Journal of Biomolecular NMR **28:** 1–10, 2004.
KLUWER/ESCOM

BACUS: A Bayesian protocol for the identification of protein NOESY spectra via unassigned spin systems*

Alexander Grishaev** & Miguel Llinás***
Department of Chemistry, Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.

Received 2 April 2003; Accepted 1 July 2003

Key words: analysis of NMR protein spectra, CLOUDS, NOESY signal identification, SPI

Abstract

NMR frequency assignments are usually considered a prerequisite for the analysis of NOESY spectra, in turn required for the calculation of biomolecular structures. In contrast, as we propose here, relatively high numbers of unambiguous NOE *identities* can be consistently achieved in an automated manner by relying only on *grouping* resonances into connected spin systems. To achieve this goal, we have developed for proteins two protocols, SPI and BACUS, based on Bayesian inference. SPI (Grishaev and Llinás, 2002c) produces a list of the ^{1}H resonance frequencies from homo- and hetero-nuclear multidimensional spectra, grouped into effective spin systems. BACUS automatically establishes probabilistic identities of NOESY cross-peaks in terms of the chemical shifts provided by SPI. BACUS requires neither assignment of resonances nor an initial structural model. It successfully copes with chemical shift overlap and does so without cycling through 3D structure calculations. The method exploits the self-consistency of the NOESY graph by taking advantage of a network of J- as well as NOE-connected 'reporter' protons sorted via SPI. BACUS was validated by tests on experimental NOESY data recorded for the col 2 and kringle 2 domains.

Abbreviations: BAF – BAckbone Finder; BACUS – Bayesian Analysis of Coupled Unassigned Spins; CLOUDS – Computed Location Of UnassigneD Spins; SIF – SIdechain Finder; SPI – SPin Identification.

Introduction

The analysis of protein NMR NOESY experiments starts by building lists of NOE-connected spin pairs that fall within suitable tolerances from each peak's chemical shift coordinates. The standard approach to deal with the ambiguity problem is to iteratively eliminate those NOE assignments suspected to be statistically incompatible with the latest, distances-based, estimated structures. Such strategy hinges on the hypothesis that correct matches, being self-consistent in the context of the underlying structure, should support each other when used as geometrical restraints; in contrast, the incorrect assignments are essentially random, hence likely to lead to inter-proton distances inconsistent with the structure.

Current methods of automated NOESY analysis are designed to filter out the ambiguous restraints via suitably designed cost functions. This is achieved by deriving probabilities of NOESY assignments either via a molecular dynamics protocol, such as in the ARIA method (Nilges, 1993, 1995; Nilges et al., 1997; Nilges and O'Donoghue, 1998; Linge et al., 2003), or via a self-correcting distance geometry method NOAH/DIAMOD/ATNOS (Güntert et al., 1993; Hänggi and Braun, 1994; Mumenthaler and Braun, 1995; Mumenthaler et al., 1997; Herrmann et al., 2002b). Both types of procedures rely on the assignment of resonance frequencies, require in-

*Extracted from the doctoral dissertation of A.G. (Carnegie Mellon University, Pittsburgh, PA, 2001). The BACUS program is available from the corresponding author upon request.

**Present address: Laboratory of Chemical Physics, NIDDK, National Institutes of Health, Building 5/B1-27, 9000 Rockville Pike, Bethesda, MD 20892-0520, U.S.A.

***To whom correspondence should be addressed. E-mail: llinas@andrew.cmu.edu

tensive 3D structure computations and are variously influenced by the initial structural models as well as by the degree of ambiguity of the input NOEs. Characteristically, with $\sim$2–10% initial unambiguous NOEs, the final unique assignments rate obtained is $\sim$60–80%, with the error rate of $\sim$3–5%, and additional 10% NOEs with multiple assignments (see, e.g., Mumenthaler and Braun, 1995).

Here, we address the following question: Assuming that one has access to a list of proton resonances specified by their chemical shifts δ_ℓ, where $\ell = 1, 2, \ldots$. N, N being the total number of hydrogen atoms or groups of magnetically equivalent hydrogen atoms, *what is the probability* $\mathcal{P}(i, j|O)$ *that an experimental NOESY cross-peak O, found at chemical shift coordinates* $\delta_p \equiv (\delta_{1,p}, \delta_{2,p})$*, can be identified as arising from protons with chemical shifts* (δ_i, δ_j)*?* In the context of this paper, by identity of a NOESY cross-peak O is meant the probabilistically significant – i.e. *unambiguous* – recognition that the NOE connects protons whose resonances occur at chemical shifts δ_i and δ_j, without prior knowledge of the specific assignment of hydrogen atoms H_i and H_j from which they originate. It is proposed that the presence of particular sets of cross-peaks can be taken advantage of in order to improve the probabilities of matches for other NOEs. Thus, our aim is to extract the correct identities by exploiting the topological self-consistency of the NOESY graph. It should be apparent that attainability of such a goal implies the possibility of refining chemical shifts-based matching probabilities without need to cycle through 3D structure calculations.

In our approach we exploit the idea that the local environment is restricted, as it is mostly constrained by covalent bonds encoded in the J-connectivities shown by COSY/TOCSY-type spectra. Thus, if two protons are proximal to each other, the rest of their J- or NOE-coupled spin systems also are likely to occur within the local neighborhood, which can be used to modify the cross-peak matching probabilities. Bayes' theorem (Kendall, 1987) provides a suitable mathematical framework for the combination of probabilities stemming from independent data sources. Based on these ideas, we have developed SPI (Grishaev and Llinás, 2002c), a procedure for computing, from homo- and hetero-nuclear multidimentional NMR experiments, the likelihood of nuclear spin resonances appearing at defined frequencies. In this paper we report on BACUS, a related heuristic protocol that generalizes the SPI approach by extending it to the sorting and identification of NOEs. As a validation, BACUS was tested on experimental NMR data for the col 2 and kringle 2 domains of matrix metalloproteinase II and plasminogen, respectively (Briknarová et al., 1999; Marti et al., 1999).

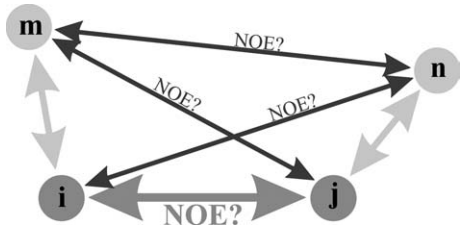

Figure 1. The concept of BACUS: For resonances m *unambiguously* connected to i, and n unambiguously connected to j, the probability to observe the NOE between resonances i and j is estimated from the observations of the NOEs between resonances (i,m), (j,n), (m,j), (n,i), and/or (m,n).

