fragments

A. Grishaev,[1] C.A. Steren,[1] B. Wu,[2] A. Pineda-Lucena,[2] C. Arrowsmith,[2] and M. Llinás[1]*
[1]*Department of Chemistry, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213*
[2]*The Ontario Cancer Institute and Department of Medical Biophysics, University of Toronto, Toronto, ON M5G 2M9, Canada*

ABSTRACT **The *ABACUS* algorithm obtains the protein NMR structure from unassigned NOESY distance restraints. *ABACUS* works as an integrated approach that uses the complete set of available NMR experimental information in parallel and yields spin system typing, NOE spin pair identities, sequence specific resonance assignments, and protein structure, all at once. The protocol starts from unassigned molecular fragments (including single amino acid spin systems) derived from triple-resonance $^{1}H/^{13}C/^{15}N$ NMR experiments. Identifications of connected spin systems and NOEs precede the full sequence specific resonance assignments. The latter are obtained iteratively via Monte Carlo-Metropolis and/or probabilistic sequence selections, molecular dynamics structure computation and *BACUS* filtering (A. Grishaev and M. Llinás, J Biomol NMR 2004;28:1–10). *ABACUS* starts from scratch, without the requirement of an initial approximate structure, and improves iteratively the NOE identities in a self-consistent fashion. The procedure was run as a blind test on data recorded on mth1743, a 70-amino acid genomic protein from *M. thermoautotrophicum*. It converges to a structure in ca. 15 cycles of computation on a 3-GHz processor PC. The calculated structures are very similar to the ones obtained via conventional methods (1.22 Å backbone RMSD). The success of *ABACUS* on mth1743 further validates *BACUS* as a NOESY identification protocol. Proteins 2005;61:36–43.**

Key words: *BACUS*; NOESY identification; *CLOUDS*; proteomics; structural genomics NMR

INTRODUCTION

A lingering problem with the determination of protein structures via NMR is the degree of subjectivity in the data interpretation process. The steps leading to the structural calculation are in many cases performed interactively, requiring times far exceeding those common when dealing with X-ray diffraction data from crystal samples. Such types of analyses also can lead to a degree of divergence among structures solved independently by different laboratories. Pinpointing causes for this divergence is often difficult to achieve, as the required depositions into public databases rarely extend to the spectral data. For these issues to be addressed, structure calculation from NMR data has to be automated to the fullest possible extent. As the bulk of the NMR structural information comes from the NOESY experiments, NOESY crosspeak assignment presents one of the most significant challenges in the NMR data analysis.

BACUS is an automated Bayesian analysis routine that aims at identifying NOE crosspeaks prior to the sequential assignments, bootstrapping identities via a set of specially designed computational tools.[1,2] While testing a *BACUS*-based protocol applicable to the analysis of $^{1}H/^{13}C/^{15}N$ heteronuclear data, we experimented with a novel procedure whereby the NOE-constrained H-atoms in the *CLOUDS* protocol [3,4] become groups of covalently linked atoms, that is, molecular fragments, that serve as building blocks to generate spatial distributions. The fragments were then assembled into a linear sequential array, which was subsequently folded under the NOE distance restraints to generate a bundle of protein structures. The latter, in turn, were iteratively tested for the self-consistency of NOE identities via *BACUS*. The procedure, named *ABACUS* (for *A*pplied *BACUS*), was validated by applying it to data from a protein of unknown structure. The blind test was run on NMR data acquired by the subgroup of authors based at the University of Toronto (UofT) on the 70-residue structural genomics target protein mth1743 from *Methanobacterium thermoautotrophicum* (mth). The UofT analysis used industry-standard protocols, including sequential spectral assignments via triple-resonance isotope-edited experiments, followed by assignment of unambiguous NOEs and subsequent assign-

Abbreviations: MD, molecular dynamics; NOE, nuclear Overhauser effect; NOESY, NOE correlation spectroscopy; RMSD, root-mean-squares deviation.

Grant sponsor: the U.S. Public Health Service; NIH; Grant numbers: GM67965 and P50 GM62513-05; Grant sponsor: The Northeast Structural Genomics Consortium; Grant sponsor: the Ontario Research and Development Challenge Fund, and Genome Canada.

A. Grishaev's present address is Laboratory of Chemical Physics, NIDDK, National Institutes of Health, Building 5, 9000 Rockville Pike, Bethesda, MD 20892.

A. Pineda-Lucena's present address is AstraZeneca, Structural Biology Laboratory, 50F23 Mereside, UK.

*Correspondence to: Miguel Llinás, Department of Chemistry, Carnegie Mellon University, Pittsburgh, PA 15213. E-mail: llinas@andrew.cmu.edu

Received 10 September 2004; Revised 1 December 2004; Accepted 13 December 2004

Published online 3 August 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20457

ment of ambiguous NOEs concurrent with structure calculation via *NOAH*.[8] Frequency lists ($^{1}H/^{13}C/^{15}N$) grouped into unassigned amino acid spin systems were made available to the Carnegie Mellon University (CMU) team prior to the public release of the sequence-specific NOE assignments and the solved NMR structure. We find that the *ABACUS* protocol is robust and rapid, and flexible enough to be readily adapted to other types of experimental heteronuclear-edited NMR data.

