Journal of Biomolecular NMR, **24:** 203–213, 2002.
KLUWER/ESCOM

Sorting signals from protein NMR spectra: SPI, a Bayesian protocol for uncovering spin systems

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

Received 14 June 2002; Accepted 18 September 2002

Key words: CLOUDS, protein NMR data sorting, resonance assignment, signal identification, spectral analysis

Abstract

Grouping of spectral peaks into J-connected spin systems is essential in the analysis of macromolecular NMR data as it provides the basis for disentangling chemical shift degeneracies. It is a mandatory step before resonance and NOESY cross-peak identities can be established. We have developed SPI, a computational protocol that scrutinizes peak lists from homo- and hetero-nuclear multidimensional NMR spectra and progressively assembles sets of resonances into consensus J- and/or NOE-connected spin systems. SPI estimates the likelihood of nuclear spin resonances appearing at defined frequencies given sets of cross-peaks measured from multi-dimensional experiments. It quantifies spin system matching probabilities via Bayesian inference. The protocol takes advantage of redundancies in the number of connectivities revealed by suites of diverse NMR experiments, systematically tracking the adequacy of each grouping hypothesis. SPI was tested on 2D homonuclear and 2D/3D^{15}N-edited data recorded from two protein modules, the col 2 domain of matrix metalloproteinase-2 (MMP-2) and the kringle 2 domain of plasminogen, of 60 and 83 amino acid residues, respectively. For these protein domains SPI identifies $\sim$ 95% unambiguous resonance frequencies, a relatively good performance vis-à-vis the reported 'manual' (interactive) analyses.

Abbreviations and Acronyms: SPI, **SP**in **I**dentification; BMRB, BioMagResBank (Madison, WI).

Introduction

The assignment of spectral cross-peaks is generally considered to be a prerequisite for biomacromolecular structure elucidation via NMR. It can be viewed as a mapping of points in chemical shifts 'space' to spin sites in the molecule according to rules encoded by experiment-specific connectivity patterns. Depending on the molecule and the available data, the procedure can be extremely time-consuming owing to (i) spurious connectivities or (ii) missing peaks, and (iii) degeneracy of resonance frequencies. In order to avoid the assignment bottleneck, we have developed CLOUDS (Grishaev and Llinás, 2002a,b), a protocol that aims at deriving protein structures starting from a rather complete set of *unassigned,* albeit *unambiguous*, NOEs. In retrospect, CLOUDS probably affords one of the most robust of the various proposed 'direct' methods of NMR structure computation (Atkinson and Saudek, 2002, and references therein). However, the necessity of a list of uniquely identified NOEs, whether assigned or not, is not restricted to CLOUDS, being a prerequisite for any distances-based NMR structure computation protocol.

Grouping is an essential early stage of the NMR data analysis as it provides the basis for resolving the chemical shift definition of the clustered spectral signals, particularly severe in the case of biomacromolecules. A variety of computational protocols have been formulated for the automation of spin system grouping in protein spectra. These approaches exploit intra- or inter-spectral redundancies in the observed connectivities. Some methods (Kleywegt et al.,

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

1991; Xu et al., 1995; Croft et al., 1997) are based on the recognition of patterns that are characteristic of amino acid residues within each experiment. They perform best with rather complete, high quality data. Others (Lukin et al., 1997; Zimmerman et al., 1997) exploit the occurrence of common (root) subsets of resonance frequencies in various triple-resonance experiments. Clearly, the comparatively higher spectral resolution characteristic of such data sets facilitates a more reliable detection of connectivites, hence of coupled spin systems.

Here we describe SPI, a procedure that combines the above two approaches in an attempt to take advantage of their best attributes. The protocol is designed to cope with those cases for which both intra- and inter-spectral redundancies, when separately considered, are insufficient to establish unique groupings, owing to ambiguities or incompleteness of the data. In the simple implementation presented here, SPI combines redundant information encoded by individual 2D homonuclear spectra with common cross-peak patterns exhibited by different $^{1}H/^{1}H$ and 2D/3D $^{1}H/^{1}H^{N}/^{15}N$ correlation experiments, for which chemical shift degeneracy can be significant. Spin systems are identified via a Bayesian statistical analysis, with probabilities reflecting both the data ambiguity and the extent of spectral information redundancy. Consensus spin systems are defined as those that satisfy various types of experimental data with reasonably high probability. SPI was tested on NMR data for the col 2 and kringle 2 domains of human matrix metalloproteinase 2 (MMP-2) and of human plasminogen, respectively (Briknarová et al., 1999; Marti et al., 1999).

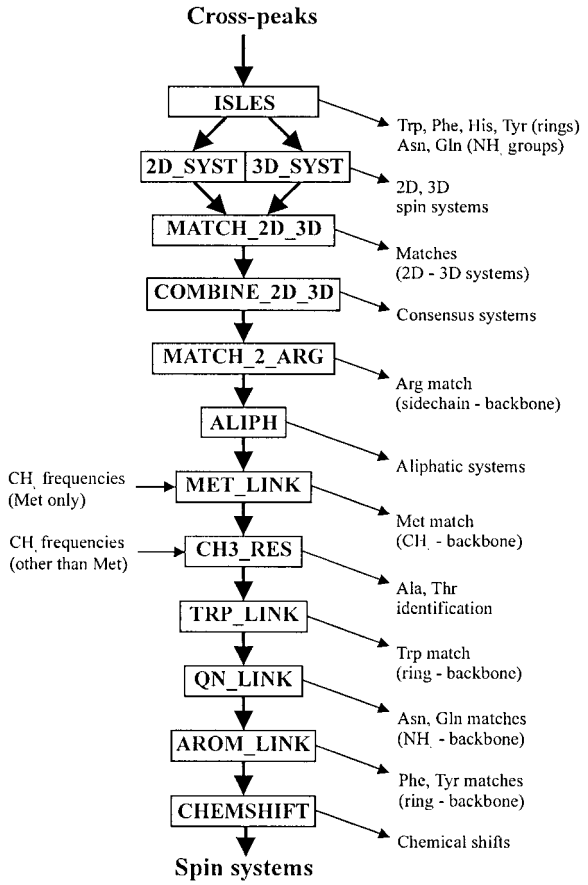

Figure 1. SPI flowchart. Stages of the program are boxed. The input consists of cross-peak chemical shift coordinates from $^{1}H/^{1}H$ COSY, TOCSY, NOESY and ^{15}N-edited HSQC, HSQC-TOCSY, HSQC-NOESY, HNHA, HNHB experiments. Execution flow is denoted by thick arrows. Input/output is indicated by thin arrows.

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 reported (Briknarová et al., 1999; Marti et al., 1999). 2D $^{1}H/^{1}H$ correlation spectra were frequency-standardized by matching common cross-peaks, each defined within $\pm$ 0.03 ppm and $\pm$ 0.04 ppm in the direct and indirect dimensions, respectively. Peak chemical shifts measured on both sides of the diagonal were averaged. Methyl resonances were identified from the intensities and relative narrowness of their diagonal peaks. All programs were written in FORTRAN77 and compiled via Digital^{TM} Visual Fortran compiler, version 6. The software is available from the corresponding author by request.

