Journal of Biomolecular NMR, **24:** 171–189, 2002.
KLUWER/ESCOM

Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS

Torsten Herrmann[a,b], Peter Güntert[a,*] & Kurt Wüthrich[a,b,**]
[a]*Institut für Molekularbiologie und Biophysik, Eidgenössische Technische Hochschule Zürich, CH-8093 Zürich, Switzerland;* [b]*The Scripps Research Institute, 10550 N. Torrey Pines Rd., La Jolla, CA 92037, U.S.A.*

Received 10 September 2002; Accepted 20 October 2002

Key words: ATNOS, automated peak picking, CANDID, DYANA, nuclear magnetic resonance, nuclear Overhauser effect, protein structure determination

Abstract

Novel algorithms are presented for automated NOESY peak picking and NOE signal identification in homonuclear 2D and heteronuclear-resolved 3D [^{1}H,^{1}H]-NOESY spectra during *de novo* protein structure determination by NMR, which have been implemented in the new software ATNOS (automated NOESY peak picking). The input for ATNOS consists of the amino acid sequence of the protein, chemical shift lists from the sequence-specific resonance assignment, and one or several 2D or 3D NOESY spectra. In the present implementation, ATNOS performs multiple cycles of NOE peak identification in concert with automated NOE assignment with the software CANDID and protein structure calculation with the program DYANA. In the second and subsequent cycles, the intermediate protein structures are used as an additional guide for the interpretation of the NOESY spectra. By incorporating the analysis of the raw NMR data into the process of automated *de novo* protein NMR structure determination, ATNOS enables direct feedback between the protein structure, the NOE assignments and the experimental NOESY spectra. The main elements of the algorithms for NOESY spectral analysis are techniques for local baseline correction and evaluation of local noise level amplitudes, automated determination of spectrum-specific threshold parameters, the use of symmetry relations, and the inclusion of the chemical shift information and the intermediate protein structures in the process of distinguishing between NOE peaks and artifacts. The ATNOS procedure has been validated with experimental NMR data sets of three proteins, for which high-quality NMR structures had previously been obtained by interactive interpretation of the NOESY spectra. The ATNOS-based structures coincide closely with those obtained with interactive peak picking. Overall, we present the algorithms used in this paper as a further important step towards objective and efficient *de novo* protein structure determination by NMR.

Abbreviations: 2D, 3D, two-, three-dimensional; NOE, nuclear Overhauser enhancement; NOESY, nuclear Overhauser enhancement spectroscopy; CANDID, program for automated NOE assignment; DYANA, torsion angle dynamics program for NMR structure calculation.

Introduction

This paper describes an initial implementation of new algorithms for NOESY peak picking and NOE peak identification in the software ATNOS. When used in combination with the software CANDID for automated NOE assignment (Herrmann et al., 2002) and a suitable algorithm for protein three-dimensional structure calculation from NMR data, for example, DYANA (Güntert et al., 1997), ATNOS not only extends the automation of the process of protein structure determination to an additional, labor-intensive step,

*Present address: RIKEN Genomics Sciences Center, W505, 1-7-22 Suehiro, Tsurumi, Yokohama, 230-0045, Japan
**To whom correspondence should be addressed. E-mail: wuthrich@mol.biol.ethz.ch

but it also enables direct feedback between the intermediate protein structures and the raw NMR data during the protein structure refinement. Thereby the list of verified NOE peaks is updated between subsequent cycles of combined NOE assignment and protein structure determination (Herrmann et al., 2002) by reference to the intermediate protein structure.

Difficulties in automated NMR signal recognition arise from inevitable mutual signal overlap and spectral distortions due to artifacts. Sophisticated algorithms are available for peak identification at the outset of a spectral analysis (Antz et al., 1995; Corne et al., 1992; Garret et al., 1991; Kleywegt et al., 1990; Koradi et al., 1998), but in practice their use in spectral regions of strong peak overlap and near noisy artifacts is limited, and manual re-inspection of the results is quite generally advised. Therefore, in present practice, NOESY peak picking is still dominantly performed with interactive computer programs (for example, Bartels et al., 1995; Neidig et al., 1995). Both automated or interactive NOE peak identification must be able to clearly distinguish between real and artifactual peaks, with the signal-to-noise ratio as the primary filter. Because of the intrinsic inverse 6th power-relationship between NOE cross peak intensity and distance between the pair of protons attributed to the cross peak, a significant fraction of the most informative 'longrange' NOE signals (Wüthrich, 1986) in a NOESY spectrum may have signal-to-noise ratios only slightly above the noise level, which emphasizes the importance of working with powerful and sophisticated filtering procedures.

The presently introduced algorithms for NOESY peak picking and NOE cross peak identification differ from the aforementioned peak identification programs by aiming at a more modest goal: From the outset of the NOESY spectral interpretation, ATNOS makes use of chemical shift lists available from previous sequence-specific resonance assignment, and in more advanced stages of the calculation also of the intermediate three-dimensional protein structure. Within the network of 'raw' NOESY spectra, chemical shift lists and intermediate three-dimensional protein structure, ATNOS achieves more extensive and reliable NOE cross peak identification than routines that rely exclusively on the information content of the NOESY spectra. In its approach, ATNOS attempts to imitate the *modus operandi* of an experienced spectroscopist, who typically chooses to combine the process of NOESY peak picking and NOE assignment. This is achieved in part within the ATNOS algorithm and in part with the combined use of ATNOS and CANDID.

A conceptual limitation of the present practice of NMR structure determination is the lack of suitable routines by which the three-dimensional protein structure can be assessed through a direct link with the raw NMR data, e.g., by calculating R-factors (Borgias and James, 1990; Gronwald et al., 2000; Nilges et al., 1991). Complete relaxation matrix calculations (Boelens et al., 1989; Borgias and James, 1988; Gronwald et al., 2000; Keepers and James, 1984; Mertz et al., 1991; Yip and Case, 1989) have been introduced for this purpose, with the aim to improve the accuracy and precision of the molecular structure by fits to the initial NOE-build-up rate (Anil-Kumar et al., 1980) in the presence of spin diffusion and internal mobility. This sophisticated approach has, however, so far been used primarily for the final stages of structure refinement rather than for *de novo* protein structure determinations. ATNOS NOE peak identification in concert with automated NOE assignment and structure calculation now affords a direct correlation between NOESY spectra and protein structure, since the lists of verified NOE peaks are updated with reference to the protein structure in each cycle of structure calculation.

Algorithms

The section describes the algorithms for automated NOESY peak picking and NOE cross peak identification contained in ATNOS, and their incorporation into a scheme for automated NMR structure determination (Figure 1). At the outset of a *de novo* structure calculation, ATNOS makes use of chemical shift lists, which must already be available from previous sequence-specific resonance assignment, and in more advanced stages of the calculation also of the intermediate protein three-dimensional structure. It thus achieves more complete and reliable peak picking than algorithms that operate on the spectral data before sequence-specific resonance assignments are available. By re-assessing the NOESY spectra in each cycle of structure calculation, ATNOS links *de novo* structure determination with the experimental NMR data in a more direct way than with the commonly used NOESY peak lists, which are typically invariant during the entire structure calculation.

In the present implementation represented by the flowchart of Figure 1, ATNOS is used in combination with CANDID and DYANA, and automated protein

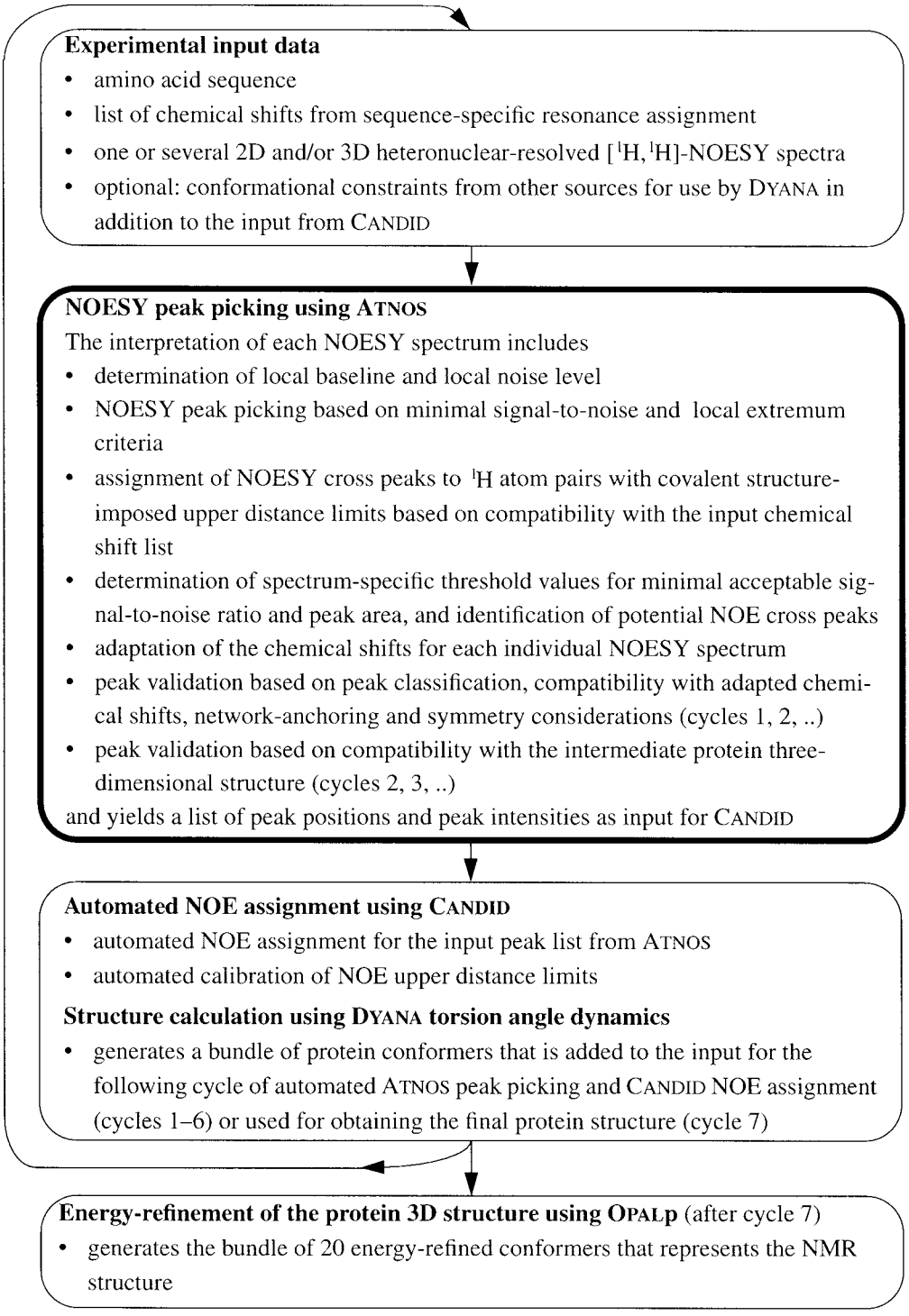

Figure 1. Flow chart of automated NMR structure determination using the new software ATNOS for NOESY peak picking and NOE cross peak identification in conjunction with the software CANDID for NOE assignment (Herrmann et al., 2002) and the program DYANA (Güntert et al., 1997) for three-dimensional protein structure calculation.