MATERIALS AND METHODS

Protein Purification

Mth1743 is a conserved and uncharacterized protein from the genome of *M. thermoautotrophicum*. The DNA coding for mth1743 (70 amino acids) from *M. thermoautotrophicum* was subcloned into the pET-15b expression vector. Uniformly $^{13}C/^{15}N$-labeled mth1743 was expressed and purified as previously described.[6] The purified protein contained the complete sequence of mth1743 plus three additional N-terminal residues (Gly-Ser-His) remaining after proteolytic cleavage of the His6 affinity tag. The concentration of protein samples ranged from 1.0 to 1.5 m*M* in an aqueous solution containing 25 m*M* sodium phosphate (pH 6.5), 550 m*M* NaCl, 1 m*M* DTT, 95% H_2O/5% D_2O.

NMR Spectroscopy and Sequence-Specific Spectral Assignments (UofT)

NMR spectra on mth1743 were collected at 25°C using a Varian Inova 600 MHz spectrometer equipped with pulsed field gradient triple-resonance probes. Chemical shifts were referenced to external DSS. Spectra were processed using the program *NMRPipe*[7] and analyzed with the program *XEASY*.[8] *SPSCAN*[9] was used to convert *NMRPipe* formatted spectra into *XEASY*. Sequence-specific assignments of HN, ^{15}N, ^{13}CO, $^{13}C^{\alpha}$, and $^{13}C^{\beta}$ resonances for all nonproline residues of mth1743 were obtained from HNCO, CBCA(CO)NH, and HNCACB spectra. Side-chain ^{1}H and ^{13}C resonances of aliphatic residues were derived from CCC-TOCSY-NNH, HCC-TOCSY-NNH, and HCCH-TOCSY spectra. Aromatic ring resonances were assigned from NOESY data. On this basis, all backbone as well as 98/99% of $^{1}H/^{13}C$ side-chain resonances were assigned at UofT.

Structure Calculation (UofT)

Distance constraints for structure calculations were derived from crosspeak intensities in a simultaneous ^{15}N- and ^{13}C-NOESY-HSQC with a mixing time of 150 ms.[10,11] The structure calculation proceeded in three stages. In the first stage, an initial fold of the protein was generated using unambiguously assigned NOEs and dihedral angle constraints derived via the program *TALOS*.[12] The calculation was performed with the program *DYANA*[5] using a torsion angle dynamics protocol. The *NOAH* module within *DYANA* was used to aid in the assignment of the remaining NOE crosspeaks. Peak analysis of the NOESY spectra were generated by interactive peak picking with the program *XEASY*.[8] In the second stage, hydrogen bond constraints were added on the basis of the structures generated in the initial stage and were restricted to those residues that were clearly in secondary structure regions as judged by NOE pattern and chemical shifts. During this stage, the 20 calculated structures from *DYANA* with the lowest target functions were used to analyze constraint violations and to assign additional NOE constraints for the following round. Several rounds of structure calculations were performed until 90% peaks in the spectra had been assigned and all violations eliminated. In the final cycle, the NMR-derived experimental constraints contained 1595 unambiguous NOEs (685 intraresidue, 259 sequential, 211 medium-range ($2 \leq |i-j| \leq 5$) and 350 long-range ($|i-j| > 5$) interproton constraints, 35 distance constraints for 17 backbone hydrogen bonds and 103 dihedral angle constraints. One hundred structures were calculated, from which the 30 structures with the lowest target functions were selected.

In the final stage, 30 selected structures were used as starting structures for further refinement in explicit water by the program *CNS*.[13] The structures were soaked with an 8-Å layer of TIP3P water molecules.[14] The 20 lowest energy structures were retained for analysis. These structures had no NOE violations >0.5 Å or dihedral angle violations >5°. The N-terminal Gly-Ser-His is disordered in solution and is not included in structural models. *PROCHECK_NMR*[15] shows that 91.5% of the ϕ and ψ angles to be in the most favored regions and 8.5% in the additional allowed regions. No bad contacts were reported. Structures were visualized using the program *MOLMOL*.[16] The structure of mth1743 consists of one α-helix and five β-strands forming a mixed β-sheet. β-Strand 1 runs parallel to β-strand 5, whereas the other β-strands are antiparallel. The amphipathic α-helix is composed of residues Ile29–Leu36 and packed against the first two antiparallel β-strands (β1 and β2). The structure has been deposited at the PDB with code 1JSB.

Structure Calculation via ABACUS (CMU)

The UofT group provided the CMU group unassigned, untyped, single amino acid spin systems containing ^{1}H, ^{13}C, and ^{15}N chemical shifts of mth1743, together with the corresponding atom types. Fragments (single amino acids) of connected spin systems were not typed or ordered according to their position in the sequence. ^{13}C- and ^{15}N-edited 3D NOESY peaks, stripped of the established manual or NOAH-derived assignments, were supplied as listings of the ^{13}C (^{15}N) and ^{1}H chemical shifts, along with the peak intensities and their corresponding estimated inter proton distance in Å. The tabulated data included a clean list of all the NOESY connectivities unambiguously assigned by the UofT group. The NOESY crosspeaks were integrated using the module *peakint* within the program *XEASY*. Upper limit distances in Å units were generated via the *DYANA* macro *calibra*. The digital resolution in ppm units, necessary for the *BACUS* operation, was also provided for each NOESY spectral dimension. The protein's primary structure was also known. Figure 1 outlines the *ABACUS* protocol (described below).

