J Biomol NMR (2017) 67:63–76
DOI 10.1007/s10858-016-0084-3

ARTICLE

Peak picking multidimensional NMR spectra with the contour geometry based algorithm CYPICK

Julia M. Würz[1] · **Peter Güntert**[1,2,3]

Received: 11 November 2016 / Accepted: 19 December 2016 / Published online: 3 February 2017

Abstract The automated identification of signals in multidimensional NMR spectra is a challenging task, complicated by signal overlap, noise, and spectral artifacts, for which no universally accepted method is available. Here, we present a new peak picking algorithm, CYPICK, that follows, as far as possible, the manual approach taken by a spectroscopist who analyzes peak patterns in contour plots of the spectrum, but is fully automated. Human visual inspection is replaced by the evaluation of geometric criteria applied to contour lines, such as local extremality, approximate circularity (after appropriate scaling of the spectrum axes), and convexity. The performance of CYPICK was evaluated for a variety of spectra from different proteins by systematic comparison with peak lists obtained by other, manual or automated, peak picking methods, as well as by analyzing the results of automated chemical shift assignment and structure calculation based on input peak lists from CYPICK. The results show that CYPICK yielded peak lists that compare in most cases favorably to those obtained by other automated peak pickers with respect to the criteria of finding a maximal number of real signals, a minimal number of artifact peaks, and maximal correctness of the chemical shift assignments and the three-dimensional structure obtained by fully automated assignment and structure calculation.

Keywords Peak picking · Peak list · Contour lines · Automated assignment · Structure calculation · CYANA

Introduction

Identifying signals in an NMR spectrum, also known as peak picking, plays a central role in biomolecular NMR studies and is a prerequisite for sequence-specific resonance assignment and structure determination. Peak lists provide an abstraction of the multidimensional spectra that contains the most essential spectral information—the position and intensity of the signals—in a form that is readily accessible by interactive or automated spectra analysis programs. The ease and reliability of spectrum analysis relies on the quality of the peak lists, which in turn depends mainly on three factors: how many of the true signals the peak lists contain, how few additional "artifact" peaks that do not correspond to true signals they contain, and how accurate they record the positions and intensities of the signals.

Peak lists do not have to be flawless to serve as a basis for chemical shift assignment and NOE assignment, followed by structure calculation. For instance, it has been shown that the automated resonance assignment algorithm FLYA can yield more than 90% correct resonance assignments even if either 60% of the true peaks are missing or five times more artifacts than real peaks are present in the input peak lists (Schmidt and Güntert 2012). Automated NOE assignment and structure calculation with CYANA (Güntert and Buchner 2015; Herrmann

Electronic supplementary material The online version of this article (doi:10.1007/s10858-016-0084-3) contains supplementary material, which is available to authorized users.

✉ Peter Güntert
guentert@em.uni-frankfurt.de

[1] Institute of Biophysical Chemistry, Center for Biomolecular Magnetic Resonance, Goethe University Frankfurt am Main, Max-von-Laue-Str. 9, 60438 Frankfurt am Main, Germany

[2] Laboratory of Physical Chemistry, ETH Zürich, Zürich, Switzerland

[3] Graduate School of Science and Engineering, Tokyo Metropolitan University, Hachioji, Tokyo, Japan

et al. 2002a) can in many cases also tolerate 30–40% missing NOESY peaks without dramatic deterioration of the resulting structures (Buchner and Güntert 2015; Jee and Güntert 2003).

Peak picking can be achieved by visual inspection of the spectra or automated methods. Along with algorithms for resonance assignment and structure calculation, the demand for automated peak picking is increasing, and various algorithms for the purpose have been proposed. Nevertheless, the task remains challenging. Reasons for this include low signal-to-noise, overlap, and artifacts such as baseline distortions, intense solvent lines, ridges, or sinc wiggles.

Most of the existing peak picking algorithms can be classified as either threshold-based methods, methods that depend on symmetry criteria, peak-shape-based methods, methods that incorporate peak picking into NMR data processing, or a combination thereof. Threshold-based methods are the most straightforward and most commonly used automated peak picking approaches. Interactive spectrum analysis programs like XEASY (Bartels et al. 1995), Sparky (Goddard and Kneller 2001), NMRViewJ (Johnson 2004; Johnson and Blevins 1994), or CcpNmr AnalysisAssign (Skinner et al. 2016; Vranken et al. 2005) (in the following abbreviated as CCPN) give the user the possibility to adjust a threshold manually and perform peak picking by finding local extrema above the threshold. These methods are particularly useful as a starting point for semi-automated peak identification, which is refined manually. WavPeak (Liu et al. 2012) employs wavelet-based smoothing of the spectrum prior to identifying peaks as local maxima. PICKY (Alipanahi et al. 2009), is a singular value decomposition (SVD)-based automated peak picking method. Machine learning and computer vision methods have also been employed for peak picking, e.g. in the CV-Peak Picker program (Klukowski et al. 2015). AUTOPSY (Koradi et al. 1998) is a sophisticated automated peak picker that includes functions to determine a local noise level and to deconvolute clusters of overlapping peaks with the help of line shapes derived from non-overlapping peaks. ATNOS (Herrmann et al. 2002b) is an automated peak picker specifically for NOESY spectra that is integrated into automated NOESY assignment and structure calculation and makes use of preliminary structural information to guide the peak picking. Peak picking can be part of NMR data processing, e.g. in the program MUNIN (Orekhov et al. 2001) that uses three-way decomposition to decompose a three-dimensional (3D) NMR spectrum into a sum of components defined as the direct product of three 1D shapes. The GAPRO peak identification algorithm (Hiller et al. 2005) establishes peak lists for high-dimensional (e.g. 4D, 5D, 6D) APSY-type spectra by picking peaks in the experimentally recorded tilted 2D projections.

The human approach to peak picking can be described as the analysis of the shape and regularity of 2D contour lines. Real signals are manifested by concentric ellipses and have common properties which artifacts do not share, e.g. regarding peak width, convexity, or similarity. However, real signals can deviate from the proposed "perfect" shape for a number of reasons, such as, for example, noise, spectral overlap, limited digital resolution, baseline instabilities, and improper phasing of the spectrum. An automated peak picking procedure should be able to handle these imperfections and shortcomings. It is thus a promising approach to automated peak picking to mimic the human way of analyzing similarity and symmetry criteria of contour lines in 2D spectral planes. This approach has first been used in the CAPP algorithm (Garrett et al. 1991). Our aim was to develop an effective and fast automated peak picking procedure within the CYANA software package (Güntert et al. 1997) that can be linked directly to automated chemical shift assignment and/or NOE assignment, followed by structure calculation. We reduced the requirements for user intervention in setting parameters as far as possible in order to increase objectivity and reproducibility in comparison with manual peak picking. In this publication we introduce CYPICK, a fully automated peak picking method implemented in CYANA.

The quality of peak lists can be evaluated in two distinct ways, both of which are used in this paper: directly, by assessing how well a peak list represents the underlying spectrum, or indirectly by using the peak lists as input for (automated) resonance assignment and structure calculation algorithms and analyzing the correctness of the resulting assignments and/or structures. A peak list can be evaluated directly by visually overlaying it on the spectrum, which is subjective and time-consuming, or by comparison with a high-quality "correct" reference peak list, which can be quantified by suitable scoring functions. For the indirect method to evaluate peak lists, we use the FLYA algorithm for automated chemical shift assignment (Schmidt and Güntert 2012) and combined automated NOESY assignment and structure calculation with CYANA (Güntert and Buchner 2015; Herrmann et al. 2002a).