structure determination proceeds in iterative cycles. The input data used for all cycles include the amino acid sequence of the protein, the chemical shift list from the previous sequence-specific resonance assignment, and one or several 2D homonuclear or 3D heteronuclear-resolved [^{1}H,^{1}H]-NOESY spectra. Parameters for the practical use of ATNOS are given in Tables 1 and 2. Each cycle of computation consists of automated NOESY peak picking and NOE cross peak identification with ATNOS, use of the resulting lists of peak positions and peak intensities as input for automated NOE assignment with CANDID, and use of a set of NOE distance constraints from CANDID as input for the structure calculation with DYANA. In the first cycle, ATNOS NOE peak validation is guided primarily by the chemical shift lists. Between subsequent cycles, information is transferred through the intermediate protein three-dimensional structure, which is used from the second cycle onward in combination with the chemical shift list to guide NOE peak validation with ATNOS and NOE assignment with CANDID. The automated NOE assignment using CANDID and the structure calculation with the DYANA torsion angle dynamics algorithm in the flow diagram of Figure 1 are performed identically and with the same parameter sets as described previously by Herrmann et al. (2002). Since the precision and accuracy of the intermediate protein structures tend to improve from cycle to cycle, the structure-based criteria for ATNOS NOE peak validation are loosened in order to facilitate identification of weaker signals, whereas the criteria for acceptance of NOE assignments and NOE upper distance bounds in CANDID are tightened in more advanced cycles.

Experimental input data

The input files for ATNOS contain the amino acid sequence, the chemical shift lists, and 2D homonuclear and/or 3D heteronuclear-resolved [^{1}H,^{1}H]-NOESY spectra. The format corresponds to the data formats of the programs XEASY for interactive spectrum analysis (Bartels et al., 1995) and DYANA (Güntert et al., 1997). The input for the complete procedure of Figure 1 may include additional conformational constraints to supplement the data from CANDID in the input for the DYANA structure calculation, for example, disulfide bond constraints (Williamson et al., 1985), spin–spin coupling constants (these are converted in each cycle in combination with the updated list of upper limit NOE distance constraints into torsion angle constraints by the grid search procedure FOUND (Güntert et al., 1998)), and dihedral angle constraints from other sources.

Local baseline determination and local noise level determination

The ability of detecting weak signals with intensities only slightly above the noise level without erroneously including also noise artifacts into the resulting NOE peak list is a crucial prerequisite for a robust and reliable peak picking algorithm. Noise and artifacts in NMR spectra are not uniform, since there are noise bands, strong solvent signals and artifacts in spectral regions close to the diagonal, which precludes the use of a constant noise level for the entire spectrum. Local noise level estimation is therefore an important part of the ATNOS peak picking algorithm, which in turn depends critically on defining regions of flat baseline in the experimental NMR spectra. To this end ATNOS makes use of elements of the FLATT algorithm (Güntert and Wüthrich, 1992), and of a technique for local noise level determination that was previously introduced in the automated peak picking algorithm AUTOPSY (Koradi et al., 1998).

(a) *Local baseline determination*

For each data point k the baseline-corrected signal intensity I^k_{bc} is obtained by subtracting from the experimentally measured intensity I^k an estimated value for the level of the baseline at the position k, I^k_b,

$$I^k_{bc} = I^k - I^k_b. \tag{1}$$

In Equation 1, the I^k values result from straightforward measurements in the experimental data, whereas a reliable determination of the local baseline level at the position of a signal peak, I^k_b, is not directly accessible and needs to be extrapolated from a detailed analysis of the 'pure-baseline regions' surrounding the peak, which is the subject of this section.

Baseline determination is performed separately for each 1D slice (rows and columns in 2D NMR spectra). First, regions of 'pure-baseline' (Güntert and Wüthrich, 1992) are identified on the assumption that a contiguous stretch of data points can be well fitted by a straight line only if it lies in a pure-baseline region. Therefore, for each data point with intensity I^k the average squared deviation from the baseline, p^k, is calculated as a fit to a straight line, $a + bl$, where l runs over a stretch of $2m + 1$ data points centered about the data point k, and m is fixed such that $2m+1$ data points

Table 1. Cycle-independent ATNOS parameters used in the structure calculations of this paper

Symbol	Parameter	Value
p_{seg}	Fraction of the total length of a 1D cross section used to evaluate the local noise amplitude (Equation 5)	5.0%
B_{min}	Minimal ratio of signal intensity to baseline level for NOESY peak picking (Equation 10)	1.5
d_{loc}	Radius of spectral region used for identification of local extrema (Equations 11 and 12)	0.01 ppm
r_{cut}	Determines the fraction of all covalent peaks used to determine the spectrum-specific signal-to-noise threshold (Equation 14)	97.5%
a_{cut}	Determines the fraction of all covalent peaks used to determine the spectrum-specific peak area threshold (Equation 15)	97.5%
R_{max}	Upper limit of the spectrum-specific threshold for the signal-to-noise ratio (Equation 16)	5.0
d_{diag}	Maximal distance from the diagonal for peaks classified to be close to the diagonal (Equation 21)	0.7 ppm (2D) 0.6 ppm (3D)
d_{solv}	Maximal distance from the solvent signal for peaks classified to be close to the solvent resonance (Equation 22)	0.05 ppm
g_{max}	Upper limit for the ratio between the minimal signal intensity along a straight line to a neighboring peak and the peak intensity (Equation 25)	0.8
f_{N}	Factor specifying a minimal intensity valley depth (Equation 26)	2.0
$\Delta\omega_1^{\mathrm{align}}$	Tolerance range for network-anchoring in the indirect ^{1}H dimension (Equation 28)	0.0025 ppm
$\Delta\omega_2^{\mathrm{align}}$	Tolerance range for network-anchoring in the direct ^{1}H dimension (Equation 28)	0.0025 ppm
$\Delta\omega_1^{\mathrm{sym}}$	Tolerance range for symmetric or transposed peak positions in the indirect ^{1}H dimension (Equation 29 and Equation 5 of Herrmann et al., 2002)	0.03 ppm
$\Delta\omega_2^{\mathrm{sym}}$	Tolerance range for symmetric or transposed peak positions in the direct ^{1}H dimension (Equation 29 and Equation 5 of Herrmann et al., 2002)	0.03 ppm
$\Delta\omega_3^{\mathrm{sym}}$	Tolerance range for symmetric or transposed peak positions in the ^{13}C or ^{15}N dimension (Equation 5 of Herrmann et al., 2002)	0.4 ppm
d_{tol}	Upper limit on acceptable violations of the maximum NOE observable distance d_{max} (Equation 30, Table 2)	0.25 Å
L_{vio}	Maximal acceptable number of violations of the upper distance limit between two atoms i and j in a bundle of L conformers (Equation 30)	$L/2$

correspond to 100 Hz (Table 1):

$$p^k = \min_{a,b} \sum_{l=-m}^{m} (I^{l+1} - a - bl)^2,\ m \geq 1. \qquad (2)$$

p^k becomes small if k is located in a pure-baseline region, and large if k is within a real or artifactual peak. Thus, all data points with p^k parameter values smaller than a threshold, p_{cut}, are considered to belong to pure-baseline regions, where the value for p_{cut} is adjusted individually for each 1D slice such that 50% of the cross section is attributed to pure-baseline regions. In addition, a maximum allowed gap width between pure-baseline regions of 5% of the length of the 1D slice is imposed. If the width of a gap exceeds this limit, the gap region will be separately searched for additional pure-baseline segments by increase of the p_{cut} value in steps of 33% of the value determined for the slice in question as described above. After identification of all pure-baseline regions in a 1D slice, the baseline levels across peak regions, I_b^{int}, are deter-

mined by linear interpolation between the signal levels of the nearestby data points in the two adjoining pure-baseline regions. If adjacent pure-baseline regions are separated by the diagonal or the solvent resonance (Figure 2), the best-fit straight line given by Equation 2 for the pure-baseline data point nearest to the perturbation is used to extrapolate the slope of the baseline in these spectral regions. The baseline value I_b^k at a given data point k in an n-dimensional spectrum is the largest one among the values of the baseline levels calculated for $i = 1, ..., n$ the slices through the data point,

$$I_b^k = \max_i(I_b^{k^i}), \tag{3}$$

with

$$I_b^{k^i} = \begin{cases} I^k & \text{if } k \text{ belongs to a pure-baseline region,} \\ I_b^{\text{int},k^i} & \text{if } k \text{ belongs to a peak region.} \end{cases} \tag{4}$$

(b) *Local noise level determination*

A noise level is determined separately for each 1D cross section (rows and columns in 2D NMR spectra). The standard deviation for the noise amplitude is calculated for each disjunctive segment, s_q, of the entire 1D cross section, S, and the size of s_q is given as a user-defined fraction of S, p_{seg} (Table 1),

$$|s_q| = p_{\text{seg}} \cdot |S| \ (q = 1, ..., m). \tag{5}$$

The minimal value of the standard deviations in all segments is accepted as the noise amplitude, δ_i, for the entire 1D slice, i (Koradi et al., 1998). The 'base noise level' of the entire spectrum, δ_b, is the minimum of the noise levels in all 1D slices,

$$\delta_b = \min_i(\delta_i). \tag{6}$$

The 'local noise level' at a given data point, N^k, is then computed as the sum of the base noise level for the entire spectrum and additional noise that may occur in the $i = 1, ..., n$ slices that pass through the data point k (Koradi et al., 1998),

$$N^k = \sqrt{\sum_{i=1}^{n} \delta_{k^i}^2 - (n-1)\delta_b^2}. \tag{7}$$

Finally, a 'global noise level' for the entire spectrum is defined as the average over all local noise levels,

$$\overline{N} = \frac{1}{M} \sum_{k=1}^{M} N^k, \tag{8}$$

where M is the number of data points in the spectrum.