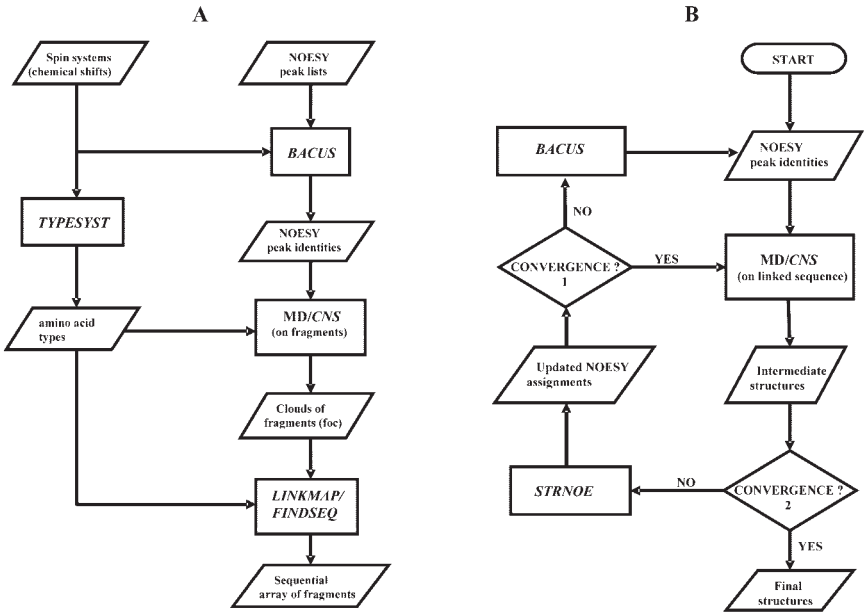

Fig. 1. *ABACUS* flow charts. (**A**) Spin system identifications and sequential assignments. (**B**) NOESY refinement and structure computation. Input to (**B**) are the NOESY identities obtained via (**A**). Convergence criteria (1) and (2) are, respectively, maximizing the number of distance restraints and minimizing the NOE pseudoenergy of the structure as obtained from *CNS*.

First Stage [Fig. 1(A)]

Residue typing

Fragment typing in terms of amino acid residues is obtained via *TYPESYST*. The program takes advantage of the fact that the experimental ^{1}H and ^{13}C chemical shift distributions such as those obtained from BMRB (Wisconsin BioMagResBank) show strong dependence on the amino acid type. For every unassigned spin system K, *TYPESYST* computes typing scores $\mathcal{P}^T(T|K)$ to attribute it to each of the 20 natural amino acid types (T). Assuming uncorrelated spin chemical shifts within the fragment, the typing probabilities are estimated from

$$\mathcal{P}^T(T|K) = \frac{\prod_{j=1}^{N_K} \frac{1}{\sigma_T^j} \exp\left[-\frac{1}{2}\left(\frac{\delta_K^j - \delta_T^j}{\sigma_T^j}\right)^2\right]}{\sum_{\{S\}}^{20} \prod_{j=1}^{N_K} \frac{1}{\sigma_S^j} \exp\left[-\frac{1}{2}\left(\frac{\delta_K^j - \delta_S^j}{\sigma_S^j}\right)^2\right]} \qquad (1)$$

where the product is over each of the N_K observed chemical shifts for the spin system K, taking into account atom type (C^j, H^i), and the sum in the denominator goes over $\{S\}$, the set of the 20 natural amino acids; δ_S^i and σ_S^j are, respectively, the average chemical shifts and their standard deviations for a type-S amino acid, as extracted from the database. *TYPESYST* removes from the list of scored types all those that are absent in the primary sequence and discards hypotheses with probabilities below the 10^{-3} threshold, until convergence. Another cycle of hypotheses pruning follows whenever the number of fragments uniquely typed as a given amino acid equals or exceeds the number of residues of that type in the sequence. In such cases, all nonunique hypothesis establishing correspondence of any fragment to those residue types are removed. After normalization, *TYPESYST* generates an output file that specifies, for each spin system, possible residue types and their associated probabilities.

NOESY crosspeak identities

NOEs are then identified, matching their coordinates to the spin systems' chemical shifts. For this purpose, the list of NOESY crosspeaks was run through the *BACUS* program,[1] adapted to incorporate $^{13}C/^{15}N$ editing. After *BACUS* convergence, uniquely identified crosspeaks were converted to NOE restraints in *CNS*/*XPLOR* format. Because these NOEs arise from protons detected via $^{13}C/^{15}N$ edited experiments, they effectively connect spatially related, but otherwise unassigned, fragments. Each of these are held together by covalent (bonds, angles, chirality) terms.

Generation of clouds of fragments

New topology definitions were created for molecular fragments of ambiguous residue typing (e.g., ASP, SER, $^{\alpha}CH—^{\beta}CH_2$, etc.) obtained via *TYPESYST*, with the fragment length composition matching the number of aliphatic carbon resonances present in the spin systems. These definitions were included within the amino acid topology file in *CNS*/*XPLOR*.