Identification of resonances and spin systems: SPI

In the minimalistic approach explored here, SPI analyzes 2D homonuclear and 3D ^{15}N-edited connectivity information to produce a list of resonances grouped into spin systems (Figure 1). Both cross-peaks and spin systems are treated as vectors in the multidimensional space of chemical shifts obtained from the available NMR experiments. The individual dimensions of such space correspond to the identities of NMR resonances (^{15}N, H^{N}, H^{α}, H^{β}, etc.). Let us assume that a peak **I** is characterized by a chemical shift coordinate x^{i}, as well as 'root' chemical shifts

$z^i_1, \ldots, z^i_n$. Namely, $\mathbf{I} \equiv (x^i, z^i_1, \ldots, z^i_n)$ of dimensionality n+1. On the other hand, let there be an m+n dimensional spin system $\mathbf{J} \equiv (y^j_1, \ldots, y^j_m, z^j_1, \ldots, z^j_n)$, which shares with **I** a projection onto the subspace of root coordinates $\mathbf{Z}_n \equiv (z_1, \ldots, z_n)$. **I** and **J** thus project onto $\mathbf{Z}_n$ to form vectors $\mathbf{z}^i \equiv (z^i_1, \ldots, z^i_n)$ and $\mathbf{z}^j \equiv (z^j_1, \ldots, z^j_n)$. Let us define a vector $\Delta\mathbf{z} \equiv \mathbf{z}^i - \mathbf{z}^j$. $\mathcal{P}(\mathbf{Z}|\mathbf{I},\mathbf{J})$, the conditional probability density of observing the experimental $\Delta\mathbf{z}$ in the $\mathbf{Z}_n$ subspace whenever **I** and **J** belong to the same spin system, is then expressed as a multi-variate Gaussian distribution:

$$\mathcal{P}(\mathbf{Z}|\mathbf{I},\mathbf{J}) = \mathrm{N}\exp\left(-\frac{1}{2}\Delta\mathbf{z}^{\dagger}\Sigma_n^{-1}\Delta\mathbf{z}\right). \quad (1)$$

Here, Σ_n is the diagonal variance matrix defined by the uncertainties σ of the resonance definitions within $\mathbf{Z}_n$, and $\mathrm{N} = \left[(2\pi)^n |\Sigma_n|\right]^{-\frac{1}{2}}$ accounts for normalization. In the following applications, the value of $\mathcal{P}(\mathbf{Z}|\mathbf{I},\mathbf{J})$ is assumed negligible whenever any component of $\Delta\mathbf{z}$ is larger than twice the corresponding variance in Σ_n. *In practice, this means that* $\mathcal{P}(\mathbf{Z}|\mathbf{I},\mathbf{J})$ *is set to zero whenever the corresponding peak is not observed within the expected chemical shift tolerances*. For example, in the case of TOCSY data (discussed below) only spin systems for which $|\delta_{2,I} - \delta_{HN,J}| < 2\sigma_{2D}$ are taken into account. Similar criteria applies to heteronuclear data.

Next, we evaluate the match between **I** and the remaining component of **J**, i.e., vector $\mathbf{y}^j \equiv (y^j_1, \ldots, y^j_m)$ within subspace $\mathbf{Y}_m$. For this, we consider other peaks $\mathbf{O}^k_\ell$, that are spanned by x^i and vectors within $\mathbf{Y}_m$ with $1 \le \ell \le m$. For a 2D peak $\mathbf{O}^k_\ell \equiv (x^k, y^k_\ell)$, we define a vector of differences $\Delta\mathbf{xy}_\ell \equiv (x^i - x^k, y^j_\ell - y^k_\ell)$. Similar to the Equation 1, the likelihood to observe the experimental vectors $\Delta\mathbf{xy}_\ell$ if **I** and **J** are within the same spin system is modeled as:

$$\mathcal{P}(\mathbf{XY}|\mathbf{I},\mathbf{J}) = \mathrm{M}\prod_{\ell=1}^{m}\exp\left(-\frac{1}{2}\Delta\mathbf{xy}^{\dagger}_\ell\Sigma_2^{-1}\Delta\mathbf{xy}_\ell\right), \quad (2)$$

where M is the normalization constant. Finally, assuming statistical independence of the spectral data within the **Z** and **XY** sub-spaces, the likelihood of the entire data conditional on **I** and **J** originating within the same spin system is written as:

$$\mathcal{P}(\mathbf{Z},\mathbf{XY}|\mathbf{I},\mathbf{J}) = \mathcal{P}(\mathbf{Z}|\mathbf{I},\mathbf{J})\cdot\mathcal{P}(\mathbf{XY}|\mathbf{I},\mathbf{J}). \quad (3)$$

While the $\mathcal{P}(\mathbf{Z}|\mathbf{I},\mathbf{J})$ term describes the quality of root overlap, $\mathcal{P}(\mathbf{XY}|\mathbf{I},\mathbf{J})$ encodes for the redundancy of spectral information.

In our protocol (described below), the initial '2D' and '3D' spin systems are the J-coupled H^N/H^α pairs from COSY and $^1H^N/^{15}N$ pairs from HSQC, respectively.

Analysis of ^{15}*N-edited data*

M_3 total roots of the 3D spin systems are obtained from the HSQC; these are denoted by (δ_{HN}, δ_N), where δ_{HN} is the chemical shift of the backbone peptidyl amide ^1H and δ_N is that of the attached ^{15}N. The ISLES subroutine (Appendix A) identifies side chain Trp $H^{\varepsilon 1}/N^{\varepsilon 1}$ and Asn/Gln $NH_2^{\delta,\varepsilon}$ groups that are not included in the roots list.

All resonances originating from HSQC-TOCSY, HNHA and HNHB spectra are probabilistically matched to the roots. For a given 3D connectivity **I** ($\delta_{H,I}$, $\delta_{HN,I}$, $\delta_{N,I}$) and a spin system **J** of root ($\delta_{HN,J}$, $\delta_{N,J}$), the probability of **I** corresponding to **J** is estimated according to the Bayes theorem (Jaynes, 1996) as:

$$\mathcal{P}(\mathbf{I},\mathbf{J}|\mathbf{Z}_2) = \frac{\mathcal{G}(\delta_{HN,I},\delta_{HN,J};\sigma_{HN})\cdot\mathcal{G}(\delta_{N,I},\delta_{N,J};\sigma_N)\mathcal{P}(\mathbf{I},\mathbf{J})}{\sum_{K=1}^{M_3}\mathcal{G}(\delta_{HN,I},\delta_{HN,K};\sigma_{HN})\cdot\mathcal{G}(\delta_{N,I},\delta_{N,K};\sigma_N)\mathcal{P}(\mathbf{I},\mathbf{K})}. \quad (4)$$

Here, probability densities $\mathcal{G}(x,y;z) \propto \exp(-0.5 \times (x-y)^2/z^2)$, σ_{HN} and σ_N are the chemical shift uncertainties of H^N and ^{15}N resonances in 3D spectra, 0.03 ppm and 0.336 ppm, respectively, and $\mathcal{P}(\mathbf{I},\mathbf{K})$ are the prior probabilities, all assumed of equal value. The likelihoods in Equation 4 are the Gaussians defined by Equation 1 with $\mathbf{Z}_2 = (H^N, ^{15}N)$.