Methods

Data processing

2D ^{1}H/^{1}H COSY, 70 ms TOCSY, 200 ms NOESY, ^{1}H/^{15}N HSQC and 3D ^{15}N-edited HSQC-NOESY, HSQC-TOCSY, HNHA and HNHB spectra of col 2 and kringle 2 were acquired and processed as previously reported (Briknarová et al., 1999; Marti et al., 1999). Spectral data were analyzed via SPI, as reported (Grishaev and Llinás, 2002c). The SPI-identified spin systems were converted into files containing the chemical shifts as well as the experimentally detected connectivity types. An example of such files is shown under Appendix A.

BACUS protocol

Whereas SPI serves to determine the grid of frequencies where NOESY crosspeaks O_{ij} can occur, BACUS establishes unambiguous identities for NOESY crosspeaks based on the spin system grouping. The protocol follows the rationale (Figure 1): If protons i and m (and/or j and n) are known to be proximal, the observation of an NOE attributable to protons j and m (and/or i and n, and/or n and m) should increase the likelihood of a NOE between i and j. In practice, 'reporter' protons m and n – crucial within the BACUS scheme – are those COSY-, TOCSY-, or NOESY-connected to i and/or j.

Consider a NOESY cross-peak p *tentatively* assigned to a pair of protons (i,j), as well as to other pairs. Let us also consider protons m and n, where

Table 1. The estimated values of probabilities $\mathcal{P}(O_{mn}|O_{ij})$ for each of the expanded sets W(m) ⊗ W(n), where the sets W = Δ, Ω_C, and Ω_T are defined in the text. The $\mathcal{P}(O_{mn}|O_{ij})$ are derived via Equations 5–9.

		n		
		Δ(j)	Ω_C(j)	Ω_T(j)
m	Δ(i)	1.0000	0.4130	0.2778
	Ω_C(i)	0.4130	0.3224	0.2440
	Ω_T(i)	0.2778	0.2440	0.2015

proton m (or n) either is identical to proton i (j), or is connected to proton i (j) via COSY or TOCSY. The pair (n,m) is sorted into one of two classes: n and m on the same residue, and n and m on different residues. The latter class is denoted by $\overline{CT}$. The intra-residue pairs are split into COSY-connected, denoted as C, and the remainder, denoted as T. It should be apparent that the C and T classes should not be interpreted to imply COSY and TOCSY connectivities only: the nature and number of connectivities can be generalized to include any type of *unambiguous* correlation available from homo- or hetero-nuclear experiments, which could include NOEs as well. We also define a class Δ, encompassing Δ(i) and Δ(j), that include solely protons i and j, respectively.

In the Bayesian jargon, the expression $\mathcal{P}(W|\mathbf{Y})$ symbolizes the probability of event W being true given event(s) **Y**, or, in other words, conditional on **Y** being true. For a pair of resonances i and j, of potential NOESY connectivity O_{ij}, and an observed NOESY cross-peak p with chemical shift coordinates $\boldsymbol{\delta}_p \equiv (\delta_{1,p}, \delta_{2,p})$ the conditional probabilities of O_{ij} matching the $\boldsymbol{\delta}_p$ chemical shifts, $\mathcal{P}(O_{i,j}|\boldsymbol{\delta}_p)$, are estimated as:

$$\mathcal{P}(O_{ij}|\boldsymbol{\delta}_p) = \frac{\mathcal{G}(\delta_{1,p}, \delta_i; \sigma) \cdot \mathcal{G}(\delta_{2,p}, \delta_j; \sigma)}{\sum_{k,\ell} \mathcal{G}(\delta_{1,p}, \delta_k; \sigma) \cdot \mathcal{G}(\delta_{2,p}, \delta_\ell; \sigma)}, \qquad (1)$$

where $\mathcal{G}(x, y; z) \equiv \exp(-0.5(x-y)^2/z^2)$ and σ is the uncertainty of the cross-peak position in each dimension. (Henceforth, whenever $\sum_{k,\ell}$ is indicated, it is defined over all possible resonances with chemical shift coordinates $\boldsymbol{\delta}_k$ and $\boldsymbol{\delta}_\ell$ that peak p can be attributed to.)

Let $\mathcal{P}(O_R|O_{ij})$ stand for the probabilities of observing *at least* one NOESY 'reporter' (O_R) cross-peak q, of chemical shift coordinates $\boldsymbol{\delta}_q \equiv (\delta_{1,q}, \delta_{2,q})$, arising from protons m and n, (reporting on i and j, respectively) conditional on (i,j) matching of NOESY cross-peak p, where p ≠ q. The $\mathcal{P}(O_R|O_{ij})$ should be distinguished from $\mathcal{P}(O_{mn}|O_{ij})$, the probabilities of observing a NOE cross-peak between a specific proton pair m and n given that a NOE cross-peak between protons i and j was observed (listed in Table 1). On this basis, the $\mathcal{P}(O_R|O_{ij})$ likelihoods can be formulated as:

$$\mathcal{P}(O_R|O_{ij}) \propto \frac{1}{NM} \sum_{m=1}^{M} \sum_{n=1}^{N} \mathcal{P}(O_{mn}|O_{ij}) \cdot \mathcal{P}(O_{mn}|\delta_q), \qquad (2)$$

where δ_q does not appear within the left side term since it is implied in O_R, by the same token that δ_p is implied in O_{ij}. $\mathcal{P}(O_{mn}|\delta_q)$ are non-zero priors for matches of the cross-peak q to resonances m and n and are computed as formulated by Equation 1 for $\mathcal{P}(O_{ij}|\delta_p)$. M and N are the numbers of reporter protons m and n for i and j, respectively.