In *CLOUDS*-like fashion,[3] molecular dynamics (MD) simulations were run on the set of unconnected, typed, and untyped, molecular fragments under the NOE distance restraints. The temperature was lowered from 3000 K to 10 K in 30 steps of 5-ps dynamics each. Quartic, strictly

repulsive nonbonded interactions were used with atomic radii scaled to 0.7 of their van der Waals values. Starting from random initial coordinates, a family of 10 or more "structures" (in effect, spatial fragment distributions or "clouds" of fragments) was generated (Fig. 2).

Sequential linking

Two protocols were developed to connect the fragments. *LINKMAP* and *FINDSEQ*. *LINKMAP* uses a stochastic sampling approach to obtain a unique sequential mapping of the individual spin systems vis-à-vis the protein's primary sequence. *FINDSEQ* relies on *ad-hoc* probabilities of the sequential neighbor position.

In *LINKMAP*, the sequence-specific placement of the fragments is established via a Monte Carlo optimization protocol reminiscent of the *MONTE* approach.[17] Let K be a fragment and K_i its arbitrary assignment to residue i in the protein. A pseudo-energy score E[{K}] is attributed to the particular placement of the spin systems {K} in the protein. E[{K}] is partioned into one- and two-body terms:

$$E[\{\mathrm{K}\}] = \sum_{\mathrm{i}}^{\mathrm{N}} E_{\mathrm{l}}(T_{\mathrm{i}}|\mathrm{K_i}) + \sum_{\mathrm{i}}^{\mathrm{N}}\sum_{\mathrm{j}}^{\mathrm{N}} \delta_{\mathrm{j,i+1}} E_2(\mathrm{seq}|\mathrm{K_i},\mathrm{K_j}) \qquad (2)$$

where N is the total number of residues in the protein. The Kroeneker delta $\delta_{\mathrm{j,i+1}}$ selects spin system pairs placed sequentially. In tune with Boltzmann distribution statistics, we define the one-body energy as

$$E_{\mathrm{l}}(T_{\mathrm{i}}|\mathrm{K_i}) \propto -\ln \mathcal{P}^{\mathrm{T}}(T_{\mathrm{i}}|\mathrm{K_i}) \qquad (3)$$

Where the $\mathcal{P}^{\mathrm{T}}(T|\mathrm{K})$ are the *TYPESYST* chemical shift-based typing probabilities (Eq. 1) and T_{i} denotes the amino acid type of residue i. The two-body energy term, $E_2(\mathrm{seq}|\mathrm{K_i},\mathrm{K_j})$, denotes a measure of confidence for the sequential linkage between the two spin systems $\mathrm{K_i}$ and $\mathrm{K_j}$, namely a potentially upstream CO-HN peptidyl link. By analogy to Eq. 3, the sequential linkage probability $\mathcal{P}(d_{\mathrm{K_iK_j}}|\mathrm{seq})$ between two spin systems is related to a pseudo-energy:

$$E_2(seq|\mathrm{K_i},\mathrm{L_j}) \propto -\ln\left[\frac{\mathcal{P}(d_{\mathrm{K_iK_j}}|\mathrm{seq})}{\sum_{m}^{N}\mathcal{P}(d_{\mathrm{K_iK_m}|\mathrm{seq}})}\right] \qquad (4)$$

where the sum extends over all the spin systems $\mathrm{K_m}$ other than $\mathrm{K_i}$.

The $\mathcal{P}(d_{\mathrm{K_iK_m}}|\mathrm{seq})$ are calculated as an overlap of the distance $d_{\mathrm{K_iK_m}}$ probability distribution between the C' and N atoms of fragments $\mathrm{K_i}$ and $\mathrm{K_m}$ over the family of fragment "structures" and the Gaussian function centered at a N—C covalent distance of 1.3 Å, with a standard deviation of 6 Å.

As alternative sequential proximity scores $E_2(\mathrm{seq}|\mathrm{K_i},\mathrm{K_j})$, we have also implemented connectivity inputs extracted directly from the processed NOESY lists, rather than from the coordinates of the 3D clouds of fragments. In this case, as measures of proximity, we have counted the numbers of interresidue $\mathrm{H^N/H^\alpha}$ and $\mathrm{H^N/H^\beta}$ NOEs that could be attributed to the spin systems $\mathrm{K_i}$ and $\mathrm{K_j}$ within the *BACUS*-processed cross-peak identity tables. The $\mathrm{H^N}$ chemical shift was that of the spin system $\mathrm{K_j}$; the resulting scores were taken as likelihoods of the upstream sequential connectivity from the spin system $\mathrm{K_i}$ to the spin system $\mathrm{K_j}$.

The routine starts with the spin systems {K} randomly assigned to residues in the protein. Fragments can be placed only at positions where the amino acid type is in agreement with the possible spin systems' typing, i.e. $\mathcal{P}^T(T_{\mathrm{i}}|\mathrm{K_i}) \neq 0$. All connectivity scores between the spin systems for which a linkage was plausible (given their type and the protein sequence), were initialized from a small non-zero prior value, rather than from zero. This effectively prevented missing a sequential connectivity between the two spin systems as, in general, the experimental data can be incomplete.

The pseudo-energy (Eq. 2) was minimized by swapping sequence positions for pairs of randomly chosen spin systems whose tentative typings overlap. Metropolis acceptance criterion was enforced: if the pseudo-energy of the system decreased as a result of such swap, it was accepted; otherwise, the move was accepted with a probability equal to the Boltzmann factor. In order to prevent trapping in local minima, a simulated annealing scheme was followed involving a cooling schedule linear in β (= $1/kT$), with β increasing from 0.1 to 10 in 100 stages, involving 10^5 attempted swaps each.