Algorithm

A simplified overview of the different steps and picking modes available for the contour peak picker CYPICK is shown in Fig. 1. The first step is to read the processed NMR spectrum. It is currently possible to read spectra in XEASY, BRUKER, UCSF, and AZARA format. After storing the spectrum in memory, an estimate of the intensity of the lowest contour level is required. This can either be the global noise level of the spectrum (CYANA command:

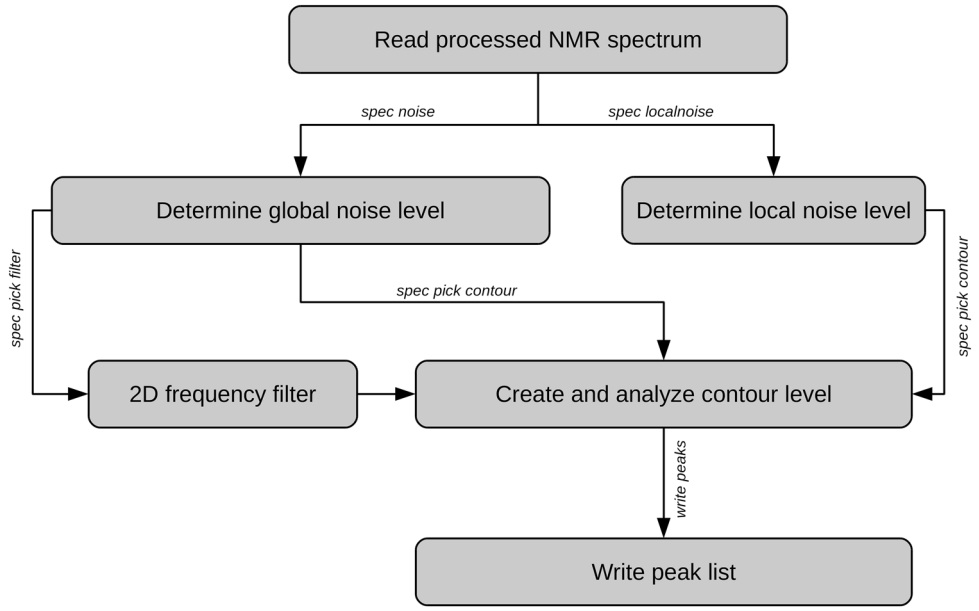

Fig. 1 Flowchart of the CYPICK peak picking algorithm implemented in CYANA. The algorithm needs as input a processed NMR spectrum. Three picking modes are available for the contour approach: First, the global noise level is determined and used as intensity of the first contour line (CYPICK command *spec pick global*). Second, a restricted peak picking with a 2D frequency filter that can be provided in the form of a peak list (CYPICK command *spec pick filter*). The position of the 2D peaks is used as a filter for local extrema which are considered in the contour approach. Third, a local noise level is determined for every data point in the spectrum (CYPICK command *spec pick local*). The local noise level is used as a first filtering step for local extrema and creation of contour lines. In the restricted peak picking mode the global noise level is needed for the selection of local extrema and the estimation of contour level intensities

spec noise) or the local noise level at each data point (command: *spec localnoise*). The following step is to find local extrema and to compute the contour lines in their vicinity. This can either be done over the complete spectrum (command: *spec pick contour*) or restricted by a frequency filter, defined by a 2D peak list (command: *spec pick filter*). The contour lines belonging to local extrema are subsequently filtered and analyzed. The remaining local extrema are stored in a peak list (command: *write peaks*). Details are explained in the following subsections.

Noise level determination

The global noise level L_{global} is determined by estimating the median of the absolute intensity values of the data points as implemented in the program PROSA (Güntert et al. 1992), which assumes that most of the data points in a multidimensional NMR spectrum are at locations not occupied by signals. The local noise level $L_{\text{local}}(\omega)$ at a given position $\omega = (\omega_1,\ldots,\omega_D)$ in the spectrum is computed by the method used in the program AUTOPSY (Koradi et al. 1998). For this purpose, each one-dimensional slice of the spectrum is subdivided into segments comprising 5% of its data points. The minimum over all segments of the standard deviation of the spectral intensities represents the noise level for a given slice. A base noise level δ_{b} is defined as the minimal noise level of any slice in the spectrum. The local noise level $L_{\text{local}}(\omega)$ is calculated from the noise levels $\delta_i(\omega)$, $i = 1,\ldots, D$, of the slices that pass through the data point ω and the base noise level δ_{b} as follows:

$$L_{\text{local}}(\omega) = \sqrt{\sum_{i=1}^{D} \delta_i(\omega)^2 - (D-1)\,\delta_b^2}$$

Determination of local extrema

A data point is checked for being a local extremum if its intensity exceeds a spectrum-specific base level, $B = \beta L$, where L denotes either the global noise level L_{global} or the local noise level L_{local} at this position, depending on the desired peak picking mode. The baseline factor β is a user-defined parameter, typically chosen between 2 and 3 (Koradi et al. 1998). Throughout this paper, we used the same value, $\beta = 3.0$, for reasons of objectivity and reproducibility. A data point is considered a local extremum if the 3^D–1 neighboring data points have an absolute intensity lower than or equal to the central data point.

Creation of contour lines

For reasons of efficiency, contour lines are computed in regions around each local extremum rather than for the entire spectrum. The size of this regions can in principle be changed by the user. However, the algorithm is insensitive to the size of the peak region, as long as it is chosen sufficiently large.

The height of successive contour levels is set by multiplying the previous contour level by a contour level factor γ, starting from the base level B, i.e. the n–th contour level is at height $B\gamma^n$. Typical values for γ are in the range of 1.2–1.4. With $\gamma = 1.4$ the contour level is approximately doubled every 2 contour lines, whereas with $\gamma = 1.2$ it is approximately doubled every 4 contour lines. We used $\gamma = 1.3$ for all calculations in this paper.

The number of contour lines encircling a local extremum depends on its absolute intensity. After having defined the peak area and the height of the contour lines, the actual positions of the points that define the contour lines, which are (open or closed) polygons, are determined by an algorithm that was first used for plotting spectra in the program PROSA (Güntert et al. 1992). It is very similar to the marching squares algorithm (Lorensen and Cline 1987) that is perfectly suited for this situation because the spectral data points are already rasterized on a regular grid. The peak area is further subdivided into sets of 4 data points (2×2 squares). The four data points of these sub-squares are checked for having an intensity higher or lower than the desired height of the contour line. 16 cases can be distinguished, which reduce to four cases under consideration of fourfold rotational symmetry. These four cases are depicted in Fig. S1. Based on these cases, 0, 1 or 2 linear pieces of contour line are determined by linear interpolation. For each local extremum above the base level, the vertices of all closed contour lines that encircle the local extremum are stored.

Scaling of contour lines

In order to achieve approximately circular contour lines for peaks, the chemical shift coordinates of the points defining the contour lines are divided by a scaling factor σ_i that is set by the user for each dimension $i = 1,\ldots,D$ of the spectrum according to the approximate line widths.

Filtering of contour lines

These contour lines are subjected to a preliminary filtering process:

- The local extremum of interest has to be inside the contour line.
- No other local extremum except the local extremum of interest may be enclosed by the contour line.
- The contour line must have at least 5 points. Shape criteria (see below) cannot be evaluated meaningfully for contour lines with fewer points.
- At least two contour lines that fulfil all preceding conditions must encircle the local maximum.

In order to test whether a local extremum is within a given closed contour line, we use the ray casting algorithm (Shimrat 1962) that relies on Jordan's polygon theorem. In this procedure, a ray is sent out from the point of interest, in this case the local extremum, and the number of intersections with the edges of the contour line are counted. Odd numbers of intersections imply that the point of interest is inside of the contour line, whereas even numbers imply that the point is outside of the contour line.

Analysis of contour lines

After filtering, the remaining contour lines that enclose a given local extremum are analyzed starting from the contour line with the highest absolute intensity. If the highest contour line does not fulfill the requirements, the next lower contour line is analyzed. At least two contour lines have to fulfill the following conditions.

The first condition to be fulfilled by a contour line belonging to a real signal is that its shape must be approximately circular. As mentioned earlier, the shape of contour lines can be described by concentric ellipses. Contour lines consist of contour points. Connecting contour points by a line, results in a polygon. The area enclosed by the contour line is determined via Gauss's area formula:

$$A = \frac{1}{2}\left|\sum_{i=1}^{n}(x_i + x_{i+1})\,(y_{i+1} - y_i)\right|$$

where n denotes the number of vertices in the contour line, x_i and y_i are the coordinates of the i-th vertex, and x_{n+1} and y_{n+1} are assumed to be identical to x_1 and y_1, respectively. The circumference of a contour line is given by

$$C = \sum_{i=1}^{n}\sqrt{\left(x_i - x_i\right)^2 + \left(y_i - y_{i+1}\right)^2}$$

For a circle, the area-to-circumference-squared ratio is

$$\frac{A}{C^2} = \frac{\pi r^2}{(2\pi r)^2} = \frac{1}{4\pi}$$

The area-to-circumference-squared ratio is determined for each contour line that survives the aforementioned filtering steps. This ratio should equal $1/4\pi$ in case of a perfect circle. We therefore define a quality factor

$$Q_{\mathrm{rad}} = e^{-750 \cdot \left(\frac{A}{C^2} - \frac{1}{4\pi}\right)^2}$$

in the range of [0,1]. The maximal value $Q_{\mathrm{rad}} = 1$ is realized by a perfect circle. Figure 2a visualizes ellipses with varying eccentricities and their corresponding Q_{rad} values.