Our goal is to calculate the probabilities $\mathcal{P}(O_{ij}|\delta_p, O_R)$. Bayes' theorem enables us to combine the prior probabilities of NOESY chemical shifts matches $\mathcal{P}(O_{ij}|\delta_p)$ with the likelihoods $\mathcal{P}(O_R|O_{ij})$, in order to produce the final, posterior probabilities $\mathcal{P}(O_{ij}|\delta_p, O_R)$, conditional on both chemical shifts and connectivities:

$$\mathcal{P}(O_{ij}|\delta_p, O_R) = \frac{\mathcal{P}(O_{ij}|\delta_p) \cdot \mathcal{P}(O_R|O_{ij})}{\sum_{k,\ell} \mathcal{P}(O_{k\ell}|\delta_p) \cdot \mathcal{P}(O_R|O_{k\ell})}. \qquad (3)$$

For protons i,j not belonging to the same residue, it should be apparent that the O_{mn} connectivities that influence $\mathcal{P}(O_R|O_{ij})$ (Equation 2) are restricted to the $\overline{CT}$ class only, while O_{im}, O_{jn} connectivities can arise from C, T, and Δ classes. In contrast, for protons i and j that belong to the same spin system, the likelihoods are specified by the probabilities of observing NOESY cross-peaks (i,j), conditioned by the observed C/T connectivities between i and j, i.e., by $\mathcal{P}(O_{ij}|X_{ij})$ where X_{ij} denotes either C_{ij} or T_{ij}. Because the (i,j) pair reports (through X = C or T) on itself, it follows that for (i,j) belonging to the same residue $\mathcal{P}(O_R|O_{ij}) \equiv \mathcal{P}(O_{ij}|X_{ij})$ which, in turn, we calculate from

$$\mathcal{P}(O_{ij}|X_{ij}) = \iint dV_{ij} dr_{ij} \mathcal{P}(O_{ij}|V_{ij}) \cdot \mathcal{P}(r_{ij}|X_{ij}) \qquad (4)$$

Here, r_{ij} is the i–j inter-proton distance and V_{ij} the volume associated to the expected NOESY cross-peak. The estimated probability densities $\mathcal{P}(V_{ij}|r_{ij})$, $\mathcal{P}(O_{ij}|V_{ij})$, and $\mathcal{P}(r_{ij}|X_{ij})$ are shown in Figures 2, 3 and 5A,B, respectively. From numerical integration of Equation 4, $\mathcal{P}(O_{ij}|C_{ij}) = 0.9998$ and $\mathcal{P}(O_{ij}|T_{ij}) = 0.6907$ were obtained.

When computing $\mathcal{P}(O_{ij}|\delta_p, O_R)$ via Equation 3, before combining with the likelihoods, $\mathcal{P}(O_{ij}|\delta_p) < 0.05$ were set to zero and the remaining priors were re-normalized. Similarly, the priors that led to posteriors < 0.05 were discarded and the remaining $\mathcal{P}(O_{ij}|\delta_p)$ re-normalized. In iterative fashion, a new set of posterior probabilities $\mathcal{P}(O_{ij}|\delta_p, O_R)$ was then calculated on the updated priors and the procedure was repeated until all posteriors were > 0.05, our convergence criterion.

Likelihoods of spectral connectivities

The crucial $\mathcal{P}(O_{mn}|O_{ij})$ distribution is derived from the statistical distance probability distributions $\mathcal{P}(r_{mn}|O_{ij})$, where r_{mn} is the distance between protons m and n, and protons i and j are spatially close enough as to generate the observed (i,j) NOESY cross-peak. Also required are $\mathcal{P}(V_{mn}|r_{mn})$, the probability distribution of cross-peak volumes V_{mn} conditional on inter-spin distances r_{mn}, and $\mathcal{P}(O_{mn}|V_{mn})$, the probability to observe a NOESY cross-peak of intensity (i.e., volume) V_{mn}. The dependence of V_{mn} on r_{mn} reflects the local intramolecular neighborhood of the pair (m,n) to the extent that the spin environment affects its dipolar relaxation in the presence of spin diffusion. $\mathcal{P}(O_{mn}|V_{mn})$ encodes for the sensitivity of cross-peak detection, as affected by spectral dimensionality, experimental noise, and line widths. From the chain rule for probability propagation, we write:

$$\mathcal{P}(O_{mn}|O_{ij}) = \iint dV_{mn} dr_{mn} \mathcal{P}(O_{mn}|V_{mn}) \cdot \mathcal{P}(V_{mn}|r_{mn}) \cdot \mathcal{P}(r_{mn}|O_{ij}). \quad (5)$$

The $\mathcal{P}(V_{mn}|r_{mn})$ are estimated on the basis of a set of distances extracted from selected protein structures in the PDB database and relaxation matrix back-calculation of NOESY cross-peak volumes at the experimental mixing times (Figure 2).

Defining σ_{noise} as the measured RMS spectral noise, the $\mathcal{P}(O_{mn}|V_{mn})$ is taken as the probability of the cross-peak maximum being $> 3\times\sigma_{noise}$. Assuming Gaussian noise, upon integration we derive:

$$\mathcal{P}(O_{mn}|V_{mn}) = 0.5 + 0.5\,\mathrm{erf}\left(\frac{\frac{V_{mn}}{2\pi\,\sigma_m\sigma_n} - 3\sigma_{noise}}{\sqrt{2}\sigma_{noise}}\right), \quad (6)$$

where erf is the error function. Here we approximate the cross-peak shape by 2D Gaussians such that σ_m and σ_n are the linewidths for resonances m and n. The $\mathcal{P}(O_{mn}|V_{mn})$ curve is shown in Figure 3.

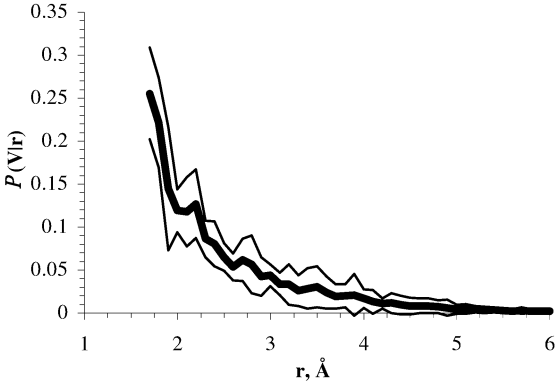

Figure 2. Mean normalized NOESY cross-peak volume versus interproton distance. Distances were obtained from the reported structures of ubiquitin (Cornilescu et al., 1998) and BPTI (Berndt et al., 1992) with NOEs backcalculated using MIDGE (Madrid et al., 1991; Grishaev and Llinás, 2002a) assuming $\tau_{mix} = 0.2$ s and $\tau_c = 3.0$ ns. Thin lines show the standard deviation about the mean (thick line), used to calculate the probability distribution $\mathcal{P}(V|r)$, assumed to be normal.