NOESY peak picking based on minimal signal-to-noise and local extremum criteria

At the outset of the spectral analysis, ATNOS performs a peak picking of the NOESY spectra with the highly permissive criteria of requiring an initial minimal signal-to-noise ratio and a minimal local extremum condition. The resulting peak list will normally contain NOE cross peaks as well as artifacts. The subsequent refined spectral analysis is focussed on identifying true NOE signals in this comprehensive peak list.

(a) *Initial minimal signal-to-noise criterion*

A data point k with intensity I^k is considered to be part of a peak if the following two conditions are fulfilled:

$$\frac{|I_{bc}^k|}{N^k} \geq R_{\text{min}} \tag{9}$$

and

$$\left|\frac{I^k}{I_b^k}\right| \geq B_{\text{min}}. \tag{10}$$

In Equations 9 and 10, I_{bc}^k is the local baseline-corrected signal intensity (Equation 1), I_b^k is the baseline level at data point k (Equation 3), N^k is the local noise amplitude (Equation 7), and B_{min} and R_{min} are user-defined parameters (see Tables 1 and 2).

(b) *Local extremum condition*

A n-dimensional frequency domain NMR spectrum represents an n-dimensional grid of data points, with the unit grid length in each of the n dimensions given by the digital resolutions $\Delta\omega_i (i = 1, ..., n)$. Accordingly, a data point k with frequency coordinates $\vec{\omega}^k = (\omega_1^k, ..., \omega_n^k)$ and intensity $I^{\vec{\omega}^k} \equiv I^k$ is accepted to represent a local extremum if it satisfies either one of the conditions of Equations 11 and 12,

$$I^{\vec{\omega}^k} \geq I^{\vec{\omega}'} \text{ for all } \vec{\omega}' \text{ with } |\vec{\omega}^k - \vec{\omega}'| \leq d_{\text{loc}}, \tag{11}$$

$$I^{\vec{\omega}^k} \leq I^{\vec{\omega}'} \text{ for all } \vec{\omega}' \text{ with } |\vec{\omega}^k - \vec{\omega}'| \leq d_{\text{loc}}, \tag{12}$$

where d_{loc} is a user-defined parameter specifying the size of a localized spectral region centered about the data point considered (Table 1). In 2D NOESY spectra the spectral region characterized by Equations 11 and 12 corresponds to a circular plane, and in 3D spectra it has a spherical shape. Both the identification of a local extremum at point k and the final, precise location

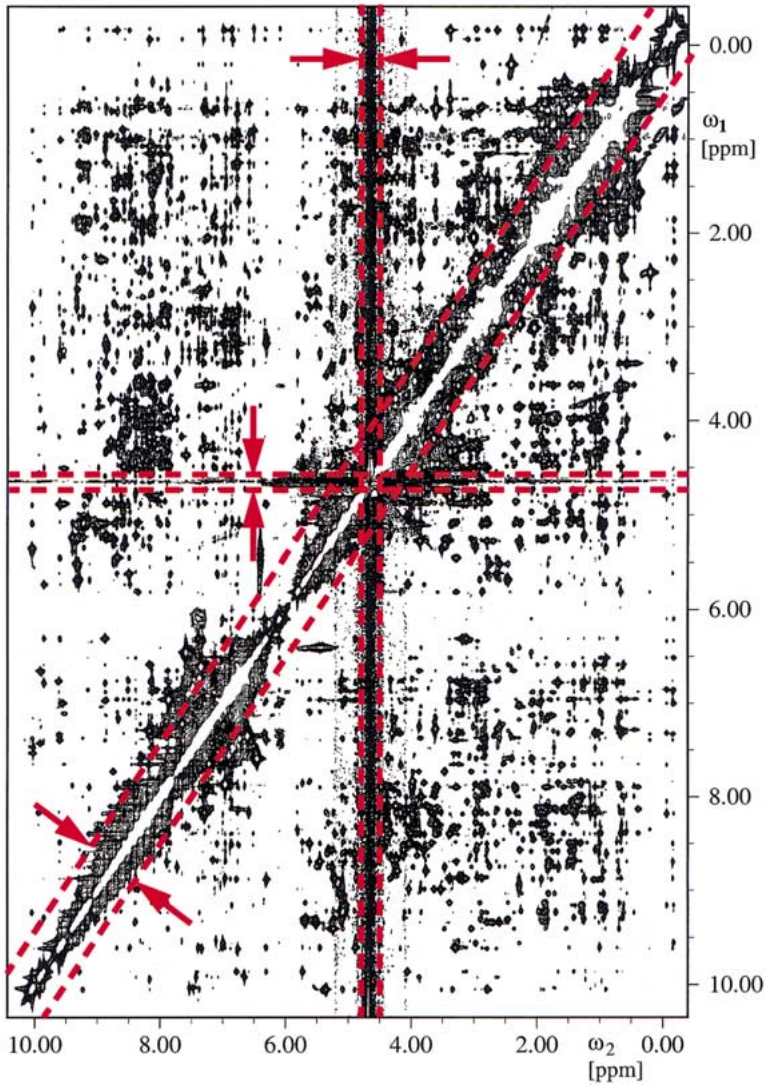

Figure 2. Illustration of the peak classification based on the location in the NOESY spectrum. Peaks located within a distance $d \leq d_{\text{diag}}$ from the diagonal, or a distance $d \leq d_{\text{solv}}$ from the solvent resonance, are discarded in the first ATNOS cycle. The interpretation of these spectral regions in the second and subsequent ATNOS cycles is then guided by reference to the intermediate protein three-dimensional structure (Figure 1).

Table 2. Cycle-dependent ATNOS parameters used in the structure calculations in this paper

Symbol	Parameter	Value in cycle i ($i = 1, ..., 7$)						
		1	2	3	4	5	6	7
R_{min}	Minimal value for signal-to-noise ratio (Equation 9)	3.5	3.25	3.0	3.0	3.0	3.0	3.0
d_{max}	Maximal value for NOE-observable ^{1}H–^{1}H distances (Equations 13, 20 and 30)	5.0	5.0	5.0	5.5	5.5	5.5	5.5
C_{min}	Minimal value of $C^p_{i,j}$ for acceptable chemical shift agreement of a peak (Equation 27)	0.9	0.8	0.8	0.8	0.8	0.8	0.8

of the extremum require higher resolution than that given by the digitalization of the spectrum. Therefore, a linear interpolation of the intensities between neighboring data points is used for the identification of the positions $\vec{\omega}'$, and a cubic spline interpolation is applied to more accurately determine the exact position of maximum signal intensity. This final interpolation step is of crucial importance, since all criteria based on chemical shift information make use of the peak position.

Identification and assignment of NOE cross peaks with covalent structure-imposed upper distance bounds

The fixed bond lengths, bond angles and chiralities of the covalent polypeptide structure impose NOE-observable upper limits on certain intraresidual and sequential ^{1}H–^{1}H distances (Güntert et al., 1998; Wüthrich, 1986; Wüthrich et al., 1983). In ATNOS, these conformation-independent upper limits, d_{ij}^{cov}, are computed analytically for atom pairs, i and j, that are separated by one or two dihedral angles. We then define 'covalent NOEs' such that the corresponding ^{1}H–^{1}H distances must satisfy Equation 13,

$$d_{ij}^{\mathrm{cov}} \leq d_{\max}, \tag{13}$$

where the parameter $d_{\max}$ is fixed by the user at a sufficiently short value so that all NOEs corresponding to $\mathrm{d}_{ij}^{\mathrm{cov}}$ should be observable in NOESY spectra (Table 2).

For all sequential and intraresidual NOEs a list of tentative chemical shift-based assignments is generated, using the same procedure as in the CANDID algorithm (Herrmann et al., 2002). All signals in the ATNOS-picked NOESY peak list which have at least one assignment to a covalent NOE are then used to derive spectrum-specific threshold values for minimal signal-to-noise ratio, R_{min}^{S} (Equation 16), and minimal peak area, A_{min}^{S} (Equation 17) for use in further discrimination of the entries in the NOESY peak list.