As a second condition, a contour line around an extremum is required to form an approximately convex polygon with all interior angles less than 180°. Nevertheless, for some of the real signals a slight deviation from perfect convexity should be tolerated. Therefore, a quality factor Q_{con} similar to Q_{rad}, within the range [0,1] is defined as $Q_{con} = \prod_{i=1}^{n} Q_{con,i}$ with

$$Q_{con,i} = \begin{cases} 1, & \alpha_i \leq \pi \\ \left(\frac{2\alpha_i}{\pi} - 3\right)^2, & \pi < \alpha_i < 3\pi/2 \\ 0, & \alpha_i \geq 3\pi/2 \end{cases}$$

where α_i denotes the interior angle at vertex i of the polygon. All convex polygons have $Q_{\mathrm{con}} = 1$. Polygons with at least one interior angle $\alpha_i \geq 270°$ have $Q_{\mathrm{con}} = 0$. Figure 2b shows different polygons with convex and concave angles α and the corresponding Q_{con} values.

On the basis of the Q_{rad} and Q_{con} values for real and erroneous signals in numerous spectra we set thresholds of $Q_{\mathrm{rad}} \geq 0.7$, $Q_{\mathrm{con}} \geq 0.7$, and $Q_{\mathrm{rad}} \cdot Q_{\mathrm{con}} \geq 0.6$ that were applied throughout this paper.

Interpolation of the local extremum.

A local extremum in a D-dimensional spectrum is accepted as a peak if it fulfills the contour-based criteria in one of the 2D planes at its location. The digital resolution of an NMR spectrum, given by the quotient of the spectral sweep width and the number of data points in the given dimension, limits the accuracy with which NMR signals can be described. Accordingly, the true position of a NMR signal is rather somewhere between data points than exactly on a measured data point. The exact peak position plays a crucial role in chemical shift assignment and NOE assignment. Therefore, it is important to determine the location of a peak as accurately as possible. In CYPICK, the exact position of the local extremum is estimated by cubic spline interpolation (Press et al. 1986) along each 1D slice passing through the local extremum of interest.

Materials and methods

Evaluation dataset

The performance of CYPICK was first analyzed on the basis of peak lists obtained for 16 different spectra of the ENTH-VHS domain At3g16270(9–135) from *Arabidopsis thaliana* (referred to as ENTH; PDB code 1VDY; BMRB code 5928, 140 residues) that are described in (López-Méndez and Güntert 2006; López-Méndez et al. 2004). We

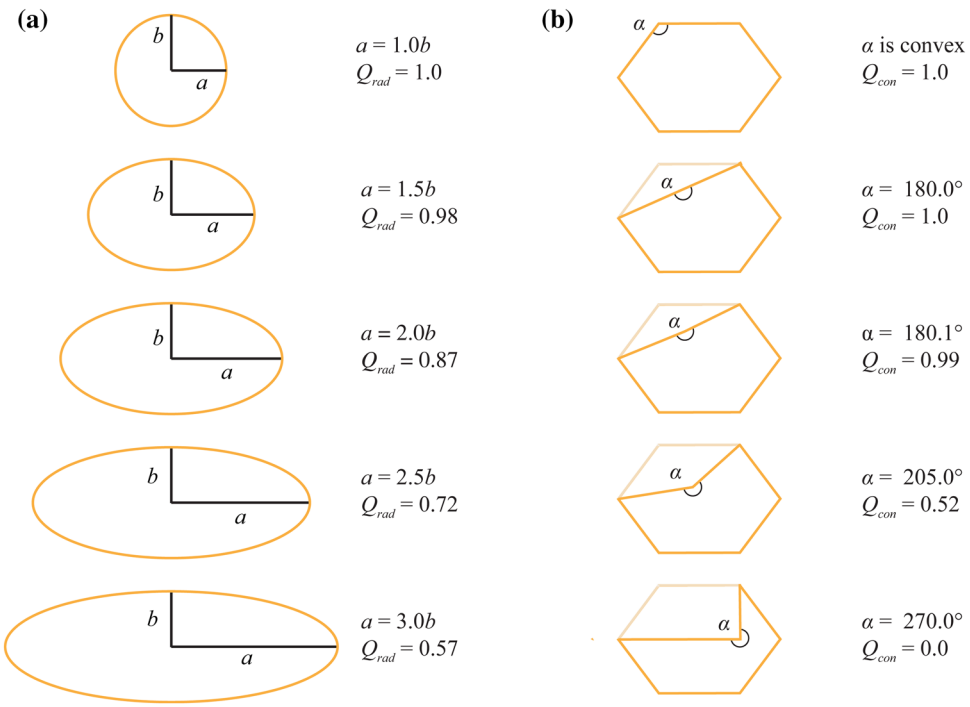

Fig. 2 **a** Influence of ellipse deformation on the Q_{rad}value. **b** Polygons with varying degree of concavity and their Q_{con}values

converted the spectra to UCSF (Sparky) format, which can be read by all peak picking programs used in this study. Manually picked peak lists and lists picked automatically by AUTOPSY were available for the ENTH spectra from an earlier study (López-Méndez and Güntert 2006). The manually picked peak lists served as the reference in determining score values for finding real peaks or artifact peaks, as well as an overall score which combines both. The exact definition of these scores is given below. The score values of the CYPICK peak lists relative to the reference peak lists were then compared to score values of peak lists determined by other programs, namely AUTOPSY, NMRViewJ, CCPN, and CV-Peak Picker.

We further used spectra for the *Arabidopsis thaliana* rhodanese domain At4g01050 (referred to as RHO; PDB code 1VEE; BMRB code 5929, 134 residues) (Pantoja-Uceda et al. 2004, 2005) and the Src homology domain from the human feline sarcoma oncogene FES (referred to as SH2; PDB code 1WQU; BMRB code 6331, 114 residues) (Scott et al. 2004, 2005) together with the ENTH data set to evaluate CYPICK. The spectra for SH2 and RHO are also summarized in (López-Méndez and Güntert 2006). Chemical shift assignments and NMR solution structures of the three proteins ENTH, RHO, and SH2 have been determined earlier by conventional techniques and their data sets have previously been used to evaluate the automated assignment algorithm FLYA (Schmidt and Güntert 2012). The available manual assignment and manually picked ^{13}C- and ^{15}N-NOESY peak lists were used to recalculate the structures with CYANA. Structure calculation was performed with 10,000 torsional angle dynamic steps and 100 starting structures. The 20 conformers with the lowest target function values were selected as the reference structure bundle for comparison with automatically determined structures. Here, we performed a completely automated protocol consisting of peak picking, chemical shift assignment, NOE assignment, and structure calculation. We evaluated the resonance assignments and structures obtained from peak lists automatically picked by the programs CYPICK, AUTOPSY (only for ENTH), NMRViewJ, CCPN, and CV-Peak Picker.

In addition, CYPICK was tested on ten data sets from the CASD-NMR project (Rosato et al. 2012, 2009): the human NFU1 iron-sulfur cluster scaffold homolog, Northeast Structural Genomics Consortium (NESG) target HR2876B (PDB code 2LTM; BMRB code 18489, 107 residues), the CTD domain of the human NFU1 iron-sulfur cluster scaffold homolog, NESG target HR2876C (PDB code 2M5O; BMRB code 19068, 97 residues), the N-terminal domain of the human mitotic checkpoint serine/threonine-protein kinase BUB1, NESG target HR5460A (PDB code 2LAH; BMRB code 17524, 160 residues), the RRM domain of the human RNA-binding protein FUS, NESG target HR6430A (PDB code 2LA6; BMRB code 17504, 99 residues), the homeobox domain of the human homeobox protein Nkx-3.1, NESG target HR6470A (PDB code 2L9R; BMRB code 17,484, 69 residues), the SANT domain of human DNAJC2, NESG target HR8254A (PDB code 2M2E; BMRB code 18909, 73 residues), a *de novo* designed protein, IF3-like fold, NESG target OR135 (PDB code 2LN3; BMRB code 18145, 83 residues) (Koga et al. 2012), a de novo designed protein, P-loop NTPase fold, NESG target OR36 (PDB code 2LCI; BMRB code 17613, 134 residues), TSTM1273 from *Salmonella typhimurium* LT2, NESG target StT322 (PDB code 2LOJ; BMRB code 18214, 63 residues), and the NifU-like protein *Saccharomyces cerevisiae*, NESG target YR313A (PDB code 2LTL; BMRB code 18487, 119 residues). For all proteins, ^{13}C-edited and ^{15}N-edited NOESY spectra were provided. The spectra were automatically picked by CYPICK. The resulting peak lists were used, together with the reference chemical shift assignments from the BMRB, as input for automated NOESY assignment and structure calculation with CYANA. The performance of CYPICK was evaluated by comparison with reference peak lists that had been prepared with the structure-based NOESY peak picker ATNOS (Guerry et al. 2015; Herrmann et al. 2002b), as well as by comparison of the structures with the reference structures deposited in the PDB.