FINDSEQ first calculates the likelihoods $\mathcal{P}(+|\mathrm{K,L})$ and $\mathcal{P}(-|\mathrm{K,L})$ that a fragment L be proximal—not necessarily sequential—upstream or downstream, respectively, of fragment K. $\mathcal{P}(+|\mathrm{K,L})$ is evaluated from the averaged distance d_{NN} separating the $\mathrm{H^N}$ atoms in fragments K and L within the family of clouds (foc)[3,4] of fragments. A Gaussian distribution centered at 3 Å, with σ = 1 Å was assumed. If the deviation $d_{\mathrm{NN}} < 3$ Å, $\mathcal{P}(+|\mathrm{K,L}) = 1$ was assumed. Similarly, $\mathcal{P}(-|\mathrm{K,L})$ is obtained by adding the individual proximal probabilities, assumed Gaussians, each calculated using the average distances between the $\mathrm{H^N}$ atom of K and $\mathrm{H^N}$, $\mathrm{H^\alpha}$ and $\mathrm{H^\beta}$ atoms of L.

Initial values are given to $\mathcal{P}(\mathrm{i}|\mathrm{K})$, the site-specific assignment probabilities that fragment K corresponds to residue at location i; $\mathcal{P}(\mathrm{i}|\mathrm{K}) = (n_{\mathrm{K}})^{-1}$ if $\mathcal{P}^T(T_{\mathrm{i}}|\mathrm{K}) > 0$, or $\mathcal{P}(\mathrm{i}|\mathrm{K}) = 0$ if $\mathcal{P}^T(T_{\mathrm{i}}|\mathrm{K}) = 0$. T_{i} indicates the amino acid type that, according to the sequence, is at location i, and n_k is the number of possible assignments for fragment K.

The initial assignment probabilities were modified according to the criterion that a particular $\mathcal{P}(\mathrm{i}|\mathrm{K})$ should increase whenever another fragment M fulfills the following two conditions: (a) M is spatially proximal to residue K, i.e. $\mathcal{P}(+|\mathrm{K,L})$ and/or $\mathcal{P}(-|\mathrm{K,L}) > 0$, and, respectively, (b) M has non-zero probability of assignment to residue i + 1 and/or i − 1, i.e. $\mathcal{P}(\mathrm{i+1}|\mathrm{M})$ and/or $\mathcal{P}(\mathrm{i-1}|\mathrm{M}) > 0$.

The $\mathcal{P}(\mathrm{i}|\mathrm{K})$ probabilities are updated via

$$\mathcal{P}(\mathrm{i}|\mathrm{K}) = \mathcal{P}(\mathrm{i}|\mathrm{K}) \times \mathcal{P}^L(\mathrm{i}|\mathrm{K}) \qquad (5)$$

where $\mathcal{P}^{\mathrm{L}}(\mathrm{i}|\mathrm{K})$ is the multiplier

Fig. 2. Stereoviews of clouds of fragments of the mth1743 domain computed via *BACUS/CNS* [Fig. 1(A)]. Fragments are labeled according to amino acid type; H-atoms and side chains are not shown. Distance restraints are estimated from the experimental NOEs. (**A**) After a first round of calculations based on tentative NOE identities. (**B**) After convergence [Fig. 1(B)], based on ABACUS-assigned NOEs.

$$\mathcal{P}^{L}(\mathrm{i}|\mathrm{K}) = \sum_{\{\mathrm{M}\}} \mathcal{P}(\mathrm{i}-1|\mathrm{M}) \times \mathcal{P}(-|K,L) + \sum_{\{\mathrm{M}\}} \mathcal{P}(\mathrm{i}+1|\mathrm{M}) \times \mathcal{P}(+|\mathrm{K},\mathrm{L}) \quad (6)$$

The sums in (6) are over {M}, all the fragments in the cloud excluding K. The $\mathcal{P}(\mathrm{i}|\mathrm{K})$ filters are then normalized,

$$\mathcal{P}(\mathrm{i}|\mathrm{K}) \rightarrow \mathcal{P}(\mathrm{i}|\mathrm{K}) / \sum_{\mathrm{j}=1}^{\mathrm{N}} \mathcal{P}(\mathrm{j}|\mathrm{K}) \quad (7)$$

where N is the number of residues in the protein. The calculations via Equations (5)–(7) are iteratively applied until the information entropy[18]

$$S = -\sum_{j,\{K\}} \mathcal{P}(j|K) \times \ln \mathcal{P}(j|K) \quad (8)$$

approaches a minimum. {*K*} is the set of fragments to be assigned, in general not equal to the number *N* of residues in the protein. Following each iteration, probabilities $\mathcal{P}(i|K) < 10^{-5}$ are set to zero and, after convergence of *S*, probabilities $\mathcal{P}(\mathrm{i}|\mathrm{K}) > 0.90$ are set to 1.