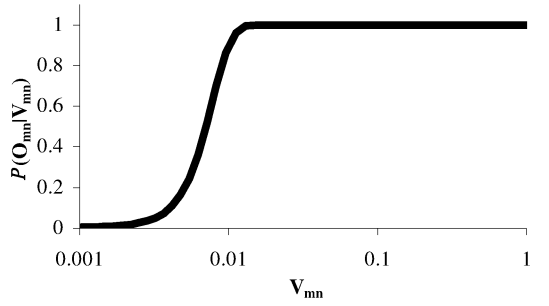

Figure 3. Probability of observation of a 2D NOESY cross-peak as a function of its volume V_{mn}. The $\mathcal{P}(O_{mn}|V_{mn})$ values were estimated via Equation 6 (see text) assuming $\sigma_m = \sigma_n = 0.012$ ppm and $\sigma_{noise} = 5 \times 10^{-5}$.

In order to calculate $\mathcal{P}(O_{mn}|O_{ij})$ (Equation 5) we also have to compute $\mathcal{P}(r_{mn}|O_{ij})$. The values of this distribution depend on the identities of m and n, as related to i and j, respectively. Correspondingly, the (m,n) reporter protons were grouped into six sets generically denoted by W: (a) Δ(i), Δ(j), include only protons i and j, respectively, (b) Ω_C(i), Ω_C(j), protons COSY-connected to protons i and j, respectively, and (c) Ω_T(i), Ω_T(j), protons in the same residue non-COSY (e.g., TOCSY) connected to i and j, respectively. The sets were combined by direct products to generate the 9 possible expanded sets $\Delta(i)\otimes\Omega_C(j)$, $\Omega_C(j)\otimes\Omega_T(i)$, etc., and the values of $\mathcal{P}(r_{mn}|O_{ij})$ were calculated for each of the expanded sets.

The starting point for the derivation of $\mathcal{P}(r_{mn}|O_{ij})$ is the inter-proton distance probability distribution $\mathcal{P}(r_{ij})$ in protein structures (Figure 4). This distri-

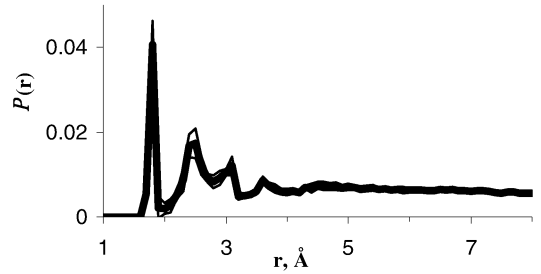

Figure 4. Inter-proton distance probability distribution, extracted from the structures of ubiquitin (Cornilescu et al., 1998), BPTI (Berndt et al., 1992), calmodulin (Ikura et al., 1992), crambin (Bonvin et al., 1993; Jelsch et al., 2000), cytochrome c (Qi et al., 1996), and human prion protein (James et al., 1997; Zahn et al., 2000). Thin lines show the standard deviation about the average (thick line).

bution, largely independent of the protein fold and size for $r_{ij} < 10$ Å, was derived from 6 high-quality structures of globular proteins representing a variety of folds. The $\mathcal{P}(r_{ij})$ *vs* r_{ij} profile outlines a repetitive pattern of spherical density shells that fade into a continuum, in a 'liquid-like' fashion, as the positional correlations decay with increasing inter-proton distances. For our purposes, the $\mathcal{P}(r_{ij})$ initially may be expressed as a sum of probability distributions according to whether the pair (H_i, H_j) belongs to the C, T or $\overline{CT}$ classes. Hence,

$$\begin{aligned}\mathcal{P}(r_{ij}) = \; & \mathcal{P}(r_{ij}|C_{ij}) \cdot \mathcal{P}(C_{ij}) + \mathcal{P}(r_{ij}|T_{ij}) \cdot \mathcal{P}(T_{ij}) \\ & + \mathcal{P}(r_{ij}|\overline{CT}_{ij}) \cdot \mathcal{P}(\overline{CT}_{ij}).\end{aligned} \quad (7)$$

The conditional distance probability distributions $\mathcal{P}(r_{ij}|C_{ij})$, $\mathcal{P}(r_{ij}|T_{ij})$, and $\mathcal{P}(r_{ij}|\overline{CT}_{ij})$, obtained for the same set of the six protein structures used to generate the $\mathcal{P}(r_{ij})$, are generally well structured (Figure 5).

We are interested in formulating an expression for the probability $\mathcal{P}(r_{ij}|O_{ij}, \overline{CT}_{ij})$ to obtain a distance r_{ij}, given that (i,j) are connected by an observable NOE, and the fact that they belong to different residues. The derivation follows Bayes theorem:

$$\mathcal{P}(r_{ij}|O_{ij}, \overline{CT}_{ij}) = \frac{\mathcal{P}(r_{ij}|\overline{CT}_{ij}) \cdot \mathcal{P}(O_{ij}|r_{ij}, \overline{CT}_{ij})}{\mathcal{P}(O_{ij}|\overline{CT}_{ij})}. \quad (8)$$

Under the assumption of isotropic rigid motion and a suitable (relaxation matrix) treatment of the NOESY data, the probability of observation of a NOE between H_i and H_j depends only on the distance r_{ij}. Hence, $\mathcal{P}(O_{ij}|r_{ij}, X_{ij}) = \mathcal{P}(O_{ij}|r_{ij}) \cdot \mathcal{P}(r_{ij}|X_{ij})$, where X refers to any of C, T, or $\overline{CT}$ classes. Therefore, ignoring normalization, Equation 8 can be recast as

$$\mathcal{P}(r_{ij}|O_{ij}, \overline{CT}_{ij}) \propto \mathcal{P}(r_{ij}|\overline{CT}_{ij}) \int dV_{ij} \mathcal{P}(O_{ij}|V_{ij}) \cdot \mathcal{P}(V_{ij}|r_{ij}), \quad (9)$$

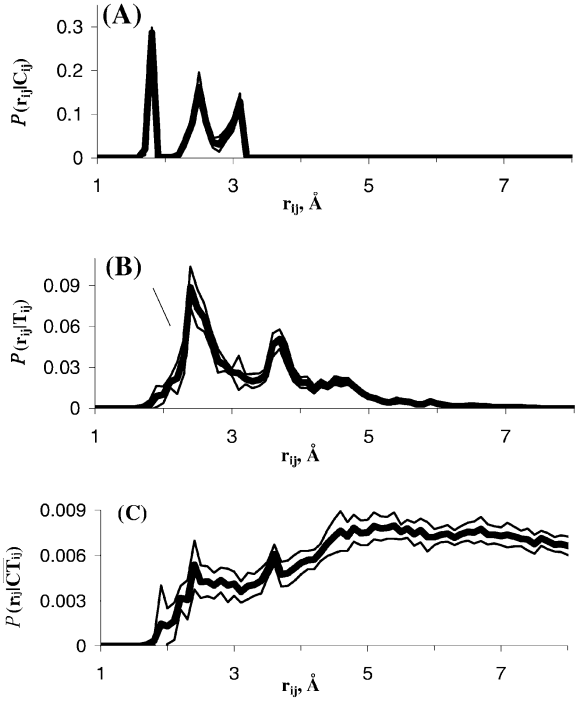

Figure 5. Inter-proton distance probability distributions for potentially COSY-connected (A), TOCSY-connected (B), and inter-residue NOE-connected (C) protons, extracted from the set of structures used for Figure 4. Thin lines show the standard deviation about the average (thick line).