Spectrum-specific threshold values derived from the covalent NOE cross peaks

For refined spectral analysis, the signal-to-noise criterion used for the NOESY peak picking (Equation 9) is substituted by threshold values for minimal signal-to-noise ratio and peak area, which are adapted individually for each NOESY spectrum. The peak area A is defined as the number of data points in a contiguous region about a local extremum that have a signal-to-noise ratio larger than R_{min} (Equation 9). If $\{R_1, ..., R_M\}$ and $\{A_1, ..., A_M\}$ are the values for the signal-to-noise ratio and the peak area in the set of M covalent peaks, respectively, then $\tilde{R} \in \{R_1, ..., R_M\}$ and $\tilde{A} \in \{A_1, ..., A_M\}$ are chosen such that

$$\left|\{R \in \{R_1, ..., R_M\} \text{ and } R \geq \tilde{R}\}\right| = r_{\mathrm{cut}} \cdot M, \tag{14}$$

and

$$\left|\{A \in \{A_1, ..., A_M\} \text{ and } A \geq \tilde{A}\}\right| = a_{\mathrm{cut}} \cdot M, \tag{15}$$

where r_{cut} and a_{cut} are user-defined fractions of M (Table 1). The spectrum-specific threshold value for the signal-to-noise ratio, R_{min}^{S}, is then computed as

$$R_{\mathrm{min}}^{S} = \min(\max(\tilde{R}, R_{\mathrm{min}}), R_{\max}), \tag{16}$$

where R_{min} is defined as in Equation 9, and $R_{\max}$ imposes an upper limit for R_{min}^{S} (Table 1). The spectrum-specific threshold value for the peak area is given by

$$A_{\mathrm{min}}^{S} = \tilde{A}. \tag{17}$$

Identification of potential NOE cross peaks based on the spectrum-specific threshold values for minimal signal-to-noise ratio and peak area

Using the spectrum-specific threshold values R_{min}^{S} and A_{min}^{S}, a list of potential NOE cross peaks is selected from the initial NOESY cross peak list by retaining only signals centered about the data points k with intensity I^k that satisfy at least one of the two conditions of Equations 18 and 19,

$$\frac{\left|I_{bc}^{k}\right|}{N^k} \geq R_{\mathrm{min}}^{S}, \tag{18}$$

$$A^k \geq A_{\mathrm{min}}^{S}. \tag{19}$$

Adapting the chemical shift values to the individual NOESY spectra

The values in the input chemical shift lists may need to be adjusted for the individual NOESY spectra, since different NMR spectra are typically used for obtaining sequence-specific resonance assignments and for the collection of conformational constraints, respectively. Therefore, an 'adapted chemical shift list' is produced for each NOESY spectrum, which represents the input for peak discrimination by ATNOS based on chemical

shift information (Figure 1), and subsequently also for the NOE assignment using CANDID (Herrmann et al., 2002).

For the chemical shift adaptation we assume in the first ATNOS cycle that the distances d_{ij} for all atom pairs, i and j, in a spectrum S are set to infinity, $d_{ij} = \infty$, unless a short distance, d_{ij}^{cov}, is imposed by the covalent polypeptide structure of the protein. In the second and subsequent ATNOS cycles, when an intermediate protein three-dimensional structure is available, d_{ij} values are in addition also calculated as the average of the distances between atom pairs i and j in the bundle of conformers, d_{ij}^{str}. Only atom pairs are retained that satisfy Equation 20,

$$d_{ij}^{\mathrm{cov}} \leq d_{\max} \text{ or } d_{ij}^{\mathrm{str}} \leq d_{\max}, \tag{20}$$

where $d_{\max}$ is a user-defined parameter (Table 2). Each of these atom pairs is then attributed to the potential NOE peak (Figure 1) that best fits its chemical shift values, whereby a Gaussian weighting factor, C_{ij}^{p}, is applied as defined in Equation 4 of Herrmann et al. (2002). If more than one atom pair is thus assigned to the same potential NOE peak, only the top-ranked atom pair is retained, using the following ranking criteria: First, atom pairs with covalent structure-imposed short upper distance limits precede other atom pairs with short distances. Second, within each of these two groups of atom pairs a ranking is made based on closeness of the chemical shift fits, as measured by C_{ij}^{p}. This discrimination in favor of assignments that should yield observable NOEs in all possible conformations of the protein corresponds to the common treatment of short-range and certain medium-range ^{1}H–^{1}H connectivities by experienced spectroscopists in the course of interactive peak identification and resonance assignment (Wüthrich, 1986). Here, the ensemble of all peaks assigned by this strategy is used to adjust the chemical shift lists to the NOESY spectrum considered.

Classification of the NOE cross peaks into classes with different validation criteria

Experience with early versions of ATNOS showed that its performance for identification of NOE cross peaks could be improved decisively by grouping the NOESY peaks into different classes, for which different criteria would then be applied for the validation of potential NOE cross peaks (Figure 1). ATNOS includes automatic routines for this peak classification that enable the use of a selection of alternative filtering procedures for the NOE validation in the different classes (following section).

Two conceptually different criteria are applied for the peak classification, of which one considers the relations to the protein structure, and the other one the location in the NMR spectrum. The first criterion distinguishes between peaks with and without a covalent structure-imposed NOE-observable upper distance limit. Among these two groups, artifacts that might erroneously be identified as NOE cross peaks with covalent structure-limited upper distance bounds would have a more limited impact on the protein structure, and therefore less stringent filtering may be applied for their validation than for the other peaks. The second criterion distinguishes between peaks located within a maximal distance either from the diagonal, d_{diag}, or the solvent resonance, d_{solv}, and all other peaks (Figure 2; see Table 1 for the parameter values used). The group of peaks close to the diagonal or the solvent line satisfies the relations 21 or 22,

$$\frac{\left|\omega_1^p - \omega_2^p\right|}{\sqrt{2}} \leq d_{\mathrm{diag}}, \tag{21}$$

$$\left|\omega_i^p - \omega^{\mathrm{solv}}\right| \leq d_{\mathrm{solv}} \; (i = 1, 2), \tag{22}$$

where $\omega_i^p (i = 1, 2)$ are the positions of peak p in the two ^{1}H dimensions, and ω^{solv} is the position of the solvent line. These spectral regions are analyzed only in the ATNOS cycles 2, 3, ..., since they usually contain an abundance of artifactual peaks.

Overall, in the first ATNOS cycle three classes of peaks are distinguished, whereby peaks close to the diagonal or the solvent (Equations 21 and 22) are discarded, and peaks located outside of these areas are divided into those representing covalent structure-limited distances, and all others. From the second ATNOS cycle onwards, the entire spectrum is used and only two classes of peaks are distinguished, i.e., peaks compatible with the intermediate three-dimensional protein structure, and all others.

Criteria for the multipass-filtering validation of potential NOE cross peaks

This section describes the filters used for the validation of potential NOE peaks (Table 3). In the ATNOS cycle 1, different multipass-filtering is applied to the covalent NOE peaks and to the other peaks, respectively, that are neither close to the diagonal nor to the solvent line. In the second and subsequent ATNOS cycles, all potential NOE peaks are accepted that are

Table 3. Multipass-filtering used for the validation of potential NOE cross peaks[a]

Criterion[b]	ATNOS cycle 1[c]		ATNOS cycles 2, 3, ...[d]
	Covalent peaks	Other peaks	All peaks
Noise band location	yes	yes	yes
Peak separation	–	yes	–
Chemical shifts	–	(yes)[e]	yes
Network-anchoring	–	(yes)[e]	–
Symmetry of NOESY	–	yes	–
Protein structure	–	–	yes

[a]To be accepted as a NOE cross peak, a peak must satisfy all the criteria indicated with 'yes'.
[b]See text for details.
[c]In ATNOS cycle 1, peaks near the diagonal or near the solvent resonance are discarded (Figure 2), and the remaining peaks are divided up into two groups, i.e., covalent NOE peaks with a covalent structure-imposed corresponding upper distance limit, and all other NOE peaks.
[d]In the second and subsequent ATNOS cycles, all potential NOE peaks that do not satisfy the three required conditions are again subjected to the treatment of cycle 1.
[e]Only one of these two criteria needs to be satisfied.

compatible with the adapted chemical shift list and the intermediate protein three-dimensional structure, and which are not part of a noise band. All other peaks are given another chance in that they are subjected again to the same multipass-filtering as in the ATNOS cycle 1. Clearly, this procedure relies critically on good quality of the intermediate protein three-dimensional structures, and on careful adaptation of the chemical shift list to the NOESY spectrum considered.

(a) *Noise band filter*
A peak p centered about the data point k is considered to belong to a noise band in a NOESY spectrum, if

$$N^k \geq 3 \cdot \overline{N}, \tag{23}$$

where N^k (Equation 7) and $\overline{N}$ (Equation 8) are the local noise level and the global noise level of the spectrum. All peaks thus attributed to a noise band are discarded from further consideration during the same cycle of calculation.

(b) *Peak separation*
In the initial NOESY peak picking (Figure 1), potential NOE cross peaks were identified as local extrema without consideration of other, nearby local extrema. This contrasts with the strategy of an experienced spectroscopist, who will make use also of information contained in the surrounding data points. Therefore, to distinguish real NOE cross peaks from artifacts which may, for example, be caused by signal distortion in the 'tail' of a real peak in an adjoining spectral plane, ATNOS now analyzes also all data points around these local extrema. To this end, the spectrum is segmented around each extremum p into 'peak areas', A^p, consisting of contiguous regions of data points with signal-to-noise ratios larger than $R_{\min}$ (Equation 9). A peak p is considered to be 'separated' if it satisfies Equation 24,

$$I^p > I^{p'_i},\ p, p'_i \in A^p; i = 1, ..., t, \tag{24}$$

where $p'_i \in A^p$ are the t additional extrema in the peak area of p that satisfy the criteria of Equations 11 or 12. Otherwise, the relation of the extremum p to all other extrema p'_i needs to be further evaluated. First, the smallest intensity along the straight line that connects the local maximum p and another local maximum p'_i, $I_{\min}(p, p'_i)$, is determined. Then the peak, p, is considered to be separated, if

$$\left|(I_{\min}(p, p'_i))/I^p\right| \leq g_{\max} \text{ for all } p'_i \in A^p; i = 1, ..., t, \tag{25}$$

and

$$\left|I^p - I_{\min}(p, p'_i)\right| \geq f_N \cdot N^k \text{ for all } p'_i \in A^p; i = 1, ..., t. \tag{26}$$

I^p is the intensity of peak p, $\{p'_i \in A^p; i = 1, ..., t\}$ are a set of t local extrema within the peak area, N^k is the local noise level of peak p at data point k (Equation 7), and $g_{\max}$ and f_N are user-defined parameters (Table 1). The condition of Equation 26 ensures that peaks with small signal-to-noise ratio are also separated by a 'minimal intensity valley depth'.