Second Stage [Fig. 1(B)]
NOE sorting and structure calculation

After data analysis via *LINKMAP* and/or *FINDSEQ*, fragments become sequence-specifically assigned. In the next step, the complete covalently linked primary structure of the protein was input to a distances-restrained MD to obtain the protein fold; no less than 20 structures were generated. Structural restraints were analyzed with *STRNOE*. Three *STRNOE*/*BACUS*/*CNS* cycles of NOE identification were performed. During these cycles, NOE identities resulting in restraints violations >1 Å were removed and followed by *BACUS* search of new identities. *STRNOE*/*CNS* was run repeatedly until no further NOE assignments were obtained via BACUS. In our study, a total of 11 cycles was run with manual assignment selection until ~95% of the NOEs were uniquely identified with residual distance violations <0.5 Å. At the final stages of calculation, backbone dihedral angle constraints derived via *TALOS*[12] and distance restraints consistent with the suggested backbone–backbone hydrogen bonds, were incorporated.

RESULTS AND DISCUSSION
Sequential Assignments via ABACUS

It is reassuring that, with data that included spin systems extending beyond C^{β}s and ^{13}C chemical shifts, *TYPESYST* was able to establish unique typing for 71% of the spin systems. The remaining 29% had an average ambiguity of 2.6 types. There were no errors for those unambiguously typed and the ambiguous cases all included the correct types. The *LINKMAP* program filtered out ~57% of all possible links as incompatible with the estimated typing, resulting in ~3.7 links per residue. The procedure yielded spin systems' sequence-specific placements that were subsequently determined to be 100% correct. Overall, it required a few minutes on a desktop PC. Figure 2 compares fragment distributions based on the final *ABACUS* structure [Fig. 2(B)] against the initial cloud [Fig. 2(A)]: a significant gain in structural organization is apparent as a result of establishing the sequence and the BACUS/CNS structure calculation.

TABLE I. Comparison of 1JSB to the Structures Calculated via *ABACUS*:[a] Backbone RMSDs

			ABACUS[c]	
		1JSB[b]	A	B
1JSB		0.23Å	1.32Å	1.22Å
ABACUS	A		0.28Å	0.35Å
	B			0.20Å

[a]Off-diagonal entries stand for the average backbone RMSD between two given set of structures; the diagonal entries stand for the average backbone RMSD to the mean within a given family of structures. The RMSD values are for segment comprising residues 5–67 (7–70 in 1JSB residue numbering).
[b]NMR structure 1JSB was calculated with NOE, dihedral angles, and H-bond restraints.
[c]NMR *ABACUS* structures were calculated with NOE and H-bond restraints (column A), and with NOE, dihedral angles and H-bond restraints (column B).

TABLE II. Comparison of 1JSB to *ABACUS* Structures, Both Calculated with Distance and Dihedral Angle Restraints: Violations of the Empirical Force Field Terms and the Experimental Restraints

	1JSB	*ABACUS*
Bonds	$(1.218 \pm 0.001) \cdot 10^{-2}$ Å	$(2.33 \pm 0.05) \cdot 10^{-3}$ Å
Angles	1.0557 ± 0.0005°	0.375 ± 0.006°
Impropers	0.3696 ± 0.0006°	0.272 ± 0.008°
Distance restraints	0.125 ± 0.005°	0.159 ± 0.001°
Dihedral angle restraints	1.70 ± 0.09 Å	0.613 ± 0.007 Å

As mentioned above, *LINKMAP* was also tested on the list of the assignment hypotheses produced by BACUS skipping the MD calculation on the unconnected molecular fragments. When using such raw output, with ~2.1 possible assignments/peak, the program converges to a sequential assignment that is completely correct. As a test, we have also input highly ambiguous NOESY assignment data that had not been processed by BACUS. Peak identification based on the chemical shift matches alone results in ~15.5 possible assignments per crosspeak. Using this data in *LINKMAP* yielded five sequential misassignments (residues 15, 21, 31, 58, 70), underscoring the positive influence of the *BACUS* step in the analysis.

As an alternative to *LINKMAP*, *FINDSEQ* can assist the sequential identification at an earlier stage of the analysis, when the spin system grouping is still somewhat incomplete. *FINDSEQ* found positions for 65 out of the 70 residues of the protein, with only one misplacement. If unambiguous typing for all the spin systems is assumed, then *FINDSEQ* identifies the correct positions of all the spin systems (70/70).

The fact that *LINKMAP* performed better than *FINDSEQ* can be explained as a result of the constraint within

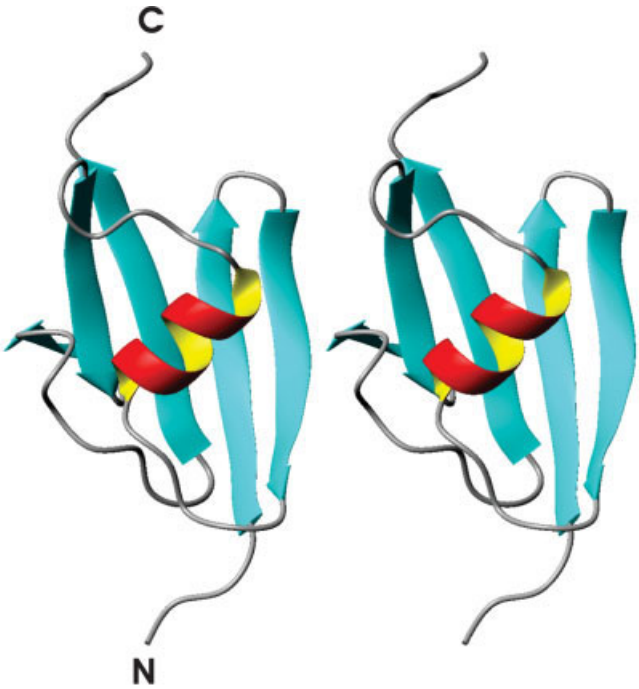

Fig. 3. Stereoview of the mth1743 domain structure obtained via *ABACUS*. NOEs and dihedral angles were used as experimental restraints. C- and N-termini are indicated.