Peak list comparison

To quantify the agreement between a trial peak list of N peaks and a reference peak list of N_0 peaks, corresponding peaks are identified with the Hungarian algorithm (Bourgeois and Lassalle 1971; Munkres 1957; Silver 1960) that finds an exact solution of the assignment problem (of combinatorial optimization; not to be confused with the problem of finding chemical shift assignments in NMR) and has a polynomial complexity of $O(n^3)$. The "cost" of assigning a peak i in the trial peak list to peak j in the reference peak list is defined as

$$C_{ij} = 1 - \exp\left(-\min(d_{ij}^2, d_{\text{cut}}^2)/2\right)$$

where

$$d_{ij}^2 = \sum_{k=1}^{D} \left(\frac{\omega_{ik} - \omega_{jk}}{\sigma_k} \right)^2$$

is the squared scaled distance between the two peak positions $(\omega_{i1},\ldots,\omega_{iD})$ and $(\omega_{j1},\ldots,\omega_{jD})$ in the D-dimensional

spectrum, scaled by the chemical shift scaling factors σ_k. The cutoff d_{cut} implements the idea that all deviations larger than a certain value d_{cut} indicate that the two peaks cannot originate from the same atoms. All such peak pairs should carry the same, high cost, regardless of the actual deviation $d_{ij} \geq d_{cut}$. We used $d_{cut} = 3$ for all calculations in this paper. The Hungarian algorithm assign each of the $M = \min(N_0,N)$ peaks in the shorter peak list to a peak in the longer peak list, such that the total cost $\sum_{k=1}^{M} C_{i_k j_k}$ is minimized. The result is a list of $k = 1,\ldots,M$ pairs (i_k,j_k) of corresponding peaks in the two peak lists. Using C_{ij} instead of the distance d_{ij} reduces drastically the computation time for the Hungarian algorithm by avoiding pointless optimizations for pairs of peaks that cannot originate from the same atoms. To further speed up the calculations, the peaks are first grouped into as small as possible clusters such that all pairs of peaks in different clusters have $d_{ij} \geq d_{cut}$, and the Hungarian algorithm is applied to each cluster separately.

The quality of peak correspondence is rated by

$$H = \sum_{k=1}^{M} exp\left(-d_{i_k j_k}^2/2\right)$$

H represents the number of corresponding peak pairs, weighted by the deviation of the peak positions; $0 \leq H \leq M$.

H can be used to define a *find score* $F = H/N_0$ and an *artifact score* $A = 1 - H/N$. Both scores take values between 0 and 1 (or 0–100%). The find score gives the fraction of "true" peaks in the reference peak list that have a corresponding peak in the trial peak list. The artifact score gives the fraction of "artifact" peaks in the trial peak list that do not have a corresponding peak in the reference peak list. If the trial and reference peak lists are identical, $F = 1$ and $A = 0$. Except for the exact definition of the number H of corresponding peak pairs, F and $1 - A$ are identical to 'recall' and 'precision' as defined by (Alipanahi et al. 2009), respectively.

It is possible to define an *overall score* $S = (H - w\,(N - H))/N_0$, given by the number of found peaks, H, minus the number of artifact peaks, $N - H$, weighted by a factor w that specifies the relative detrimental effect of artifacts in comparison to real peaks. In this paper, we used $w = 0.2$, assuming that 5 additional artifact peaks are as severe as one missing true peak, as suggested by observations on their effect on automated resonance assignment (Schmidt and Güntert 2012) and structure calculation (Buchner and Güntert 2015). The overall score combines the found and artifact scores according to $S = F - w\,(N/N_0)\,A$ and reaches a maximum value of 1 in the ideal case of identical peak lists. It never exceeds the find score and can become negative if very many artifacts are present.

The calculation of these scores has been implemented in the new CYANA command *peaks compare*.

Automated peak picking

Automated peak picking by CYPICK was performed as described in the "Algorithm" section. Parameters used for the individual spectra are given in Table S1 for ENTH, RHO, and SH2, and Table S2 for the CASD-NMR data sets. Parameters not given in Tables S1 and S2 were kept at their aforementioned constant values.

CYPICK results were compared to those from other well established peak picking programs. In case of the proteins ENTH, RHO, and SH2, we used the automated peak picking routines of the CCPN (Table S3) and NMRViewJ (Table S4) software packages, for which the user has to define a global noise level that is used as threshold for peak picking. All local extrema above the specified cutoff value are picked and stored. NMRViewJ additionally comprises methods to determine local thresholds, which allow e.g. the exclusion of solvent lines. We further employed the CV-Peak Picker as an example of a sophisticated algorithm based on image recognition techniques (Table S5) to pick peaks in the ENTH, RHO, and SH2 data sets. For ENTH, automatically picked peak lists from the program AUTOPSY and manually picked ENTH lists were available from an earlier study. Spectral unfolding and removal of peaks in a narrow region (4.70 ± 0.08 ppm) around the water line was performed equally for the peak lists from all programs.

Chemical shift assignment calculations

Automated chemical shift assignment was performed by the FLYA algorithm (Schmidt and Güntert 2012). The tolerances for chemical shift matching and comparison with the reference chemical shift assignment were set to 0.03 ppm for ^{1}H and 0.4 ppm for ^{13}C and ^{15}N. The population size of the evolutionary algorithm was 50 except when performing solely NOESY-based chemical shift assignment, for which it was increased to 200. The number of local optimization steps per generation of the evolutionary algorithm was limited to 15,000. The chemical shift assignment was consolidated from 20 independent runs. The side-chain terminal amide groups of arginine and lysine were excluded from the assignment calculations.

Structure calculation

The chemical shift assignment established by FLYA or the reference assignment was used to obtain torsional angle restraints using TALOS+ (Shen et al. 2009). Combined automated NOE assignment and structure calculation was performed by the standard CYANA protocol (Güntert and Buchner 2015), using as input the protein sequence, assigned chemical shifts, torsional angle restraints, and

unassigned NOESY peak lists. Tolerances for chemical shift and peak position matching were set to 0.03 ppm for ^{1}H and 0.4 ppm for ^{13}C and ^{15}N. NOESY peak intensities were converted into upper distance limits according to a $1/r^6$ dependence. Structure calculation was performed starting from 200 conformers using 10,000 torsion angle dynamic steps. The 20 best conformers in terms of CYANA target function were selected to represent the structure bundle. No energy refinement was performed on the resulting structures.

The accuracy of the structure was evaluated on the basis of the RMSD bias to the reference structure (Güntert 1998). To this end, the conformers of the structure bundle were first superimposed for lowest RMSD within their ordered regions, which were determined by CYRANGE (Kirchner and Güntert 2011) in case of ENTH, RHO, and SH2, or applied as specified in the CASD-NMR publication (Rosato et al. 2015). Then the average structure is obtained by averaging, for each atom, the coordinates in the superimposed conformers. The backbone RMSD between this average structure and the (mean) reference structure yields the RMSD bias. The precision of calculated structure bundles is expressed by the RMSD radius (Güntert 1998), i.e. the average RMSD between individual conformers and the mean coordinates of the structure.