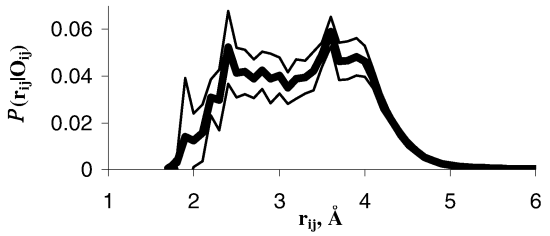

Figure 6. Distance probability distribution for potentially NOE-connected protons i and j, which belong to different residues, extracted from the set of model structures listed in the caption to Figure 4. Thin lines show the standard deviation about the average (thick line).

where the integral accounts for $\mathcal{P}(O_{ij}|r_{ij})$. The resulting $\mathcal{P}(r_{ij}|O_{ij}, \overline{CT}_{ij})$ distribution is illustrated in Figure 6.

The probability distributions $\mathcal{P}(r_{im}|W_{im})$, $\mathcal{P}(r_{jn}|W_{jn})$, and $\mathcal{P}(r_{ij}|O_{ij})$ were input to Monte Carlo simulations to generate random spatial arrangements of protons i, j, m, and n for each of the previously defined direct product expanded sets, and W stands for any of the sets of reporter protons defined above. Moves that placed proton pairs (m,n), (i,n), or

(j,m) at < 1.7 Å, were rejected; the resulting probability distributions $\mathcal{P}(r_{mn}|O_{ij})$ were accumulated over $\sim 10^5$ configurations. Values of $\mathcal{P}(O_{mn}|O_{ij})$ for each of the expanded sets, calculated from $\mathcal{P}(r_{mn}|O_{ij})$ via Equation 5, are those listed in Table 1.

Results and discussion

The performance of BACUS was monitored through the information entropies S_p (Shannon, 1948) of the set of cross-peak prior matches for all peaks p:

$$S_p = -\sum_{k,\ell} \mathcal{P}(O_{k\ell}) \ln \mathcal{P}(O_{k\ell}). \quad (10)$$

After convergence of S_p to a minimum, the output matches were analyzed in automated fashion. Each peak was associated with its matching odds, defined as the ratio of its largest matching probability to the next largest. For the analysis of col 2 data, the peaks were sorted into three classes: (1) With infinite matching odds (only a single, unique match); (2) with matching odds between infinity and 2.0; (3) with matching odds between 2.0 and 1.0. Within the second class, the matches were singled out by choosing, for each cross-peak, the one with the highest probability, while matches within the third class were left ambiguous. Thus, all cross-peaks within the first and second classes became uniquely matched.

In the case of kringle 2, an identity hypothesis (i,j) was rejected if a higher probability hypothesis (k,ℓ) existed, such that the posterior odds, defined as the ratio $\mathcal{P}(O_{k\ell}|\delta_p, O_R)/\mathcal{P}(O_{ij}|\delta_p, O_R)$, was > 2.0. The value of the cutoff, optimized for protocol's performance, was found to provide a satisfactory compromise between resolving power and accuracy. It was observed that cutoff values within the 2–3 range are reasonable, while higher values lead to a considerable decrease in the number of peaks that are uniquely identified by the program.

Using kringle 2, we also have explored the dependence of BACUS on σ, a main adjustable parameter. Our tests show that for σ within the 0.0025–0.0400 ppm range the algorithm converges in 20–50 iterations (< 1 min). We have resorted to four indicators to evaluate the protocol's performance (0.0025 ppm step size for σ). First, we address the 'resolving power' of the algorithm. One parameter is the protocol's 'compression ratio' (Figure 7A), defined as total number of input chemical shift-based identities divided by the number of output identities. A higher compression ratio signifies a better protocol performance. Another

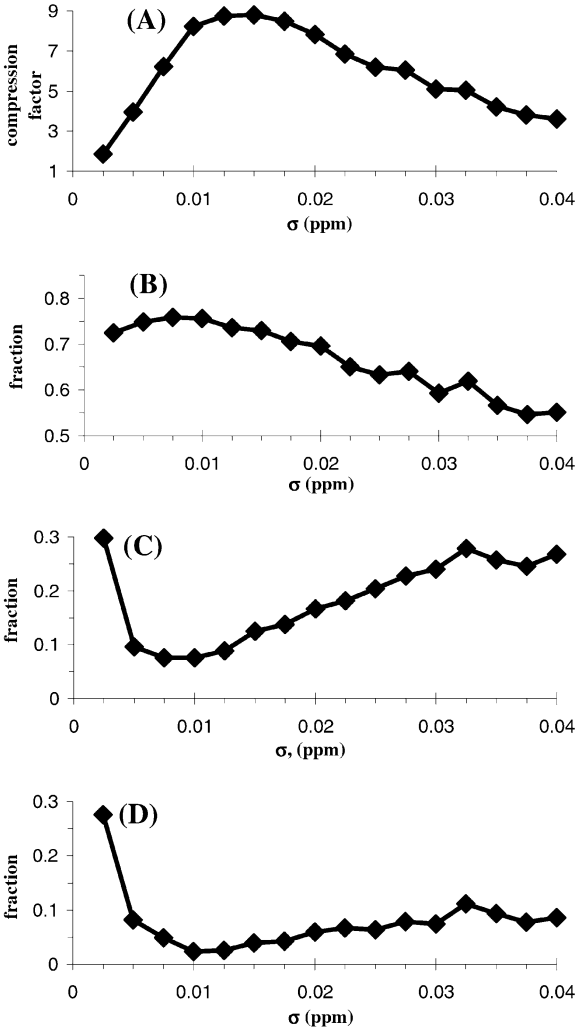

Figure 7. BACUS performance parameters for kringle 2 as functions of σ the chemical shift uncertainty. (A) Compression ratio; (B) fraction of uniquely assignable peaks; (C) fraction of unique assignments that are different from those reported; (D) fraction of unique assignments which differ from the reported by > 6 Å.

useful parameter is the fraction of peaks for which BACUS obtains unique identities (Figure 7B). Again, our goal is to maximize this number. The remaining two parameters describe the protocol's accuracy by reference to the reported manual assignments and protein structures (Briknarová et al., 1999; Marti et al., 1999). One of these is the fraction of the unambiguously determined identities that differ from the published assignments (Figure 7C). Since some of the NOESY cross-peaks could correspond to more than one proton pair, an extra measure of accuracy is the fraction of the uniquely identified cross-peaks that map to distances in the structure > 6 Å (Figure 7D).