LINKMAP that all the spin systems belong to the protein. However, this constraint can destabilize its performance if the requirement is not met. For such cases, *FINDSEQ* ought to be useful. As a test, 10 amino acids were eliminated from the sequence either at the N-terminus or at the C-terminus, yielding proteins of 60 amino acids each, while 70 residues were input to *FINDSEQ*. When residues were eliminated from the C-terminus, positions for 53 spin systems were found with three misplacements. For the case of residues missing at the N-terminus, *FINDSEQ* found positions for 55 spin systems with one misplacement. Considering the stringency of the test, the results indicate robust performance for the program.

NOESY Assignments and Structure Calculation

The NOESY peak list generated by the UofT group contained 2792 NOEs (765 in ^{15}N-edited and 2027 in the ^{13}C-edited NOESYs). From the *ABACUS* input a total of 350 peaks—assumed to be diagonal as identified whenever three or fewer digital points separated the two ^{1}H frequencies—were removed, resulting in 2557 peaks for further identification. The subsequent analysis showed that among the 350 discarded peaks, 15 had been assigned by the UofT team, to be off-diagonal.

For 2225 NOESY peaks, the assignments obtained at UofT via standard (i.e., assignments-based) protocols and at CMU via *ABACUS* were identical. One hundred thirty-nine peaks assigned uniquely at UofT yielded multiple *ABACUS* assignments, one of which corresponds to the above-mentioned unique assignment. Ignoring the latter, the number of peaks whose assignments differed becomes

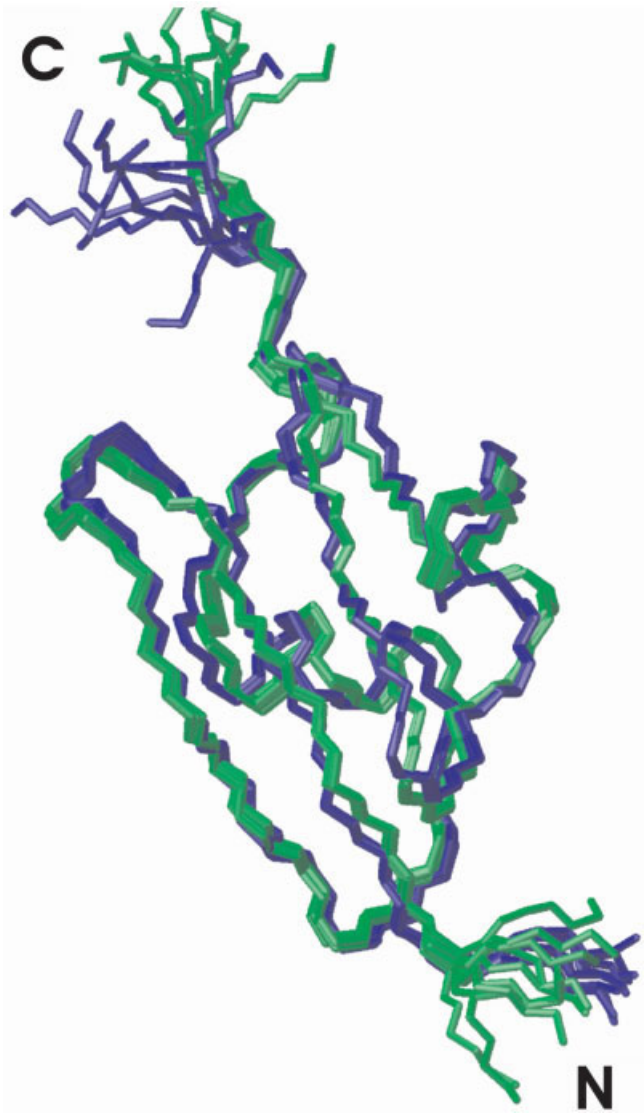

Fig. 4. Comparison of NMR structural bundles for mth1743: 1JSB (green) and *ABACUS*-based (blue). Polypeptide NMR backbone structures of residues 1–70 are shown; 10 each. C- and N-termini are indicated. The view shown is rotated along the y-axis by 180° relative to that depicted in Figure 3.

95 (3.8% of the total). Out of these, eight had no possible assignments via *ABACUS*, three had multiple assignments, all of which differed from the one obtained at the UofT, and 83 had unique assignments different from those obtained at UofT. For 13 of these, the differing atom was on the same residue as in the UofT assignments. From this analysis, it is suggested that the number of significantly different assignments obtained at UofT and at CMU via *ABACUS* is unlikely to encompass more than ~3% of the total number of NOESY crosspeaks.