Results and discussion

Dependence of peak picking scores on the number of peaks

Automated peak picking involves to some extent a trade-off between finding a maximal number of true peaks and minimizing the number of artifact peaks. As an example, Fig. 3 illustrates the behavior of the scores, used to evaluate peak picking, upon varying the number of real signals and artifacts for ^{13}C-resolved NOESY peak lists of the protein ENTH. Peak lists with different numbers of real and artifact peaks where produced with CYPICK by varying the baseline factor, which determines the height of the lowest contour line, over a wide range. Decreasing the lowest contour line leads to a steep increase in the number of picked peaks, N, and a—much smaller—increase in the number of picked real peaks, H (Fig. 3a). At low baseline factors, the number of picked peaks exceeds by far the number of reference peaks, indicating that most of the peaks that are picked close to the noise are in fact artifacts. Consequently, the artifact score $A = 1 - H/N$ and the find score $F = H/N_0$ increase with decreasing noise level (Fig. 3b). Both scores approach but do not reach their ideal values of $A = 0$ and $F = 1$, i.e. some strong artifacts are always picked and a small fraction of the manually identified peaks can never be found by the algorithm. Strong artifacts can be attributed

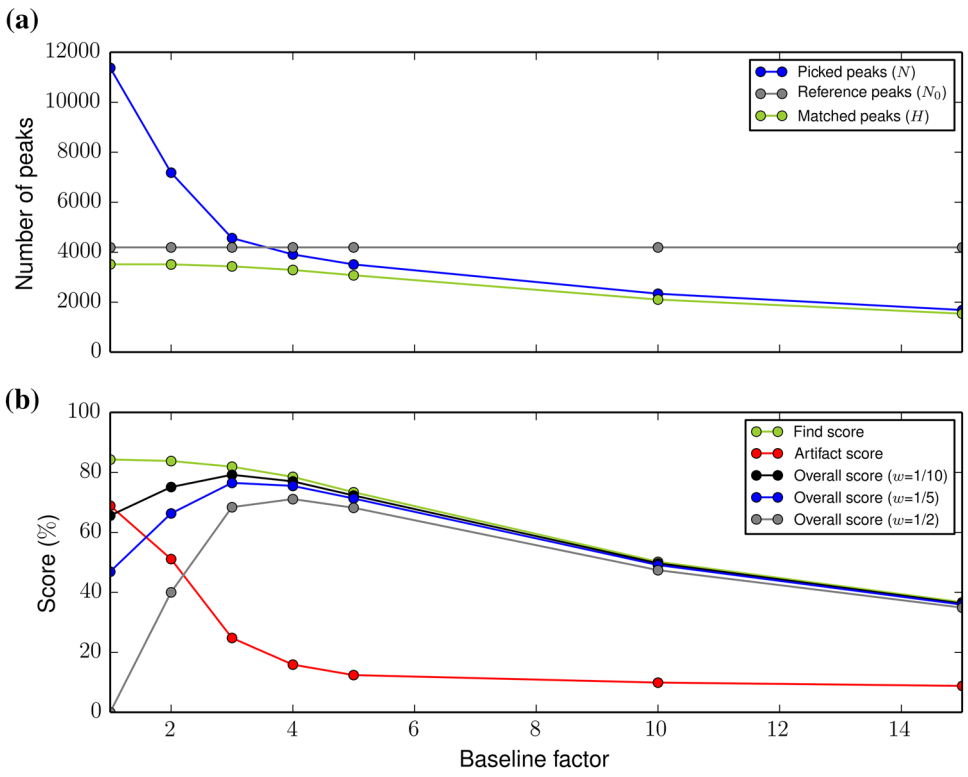

Fig. 3 Influence of the baseline factor on the number of peaks picked by CYPICK in the 3D ^{13}C-edited NOESY spectrum of the protein ENTH. **a** Number of peaks picked by CYPICK (*N, blue*), constant number of manually picked reference peaks (N_0, *grey*), and weighted number of matches between the two peak lists (*H, green*). **b** Find score (*green*), artifact score (*red*), and overall scores (*black, blue, grey*). The latter are shown for different values of the weighting factor *w*

mainly to truncation artifacts of strong peaks, axial peaks, and a few putative real peaks missing in the manually prepared reference peak list.

The overall score $S = (H - w\ (N - H))/N_0 = F - w\ (N/N_0)A$ combines the find and artifact scores in order to express the quality of a peak list in a single number, which takes into account that one strives towards maximizing the number of real peaks while minimizing the number of artifacts. The overall score includes a weighting factor, w, to account for the fact that a missing real peak has a more detrimental effect than an incorrect peak on the resonance assignment (Schmidt and Güntert 2012) and structure calculation (Buchner and Güntert 2015). Lower values of w reduce the weight of artifacts in the overall score (Fig. 3b). Based on earlier observations (Buchner and Güntert 2015; Schmidt and Güntert 2012), we used $w = 0.2$ throughout this paper. For this value of w, the overall score has its maximum at a baseline factor of 3.0, which was applied throughout this paper.

Scores for automated peak picking of spectra for the protein ENTH

Automated peak picking of the 16 different spectra of various types that are available for the protein ENTH was performed with CYPICK, AUTOPSY, NMRViewJ, CCPN, and CV-Peak Picker. The computation times for CYPICK varied between 1 s for the ^{15}N-HSQC and 31 s for the ^{13}C-resolved NOESY spectrum on a standard desktop computer. The resulting peak picking scores are shown in Table 1 and Fig. S2. Over all spectra, the average find scores for the different algorithms were similar, ranging from 72 to 76%, whereas the average artifact scores displayed a higher degree of variation, from 29 to 49% (Table 1). Also the average overall scores vary appreciably from 55 to 68% for the different algorithms. CYPICK obtained the highest values.

Considering only the two 2D HSQC spectra, CYPICK produced peak lists with an acceptable find score and one of the lowest artifact scores, together with CV-Peak Picker, both share similar overall scores. CCPN peak lists achieved the highest find and artifact scores, indicative of an underestimation of the noise threshold. Consequently, their overall score was lowest. The standard deviations within the '2D' group was relatively high on account of the significant differences in resolution and overlap of the ^{15}N-HSQC and ^{13}C-HSQC spectra. Automatic peak picking on the ^{15}N-HSQC spectrum is in general performed more accurately with find scores around 90% (Fig. S2). This can be explained by the fact that the ^{15}N-HSQC is among the most sensitive experiments with well resolved peaks and very few artifacts. Automatic peak picking

Table 1 Peak picking scores [%] for automated peak picking relative to manually prepared reference peak lists for the protein ENTH

	CYPICK	AUTOPSY	NMRViewJ	CCPN	CV-Peak Picker
All available peak lists					
Find score	75 ± 14	72 ± 15	76 ± 12	74 ± 14	72 ± 14
Artifact score	29 ± 11	49 ± 15	44 ± 9	49 ± 18	35 ± 19
Overall score	68 ± 14	56 ± 18	63 ± 11	55 ± 17	61 ± 19
2D: 2D [^{15}N,^{1}H]- and [^{13}C,^{1}H]-HSQC					
Find score	73 ± 18	65 ± 1	76 ± 16	79 ± 13	71 ± 18
Artifact score	19 ± 1	33 ± 8	36 ± 4	54 ± 37	14 ± 5
Overall score	69 ± 17	58 ± 1	67 ± 13	48 ± 46	69 ± 18
Backbone: CBCANH, CBCA(CO)NH, HNCA, HN(CO)CA, HNCO, HN(CA)CO					
Find score	84 ± 14	86 ± 10	87 ± 10	84 ± 13	83 ± 15
Artifact score	28 ± 10	40 ± 12	46 ± 8	49 ± 18	26 ± 9
Overall score	77 ± 10	73 ± 9	71 ± 8	63 ± 11	76 ± 12
Side-chain: HBHA(CO)NH, (H)CC(CO)NH, H(CCCO)NH, HCCH-COSY, (H)CCH-TOCSY, HCCH-TOCSY					
Find score	65 ± 9	61 ± 13	66 ± 8	66 ± 7	60 ± 7
Artifact score	38 ± 9	62 ± 9	45 ± 12	50 ± 10	46 ± 15
Overall score	56 ± 10	39 ± 14	55 ± 11	50 ± 12	47 ± 11
NOESY: 3D ^{13}C-edited and ^{15}N-edited NOESY					
Find score	79 ± 4	72 ± 5	74 ± 10	67 ± 20	73 ± 1
Artifact score	19 ± 9	51 ± 2	41 ± 4	41 ± 29	53 ± 32
Overall score	75 ± 2	56 ± 2	64 ± 7	52 ± 4	46 ± 27

Mean and standard deviation are calculated over the score values of the given sets of peak lists

of ^{13}C-HSQC spectra, on the other hand, is much more demanding due to the high degree of overlap usually being present (Fig. S3).