Table 2. Statistics of BACUS protocol for col 2 data

	Unique matches (Class 1)	Matches with odds > 2.0 (Class 2)	Matches with odds < 2.0 (Class 3)
Number of peaks in the class	751 (71.6%)	178 (17.0%)	120 (11.4%)
Disagreements with manual NOESY assignment	1	4	26
Disagreements within the class (%)	0.13%	2.2%	22%

From inspection of Figure 7, it is apparent that the protocol exhibits best performance for σ within the 0.0075–0.0150 ppm range. Since the responses of all four parameters are similar, for a novel protein of unknown fold the result suggests that σ can be optimized by monitoring just the first pair of parameters, for which neither knowledge of the assignments nor the structure, is necessary. The optimal range found for σ is in line with the data at hand, as it corresponds to the expected uncertainty of the resonance and cross-peak definition (1–2 points in our 2D spectra). On this basis, we chose $\sigma = 0.01$ ppm for the col 2 and kringle 2 data sets.

A total of 1049 cross peaks were identified in the combined set of 2D NOESY spectra for col 2. The distribution of the entropies of the prior matching probabilities (Figure 8) shows that the procedure would have little chance of success if it were based exclusively on chemical shift matching, as only $\sim 9\%$ of the prior matches are defined uniquely. Posterior matching probabilities converged after 18 iterations in < 5 min with a 300 MHz Pentium II processor. Out of 1049 input cross-peaks, 88% were matched with odds better than 2:1. Among these, 0.5% were in disagreement with the reported manual assignments (Briknarová et al., 1999). The statistics of the final matching probabilities are listed in Table 2. Vis-à-vis the reported structure (Briknarová et al., 1999), all mismatches in classes 1 and 2 were found compatible, and only 7 (16%) of the mismatches from the class 3 in disagreement.

The fraction of NOESY cross-peaks as function of the effective number of possible identities for kringle 2, before and after BACUS analysis, is shown in Figure 9. Out of the 1354 input peaks, 1023 (75.6%) were identified uniquely. Of these, 78 (7.6%) were in disagreement with the manual assignments. However, only 24 peaks (2.3% of unique identities) corresponded to proton pairs distanced > 6 Å apart in the structure, likely to represent true errors due to the algorithm. These peaks, along with their reported manual assignments, BACUS identities, and interproton distances in the kringle 2 structure are listed in Table 3.

The first eight entries in Table 3 arise from resonances that are missing in the output of SPI. The next two are due to the errors in the peak positions. The assignment of entry 11 is ambiguous – it is not clear which, manual or automated, assignment is correct, as both distances are quite large. Entries 12 and 13 are due to incompleteness of spin systems via SPI. Entries 14, 19–24 are due to unfavorable contact geometry of the involved residues, making the subsets of useful reporter protons very limited. Entries 15 and 16 reflect an overestimation of the intra-residual likelihoods with respect to the intra-residual likelihoods in the BACUS protocol, due to the *ad hoc* nature of Equation 2. Entries 17 and 18 are due to normalization of the probabilities in Equation 2 by the product of the numbers of the reporter protons. The latter favors the assignments (in these cases, erroneous) of the cross-peaks to the shorter spin system over the longer spin system if the numbers of the cross-peaks reporting to both residues are similar.

It is apparent from Table 3 that misassignments are most dramatic in terms of violation distances when the correct identity hypotheses are excluded from the start. Despite the noted shortcomings, in most other cases the distance differences between the manual and automated assignments are not as large.

Overall, the output indicates that $\sim 4\%$ of the input cross-peaks are left unrefined with two or more identifications corresponding to *intra-residue* connectivities. Comparison of these identifications against the previously reported structures showed that most of them are likely to be correct. Since in typical NOESY

Table 3. SPI/BACUS peformance for kringle 2[a]

peak #	Peak coordinates	Manual assignment assignment	Distance, manual	Automated assignment	Distance, automated
1	4.92/3.22	R59:H^{α}/P61:$H^{\delta 3}$	2.43	R59:H^{α}/C22:$H^{\beta 2}$	13.02
2	4.07/3.22	P61:$H^{\delta 2}$/P61:$H^{\delta 3}$	1.76	A24:H^{α}/C22:$H^{\beta 2}$	6.11
3	3.22/1.40	P61:$H^{\delta 3}$/P61:$H^{\beta 2}$	4.04	C22:$H^{\beta 2}$/P61:$H^{\beta 2}$	8.91
4	3.22/2.18	P61:$H^{\delta 3}$/R59:$H^{\beta 2}$	4.73	C22:$H^{\beta 2}$/R59:$H^{\beta 2}$	13.38
5	3.22/2.03	P61:$H^{\delta 3}$/P61:$H^{\gamma 2}$	3.03	C22:$H^{\beta 2}$/P61:$H^{\gamma 2}$	9.13
6	2.22/1.23	P30:$H^{\beta 3}$/P30:$H^{\beta 2}$	1.77	P30:$H^{\beta 3}$/T81:$H^{\gamma 2}$	32.89
7	1.57/1.22	P30:$H^{\gamma 2}$/P30:$H^{\beta 2}$	2.34	P30:$H^{\gamma 2}$/T81:$H^{\gamma 2}$	30.18
8	7.13/0.48	W25:$H^{\delta 1}$/L46:$H^{\beta 3}$	3.42	W25:$H^{\delta 1}$/L46:$H^{\delta 2}$	6.68
9	4.94/2.42	R52:H^{α}/C51:$H^{\beta 3}$	4.64	W25:$H^{\eta 2}$/D10:$H^{\beta 2}$	11.92
10	4.59/3.11	C80:H^{α}/C1:$H^{\beta 3}$	3.61	C80:H^{α}/D55:$H^{\beta 2}$	21.03
11	3.52/1.41	G19:$H^{\alpha 3}$/K70:$H^{\beta 2}$	6.78	G19:$H^{\alpha 3}$/K15:H^{γ}	7.76
12	6.81/1.02	Y9:H^{ε}/I13:$H^{\delta 1}$	2.32	Y50:H^{ε}/K70:$H^{\gamma 3}$	17.95
13	7.49/2.58	L74:H^{N}/M17:$H^{\gamma 2}$	3.81	L74:H^{N}/W72:$H^{\beta 3}$	7.30
14	5.15/4.90	S14:H^{α}/Y50:H^{α}	2.89	S14:H^{α}/W62:$H^{\zeta 3}$	10.60
15	7.43/6.43	W62:$H^{\delta 1}$/W72:$H^{\zeta 3}$	4.75	W72:H^{N}/W72:Hζ3	7.12
16	7.14/6.46	F64:H^{ε}/W25:$H^{\zeta 3}$	2.52	W25:$H^{\delta 1}$/W25:$H^{\zeta 3}$	6.68
17	7.84/4.97	K70:H^{N}/D67:H^{α}	4.54	T65:H^{N}/D67:H^{α}	7.34
18	7.84/2.62	K70:H^{N}/D67:$H^{\beta 2}$	4.36	T65:H^{N}/D67:$H^{\beta 2}$	6.87
19	7.13/4.07	W25:$H^{\delta 1}$/A24:H^{α}	4.00	N49:$H^{\delta 22}$/A24:H^{α}	6.04
20	3.16/2.43	C75:$H^{\beta 2}$/C51:$H^{\beta 3}$	2.80	Y9:$H^{\beta 2}$/D10:$H^{\beta 2}$	7.46
21	2.72/1.34	D26:$H^{\beta 3}$/A24:H^{β}	4.38	C22:$H^{\beta 3}$/A24:H^{β}	6.45
22	2.59/0.96	E57:H^{γ}/L74:$H^{\delta 1}$	4.39	W72:$H^{\beta 3}$/L74:$H^{\delta 1}$	10.12
23	7.19/1.99	K15:H^{N}/I13:H^{β}	4.09	K15:H^{N}/L20:$H^{\beta 2}$	8.38
24	8.50/1.25	C22:H^{N}/T16:$H^{\gamma 2}$	4.16	C22:H^{N}/Q23:$H^{\beta 3}$	6.67