Structural statistics on 1JSB and the structures calculated via *ABACUS* are summarized in Tables I and II. Overall, the *ABACUS*-derived protein fold is essentially the same as that for 1JSB. Although MTH1743 shares less than 20% sequence homology to This, MoaD, and ubiq-

uitin, it exhibits a ubiquitin fold. RMSD values were calculated for backbone (N, C^{α}, C′) atoms of residues 3–67. In view of the type and amount of data used, the extent of agreement between the sets of structures calculated via the two approaches is reassuring. As is apparent from inspection of Table I, the difference between the bundles of the UofT and CMU-calculated structures (~1.2 Å) significantly exceeds the RMSD within each of the structural bundles (~0.2–0.3 Å). This mismatch exemplifies a common situation in the NMR structure refinement with the precision of the structural bundle often exceeding its accuracy. The structures of mth1743 obtained via *ABACUS* are shown in Figures 3 and 4.

CONCLUSIONS

The mth1743 analyses performed independently by two groups of researchers at UofT and CMU, following significantly different computational/analytical protocols, but starting from the same experimental NMR data, converged to structures and NOE assignments that are highly similar: 97% of NOESY assignment similarity and ~1.2 Å backbone RMSD (Fig. 4). These results effectively validate a structure calculation approach based on the *ABACUS* procedure, a direct extension of the *CLOUDS* protocol that places emphasis on establishing NOESY identities prior to the full sequential resonance assignment.

Most importantly, *ABACUS* exhibits fast convergence from the unassigned input data to the final structures. *ABACUS* also affords a basis for an unbiased (no prior structure or fold is assumed) iterative procedure to refine the subset of ambiguous NOEs. Finally, the *ABACUS* code can be adapted to the chosen experimental protocols since the paradigm behind the approach, namely its emphasis on the NOE constraints, is of general applicability, not being restricted to the type of $^{13}C/^{15}N$-editted data set exemplified by the test case investigated in this study. As presented here, *ABACUS* suggests itself as a robust "direct" approach for high-throughput bio-molecular structure determination via NMR.

Accession Numbers

The UofT chemical shifts have been submitted to the BMRB (accession #5106), and the structure ensemble and NOE constraint file has been submitted to the PDB (accession #1JSB).

ACKNOWLEDGMENTS

The authors thank Dr. A. Lemak for comments on the manuscript.

REFERENCES

1. Grishaev A, Llinás M. BACUS, a Bayesian protocol for the identification of protein NOESY spectra via unassigned spin systems. J Biomol NMR 2004;28:1–10.
2. Grishaev A, Llinás M. Sorting signals from protein NMR spectra: SPI, a Bayesian protocol for uncovering spin systems. J Biomol NMR 2002;24:203–213.
3. Grishaev A, Llinás M. CLOUDS, a protocol for deriving a molecular proton density via NMR. Proc Natl Acad Sci USA 2002;99:6707–6712.
4. Grishaev A, Llinás M. Protein structure elucidation from NMR proton density. Proc Natl Acad Sci USA 2002;99:6713–6718.
5. Güntert P, Mumenthaler C, Wüthrich KJ. Torsion angle dynamics for NMR structure calculation with the new program DYANA. J Mol Biol 1997;273:283–298.
6. Yee A, Chang X, Pineda-Lucena A, Wu B, Semesi A, Le B, et al. An NMR approach to structural proteomics. Proc Natl Acad Sci USA 2002;99:1825–1830.
7. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol NMR 1995;6:277–293.
8. Bartels C, Xia T, Billeter M, Günter P, Wüthrich, K. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. J Biomol NMR 1995;6:1–10.
9. Galser R, Wüthrich K. SPSCAN—software for the automatic evaluation of high-resolution NMR spectra of biological macromolecules. http://www.molebio.uni-jena.de/~rwg/spscan, 1997.
10. Pascal SM, Muhandiram DR, Yamazaki T, Forman-Kay JD, Kay LE. Simultaneous acquisition of ^{15}N-edited and ^{13}C-edited NOE spectra of proteins dissolved in H_2O. J Magn Reson B 1994;103:197–201.
11. Xia Y, Yee A, Arrowsmith CH, Gao X. ^{1}HC and ^{1}HN total NOE correlations in a single 3D NMR experiment. ^{15}N and ^{13}C time-sharing in t_1 and t_2 dimensions for simultaneous data acquisition. J Biomol NMR 2003;27:193–203.
12. Cornilescu G, Delaglio F, Bax A. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J Biomol NMR 1999;13:289–302.
13. Brünger AT, Adams PD, Clore GM, Delano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, et al. Crystallography and NMR system (CNS): a new software suite for macromolecular structure determination. Acta Crystallogr D 1998;54:905–921.
14. Linge JP, Williams MA, Spronk AEM, Bonvin AMJJ, Nilges M. Refinement of protein structures in explicit solvent. Proteins 2003;50:496–506.
15. Laskowski RA, Rullmann JA, MacArthur MW, Kaptein R, Thornton JM. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 1996;8:477–496.
16. Koradi R, Billeter M, Wüthrich K. MOLMOL: a program for display and analysis of macromolecular structures. J Mol Graphics 1996;14:51–55.
17. Hitchens TK, Lukin JA, Zhan Y, McCallum SA, Rule GS. MONTE: an automated Monte Carlo based approach to nuclear magnetic resonance assignment of proteins. J Biomol NMR 2003;25:1–9.
18. Shannon CE. A mathematical theory of communication. Bell Syst Tech J 1948;27:379–423.