Automatic peak picking of the triple resonance 'backbone' spectra for backbone assignment yielded uniformly high average find scores of 83–87% (Table 1). Within this group CYPICK and CV-Peak Picker achieved the lowest average artifact score of 28 and 26%, respectively, compared to 40–49% for the other algorithms. Due to their higher sensitivity and better resolution, backbone assignment spectra are in general more straightforward to pick than side-chain experiments. HNCO and HN(CO)CA spectra were picked with find scores close to 100% by CYPICK (Fig. S2), which reflects the high sensitivity and resolution of these spectra. In HNCA, HN(CA)CO and CBCANH, CYPICK missed some weak peaks (approximately 50 peaks in HNCA and HN(CA)CO, and 150 peaks in CBCANH) that are buried in noise and show irregular peak shapes. Artifacts within these lists from CYPICK can be attributed mainly to sinc wiggles.

'Side-chain' spectra peak picking was performed with average find scores of 60–66% and relatively high average artifact scores of 38–62% (of these 15–25% can be attributed to solvent signals that were not filtered out), which in case of AUTOPSY did even exceed the find score (Table 1). The highest average find score was achieved by the NMRViewJ, CCPN, and CYPICK peak lists. CYPICK peak lists showed the lowest average artifact score among the programs, resulting again in the highest overall score of 56%, closely followed by NMRViewJ, whereas the other algorithms have overall scores that are 5–17% lower. TOCSY- and COSY-type sidechain assignment spectra usually exhibit a high degree of overlap, which makes automatic peak picking challenging and leads to the omission of many real signals by CYPICK because these peaks show deviations from the expected peak shape.

Automatic peak picking of the 3D NOESY spectra of ENTH was best performed by CYPICK which produced the highest mean find score of 79 vs. 67–74% for the other programs, as well as the lowest average artifact score (19 vs. 41–53%), leading to a significantly higher average overall score (75 vs. 46–64%).

When comparing automatic peak picking by CYPICK to the other programs, the higher robustness manifested by consistently highest overall scores is mainly due to the fact that CYPICK picks considerably fewer artifacts than other methods (Table 1, Fig. S2). Approximately 20% of the CYPICK artifacts in the ^{13}C-NOESY spectrum were localized in a region of 4.7 ± 0.5 ppm and can accordingly be attributed to solvent signals. The find scores are more uniform; those from CYPICK are usually among the highest. CYPICK performs particularly well in the automated peak picking of NOESY spectra, which is promising for NOE distance restraint-based structure calculation and for the solely NOESY-based chemical shift assignment procedure in FLYA (Schmidt and Güntert 2013). The individual scores for each spectrum and program in Fig. S2 show a stable performance of CYPICK without outliers for individual spectra.

Automated resonance assignment, NOE assignment and structure calculation

Automatically established peak lists for the proteins ENTH, RHO, and SH2 were used as input for automated chemical shift assignment with FLYA, followed by combined NOE assignment and structure calculation with CYANA.

Table 2 summarizes the assignment and structure calculation results obtained using all available peak lists as input for FLYA. Structure bundles are presented in Fig. S4 a–c. Despite the abovementioned variations in the peak picking scores, the overall correctness of the chemical shift assignments by FLYA was relatively uniform over the different peak picking methods that were used to prepare the input peak list: 86–90% for ENTH, 87–91% for RHO, and 87–88% for SH2. In all cases, the CYPICK peak lists yielded a result within 1% of the best assignment.

For ENTH, the assignment correctness was best for AUTOPSY and CYPICK, and about 4% lower for CCPN, which is in line with the CCPN peak lists showing the lowest overall score (Table 1). On the other hand, the fact that AUTOPSY yielded the most correct assignment could not have been discerned from the peak picking score values. The correctness of the resonance assignment was reflected in the structural statistics. The backbone RMSD to the reference was 0.90 Å for the structure obtained using the CYPICK peak lists, and 0.99 Å for AUTOPSY, whereas NMRViewJ, CV-Peak Picker, and CCPN yielded RMSD bias values well above 1 Å.

For RHO, NMRViewJ, CYPICK and CV-Peak Picker achieved a similar overall chemical shift correctness of 89–91%, whereas CCPN yielded 87%. The resulting structures were closest to the reference for CYPICK with a backbone RMSD of 1.35 Å, followed by NMRViewJ and CV-Peak Picker with RMSD bias below 1.75 Å. In case of CCPN, however, the structure calculation converged to an incorrect structure bundle. This can be explained by a lack of structural information that could be deduced from the NOESY peak lists. Automated NOE assignment based on the CYPICK peak lists led to 2392 distance restraints, of which 725 were long-range, whereas automated NOE assignment with CCPN peak lists resulted in a significantly lower number of 1192 distance restraints, of which only 214 were long-range.

For SH2, the chemical shift assignment accuracy was essentially the same with the peak lists from all programs,

Table 2 Results of FLYA automated chemical shift assignment using all available peak lists and CYANA structure calculation

	CYPICK	AUTOPSY	NMRViewJ	CCPN	CV-Peak Picker
ENTH					
Backbone (%)	95.4	96.0	94.9	92.7	94.9
Side-chain (%)	85.2	85.3	83.5	80.7	82.8
All atoms (%)	89.4	89.7	88.2	85.5	87.7
RMSD radius (Å)	0.48	0.33	0.41	0.77	0.70
RMSD bias (Å)	0.91	0.99	1.20	1.78	1.55
RHO					
Backbone (%)	96.4	–	95.0	92.6	95.3
Side-chain (%)	86.2	–	88.5	85.5	84.3
All atoms (%)	90.6	–	91.3	87.4	89.1
RMSD radius (Å)	0.27	–	0.35	1.49	0.37
RMSD bias (Å)	1.35	–	1.61	6.41	1.74
SH2					
Backbone (%)	96.1	–	91.6	97.1	97.1
Side-chain (%)	81.4	–	83.4	81.4	81.6
All atoms (%)	87.3	–	86.7	87.7	87.9
RMSD radius (Å)	0.21	–	0.22	0.22	0.31
RMSD bias (Å)	0.98	–	0.91	1.23	1.07

'Backbone', 'Side-chain' and 'All atoms' refers to the chemical shift assignment correctness with respect to a manual chemical shift assignment. 'Backbone' includes the atoms N, H^N, C', C^α, and C^β, 'Side-chain' includes all atoms except 'Backbone' atoms, 'All atoms' includes all atoms. RMSD radius is the average backbone RMSD of the 20 individual conformers to their mean coordinates. RMSD bias is the backbone RMSD between the mean coordinates of the structure bundle and the reference structure. Residue ranges for RMSDs calculation, determined with CYRANGE (Kirchner and Güntert 2011): 9–102 and 113–130 of ENTH, 6–125 of RHO, and 8–109 for SH2

showing only 1.2% variation. The structural accuracy was also very similar. RMSD bias values below 1 Å were achieved with CYPICK and NMRViewJ peak lists, whereas CV-Peak Picker and CCPN yielded RMSD bias values slightly above 1.0 Å.

We also tried to obtain the resonance assignments by automated chemical shift assignment with FLYA using as input exclusively the 3D NOESY spectra. This approach is generally challenging for FLYA and requires good input NOESY peak lists (Ikeya et al. 2011; Schmidt and Güntert 2013). Using the NOESY peak lists from CYPICK, 77–80% correct assignments could be achieved for the three proteins ENTH, RHO, and SH2 (Table 3). Structure bundles are presented in Fig. S4 (d)-(f). The peak lists from

Table 3 Results of FLYA automated chemical shift assignment using exclusively ^{13}C-edited and ^{15}N-edited NOESY peak lists and CYANA structure calculation

	CYPICK	AUTOPSY	NMRViewJ	CCPN	CV-Peak Picker
ENTH					
All atoms (%)	79.3	71.8	75.4	66.0	73.6
RMSD radius (Å)	0.51	0.61	0.44	4.39	2.98
RMSD bias (Å)	1.43	3.58	2.40	10.31	4.86
RHO					
All atoms (%)	79.5	–	76.1	72.2	78.7
RMSD radius (Å)	0.29	–	0.55	4.60	0.43
RMSD bias (Å)	2.11	–	4.46	8.95	3.49
SH2					
All atoms (%)	77.0	–	70.8	80.3	79.0
RMSD radius (Å)	0.29	–	0.31	0.38	0.41
RMSD bias (Å)	1.56	–	1.73	1.50	2.20

See Table 2 for details on RMSD calculation. AUTOPSY peak lists were available only for ENTH from an earlier study (López-Méndez and Güntert 2006). For technical reasons, the program could not be run for the other proteins

the other programs yielded in general fewer correct assignments, except for CCPN in the case SH2, where 80% correct assignments were achieved, as compared to 77% for CYPICK. This is reflected also in the accuracy in the structures obtained by automated NOESY assignment based on the FLYA chemical shifts. CYPICK yielded backbone RMSDs to the reference structure of 1.4–2.1 Å, i.e. for all three proteins an essentially correct structure, whereas most of the structures obtained for ENTH and RHO using the peak lists from the other programs were incorrect with RMSD bias values of 2.4–10.3 Å (Table 3). Only for the smaller SH2 protein the peak lists from all programs were sufficient to yield a structure with 1.5–2.2 Å backbone RMSD to the reference. For ENTH, these results can be compared with the NOESY peak list scores of Table 1. The best overall scores for the NOESY peak lists were achieved with CYPICK (75%), followed by NMRViewJ (64%), and the other programs (46–56%). The different quality of these NOESY peak lists is clearly reflected in Table 3: CYPICK peak lists yielded the highest assignment correctness (79%) and lowest RMSD bias (1.4 Å), followed by NMRViewJ (75%/2.4 Å), and the other programs (66–74%/3.6–10.3 Å).