(c) *Chemical shift compatibility filter*

Consider that a grid spanned by the chemical shifts of all atoms with sequence-specific assignments is overlaid onto the NOESY spectrum. Potential NOE cross peaks that coincide closely with a grid point are then more likely to be real peaks than those that are located between grid points. Peak discrimination in favor of peaks close to a chemical shift grid point mimics the typical approach of an experienced spectroscopist to perform interactive peak picking in conjunction with resonance assignment. In ATNOS, the agreement between the position of a peak, p, and the chemical shift grid is quantified by the aforementioned $C_{i,j}^{p}$ value, where peaks that satisfy Equation 27 are considered to be compatible with the chemical shift list,

$$C_{i,j}^{p} \geq C_{\min}. \tag{27}$$

$C_{i,j}^{p}$ is a Gaussian weighting factor that has a value of 1.0 for a perfect fit (Herrmann et al., 2002), and $C_{\min}$ is a user-defined parameter (Table 2).

(d) *Network-anchoring filter*

The concept of network-anchoring as originally introduced for discriminating between multiple initial assignments of a NOE cross peak (Herrmann et al., 2002) is based on the consideration that the correctly assigned NOE distance constraints form a self-consistent network that is compatible with the protein three-dimensional structure. ATNOS uses a simplified form of network-anchoring for the validation of potential NOE cross peaks. Thereby a peak p of a 2D NOESY spectrum is considered to be network-anchored if there are at least two other potential NOE peaks, p' and p'', in both proton dimensions i such that

$$\left|\omega_i^{p} - \omega_i^{p'}\right| \leq \Delta\omega_i^{\text{align}} \text{ and } \left|\omega_i^{p} - \omega_i^{p''}\right| \leq \Delta\omega_i^{\text{align}} \ (i = 1, 2), \tag{28}$$

where $\Delta\omega_i^{\text{align}}$ is a user-defined tolerance range (Table 1).

In 3D heteronuclear-resolved [^{1}H,^{1}H]-NOESY spectra, the additional peak separation afforded by the ^{13}C or ^{15}N frequency dimension is used to define a more stringent alignment criterion than is possible for 2D [^{1}H,^{1}H]-NOESY spectra: A peak p is considered to be network-anchored only if it is aligned with at least two other peaks along the direct proton dimension within the same heavy atom plane.

(e) *Symmetry filter*

Standard NOESY spectra are intrinsically symmetric with regard to their diagonal (Anil-Kumar et al., 1980), so that detecting pairs of symmetry-related peaks on both sides of the diagonal represents support for correct NOE cross peak identification. Since in practice the peaks in symmetry-related positions may have significantly different intensities, a permissive symmetry-related filter is used for peak validation. In a 2D NOESY spectrum, a symmetry-related NMR signal at data point k within a tolerance range about the mirrored position p^T of peak p, $\Delta\omega_i^{\text{sym}}$ ($i = 1, 2$), is accepted if the data point k satisfies the Equations 9 and 29,

$$\left|\omega_i^{p^T} - \omega_i^{k}\right| \leq \Delta\omega_i^{\text{sym}} \ (i = 1, 2), \tag{29}$$

where $\Delta\omega_i^{\text{sym}}$ are user-defined parameters (Table 1).

In 3D heteronuclear-resolved [^{1}H,^{1}H]-NOESY spectra, a similar symmetry filter can be applied by using an additional tolerance range for the ^{13}C or ^{15}N chemical shifts, $\Delta\omega_3^{\text{sym}}$ (Table 1), to determine the position of the transposed peak (Equation 5 of Herrmann et al., 2002).

(f) *Protein three-dimensional structure compatibility filter (cycles 2, 3, ...)*

Compatibility with the corresponding ^{1}H–^{1}H distances in the intermediate protein three-dimensional structure is a critical criterion for NOE cross peak validation, which is implicitly related to the aforementioned network-anchoring. A peak is considered to be compatible with the intermediate bundle of L conformers if for at least one initial NOE assignment to an atom pair, i and j, the condition of Equation 30 is satisfied:

$$\sum_{l=1}^{L} \Theta(d_{i,j}^{l} - (d_{\max} + d_{\text{tol}})) \leq L_{\text{vio}}. \tag{30}$$

$\Theta\ (x)$ is the Heavyside function, which has the values 0 for $x < 0$ or 1 for $x \geq 0$. $d_{i,j}^{l}$ is the distance of an atom pair, i and j, in conformer l, $d_{\max}$ is a user-defined upper distance bound (Table 2), d_{tol} is a user-defined additional tolerance distance (Table 1), and L_{vio} is a user-defined parameter specifying a fraction of all conformers (Table 1). Equation 30 requires that the number of conformers for which the upper distance limit for observable NOEs between two atoms i and j, $d_{\max}$, is violated by more than d_{tol} does not exceed the predetermined value of L_{vio}.

Automated combined NOE assignment and protein three-dimensional structure calculation using CANDID and DYANA

This part of the protein structure determination (Figure 1) follows exactly the recent publication on the software CANDID (Herrmann et al., 2002). In each cycle the updated NOE peak list and the adjusted chemical shift list resulting from the ATNOS analysis of the NOESY spectra are used as input for the CANDID algorithm. CANDID performs automated NOE assignment and distance calibration of NOE intensities, and thus generates an updated input of NOE upper distance constraints for the next structure calculation with DYANA (Güntert et al., 1997). This input can be further supplemented with additional conformational constraints (Figure 1).

Experimental methods

Data used for the validation of automated ATNOS peak picking of NOESY spectra and NOE cross peak identification

For the evaluation of the performance of ATNOS we used experimental NMR data sets of three proteins (Table 4) for which high-quality NMR structures had previously been determined (Protein Data Bank entries: CopZ, 1CPZ; WmKT, 1WKT; $BmPBP^A$, 1GM0) using interactive NOESY peak picking. For all three proteins nearly complete sequence-specific resonance assignments for the backbone and the side-chains are available (BioMagResBank accession codes: CopZ, 4344; WmKT, 5255; $BmPBP^A$, 4849). All NOESY spectra used for the previous structure determinations were also used here (Table 4).

The present validation of NOESY peak picking and NOE cross peak identification by ATNOS is based on using the resulting peak lists as input for automated NOE assignment with CANDID (Herrmann et al., 2002), which in turn generates an input of NOE distance constraints for the program DYANA for protein structure calculation (Güntert et al., 1997). The calculations with the softwares CANDID and DYANA were performed identically as in Herrmann et al. (2002).

For CopZ and WmKT, experimentally determined ^3J-coupling constants were used as supplementary input for DYANA in the same way as in the reference structure determinations. In each ATNOS/CANDID/DYANA cycle these scalar coupling constants were combined with the updated list of upper limit intraresidual and sequential NOE distance constraints and converted into torsion angle constraints by the grid search procedure FOUND (Güntert et al., 1998). Stereospecific assignments of diastereotopic pairs of protons or methyl groups from the reference structure determinations were not included into the input for the new ATNOS/CANDID/DYANA structure determination. Each disulfide bridge was constrained by a standard set of three upper and three lower distance constraints (Williamson et al., 1985), which were added in all cycles to the input for DYANA.

The structures obtained with ATNOS are compared with reference structures based on interactive NOE cross peak identification. Otherwise, identical protocols of automated NOE assignment and protein structure calculation were used for all the structure determinations. These reference structures are the result of either the original *de novo* structure determination with CANDID and DYANA ($BmPBP^A$; Horst et al., 2001), or recalculations of the structure from the original input data using CANDID and DYANA (CopZ and WmKT; Herrmann et al., 2002).

Standard protocol used for automated structure determination using ATNOS

The calculations comprised seven iterative cycles of NOESY peak picking and NOE cross peak identification with ATNOS, automated NOE assignment with CANDID (Herrmann et al., 2002), 3D structure calculation with DYANA (Güntert et al., 1997), and energy-refinement with OPALp (Luginbühl et al., 1996; Koradi et al., 2000). This protocol corresponds to the standard protocol for CANDID and DYANA as described in detail by Herrmann et al. (2002), except that the cycle-invariant chemical shift and peak lists in the input for CANDID are replaced by ATNOS chemical shift and peak lists that are updated in each cycle by a new, protein structure-guided search of the experimental NOESY spectra.

Computations were performed on shared-memory multiprocessor SGI computers using four R12000 processors in parallel for the structure calculations. The computation time for a complete automated structure determination with ATNOS, CANDID and DYANA ranged from 3.9 h for CopZ to 8.3 h for $BmPBP^A$ on a single processor, and was spent predominantly with the DYANA structure calculations of,

Table 4. Experimental chemical shift assignments and NOESY spectra recorded to obtain the conformational constraints for the structure determinations of three proteins that have been used in this paper to validate automated ATNOS peak picking and NOE identification

Protein[a]	Size (residues)	Assigned chemical shifts (%)[b]	NOESY spectra[c]	Digital resolution (Hz)[d]
CopZ	68	94.9	2D, H_2O	8.2; 2.0
			3D (^{15}N), H_2O	19.1; 9.5; 26.1
WmKT	88	97.0	2D, H_2O	4.9; 2.4
$BmPBP^A$	142	97.1	3D (^{15}N), H_2O	18.4; 4.6; 35.6
			3D (^{13}C), H_2O	35.1; 4.4; 101.9
			3D ($^{13}C^{arom}$), D_2O	17.6; 4.6; 36.8

[a]CopZ: Apo-form of the copper chaperone Z (Wimmer et al., 1999); WmKT: Killer toxin from the yeast *Williopsis mrakii* (Antuch et al., 1996); $BmPBP^A$: A-form of the pheromone-binding protein from the silkworm *Bombyx mori* (Horst et al., 2001).
[b]Percent of the total number of non-labile hydrogen atoms and backbone amide protons for which the chemical shifts are known from the sequence-specific assignment. Pairs of diastereotopic protons or methyl groups are considered to be assigned when at least one of the two ^{1}H chemical shifts is known.
[c]Notation used: 2D, two-dimensional [^{1}H,^{1}H]-NOESY; H_2O, solvent of 95% H_2O / 5% D_2O; D_2O, solvent of 100% D_2O; 3D(^{15}N), 3D ^{15}N-resolved [^{1}H,^{1}H]-NOESY; 3D(^{13}C), 3D ($^{13}C^{arom}$), three-dimensional ^{13}C-resolved [^{1}H,^{1}H]-NOESY with the ^{13}C carrier frequency in the aliphatic or aromatic region, respectively.
[d]The first two numbers give the digital resolution in the indirect and direct proton dimensions, respectively. The third number gives the digital resolution in the ^{13}C or ^{15}N dimension. All NMR spectra were recorded at a ^{1}H frequency of 750 MHz, except for the 2D [^{1}H,^{1}H]-NOESY spectrum of CopZ, which was measured at a proton frequency of 600 MHz.

in total, 460 conformers per structure determination, i.e., 80, 80, 60, 60, 60, 60 and 60 in the cycles 1 to 7.