[a]Total NOESY cross-peak misassignments: 24/1023 (2.3%).

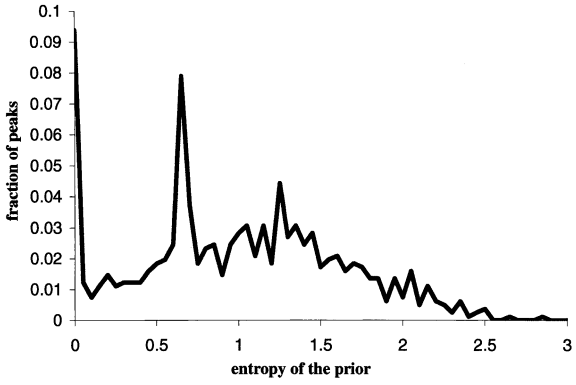

Figure 8. Distribution of the information entropies for the prior assignment probabilities per peak for col 2.

spectra the fraction of cross-peaks that arise from intra-residue connectivities is $\sim$40%, we assess that there are at least another $\sim$4% of cross-peaks that originate from several (intra- and inter-residue) pairs of resonances, some of which could not be refined by BACUS. Therefore, circa $< 8\%$ of cross-peaks should be expected to lead to multiple identities, which means that the apparent numbers of unique identifications may underestimate the intrinsic refining power of the method.

Conclusions

In the standard protocol of protein NMR data analysis, the emphasis is the assignment of resonance frequencies to specific spins or groups of magnetically equivalent nuclear spins (Wüthrich, 1986). In order to derive the structure, this is followed by the full assignment and quantification of the NOESY experiment. However, the assignment of cross-peaks in protein NOESY spectra can be extremely time-consuming, especially in the absence of data from approximate or homologous structural models. The complexity of the

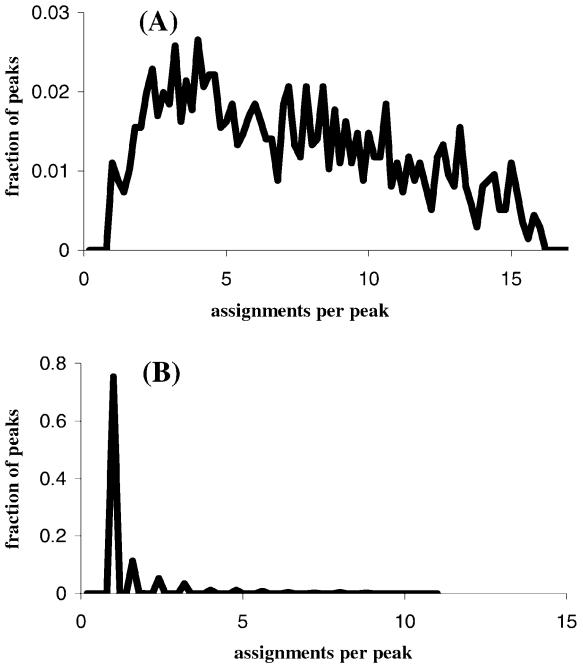

Figure 9. Distributions of the numbers of assignment per peak for kringle 2 before (A) and after (B) BACUS procedure. The numbers of assignment for each peak were calculated as exponents of the corresponding information entropies, thus the appearance of non-integer numbers of assignments per peak.

process is mainly due to the degeneracy of the NMR resonance frequencies, a factor that aggravates with increasing size of the macromolecule.

The CLOUDS approach (Grishaev and Llinás, 2002a,b) derives protein structures starting *solely* from inter-proton distances obtained from a relaxation matrix analysis of the NOESY. As is the case for the standard protocol, CLOUDS demands an input file of *unambiguous* NOEs. This requirement motivated us to develop two programs, SPI as a tool to sort the resonances' chemical shifts and, as reported here, BACUS, a protocol that maps the NOESY cross-peak to its source proton frequencies, in turn probabilistically determined via SPI. Thus, by exploiting Bayesian inference at several stages (BAF, SIF, SPI, BACUS), the complete CLOUDS protocol intrinsically differs from the by now well established stochastic schemes of NMR data optimization (e.g., Herrmann et al., 2002b; Linge et al., 2003) in that its emphasis is in the *unassigned* proton signals as the main source of structural information.