Structure calculation of CASD-NMR proteins using NOESY peak lists from CYPICK

Critical Assessment of automated Structure Determination of proteins by NMR (CASD-NMR) is a project for the blind testing of routine, fully automated determination of protein structures from NMR data (Rosato et al. 2009, 2012). From the most recent round of CASD-NMR, NMR data sets are available for ten proteins (Rosato et al. 2015), comprising NOESY spectra, NOESY peak lists, manually determined reference chemical shift assignments, and reference structures. We performed automated peak picking using CYPICK with default parameters for all these NOESY spectra. Together with the protein sequence, the reference chemical shift assignment, and torsion angles restraints derived from the reference assignment, the peak lists from CYPICK were then used as input for combined automated NOE assignment and structure calculation by CYANA. Results are given in Table 4. Structure bundles are presented in Fig. S5. Automatic peak picking of the CASD-NMR NOESY spectra led to overall scores of 53–84% with respect to the structure-based ATNOS cycle 7 peak lists (Guerry et al. 2015) that were used as a reference. In most cases, these scores were lower than those observed above for the NOESY peak lists of the protein ENTH (Table 1). One reason for this are the significantly higher artifact scores of the CYPICK peak lists. These were computed with respect to the final ATNOS peak lists, which were filtered based on the known chemical shift assignment and the 3D structure and thus contain very few artifacts. However, CYPICK and ATNOS peak lists share to a large extent the same peaks, as expressed by high find scores ranging from 70 to 93%. It is also possible that CYPICK identified true peaks that ATNOS peak lists lack. In most cases, scores for the ^{15}N-resolved NOESY peak list are better than for the ^{13}C-resolved NOESY, which is complicated by a high degree of signal overlap.

For five out of ten proteins, i.e. HR2876B, HR2876C, HR6430A, HR6470A, and OR135, structure calculation with automatic picked NOESY peak lists by CYPICK was successful, yielding structures with a backbone RMSD to the reference structure of 0.6–1.1 Å (Table 4). Also for the other proteins correctly folded structures were found, albeit

Table 4 Peak picking and structure calculation results for CASD-NMR proteins

Protein	Residues	Scores [%] of CYPICK versus ATNOS cycle 7 (^{13}C-/^{15}N-resolved NOESY)			Backbone RMSD to reference [Å]		
		Find	Artifact	Overall	CYPICK	Raw	Refined
HR2876B	107	76.7/91.1	46.8/35.0	63.2/81.3	0.89	0.95	0.79
HR2876C	97	85.5/93.4	52.0/68.8	67.0/52.5	0.98	0.88	0.71
HR5460A	160	77.5/88.8	54.2/48.1	59.2/72.4	2.95	3.38	1.38
HR6430A	99	72.3/86.9	47.3/35.5	59.4/77.4	1.07	1.15	0.92
HR6470A	69	70.3/82.5	49.7/42.8	56.4/70.1	0.60	0.61	0.37
HR8254A	73				1.95	7.43	0.77
OR135	83	82.3/87.3	46.8/58.7	67.8/62.5	0.95	1.13	0.89
OR36	134	88.0/87.1	58.7/35.7	63.0/77.5	3.02	1.03	0.98
StT322	63				2.08	6.73	1.49
YR313A	119	73.9/89.7	43.8/24.4	62.4/83.9	3.22	1.64	1.59

Residues ranges for RMSD calculation (Rosato et al. 2015): 13–105 for HR2876B, 17–91 for HR2876C, 14–25 and 33–158 for HR5460A, 14–99 for HR6430A, 15–56 for HR6470A, 554–608 for HR8254A, 4–74 for OR136, 2–46 and 53–125 for OR36, 23–63 for StT322, and 17–41 and 45–115 for YR313A. ATNOS peak lists are not available for HR8254A and StT322 (Guerry et al. 2015)

with slightly higher RMSD biases of 2.0–3.2 Å. For comparison, the RMSD bias of the structures obtained by the same approach but based on refined manual peak lists was 0.4–1.6 Å (Table 4). In addition to manually refined final peak lists, the CASD-NMR data sets include also uncurated, "raw" peak lists from earlier stages of the original structure determination. These "raw" peak lists yielded structures with RMSD bias values of 1.0–7.4 Å (Table 4). In general, the peak lists from CYPICK thus yielded structures with an accuracy between those obtained from the manually curated and uncurated peak lists provided by CASD-NMR.

Computation time

The computation time for CYPICK is short and depends mainly on the size of the spectrum and the number of local extrema that are analyzed. For instance, peak picking of the ENTH spectra of Table 1 required between 1 s for the ^{15}N-HSQC spectrum and 31 s for the ^{13}C-resolved NOESY spectrum on an Intel Core i7 3.2 GHz processor.

Conclusions

In this paper we introduced the CYPICK algorithm as an automated peak picking method that analyzes geometric criteria for contour lines. The approach uses only local spectral information to identify peaks at a given location in a multidimensional spectrum. This makes it universally applicable to any type of multidimensional NMR spectrum, requiring as input only the spectrum of interest. In general, the relative scaling factors for the spectral dimensions are the only parameters to be set by the user. Despite the straightforward mode of use of CYPICK, the results are in the majority of cases comparable or better than those from other, also more sophisticated peak picking algorithms. CYPICK achieves a good balance between picking real signals and rejecting artifacts, and the resulting peak lists are sufficiently good to determine the resonance assignments and 3D structures of proteins by a fully automatic approach.

Experience with these applications indicates several future directions to improve the reliability of CYPICK: (i) Peak picking by CYPICK requires a local extremum as start point for peak identification. Signals that do not present a local extremum, such as "shoulders" located on the slope of a stronger, overlapping peak, are currently discarded and not further analyzed. Relaxing the requirement for a local extremum can improve the completeness of peak lists for crowded spectra, such as ^{13}C-HSQC, HCCH-TOCSY and NOESY. (ii) Very weak signals that do not sufficiently exceed the noise level are currently discarded. Refined criteria on the regularity of peak contours may enable the identification of very weak but "well-shaped" signals without unduly increasing the picking of artifacts. (iii) Most of the picked artifacts originate from small regions of the spectrum, typically narrow bands. Their number may be reduced significantly by a better recognition and exclusion of these problem regions. (iv) Peak picking by CYPICK does not take into account other information than the local features of the spectrum at and near the location of interest. It has been shown that especially the number of artifact peaks can be reduced by considering self-consistency within a spectrum or between spectra (Hiller et al. 2005), or by guiding peak picking by external information, such as known chemical shift assignments or a known 3D structure (Herrmann et al. 2002b). (v) In situations of strong overlap more real signals could be identified by visual inspection than by CYPICK. Deconvolution methods for overlapping peaks may improve this situation. In addition, it is conceivable to make use of the contour-based quality factors Q_{rad} and Q_{con} in automated resonance assignment and NOESY assignment in order to treat reliable and tentative peaks differently in these algorithms.

In conclusion, with CYPICK a stable and versatile automated peak picking method has been integrated into the CYANA software package that removes a bottleneck in its otherwise fully automated pipeline of resonance assignment, NOESY assignment, and structure calculation.

Acknowledgements We thank Torsten Herrmann for providing NOESY spectra and peak lists produced by ATNOS for the CASD-NMR proteins, Piotr Klukowski for providing peak lists produced by the CV-Peak Picker software, and Fred Damberger for helpful discussions. We gratefully acknowledge financial support by the Lichtenberg program of the Volkswagen Foundation, a Grant-in-Aid for Scientific Research of the Japan Society for the Promotion of Science (JSPS), and a Eurostars grant by the Swiss Confederation.