Structure analysis and comparison

Root mean square deviation (RMSD) values are used for two different types of comparisons: The RMSD of a bundle of n conformers is the average of the n RMSD values between the individual conformers and their mean coordinates, which are obtained by superimposing conformers 2, ..., n onto the first conformer for minimal RMSD of the backbone atoms N, C^α and C', and subsequent calculation of the arithmetic average of the Cartesian coordinates. The RMSD between two mean structures is the RMSD value between the mean coordinates of two bundles of conformers, for example, corresponding bundles obtained using NOE cross peak identification either by ATNOS or by an interactive approach. Both types of RMSD values were calculated for the well-defined polypeptide segments identified in the original structure determinations and used also for the reference structure determinations (Table 5). The program MOLMOL (Koradi et al., 1996) was used to visualize the three-dimensional protein structures and for the calculation of RMSD values.

Implementation of ATNOS

The software ATNOS was written in standard Fortran-77 as an independent module within the data structures and the framework of the user interface of the program DYANA (Güntert et al., 1997). Multidimensional NMR spectra are read using the input routine from the program PROSA (Güntert et al., 1992). In the present implementation it is used in combination with the softwares CANDID and DYANA. Future plans are to combine ATNOS and CANDID into one autonomous software package for use in conjunction with a selection of the commonly used structure calculation algorithms, such as XPLOR, CNS and DYANA (Brünger, 1992; Brünger et al., 1998; Güntert et al., 1997).

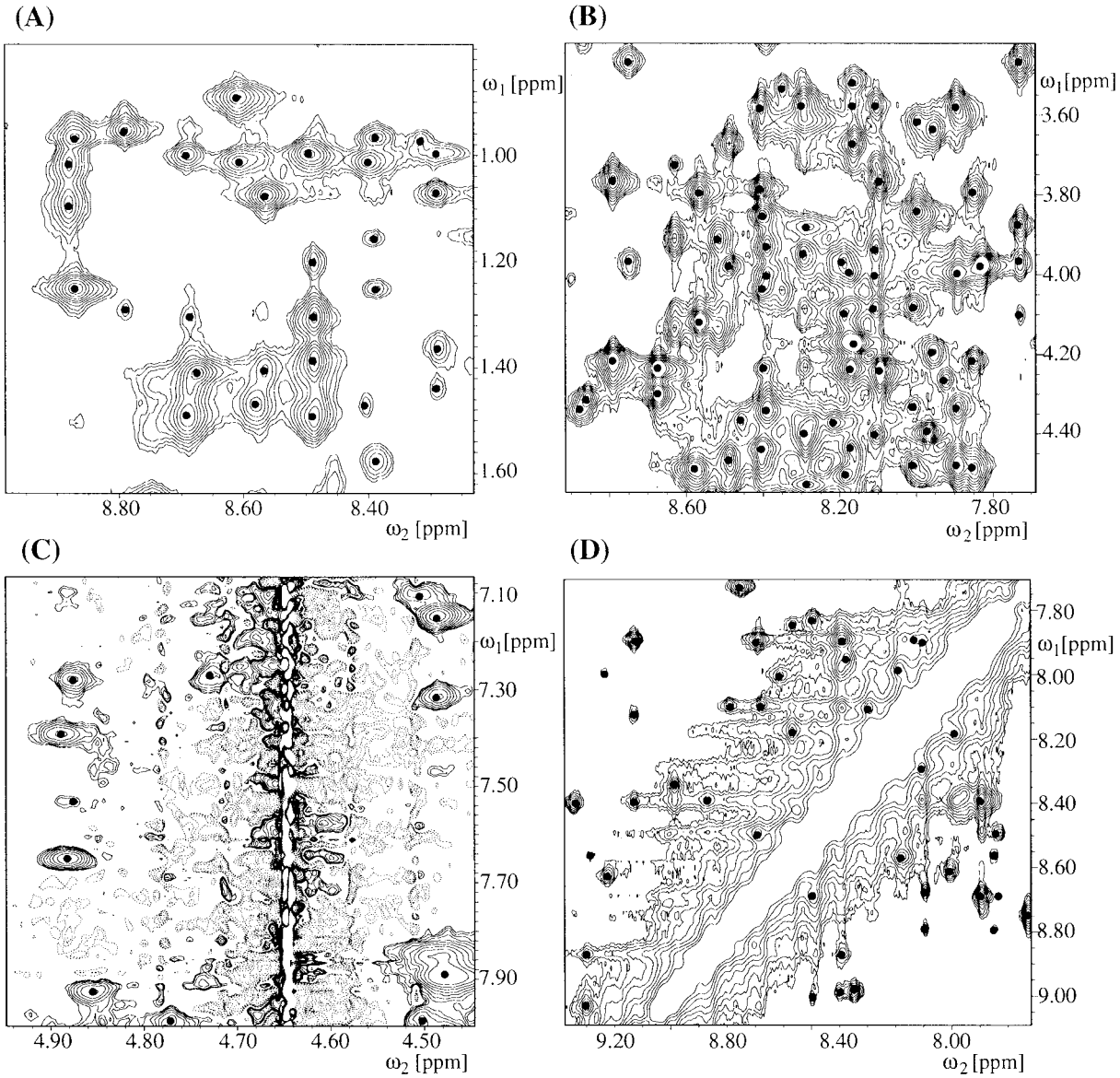

Figure 3. Representative regions of the 2D [^{1}H,^{1}H]-NOESY spectrum of the protein WmKT. Automatically picked NOE cross peaks for which a unique assignment has been retained after the ATNOS/CANDID/DYANA cycle 7 are marked with dots. (A) Region with little signal overlap; (B) region with high density of NOE cross peaks but without major perturbations by the solvent or by diagonal peaks; (C) region including the water line at $\omega_2 = 4.65$ ppm; (D) region including the diagonal.

Results

Validation of ATNOS with experimental NMR data sets

To assess the potential of the presently introduced ATNOS approach for automation of NOESY peak picking and NOE cross peak identification, we used the experimental NMR data of three proteins for which high-quality structures had previously been obtained with interactive NOE cross peak identification and otherwise identical protocols for the structure determination (Table 4; see also Experimental methods). The three proteins represent different molecular sizes and different secondary structure types. Different isotope labeling strategies had been used for the three structure determinations, i.e., natural isotope abundance, uniform ^{15}N-labeling, or uniform ^{13}C,^{15}N-labeling, and

the conformational constraints were collected from different types of homonuclear 2D and heteronuclear-resolved 3D NOESY experiments carried out at different ^{1}H frequencies (see Table 4 and Experimental methods).

For all test calculations described in this paper, the same NOESY spectra were used as for the reference structure determinations. Except in cycle 1 (see Equations 21 and 22) the entire NOESY spectra were evaluated with ATNOS. Figure 3 shows representative spectral regions from the 2D [^{1}H,^{1}H]-NOESY spectrum of CopZ with indication of the peaks identified by ATNOS and assigned by CANDID in cycle 7 (similar results were obtained for the other spectra of Table 4). The peak picking algorithm worked reliably in spectral regions with moderate (Figure 3A) and strong signal overlap (Figure 3B). Near the waterline (Figure 3C) and the diagonal (Figure 3D), ATNOS correctly identified NOE cross peaks without including artifactual perturbations into the peak lists, but peak picking in these areas was started only with cycle 2 of the ATNOS/CANDID/DYANA calculations and relied heavily on reference to the intermediate protein structure from the previous cycle (see Equations 21 and 22, and the text preceding these equations). The results of the structure calculations are listed in Table 5. For all three test proteins a low final DYANA target function value and a small RMSD value for the final bundle of 20 conformers were obtained. The evolution of characteristic output data from ATNOS/CANDID/DYANA is depicted in Figure 4. The increase of the number of NOE cross peaks identified and of the conformationally meaningful NOE upper distance constraints between the first and second cycles (Figures 4A and 4B) reflects that the regions near the diagonal and the solvent line (Figure 2) were added for the spectral analysis in cycle 2, and that additional protein structure-based information was available in cycle 2 to guide the analysis of the NOESY spectra (Figure 1). These numbers are nearly constant after the second cycle (Figures 4A and 4B), reflecting that the correct protein fold had already been found after the first cycle of calculation (Figure 5A), whereby the slight decrease of the number of NOE distance constraints during the later cycles is caused by the increasing stringency of the filtering criteria for NOE assignment with CANDID (Herrmann et al., 2002). The variation of the residual DYANA target function values from cycles 1 to 7 (Figure 4C) is indicative of a high-quality performance of the automated ATNOS NOE cross peak identification in conjunction