Analysis of homonuclear 2D data

In order to build 2D spin systems, the amide-aliphatic TOCSY connectivities are matched to the H^N/H^α COSY roots (total number M_2). Similar to the 3D case, aromatic and Asn/Gln NH_2 peaks, identified via ISLES (Appendix A), are excluded from COSY and TOCSY peak lists. The initial 'H^N-based probabilities' are written as functions of H^N frequencies only. The probability that a peak **I** of coordinates ($\delta_{1,I}$, $\delta_{2,I}$) and a spin system **J** of coordinates ($\delta_{HN,J}$, $\delta_{H\alpha,J}$) originate from the same residue is estimated from likelihoods via Equation 1 with $\mathbf{Z}_1 = (H^N)$:

$$\mathcal{P}(\mathbf{I},\mathbf{J}|\mathbf{Z}_1) = \frac{\mathcal{G}(\delta_2,\delta_{HN,J};\sigma_{2D})\mathcal{P}(\mathbf{I},\mathbf{J})}{\sum_{K=1}^{M_2}\mathcal{G}(\delta_2,\delta_{HN,K};\sigma_{2D})\mathcal{P}(\mathbf{I},\mathbf{K})}, \quad (5)$$

Table 1. Differences between the manual assignments and SPI output

Site	Appearance	Possible reason
(A) col2		
Leu3 H^{γ}	missing	overlap with Leu3 $H^{\beta 3}$
Phe4 H^{δ}, H^{ε}, H^{ζ}	missing	all chemical shifts degenerate
Met6 H^{ε}	missing	overlap with Met6 $H^{\beta 3}$
Phe21 H^{ζ}	missing	overlap with Phe21 H^{ε}
Gln22 $H^{\gamma 3}$	missing	overlap with Gln22 $H^{\beta 2}$
Ser25 $H^{\beta 2}$	missing	overlap with Ser25 $H^{\beta 3}$
Tyr38 $H^{\beta 2}$	missing	overlap with Tyr38 $H^{\beta 3}$
Thr44 H^{β}	missing	overlap with Thr44 H^{α}
Pro57 H^{γ}	missing	overlap with Pro57 $H^{\beta 2}$
(B) kringle 2		
Ser2 H^{β}, 3.92 ppm	additional found	chemical shift close to the reported
Met6 system	broken into (H^{N}, H^{α}), and (H^{α}, H^{β}'s, H^{γ}'s and H^{ε})	missing J-connectivities
Ile17 $H^{\gamma 12}$	missing	missing J-connectivities
Ile17 $H^{\gamma 13}$	missing	overlap with Ile17 $H^{\gamma 2}$
Pro34 $H^{\beta 2}$	missing	missing J-connectivities
Pro41 $H^{\gamma 2}$ and H^{δ}	missing	missing J-connectivities
Pro45 $H^{\gamma 2}$	missing	missing J-connectivities
Leu49 $H^{\beta 3}$	missing	overlap with L49 $H^{\delta 2}$
Lys50 $H^{\gamma 3}$, H^{δ}, $H^{\varepsilon 2}$ and $H^{\varepsilon 3}$	missing	missing J-connectivities
Pro57 system	broken into (H^{α} and H^{β}'s) and (H^{γ}'s and H^{δ}'s)	missing J-connectivities
Pro63 H^{α}, $H^{\delta 2}$ and $H^{\delta 3}$	missing	missing J-connectivities
Pro70 $H^{\delta 2}$	missing	missing J-connectivities
Pro70 $H^{\delta 3}$	missing	overlap with Pro70 H^{α}
Pro70 $H^{\gamma 3}$	missing	overlap with Pro70 $H^{\beta 2}$
Arg73 $H^{\beta 3}$ and H^{δ}	missing	missing J-connectivities
Arg81 H^{γ}	missing	missing J-connectivities

where, again, we assume equal priors $\mathcal{P}(\mathbf{I},\mathbf{J})$ for all **J**'s, $\mathcal{G}(x,y;z)$ are Gaussian probability densities as described for Equation 4 with σ_{2D} ~0.015 ppm, the resonance position uncertainty in the 2D spectra.

Next, the root-based probabilities obtained from Equation 5 are used to recalculate spin system membership probabilities by cross-referencing other observed J-connectivities to the H^{α} chemical shifts, $\delta_{H\alpha,I}$, of the 2D spin systems. For every ambiguous match between a resonance δ_1, and the spin system **J** with coordinates ($\delta_{HN,J}$, $\delta_{H\alpha,J}$) obtained as above, the program searches for a COSY/TOCSY peak that is the closest to the coordinates (δ_1,$\delta_{H\alpha,J}$) within $2\sigma_{2D}$. If such peak with coordinates ($\delta_{1,J}$, $\delta_{2,J}$) is found, in the Bayesian context, the 'H^{N}/H^{α}-based probabilities' are estimated via:

$$\mathcal{P}(\mathbf{I},\mathbf{J}|\mathbf{Z}_1,\mathbf{XY}) = \frac{\mathcal{G}(\delta_1,\delta_{1,J};\sigma_{2D})\cdot\mathcal{G}(\delta_{H,\alpha,J},\delta_{2,J};\sigma_{2D})\cdot\mathcal{P}(\mathbf{I},\mathbf{J}|\mathbf{Z_1})}{\sum_{K=1}^{M_2}\mathcal{G}(\delta_1,\delta_{1,K};\sigma_{2D})\cdot\mathcal{G}(\delta_{H\alpha,K},\delta_{2,K};\sigma_{2D})\cdot\mathcal{P}(\mathbf{I},\mathbf{K}|\mathbf{Z_1})}. \quad (6)$$

By reference to Equation 3, $\mathbf{Z}_1 = (H^{N})$, $\mathbf{Y} = (H^{\alpha})$ and the priors $\mathcal{P}(\mathbf{I},\mathbf{J}|\mathbf{Z}_1)$ are the posteriors obtained through Equation 5.