BACUS sorts NOEs by searching for self-consistency of the overall NOESY (Figure 1). The use of 'reporter' protons reduces the impact of chemical shift degeneracy and compresses the number of possible NOE identities by a factor of 6–9, yielding, on the test cases, unambiguous matches for 75–88% of the input cross-peaks. Overall, the performance of BACUS when dealing with both col 2 and kringle 2 data is comparable to those published for ARIA or NOAH/DIAMOD when tested against other proteins' data sets, both in terms of relative number of uniquely identified cross-peaks (75–88%, starting from 1–8%), and of error level ($\sim 5\%$).

The BACUS' performance does not hinge on prior knowledge of an approximate protein fold. Moreover, the entire BACUS formalism is tunable as the computed probabilities explicitly depend on spectral parameters such as mixing time and dimensionality, as well as sample conditions, e.g., signal-to-noise and molecular rotational correlation time. Because it resorts to reporter, neighboring spins, BACUS (Figure 1) is reminiscent of the 'network-anchoring' protocol recently developed by Herrmann et al. (2002a,b) for the ATMOS protocol. However, in contrast with the latter, BACUS is entirely probabilistic in nature and, in its current implementation, does not rely on sequence-specific resonance assignments, running in seconds without iterative structure calculations.

As a protocol based on statistical criteria, BACUS cannot guarantee correct identifiations of the complete set of NOEs. Furthermore, some of the approximations made during the derivations may have to be re-assessed in order to improve the level of rigor. Notwithstanding these caveats, our tests suggest that the BACUS protocol is robust and reliable. It is our hope that such an approach as presented here will serve to complement, conceptually and in practice, key components of the presently available methods of protein NMR data analysis and molecular structure elucidation.

Acknowledgement

This research was sponsored by the U.S. Public Health Service, NIH Grant HL-29409.

Appendix A: Examples of input file formats for BACUS

The lines encode for the Glu4 spin system of kringle 2: H^N, 9.491 ppm; H^α, 4.824 ppm; $H^{\beta 2}$, 2.248 ppm; $H^{\beta 3}$, 1.781 ppm; and H^γ 2.356 ppm.

Cosy-type connectivities; 'cylink.inp' file

Resonance ID	Chemical shift	# of resonances linked to	Their IDs		
1	8.491	1	2		
2	4.824	3	1	3	4
3	2.248	3	2	4	5
4	1.781	3	3	4	5
5	2.356	2	3	4	

Tocsy-type connectivities (extra to those in 'cylink.inp', same format); 'tylink.inp' file

1	8.491	3	3	4	5
2	4.824	1	5		
3	2.248	1	1		
4	1.781	1	1		
5	2.356	2	1	2	

References

Berndt K.D., Güntert, P., Orbons, L.P. and Wüthrich, K. (1992) *J. Mol. Biol.*, **227**, 757–775.

Bonvin, A., Rullmann, J., Lamerichs, R., Boelens, R. and Kaptein R. (1993) *Proteins*, **15**, 385–400.

Briknarová, K., Grishaev, A., Bányai, L., Tórdai, H., Patthy, L. and Llinás, M. (1999) *Structure*, **7**, 1235–1245.

Cornilescu, G., Marquardt, J.L., Ottiger, M. and Bax, A. (1998) *J. Am. Chem. Soc.*, **120**, 6836–6837.

Grishaev, A. and Llinás, M. (2002a) *Proc. Natl. Acad. Sci. USA*, **99**, 6707–6712.

Grishaev, A. and Llinás, M. (2002b) *Proc. Natl. Acad. Sci. USA*, **99**, 6713–6718.

Grishaev, A. and Llinás, M. (2002c) *J. Biomol. NMR*, **24**, 203–213.

Güntert, P., Berndt, K.D. and Wüthrich, K. (1993) *J. Biomol. NMR*, **3**, 601–606.

Hänggi, G. and Braun, W. (1994) *FEBS Lett.*, **344**, 147–153.

Herrmann, T., Güntert, P. and Wüthrich, K. (2002a) *J. Mol. Biol.*, **319**, 209–227.

Herrmann, T., Güntert, P. and Wüthrich, K. (2002b) *J. Biomol. NMR*, **24**, 171–189.

Ikura, M., Clore, G.M., Gronenborn, A.M., Zhu, G., Klee, C.B. and Bax, A. (1992) *Science*, **256**, 632–638.

James, T.L., Liu, H., Ulyanov, N.B., Farr-Jones, S., Zhang, H., Donne, D.G., Kaneko, K., Groth, D., Mehlhorn, I., Prusine, S.B. and Cohen, F.E. (1997) *Proc Natl. Acad. Sci. USA*, **94**, 10086–10091.

Jelsch, C., Teeter, M.M., Lamzin, V., Pichon-Lesme, V., Blessing, B. and Lecomte, C. (2000) *Proc. Nat. Acad. Sci. USA*, **97**, 3171–3176.

Kendall, M. (1987) *Advanced Theory of Statistics*, Vol. 2, Oxford University Press, New York, NY.

Linge, J.P., Habeck, M., Rieping, W. and Nilges, M. (2003) *Bioinformatics*, **19**, 315–316.

Madrid, M., Llinás, E. and Llinás, M. (1991) *J. Magn. Reson.*, **93**, 329–346.

Marti, D., Schaller, J. and Llinás, M. (1999) *Biochemistry*, **38**, 15741–15755.

Mumenthaler, C. and Braun, W. (1995) *J. Mol. Biol.*, **254**, 465–480.

Mumenthaler, C., Güntert, P., Braun, W. and Wüthrich, K. (1997) *J. Biomol. NMR*, **10**, 351–362.

Nilges, M. (1993) *Proteins*, **17**, 297–309.

Nilges, M. (1995) *J. Mol. Biol.*, **245**, 645–660.

Nilges, M. and O'Donoghue, S. I. (1998) *Prog. NMR Spectrosc.*, **32**, 107–139.

Nilges, M., Macias, M.J., O'Donoghue, S.I. and Oshkinat, H. (1997) *J. Mol. Biol.*, **269**, 408–422.

Qi, P.X., Beckman, R.A. and Wand, A.J. (1996) *Biochemistry*, **35**, 12275–12286.

Shannon, C.E. (1948) *Bell Syst. Tech. J.*, **27**, 379–423.

Wüthrich, K. (1986). *NMR of Proteins and Nucleic Acids*, Wiley & Sons, New York, NY, pp. 117–199.

Zahn, R., Liu, A., Luhrs, T., Calzolai, L., Von Schroetter, C., Garcia, F.L., Riek, R., Wider, G., Billeter, M. and Wüthrich, K. (2000) *Proc. Nat. Acad. Sci. USA*, **97**, 145–150.