References

Alipanahi B, Gao X, Karakoc E, Donaldson L, Li M (2009) PICKY: a novel SVD-based NMR spectra peak picking method. Bioinformatics 25:i268–i275

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

Bourgeois F, Lassalle JC (1971) An extension of the Munkres algorithm for the assignment problem to rectangular matrices. Commun ACM 14:802–804

Buchner L, Güntert P (2015) Systematic evaluation of combined automated NOE assignment and structure calculation with CYANA. J Biomol NMR 62:81–95

Garrett DS, Powers R, Gronenborn AM, Clore GM (1991) A common sense approach to peak picking two-, three- and four-dimensional spectra using automatic computer analysis of contour diagrams. J Magn Reson 95:214–220

Goddard TD, Kneller DG (2001) Sparky 3. University of California, San Francisco

Guerry P, Duong VD, Herrmann T (2015) CASD-NMR 2: robust and accurate unsupervised analysis of raw NOESY spectra and protein structure determination with UNIO. J Biomol NMR 62:473–480

Güntert P (1998) Structure calculation of biological macromolecules from NMR data. Q Rev Biophys 31:145–237

Güntert P, Buchner L (2015) Combined automated NOE assignment and structure calculation with CYANA. J Biomol NMR 62:453–471

Güntert P, Dötsch V, Wider G, Wüthrich K (1992) Processing of multidimensional NMR data with the new software PROSA. J Biomol NMR 2:619–629

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

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

Herrmann T, Güntert P, Wüthrich K (2002b) Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. J Biomol NMR 24:171–189

Hiller S, Fiorito F, Wüthrich K, Wider G (2005) Automated projection spectroscopy (APSY). Proc Natl Acad Sci USA 102:10876–10881

Ikeya T, Jee J-G, Shigemitsu Y, Hamatsu J, Mishima M, Ito Y, Kainosho M, Güntert P (2011) Exclusively NOESY-based automated NMR assignment and structure determination of proteins. J Biomol NMR 50:137–146

Jee J, Güntert P (2003) Influence of the completeness of chemical shift assignments on NMR structures obtained with automated NOE assignment. J Struct Funct Genom 4:179–189

Johnson BA (2004) Using NMRView to visualize and analyze the NMR spectra of macromolecules. Meth Mol Biol 278:313–352

Johnson BA, Blevins RA (1994) NMR View - a computer program for the visualization and analysis of NMR data. J Biomol NMR 4:603–614

Kirchner DK, Güntert P (2011) Objective identification of residue ranges for the superposition of protein structures. BMC Bioinformatics 12:170

Klukowski P, Walczak MJ, Gonczarek A, Boudet J, Wider G (2015) Computer vision-based automated peak picking applied to protein NMR spectra. Bioinformatics 31:2981–2988

Koga N, Tatsumi-Koga R, Liu GH, Xiao R, Acton TB, Montelione GT, Baker D (2012) Principles for designing ideal protein structures. Nature 491:222–227

Koradi R, Billeter M, Engeli M, Güntert P, Wüthrich K (1998) Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. J Magn Reson 135:288–297

Liu Z, Abbas A, Jing BY, Gao X (2012) WaVPeak: picking NMR peaks through wavelet-based smoothing and volume-based filtering. Bioinformatics 28:914–920

López-Méndez B, Güntert P (2006) Automated protein structure determination from NMR spectra. J Am Chem Soc 128:13112–13122

López-Méndez B, Pantoja-Uceda D, Tomizawa T, Koshiba S, Kigawa T, Shirouzu M, Terada T, Inoue M, Yabuki T, Aoki M, Seki E, Matsuda T, Hirota H, Yoshida M, Tanaka A, Osanai T, Seki M, Shinozaki K, Yokoyama S, Güntert P (2004) NMR assignment of the hypothetical ENTH-VHS domain At3g16270 from *Arabidopsis thaliana*. J Biomol NMR 29:205–206

Lorensen WE, Cline HE (1987) Marching cubes: a high resolution 3D surface construction algorithm. Comp Graph 21:163–169

Munkres J (1957) Algorithms for the assignment and transportation problems. J Soc Indust Appl Math 5:32–38

Orekhov VY, Ibraghimov IV, Billeter M (2001) MUNIN: a new approach to multi-dimensional NMR spectra interpretation. J Biomol NMR 20:49–60

Pantoja-Uceda D, López-Méndez B, Koshiba S, Kigawa T, Shirouzu M, Terada T, Inoue M, Yabuki T, Aoki M, Seki E, Matsuda T, Hirota H, Yoshida M, Tanaka A, Osanai T, Seki M, Shinozaki K, Yokoyama S, Güntert P (2004) NMR assignment of the hypothetical rhodanese domain At4g01050 from *Arabidopsis thaliana*. J Biomol NMR 29:207–208

Pantoja-Uceda D, López-Méndez B, Koshiba S, Inoue M, Kigawa T, Terada T, Shirouzu M, Tanaka A, Seki M, Shinozaki K, Yokoyama S, Güntert P (2005) Solution structure of the rhodanese homology domain At4g01050(175–295) from *Arabidopsis thaliana*. Protein Sci 14:224–230

Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1986) Numerical recipes. The art of scientific computing. Cambridge University Press, Cambridge

Rosato A, Bagaria A, Baker D, Bardiaux B, Cavalli A, Doreleijers JF, Giachetti A, Guerry P, Güntert P, Herrmann T, Huang YJ, Jonker HRA, Mao B, Malliavin TE, Montelione GT, Nilges M, Raman S, van der Schot G, Vranken WF, Vuister GW, Bonvin AMJJ (2009) CASD-NMR: critical assessment of automated structure determination by NMR. Nat Methods 6:625–626

Rosato A, Aramini JM, Arrowsmith C, Bagaria A, Baker D, Cavalli A, Doreleijers JF, Eletsky A, Giachetti A, Guerry P, Gutmanas A, Güntert P, He YF, Herrmann T, Huang YPJ, Jaravine V, Jonker HRA, Kennedy MA, Lange OF, Liu GH, Malliavin TE, Mani R, Mao BC, Montelione GT, Nilges M, Rossi P, van der Schot G, Schwalbe H, Szyperski TA, Vendruscolo M, Vernon R, Vranken WF, de Vries S, Vuister GW, Wu B, Yang YH, Bonvin AMJJ (2012) Blind testing of routine, fully automated determination of protein structures from NMR data. Structure 20:227–236

Rosato A, Vranken W, Fogh RH, Ragan TJ, Tejero R, Pederson K, Lee HW, Prestegard JH, Yee A, Wu B, Lemak A, Houliston S, Arrowsmith CH, Kennedy M, Acton TB, Xiao R, Liu GH, Montelione GT, Vuister GW (2015) The second round of critical assessment of automated structure determination of proteins by NMR: CASD-NMR-2013. J Biomol NMR 62:413–424

Schmidt E, Güntert P (2012) A new algorithm for reliable and general NMR resonance assignment. J Am Chem Soc 134:12817–12829

Schmidt E, Güntert P (2013) Reliability of exclusively NOESY-based automated resonance assignment and structure determination of proteins. J Biomol NMR 57:193–204

Scott A, Pantoja-Uceda D, Koshiba S, Inoue M, Kigawa T, Terada T, Shirouzu M, Tanaka A, Sugano S, Yokoyama S, Güntert P (2004) NMR assignment of the SH2 domain from the human feline sarcoma oncogene FES. J Biomol NMR 30:463–464

Scott A, Pantoja-Uceda D, Koshiba S, Inoue M, Kigawa T, Terada T, Shirouzu M, Tanaka A, Sugano S, Yokoyama S, Güntert P (2005) Solution structure of the Src homology 2 domain from the human feline sarcoma oncogene Fes. J Biomol NMR 31:357–361

Shen Y, Delaglio F, Cornilescu G, Bax A (2009) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 44:213–223

Shimrat M (1962) Algorithm-112. Position of point relative to polygon. Commun ACM 5:434–434

Silver R (1960) An algorithm for the assignment problem. Commun ACM 3:605–606

Skinner SP, Fogh RH, Boucher W, Ragan TJ, Mureddu LG, Vuister GW (2016) CcpNmr AnalysisAssign: a flexible platform for integrated NMR analysis. J Biomol NMR 66

Vranken WF, Boucher W, Stevens TJ, Fogh RH, Pajon A, Llinas M, Ulrich EL, Markley JL, Ionides J, Laue ED (2005) The CCPN data model for NMR spectroscopy: development of a software pipeline. Proteins 59:687–696