At this point, a number of side chain resonances result uniquely matched to their spin systems and this knowledge can be taken advantage of to refine the remaining matches. Let us consider a resonance at δ_1 *ambiguously* matched to spin system **J**, and a spin H with a chemical shift $\delta_{H,J}$, *unambiguously* matched to

J. The program searches for the COSY/TOCSY peak that, within $2\sigma_{2D}$, most closely matches (δ_1,$\delta_{H,J}$). If such peak ($\delta'_{1,J}$,$\delta'_{2,J}$) is found, the 'H^N/H^α/H-based probabilities' are written via Equation 3 with $\mathbf{Y}' = (H^\alpha,H)$ as:

$$\mathcal{P}(\mathbf{I}, \mathbf{J}|\mathbf{Z_1}, \mathbf{XY}') = \frac{\mathcal{G}(\delta_1, \delta'_{1,J}; \sigma_{2D}) \cdot \mathcal{G}(\delta_{H,J}, \delta'_{2,J}; \sigma_{2D}) \cdot \mathcal{P}(\mathbf{I}, \mathbf{J}|\mathbf{Z_1}, \mathbf{XY})}{\sum_{K=1}^{M_2} \mathcal{G}(\delta_1, \delta'_{1,K}; \sigma_{2D}) \cdot \mathcal{G}(\delta_{H,J}, \delta'_{2,K}; \sigma_{2D}) \cdot \mathcal{P}(\mathbf{I}, \mathbf{K}|\mathbf{Z_1}, \mathbf{XY})}. \quad (7)$$

Here, the priors $\mathcal{P}(\mathbf{I},\mathbf{J}|\mathbf{Z}_1, \mathbf{XY})$ are the posteriors obtained via Equation 6. Equations 3 and 7 are then applied with the full set of H resonances spanning the entire **Y** subspace. Once the procedure converges, SPI detects and links Gly and Arg spin systems (see Appendix A).

Matching, combination and extension of 2D and 3D spin systems

Subroutine MATCH2D3D probabilistically matches the independently identified 2D and 3D spin systems on the basis of common chemical shifts. Initially, ambiguous matches of the H^α atoms (from HNHA spectrum) are associated with their most probable 3D spin system roots, producing uniquely defined N/H^N/H^α triads. The latter are frequency-matched to the roots of the 2D spin systems with probabilities according to Equation 1 with $\mathbf{Z'_2} = (H^N,H^\alpha)$, where instead of matching a resonance **I** to a spin system **J**, we are dealing with spin systems $\mathbf{J}_1$ and $\mathbf{J}_2$:

$$\mathcal{P}(\mathbf{J}_1, \mathbf{J}_2|\mathbf{Z'_2}) \propto \mathcal{G}(\delta_{HN,J1}, \delta_{HN,J2}; \sigma_{HN,3D}) \cdot \mathcal{G}(\delta_{H\alpha,J1}, \delta_{H\alpha,J2}; \sigma_{H,3D}). \quad (8)$$

Here, ($\delta_{HN,J1}$,$\delta_{H\alpha,J1}$) is the root of the 2D spin system $\mathbf{J}_1$, and ($\delta_{N,J2}$, $\delta_{HN,J2}$, $\delta_{H\alpha,J2}$) of the 3D spin system $\mathbf{J}_2$; $\sigma_{H,3D}$ is the point size in the indirect (H^{other}) dimension of the 3D spectra. The tolerances of $3\sigma_{HN,3D}$ and $3\sigma_{H,3D}$ are used for the H^N and H^{other} dimensions, respectively. In those cases where the placement of a spin is unambiguous in only one of the uniquely matched systems, the latter are used for the intra-spin system identification.

Ambiguous matches between the 2D and 3D spin systems are refined by considering additional spins shared by the two spin systems. The program searches for overlap between the uniquely placed side chain resonances in the 2D and 3D systems. Whenever a match is found, the probabilities from Equation 8 are multiplied by the corresponding Gaussian factor, reflecting recurring updates of the common root subspace. Finally, the matches are selected in order of decreasing probabilities.

The 2D and 3D spin systems are then combined via subroutine COMBINE2D3D. For this, the program generates a list of 'consensus' resonances, i.e., present in both 2D and 3D spin systems. Those resonances that are found in 3D systems only are inserted into such consensus system whenever they exhibit at least two COSY or TOCSY connectivities to resonances of the consensus system. Consistently, resonances that belong to 2D systems only are promoted to the consensus system whenever they exhibit at least two J-connectivities to the latter.

Consensus spin systems are then extended into the aliphatic area ($\delta < 6$ ppm). In order for an aliphatic resonance to become assigned to a consensus spin system it should exhibit at least three COSY/TOCSY connectivities for a system of four or more resonances (or two connectivities for a 3-resonance system). As consensus spin systems grow in size, further resonances are added until convergence.

There are aliphatic cross-peaks whose resonances do not meet criteria to belong to the consensus systems thus far established, e.g., Pro systems which lack H^N. Such COSY cross-peaks become roots which the program attempts to combine via *all* COSY/TOCSY connectivities in the aliphatic region. For these, subroutine ALIPH creates a separate peak list as well as a list of the derived fragments with three or more members. Finally, residues Phe, Tyr, Trp, His, Met, Asn and Gln contain spin sub-systems that do not connect each other via standard ^{15}N-edited or ^{1}H/^{1}H J-correlated spectra. Others, such as Arg, often exhibit two parallel spin systems, one originating from the backbone H^N atom, the other from the nitrogen-bonded H^ε. Subroutines MATCH2ARG, METLINK, TRPLINK, QNLINK, and AROMLINK aim at matching such sub-systems exploiting the observed NOESY connectivities (Appendix A).

Subroutine CHEMSHIFT averages chemical shifts of each spin system over the complete set of cross-peaks, and outputs the final listing. The detailed program architecture is outlined in Figure 2.

Results and discussion

The SPI analysis of col 2 spectral data proved to be relatively straightforward, yielding 310 unambiguous proton frequencies compared to 321 obtained via man-