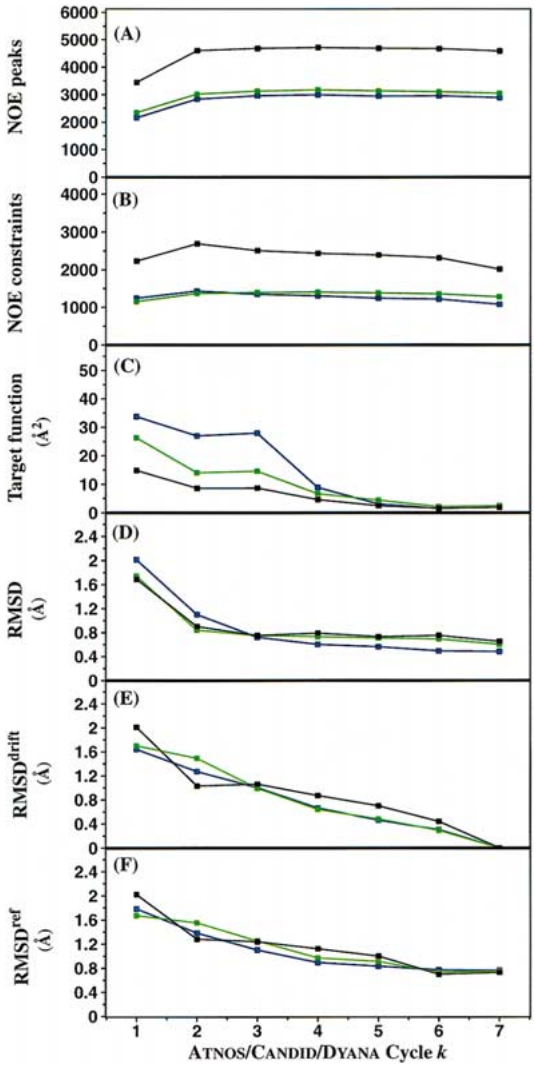

Figure 4. Evolution of characteristic parameters for the automated NMR structure determinations with ATNOS peak picking of the NOESY spectra and NOE cross peak identification, CANDID NOE assignment and DYANA structure calculation for the three proteins CopZ (blue), WmKT (green), and BmPBP^A (black). (A) Number of NOE cross peaks identified by ATNOS and assigned to non-diagonal proton pairs by CANDID. (B) Number of NOE upper distance constraints in the input for the structure calculation. (C) Average final target function value for the bundle of conformers representing the result of the DYANA structure calculation. (D) RMSD, calculated as the average of the RMSD values between the individual conformers in the bundles and their mean coordinates. (E) RMSD^{drift}, calculated as the RMSD between the mean coordinates of the bundles of conformers obtained after the k-th and the seventh cycle. (F) RMSD^{ref}, calculated between the mean coordinates of the bundle of conformers obtained after the k-th cycle and the bundle of conformers used to represent the result of the reference structure determination. All RMSD values are calculated for the backbone atoms N, C^{α} and C' of the well defined polypeptide segments identified in Table 5, footnote d.

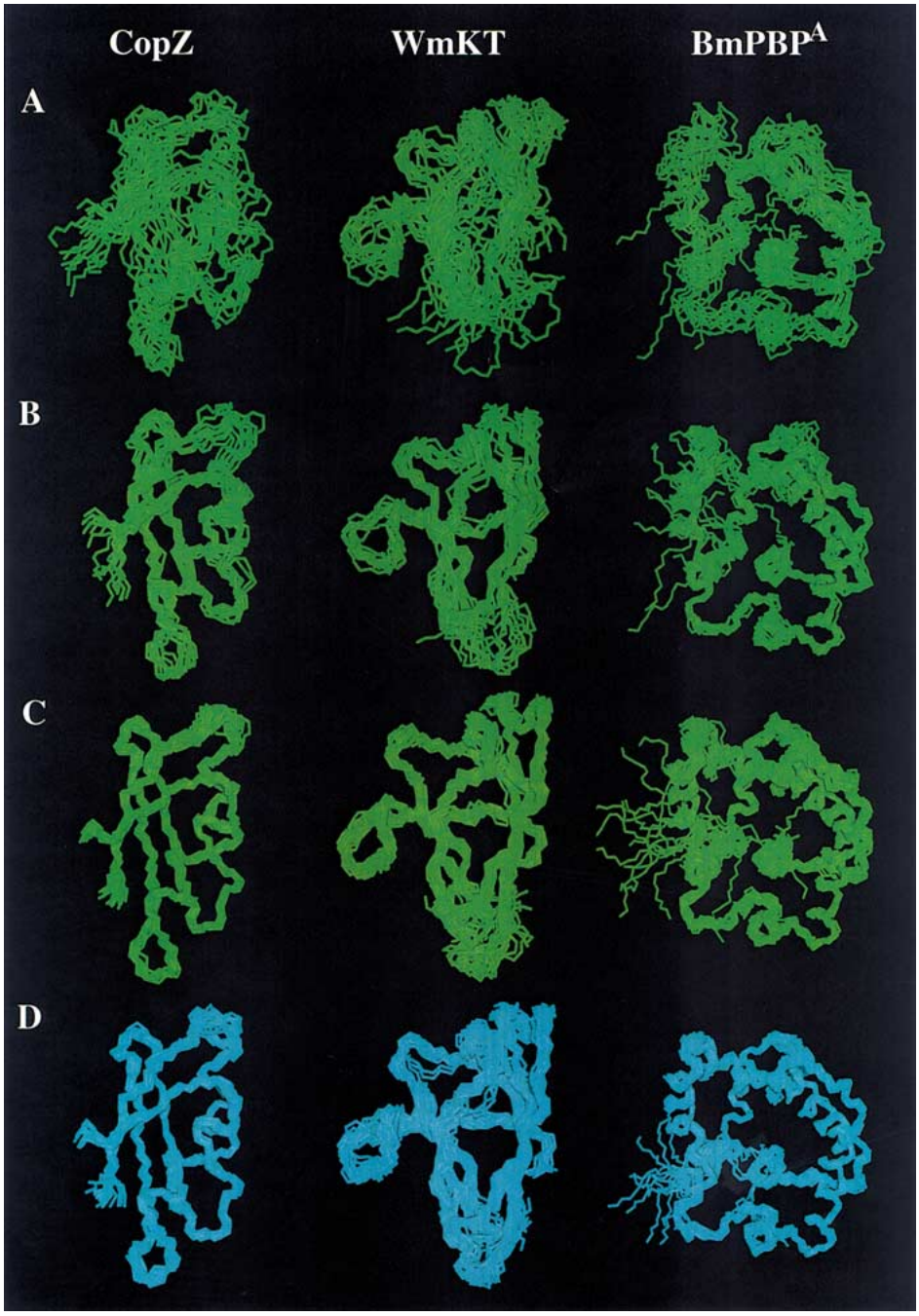

Figure 5. Bundles of conformers of the three proteins used for the validation of ATNOS peak picking. (A) Result of cycle 1 (10 conformers). (B) Result of cycle 2 (10 conformers). (C) Final structure after cycle 7 (20 conformers after energy-refinement). (D) reference structure determination based on interactive NOE cross peak identification (20 conformers after energy-refinement).

Table 5. Experimental input for the final structure calculations of the three proteins of Table 4 and statistics of the structure determinations based on NOE cross peak identification either by ATNOS or by an interactive approach

Quantity	CopZ	WmKT	$BmPBP^A$
ATNOS peak picking:			
NOE cross peaks assigned[a]	2094	3049	1236
	795		2971
			389
NOE upper distance limits[b]	1074	1274	2012
Residual DYANA target function value ($Å^2$)[c]	1.99	2.40	1.88
RMSD (Å)[d]	0.48	0.61	0.66
Interactive peak picking:			
NOE cross peaks assigned[a]	1025	1865	1137
	873		3133
			232
NOE upper distance limits[b]	937	1223	2109
Residual DYANA target function value ($Å^2$)[c]	1.56	1.74	1.27
RMSD (Å)[d]	0.53	0.76	0.46
Comparisons of ATNOS-based structures with reference structures:			
RMSD between mean structures (Å)	0.72	0.66	0.70
Ramachandran plot statistics:[e]			
Most favored regions (%)	76 / 85	64 / 65	79 / 84
Additional allowed regions (%)	20 / 12	30 / 28	17 / 15
Generously allowed regions (%)	3 / 2	5 / 5	2 / 1
Disallowed regions (%)	1 / 1	2 / 2	2 / 0

[a]Based on the experimental input data of Table 4. From top to bottom the number of assigned NOE cross peaks given for each protein corresponds to the NOESY spectra in Table 4. Counted are NOE cross peaks that have been identified by ATNOS and assigned to a specified non-diagonal proton pair by CANDID.
[b]Number of NOE upper distance limits that represent conformational restraints on the polypeptide fold.
[c]The residual DYANA target function value is the average for the bundles of conformers representing the NMR structure. The target function values before energy minimization are given.
[d]The RMSD is the average of the RMSD values between the individual conformers in the bundle and their mean coordinates for the backbone atoms N, C^α and C' of residues 2–67 for CopZ, 4–39 and 47–87 for WmKT, and 10–140 for $BmPBP^A$. The RMSD values after energy minimization are given.
[e]As determined by PROCHECK. The first number indicates the value calculated by PROCHECK (Morris et al., 1992) for the ATNOS-based structure, and the second number is the value for the reference structure.

with the automated CANDID NOE assignment. The evolution of the RMSD values for the bundle of conformers (Figure 4D) shows that the structure obtained after cycle 2 is nearly as precisely defined as the final structure (Figures 5B and 5C), which in turn is only possible if the first cycle already leads to the correct fold (Figures 5A and 5B). Indeed, the further guiding of the NOESY peak picking with the intermediate structure bundle from cycle 2 onwards leads to results that show the structures of cycle 2 and the final cycle to coincide closely also in terms of accuracy (Figures 5B–D). The ATNOS procedure is also reliably stable in the sense that the $RMSD^{drift}$ values decrease monotonously towards the final structure during the seven cycles of calculation (Figure 4E). Furthermore, for all three proteins of Table 4, the RMSD between the mean structure after cycle 1 and the mean reference structure are below 2.0 Å (Figure 4F). That the correct fold of the protein is obtained in the first cycle is crucial for reliable and robust automated structure

determination, since in the later cycles the analysis of the NOESY spectra with ATNOS and the automated NOE assignments with CANDID (Herrmann et al., 2002) are both dominantly driven by reference to the intermediate 3D protein structure.