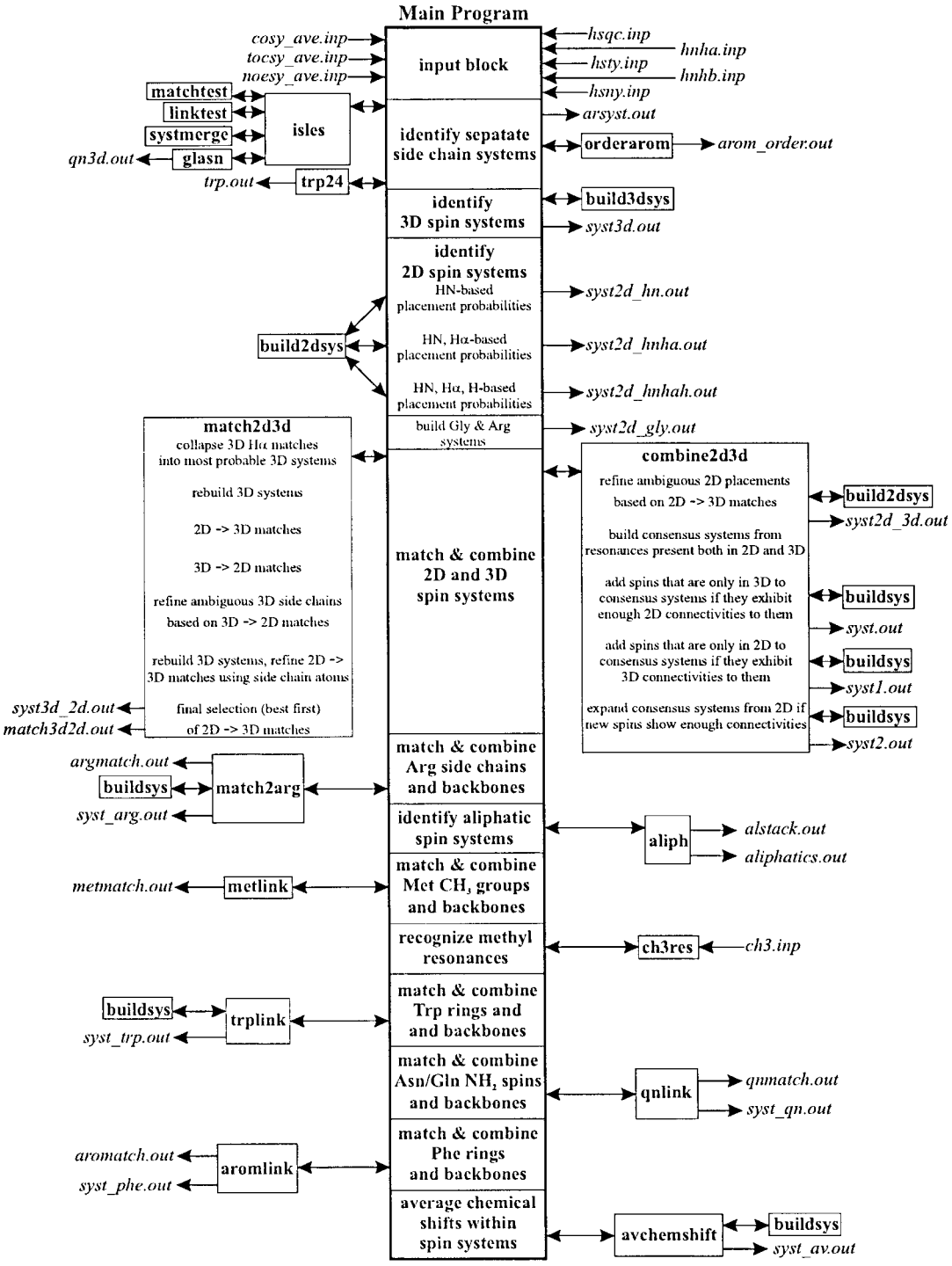

Figure 2. SPI: Architecture of the program. Subroutines are boxed, in bold, with calls indicated by doubly-pointed arrows. Files are in bold/italics, with input/output indicated by singly-pointed arrows. The program is executed from top to bottom.

Table 2. Differences between the ideal peak lists and SPI input.

		False negatives	False positives
Col 2	COSY	0.22	0.13
	TOCSY	0.39	0.19
	HSQC	0.00	0.12
	HSQC-TOCSY	0.21	0.07
	HNHA	0.02	0.12
	HNHB	0.06	0.28
Kringle 2	COSY	0.26	0.20
	TOCSY	0.44	0.17
	HSQC	0.00	0.42
	HSQC-TOCSY	0.29	0.39
	HNHA	0.14	0.10
	HNHB	0.38	0.22

False negatives are defined as the number of peaks missing from a given multi-dimensional spectrum divided by the number of cross-peaks expected from the amino acid sequence, taking into account the degree of degeneracy of the contributing frequencies. False positives are defined as the number of unexplained peaks divided by the total number of peaks in the multi-dimensional spectrum.

ual analysis (Briknarová et al., 1999). By reference to the latter, SPI matching of Phe and Trp side chains to the corresponding backbone spin systems produced no errors. Spin systems corresponding to Gly, Arg, Trp, and various Asn, Gln, Met and Phe residues were identified. For the Met6 CH_3^{ε} however, whose chemical shift coincides with that of its $H^{\beta 3}$, unambiguous matching was based, as with other methyls, on linewidth criteria. In future implementations, such uncertainties can be narrowed by resorting to ^{13}C data, e.g., from a $^{1}H/^{13}C$ HSQC experiment, where the carbon chemical shifts of the various methyl groups are readily identifiable and better resolved. Differences between reported col 2 assignments and identities obtained via SPI are summarized in Table 1A, most of them stemming from chemical shift degeneracy within their respective spin systems.

For kringle 2, a larger molecule, the SPI analysis proved to be more challenging. Extra 3D systems arising from partially unfolded ^{15}N protein sample were identified and discarded, as they did not lead to satisfactory matches to the 2D spin systems. A total of 93 amino acid systems remained after combining the 2D and 3D data. These consensus systems contained 458 protons, compared to 478 protons obtained via manual analysis (Marti et al., 1999). The NMR data for kringle 2 were recorded on a sample complexed to AMCHA, *trans*-(aminomethyl)cyclohexanecarboxylic acid. This posed no problem as SPI correctly identified the complete ligand spin system. Differences between the reported spin systems of kringle 2 and the ones determined via SPI are due to (a) missing cross-peaks and (b) intra-spin system chemical shift degeneracy (Table 1B).

By reference to the published kringle 2 data (Marti et al., 1999), all matchings (Figure 3) of Asn and Gln NH_2 groups and Trp and Phe rings to their respective backbone spin systems were found to be correct. Matching of Arg backbone and side chain spin systems was also stable (Figures 3C and 3D). His and Tyr ring systems were all identified as well. From our experience with both kringle 2 and col 2 data, matching of these systems is unstable in the absence of H^{δ} identities. The distinction was not always unambiguous from the experimental data, as the statistical H^{ε} and H^{δ} chemical shift distributions (BioMagResBank) are not sufficiently well-resolved. For example, when attempting to link the His and Tyr spin systems in kringle 2 on the basis of NOEs from $H^{\delta,\varepsilon}$, ca. 30% turned out to be incorrect. As a result, all two-member His and Trp ring spin systems were left floating, unmatched to the backbone. In the context of the CLOUDS protocol, this does not invalidate the approach since the final spatial locations of spin systems depend on the global positioning of the *identified* protons according to distances estimated from the NOESY spectrum. In other words, *unambiguity of chemical shifts is more important than completeness of the linked fragments SPI provides*.

For Pro residues, analysis of the connectivities in the 2D data yielded $\sim$ 60% of the expected resonances. Complicating factors when identifying aliphatic side chain systems are: resonance degeneracies, crowding of peaks in a relatively small spectral area, and excitation profile in the H^{α}/H^{β} region associated with the implemented WATERGATE water suppression scheme (Piotto et al., 1992). Such problems can be alleviated in principle by recourse to higher-dimensionality and/or ^{13}C-edited experiments.

SPI performance

SPI mimics the standard interactive analysis of NMR spectra, based on subconscious probabilistic assessments at each step of the process, while expediting it by taking advantage of Bayesian inference which quantifies the decision making steps. An important feature is its search for self-consistency when analyzing experimental J-connectivity data. By cycling

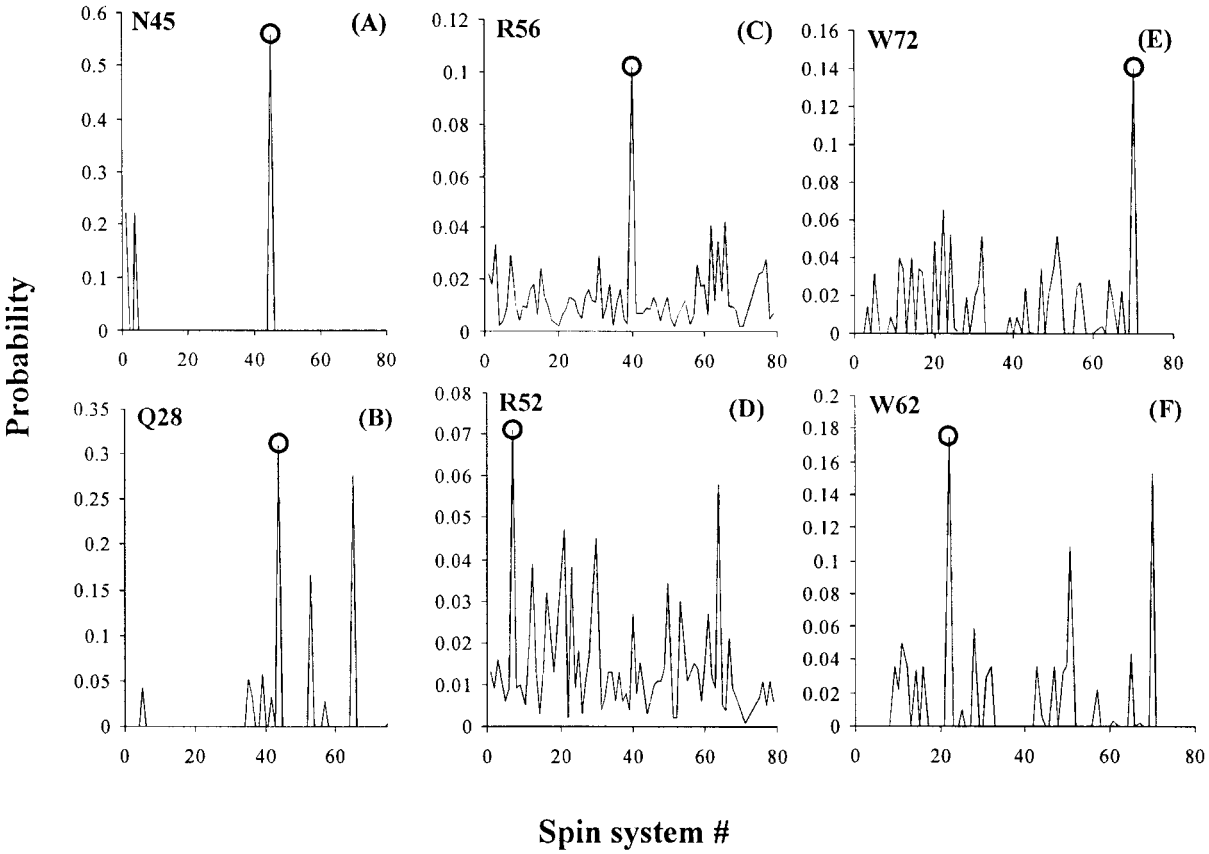

Figure 3. Examples of matching backbone/side chain spin systems probabilities for kringle 2. Top row, the least ambiguous matches; bottom row, the most ambiguous matches. Correct choices are circled. Totals of 7 Gln/Asn, 5 Arg, and 3 Trp were considered.

Table 3. SPI: degeneracies of resonance frequencies within tolerances

		Degeneracy									
		1	2	3	4	5	6	7	8	9	>9
Col2	H^N (2D)	0.38	0.55	0.07							
	H^α (2D)	0.26	0.47	0.25	0.02						
	H (2D)	0.21	0.37	0.21	0.12	0.01	0.05	0.02	0.01	0.07	
	$H^N/^{15}N$ (3D)	0.88	0.07	0.05							
	H^α (3D)	0.06	0.09	0.12	0.16	0.20	0.09	0.13	0.07	0.06	0.02
	H (3D)	0.05	0.08	0.03	0.14	0.12	0.07	0.12	0.12	0.07	0.20
Kringle 2	H^N (2D)	0.27	0.34	0.23	0.12	0.04					
	H^α (2D)	0.14	0.42	0.22	0.08	0.01	0.02	0.02	0.09		
	H (2D)	0.28	0.17	0.16	0.13	0.11	0.10	0.03	0.01	0.01	
	$H^N/^{15}N$ (3D)	0.78	0.17	0.04	0.01						
	H^α (3D)	0.01	0.01	0.03	0.03	0.06	0.03	0.11	0.17	0.10	0.45
	H (3D)	0.15	0.08	0.03	0.03	0.05	0.04	0.04	0.06	0.08	0.44

The numbers indicate the fraction of proton resonances degenerate by reference to reported assignments (Briknarová et al., 1999; Marti et al., 1999) within each type: H^N, backbone amide; H^α, backbone alpha; and H, any side chain hydrogen. Tolerances of 2σ apply in each dimension.

the execution and subsequent peak list editing steps, the program helps to compensate for missing connectivities from the initial input files, thus completely assembling most of the tested proteins spin systems.

Tables 2 and 3 list, respectively, quality attributes of the SPI input, as compared to the ideal peak lists, and degree of resonance degeneracies within the tolerances specified for the various stages of the protocol. Expectedly, the HSQC yields both better resolution and peak list quality (Tables 2 and 3). Similarly, joint treatment of the ^{15}N-edited and 2D homonuclear data reduces ambiguities while increasing the number of identified resonances.

The number of SPI-established resonances are somewhat less than those reported (Briknarová et al., 1999; Marti et al., 1999). Primary reasons are: (i) Low intensity cross-peaks (Table 2), and (ii) resonance overlap (Table 3). The first problem may be overcome, as is routine in the manual procedure, by substituting NOEs for the missing J-connectivities. Because of the inherent risk of selecting inter-residue cross-peaks, especially in the absence of sequence-specific assignments, we chose not to pursue this option. In the spirit of the CLOUDS approach, missing H-atoms eventually fall into place, as guided by the NOEs among the *identified* protons, in the final structure generation. In order to deal with the second problem, SPI attempts at maximally exploiting redundancies within the experimental J-connectivity map.