Comparison with the reference structure determinations

The present validation of automated NOESY peak picking with ATNOS relies largely on comparisons of the resulting protein structures with reference structure determinations. Previously, we had shown that protein structures resulting from interactive NOESY peak picking and either interactive NOE assignment or automated NOE assignment with CANDID coincide closely both in terms of accuracy and precision (Herrmann et al., 2002). Therefore, we are now in a position to separately assess the effect on the outcome of a structure determination using either interactive or automated ATNOS analysis of the NOESY spectra.

The numbers of assigned NOE cross peaks as well as the numbers of NOE upper distance constraints that resulted either from the interactive or the automated approach show only small differences (Table 5), indicating that the previous interactive work and ATNOS made similar use of the spectral information. (Since the ATNOS interpretation considers the complete data set, the number of peaks identified in the 2D [^{1}H,^{1}H]-NOESY spectra (Table 5, CopZ and WmKT) is about twice that from the interactive approach, which analyzes only one half of the diagonally symmetric spectra.) The agreement between the results of the two different approaches for the analysis of the NOESY spectra carries over into structures that are very similar in terms of precision and accuracy.With both approaches, the residual DYANA target function values are all below 2.5 Å^2, and the global RMSD values are in the range 0.5 to 0.8 Å. The RMSD value between the mean reference structure and the corresponding result based on the automated ATNOS approach is approximately 0.7 Å for all three proteins, and throughout it is smaller than the sum of the RMSD values of the corresponding bundles (Table 5). The stereochemical qualities measured with the program PROCHECK (Morris et al., 1992) are closely similar in the two sets of structure determinations (Table 5). The reference structures have between 93% and 99% of the residues in the 'most favoured' and 'additional allowed' regions of the Ramachandran plot, as defined by PROCHECK, whereas the corresponding values of the ATNOS structures are in the range from 94% to 96%. In agreement with these numerical data, visible deviations between the structures obtained with automated or interactive peak picking (Figures 5C and 5D) are seen exclusively in the precision of surface loop regions, some of which are better defined with one approach, and others with the other one.

Discussion and outlook

This paper introduces new concepts for NOESY peak picking and NOE cross peak validation. At the outset of the spectral analysis, ATNOS uses highly permissive criteria to identify a comprehensive set of peaks in the NOESY spectra, which includes all NOE signals that are present with sufficient intensity as well as artifacts. The knowledge about the covalent polypeptide structure serves as a reference for deriving spectrum-specific threshold values for critical spectral parameters, which are then used to identify a set of potential NOE cross peaks. During further refined spectral analysis these potential NOE peaks are subjected to a multipass-filtering process (Table 3) for the final NOE cross peak validation. Thereby, the chemical shift database and the intermediate protein three-dimensional structure represent the key references for extensive and reliable NOE cross peak identification.

Proof of principle for the new, automated approach was established by comparing the resulting protein structures with those obtained based on interactive peak picking of the NOESY spectra. To this end, ATNOS was combined with CANDID for automated NOE assignment and DYANA for structure calculation. Overall, NOESY peak picking and NOE cross peak validation with ATNOS yielded similar interpretations of the NOESY spectra and nearly identical protein three-dimensional structures to those obtained using interactive NOESY analysis. The successful generation of the correct polypeptide fold after the first computation cycle is a crucial intermediate result in the iterative ATNOS/CANDID schedule with DYANA structure calculation, and plays the key role in establishing direct feedback between the raw NMR data and the protein three-dimensional structure.

The experience gained during the development and testing of the present implementation of ATNOS indicates that the following two conditions have to be met for proper performance of the ATNOS/CANDID procedure with DYANA structure calculation.

(a) ATNOS must validate NOE signals for at least 85% of all pairwise combinations of protons i and j ($i \neq j$) for which sequence-specific NMR assignments are available, and which have covalent structure-imposed upper distance limits shorter than d_{max} (Equation 13). (b) The conditions previously given for proper performance of automated NOE assignment with CANDID and structure calculation with DYANA must be satisfied (Herrmann et al., 2002).

The condition (a) requires high quality of the NOESY spectra and completeness of the chemical shift database derived from the previous sequence-specific resonance assignment. A low percentage of validated covalent NOE cross peaks typically results when the signal-to-noise ratio is too poor for automated spectral analysis, the chemical shift database is incomplete, or the chemical shifts are not sufficiently precisely adapted to the NOESY spectrum considered. In this situation, the input data need to be critically reevaluated prior to a next attempt of automated interpretation, in particular the adaptation of the chemical shift lists to the NOESY spectra used. Proper chemical shift adaptation is generally a concern in NMR structure determinations using different NMR spectra for obtaining sequence-specific resonance assignments, and for the collection of conformational constraints, respectively. Clearly, whenever the difference between the NOE cross peak positions and the chemical shift values of a given atom pair is larger than the predetermined tolerance range, then ATNOS will fail to identify the corresponding NOE cross peak. In structure determinations with homonuclear ^{1}H NMR, where the resonance assignments are based on sequential NOEs observed in the same data sets that are also used for the collection of conformational constraints (Wüthrich, 1986), no adaptation of the chemical shift database is usually needed.

Acknowledgements

Financial support by the Schweizerischer Nationalfonds (projects 31.49047.96 and 31.66427.01), and the use of the high-performance computing facilities of ETH Zürich and The Scripps Research Institute are gratefully acknowledged.

References

Anil-Kumar, Ernst, R.R. and Wüthrich, K. (1980) *Biochem. Biophys. Res. Commun.*, **95**, 1–6.

Antuch, W., Güntert, P. and Wüthrich, K. (1996) *Nat. Struct. Biol.*, **3**, 662–665.

Antz, C., Neidig K.P. and Kalbitzer H. R. (1995) *J. Biomol. NMR*, **5**, 287–296.

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

Boelens, R., Koning, T.M.G., van der Marel, G.A., van Boom, J.H. and Kaptein, R. (1989) *J. Magn. Reson.*, **82**, 290–308.

Borgias, B.A. and James, T.L. (1988) *J. Magn. Reson.*, **79**, 493–513.

Borgias, B.A. and James, T.L. (1990) *J. Magn. Reson.*, **87**, 475–487.

Brünger, A.T. (1992) *X-PLOR, Version 3.1. A System for X-Ray Crystallography and NMR*, Yale University Press, New Haven, USA.

Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T. and Warren, G.L. (1998) *Acta Crystallogr.*, **D54**, 905–921.

Corne, S.A. and Johnson, P. (1992) *Neural Networks*, **6**, 1023–1032.

Garret, D.S., Powers, R., Gronenborn, A.M. and Glore, G.M. (1991) *J. Magn. Reson.*, **95**, 214–220.

Gronwald, W., Kirchhöfer, R., Görler, A., Kremer, W., Ganslmeier, B., Neidig, K.-P. and Kalbitzer, H.R. (2000) *J. Biomol. NMR*, **17**, 137–151.

Güntert, P. and Wüthrich, K. (1992) *J. Magn. Reson.*, **96**, 403–407.

Güntert, P., Billeter, M., Ohlenschläger, O., Brown, L.R. and Wüthrich, K. (1998) *J. Biomol. NMR*, **12**, 543–548.

Güntert, P., Braun, W. and Wüthrich, K. (1991) *J. Mol. Biol.*, **217**, 517–530.

Güntert, P., Dötsch, V., Wider, G. and Wüthrich K. (1992) *J. Biomol. NMR*, **2**, 619–629.

Güntert, P., Mumenthaler, C. and Wüthrich, K. (1997) *J. Mol. Biol.*, **273**, 283–298.

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

Horst, R., Damberger, F., Luginbühl, P., Güntert, P., Peng, G., Nikonova, L., Leal, W.S. and Wüthrich, K. (2001). *Proc. Natl. Acad. Sci. USA*, **98**, 14374–14379.

Keepers, J.W. and James, T.L. (1984) *J. Magn. Reson.*, **57**, 404–426.

Kleywegt, G.J., Boelens, R. and Kaptein, R. (1990) *J. Magn. Reson.*, **88**, 601–608.

Koradi, R., Billeter, M., Engeli, M., Güntert, P. and Wüthrich, K. (1998) *J. Magn. Reson.* **135**, 288–297.

Koradi, R., Billeter, M. and Güntert, P. (2000) *Comput. Phys. Commun.*, **124**, 139–147.

Koradi, R., Billeter, M. and Wüthrich, K. (1996) *J. Mol. Graph.* **14**, 51–55.

Luginbühl, P., Güntert, P., Billeter, M. and Wüthrich, K. (1996) *J. Biomol. NMR*, **8**, 136–146.

Mertz, J.E., Güntert, P., Wüthrich, K. and Braun, W. (1991) *J. Biomol. NMR*, **1**, 257–269.

Morris, A.L., MacArthur, M.W., Hutchinson, E.G. and Thornton, J.M. (1992) *Proteins*, **12**, 345–364.

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

Nilges, M., Habazettl, J., Brünger, A.T. and Holak, T.A. (1991) *J. Mol. Biol.*, **219**, 499–510.

Williamson, M., Havel, T. F. and Wüthrich, K. (1985) *J. Mol. Biol.*, **155**, 311–319.

Wimmer, R., Herrmann, T., Solioz, M. and Wüthrich, K. (1999) *J. Biol. Chem.*, **274**, 22597–22603.

Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York, NY.

Wüthrich, K., Billeter, M. and Braun, W. (1983) *J. Mol. Biol.*, **169**, 949–961.

Yip, P. and Case, D.A. (1989) *J. Magn. Reson.*, **83**, 643–648.