The SPI approach is based on the *existence* of J-connectivities, not on the J-split cross-peak multiplicity pattern. The latter provides valuable information for the identity of specific proton pairs within spin systems and can be incorporated to further reduce ambiguities. On the other hand, since the basic concept behind SPI is not data specific, the protocol can be adapted readily to other types of J-connectivity data not exploited in this study (^{13}C-edited, triple-resonance, etc). Moreover, novel data processing approaches, such as the Filter Diagonalization Method (Chen et al., 2000), might provide additional advantages as they push farther the limits of the attainable spectral resolution, potentially enhancing the definition of resonance frequencies.

The SPI processing of col-2 and kringle-2 data yields ∼95% of the manually established resonance frequencies. Moreover, the $H^N/H^\alpha/H$ resolving power of SPI when dissecting homonuclear 2D spectra, indicates that the method approaches the intrinsic resolution afforded by ^{15}N-edited 3D experiments. Since SPI reduces the number of NMR experiments required for the identification of spin frequences, it suggests itself as an attractive starting point for developing a fully automated, high-throughput macromolecular structure computation protocol.

Acknowledgement

This work was sponsored by the U.S. Public Health Service, National Institutes of Health Grant HL-29409.

Appendix A. Residue-specific subroutines of SPI

Aromatic spin systems

The ISLES subroutine deals with Phe, Tyr, Trp, and His ring spin systems. It starts by identifying Trp $H^{\delta1}$ and $H^{\varepsilon1}$ resonances based on COSY and HNHA cross-peaks, within ranges 5–8 ppm for $H^{\delta1}$, 9–12 ppm for $H^{\varepsilon1}$ and 125–135 ppm for $N^{\varepsilon1}$. Remaining COSY connectivities in the 6–9 ppm range are considered roots of additional, non-Trp, ring spin systems. Whenever a pair of such roots exhibits a subset of resonances that lie within $2\sigma_{2D}$, subroutine LINKTEST checks for COSY, TOCSY connectivities among the remaining spins. When the connectivity matrix is complete, the two partial spin systems are joined via subroutine SYSTMERGE. The program also checks for TOCSY/COSY linkages to the thus generated spin systems, stemming from resonances between 4.8–6.0 ppm (extended aromatic area). If any such resonance is completely connected to the aromatic spin system, it is added to it. Clearly, the chemical shift ranges we have selected can be enlarged to include outliers.

NH_2 groups of Gln and Asn side chains are discriminated via subroutine GLASN. It starts by detecting HSQC cross-peak pairs within the range of 100–120 ppm that differ by < 2 digital points (~ 0.01 ppm) in the ^{15}N dimension. Further observation of NH_2 geminal pair cross-peaks in the HSQC-NOESY identifies its origin as Gln or Asn.

Subroutine ORDERAROM organizes aromatic spin systems according to COSY connectivities. Three-spin systems are assumed to be Phe $H^{\delta}/H^{\varepsilon}/H^{\zeta}$ and four-spin systems Trp $H^{\zeta2}/H^{\eta2}/H^{\zeta3}/H^{\varepsilon3}$. The subroutine TRP24 matches previously identified $H^{\delta1}/H^{\varepsilon1}$ Trp pairs to the $H^{\zeta2}/H^{\eta3}/H^{\zeta3}/H^{\varepsilon3}$. The $H^{\varepsilon1}/H^{\eta2}$ connections are established from HSQC-NOESY signals

at the corresponding $H^{\varepsilon 1}/N^{\varepsilon 1}$ frequencies. Trp $H^{\zeta 2}$ and $H^{\varepsilon 3}$, both of which exhibit two COSY connectivities, are distinct from $H^{\zeta 3}$ and $H^{\eta 2}$, which show only one. Therefore, TRP24 explores NOE connectivities of $H^{\varepsilon 1}$ to both $H^{\zeta 2}$ and $H^{\varepsilon 3}$ (thus far unidentified). The found NOE identifies $H^{\zeta 2}$, hence the entire $H^{\zeta 2}/H^{\eta 2}/H^{\zeta 3}/H^{\varepsilon 3}$ string.

Derivation of Gly and Arg spin systems from 2D data

Based on the 2D data, SPI detects and links Gly and/or Arg spin systems whenever they exhibit two distinct H^{α} atoms COSY-connected to the H^{N} (Gly), or two H^{δ} atoms COSY-connected to the H^{ε} (Arg). For example, in the case of Gly, given the two roots ($\delta_{HN,J1}$, $\delta_{H\alpha,J1}$) and ($\delta_{HN,j2}$, $\delta_{H\alpha,J2}$), with $|\delta_{HN,J1} - \delta_{HN,J2}| < 2\sigma_{2D}$, the linking probability is:

$$\mathcal{P}(\mathbf{J}_1, \mathbf{J}_2|\mathbf{Z}, \mathbf{XY}) \propto \mathcal{G}(\delta_{HN,J1}, \delta_{HN,J2}; \sigma_{2D}) \cdot \mathcal{G}(\delta_{H\alpha,J1}, \delta_{1,K}; \sigma_{2D}) \cdot \mathcal{G}(\delta_{H\alpha,J2}, \delta_{2,K}; \sigma_{2D}), \quad (1)$$

where ($\delta_{1,k}$, $\delta_{2,k}$) is the COSY/TOCSY cross-peak closest to coordinates ($\delta_{H\alpha,J1}$, $\delta_{H\alpha,J2}$), within $2\sigma_{2D}$. The matches are derived best-first, i.e. in the order of decreasing $\mathcal{P}(\mathbf{J}_1,\mathbf{J}_2|\mathbf{Z},\mathbf{XY})$ values. Gly residues are differentiated from the Arg side chains based on spin system size.

Combination of backbone and side chain spin systems

Matching of Arg strings is accomplished via subroutine MATCH2ARG. Arg side chains, originating from H^{ε}, are recognized based on the $^{15}N^{\varepsilon}$ chemical shifts (70–90 ppm). The program tests for matches between Arg side chains and those that are not. MATCH2ARG calculates a score that combines frequency matches between the two parallel systems, with the COSY/TOCSY/NOESY connectivity network linking remaining unmatched frequencies. The probability of integration for parallel Arg side chain system *i* (starting from H^{ε}) and system *j* (starting from H^{N}) is:

$$\mathcal{P}_{\mathrm{Arg}}(i, j) = \frac{\sum_{k=1}^{N_{ij}} \mathcal{G}(\delta_k^i, \delta_k^j; \sigma_{2D})}{N_{max}} + \left(1 - \frac{N_{ij}}{N_{max}}\right) \sum_{l=1}^{N_i - N_{ij}} \sum_{m=1}^{N_j - N_{ij}} \frac{\mathcal{G}(\delta_l^i, \delta_{lm}^1; \sigma_{2D}) \cdot \mathcal{G}(\delta_m^j, \delta_{lm}^2; \sigma_{2D})}{(N_i - N_{ij}) \cdot (N_j - N_{ij})}, \quad (2)$$

where N_i and N_j are the numbers of resonances in strings *i* and *j*, respectively, N_{ij} is the number of resonances common to *i* and *j*, as found by the program, $N_{max} = \max(N_i, N_j)$, and (δ_{lm}^1, δ_{lm}^2) is the NOESY/TOCSY/COSY cross-peak that is the closest to the (δ_l^i, δ_m^j) coordinates within $2\sigma_{2D}$.

The Met methyl groups can be recognized from their chemical shifts ($\sim$ 2.0 ppm) and extremely sharp, singlet-like diagonal peaks in TOCSY or NOESY. Their matching to the backbone spin systems is achieved via subroutine METLINK. The latter accesses a list of such Met H^{ε} frequencies and tests connectivities between those frequencies and spin systems of unidentified type that have fewer than seven resonances. The matching score is:

$$\mathcal{P}_{\mathrm{Met}}(i, \delta_j^{CH_3}) = \frac{\sum_{m=1}^{N_i} \mathcal{G}(\delta_m^i, \delta_{ijm}^1; \sigma_{2D}) \cdot \mathcal{G}(\delta_j^{CH_3}, \delta_{ijm}^2; \sigma_{2D})}{\sum_{k=1}^{N_s} \sum_{m=1}^{N_k} \mathcal{G}(\delta_m^k, \delta_{kjm}^1; \sigma_{2D}) \cdot \mathcal{G}(\delta_j^{CH_3}, \delta_{kjm}^2; \sigma_{2D})}, \quad (3)$$

where the summations Σ_m's go over all spins in the consensus spin systems and Σ_k over all N_s spin systems. Here $\delta_j^{CH_3}$ is the CH_3 resonance in question and (δ_{ijm}^1, δ_{ijm}^2) is the NOESY cross-peak that is closest to the (δ_m^i, $\delta_j^{CH_3}$) coordinates within $2\sigma_{2D}$.

Trp rings, identified via ISLES, are linked to their respective backbone spin systems by subroutine TRPLINK. The program tests for NOESY matches between the established $H^{\delta 1}$ and $H^{\varepsilon 3}$ of the ring system *i* and the H^{α} and H^{β}'s of the prospective Trp backbone system *j* that exhibit 3 or 4 ^1H resonances. The program takes advantage of the chemical shift of Trp H^{β} atoms (3.21 ± 0.34 ppm), obtained from BMRB. Only backbone spin systems for which H^{β} chemical shifts are within 3 standard deviations from the average, are considered. The probability expression is similar to Equation 3:

$$\mathcal{P}_{\mathrm{Trp}}(i, j) = \frac{\sum_{k=1}^{2} \sum_{l=1}^{N_j - 1} \mathcal{G}(\delta_k^i, \delta_{ijkl}^1; \sigma_{2D}) \cdot \mathcal{G}(\delta_l^j, \delta_{ijkl}^2; \sigma_{2D})/(N_j - 1)}{\sum_{n=1}^{N_s} \sum_{k=1}^{2} \sum_{l=1}^{N_n - 1} \mathcal{G}(\delta_k^i, \delta_{inkl}^1; \sigma_{2D}) \cdot \mathcal{G}(\delta_l^n, \delta_{inkl}^2; \sigma_{2D})/(N_n - 1)}. \quad (4)$$

Here, Σ_k's run over the $H^{\delta 1}$ and $H^{\varepsilon 3}$ resonances of the Trp rings spin system *i*, Σ_l's over H^{α} and H^{β}'s of the candidate backbone spin systems, and Σ_n's over all candidate spin systems. As before, (δ_{ijkl}^1, δ_{ijkl}^2) is the 2D NOESY cross-peak that is closest to the (δ_k^i, δ_l^j) coordinates within $2\sigma_{2D}$. Notice normalization by the factors ($N_n - 1$) that is absent in Equation 3. Here we have already pre-selected candidate Trp backbone systems using both the number and the chemical shifts of

their H^{β} atoms. Normalization of individual terms in Equation 4 is meant to remove the number-of-matches bias inside this subgroup. However, with Met H^{ε} no such criteria were applied prior to Equation 3. In this case, it is the functional form of the equation itself that biases the selection towards those with larger number of cross-peak frequency matches.

Spin groups previously identified as Asn and Gln side chain $^{15}N^{1}H_2$ triads are probabilistically matched to the backbone parts of the spin systems via subroutine QNLINK based on NOESY connectivity scores. The candidate backbone spin systems are required to have 1–4 side chain ^{1}H resonances. The expression for the matching probabilities is similar to Equation 4, encoding for NOESY connectivities from the pair of H atoms of the NH_2 spin system i, to the H^{α}, H^{β} and H^{γ} atoms of backbone spin system j. The NOESY cross-peak positions are initially taken from 3D HSQC-NOESY spectra and subsequently refined based on 2D NOESY data. Systems with ≤ 2 side chain resonances prior to the match are categorized as Asn or Gln, and those with > 2 as Gln.

Phe side chains are matched to their backbone systems (3–4 resonances) via subroutine AROMLINK. Similar to the Trp case, the backbone systems are screened according to the average chemical shifts of the H^{β} atoms of Phe (3.01 ± 0.31 ppm), also taken from BMRB. Those containing H^{β}'s outside 3 standard deviations from the average are ignored. Matching probabilities are similar to Equation 4, where the Σ_k's run over the H^{δ} resonances only. The systems with ambiguous H^{δ} resonances are left floating, as they can be identified in the later stages of analysis.

References

Atkinson, R.A. and Saudek, V. (2002) *FEBS Lett.*, **510**, 1–4.

BioMagResBank, http://www.bmrb.wisc.edu/

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

Chen, J., Mandelstham, V.A. and Shaka, A.J. (2000) *J. Magn. Reson.*, **146**, 363–368.

Croft, D., Kemmink, J., Neidig, K.P. and Oschkinat, H. (1997) *J. Biomol. NMR*, **10**, 207–219.

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.

Jaynes, E.T. (1996) *Probability Theory – The Logics of Science*. USA, http://omega.albany.edu:8008/JaynesBook.html

Kleywegt, G.J., Boelens, R., Cox, M., Llinás, M., and Kaptein, R. (1991) *J. Biomol. NMR*, **1**, 23–47.

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

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

Piotto, M., Saudek, V. and Sklenar, V. (1992) *J. Biomol. NMR*, **2**, 661–665.

Xu, J., Weber, P.L. and Borer, P.N. (1995) *J. Biomol. NMR*, **5**, 183–192.

Zimmerman, D.E., Kulikowski, C.A., Huang, Y.P., Feng, W.Q., Tashiro, M., Shimotakahara, S., Chien, C.Y., Powers, R. and Montelione, G.T. (1997) *J. Mol. Biol.*, **269**, 592–610.