Automation of peak-tracking analysis of stepwise perturbed NMR spectra

J Biomol NMR manuscript No.
(will be inserted by the editor)

Automation of peak-tracking analysis of stepwise perturbed NMR spectra

Tommaso Banelli[1] · **Marco Vuano**[1] · **Federico Fogolari**[2,3] · **Andrea Fusiello**[4] · **Gennaro Esposito**[2,3,5] · **Alessandra Corazza**[1,2]

Received: date / Accepted: date

Abstract We describe a new algoritmic approach able to automatically pick and track the NMR resonances of a large number of 2D NMR spectra acquired during a stepwise variation of a physical parameter. The method has been named TINT (Trace in Track), referring to the idea that a gaussian decomposition traces peaks within the tracks recognised through 3D matematical morphology. It is capable of determining the evolution of the chemical shifts, intensity and linewidths of each tracked peak.

The performances obtained in term of track reconstruction and correct assignment on realistic synthetic spectra were high above 90% when a noise level similar to that of experimental data were considered. TINT was applied successfully to several protein systems during a temperature ramp in isotope exchange experiments. A comparison with a state-of-the-art algorithm showed promising results for great numbers of spectra and low signal to noise ratios, when the graduality of the perturbation is appropriate.

TINT can be applied to different kinds of high throughput chemical shift mapping experiments, with quasi-continuous variations, in which a quantitative automated recognition is crucial.

[1] Dipartimento di Scienze Mediche e Biologiche, Università di Udine - P.le Kolbe, 4, 33100 Udine, Italy
[2] INBB - Viale Medaglie d'Oro, 306, 00136 Roma, Italy.
[3] Dipartimento di Scienze Matematiche Informatiche e F isiche, Università di Udine - Via delle Scienze, 206, 33100 Udine, Italy.
[4] Dipartimento Politecnico di Ingegneria e Architettura - Università di Udine, Via delle Scienze, 208; 33100 Italy.
[5] Science&Math Division, New York University Abu Dhabi, Saadiyat Campus, PO Box 129188, Abu Dhabi, UAE
E-mail: alessandra.corazza@uniud.it

Keywords peak picking · peak tracking · 2D NMR · mathematical morphology · isotopic exchange · noise estimation

1 Introduction

With the growing storage capability of the modern computers and their decreasing cost, it is more and more feasible to acquire series of stepwise perturbed 2D NMR spectra in various kind of experiments designed to characterize molecular behavior during variation of physical and chemical conditions. In fact it is possible to monitor modifications in NMR spectra following structural and functional changes of molecules in solution due to thermal, pH or external pressure variation, chemical reactions, solvation, complex formation or ligand binding.

What is observed is a strong correlation between the chemical and physical changes of the systems under study and the features of the NMR spectrum, i.e. chemical shifts, intensities, peak multiplicity, peak onset and/or loss. It is theoretically and experimentally confirmed that the major alterations are experienced by those peaks that are correlated to residues that are involved in the process under consideration (see reviews [35,36,49] and references therein) which makes high resolution NMR a so powerful and informative technique in chemistry, physics and biophysics.

In particular the displacement of the chemical shift and the modification of signal linewidth and intensity are used to probe relevant events in experiments such as drug screening [9], SAR by NMR [22,47], protein-protein interaction [32,33,17,13,10], chemical shift covariance analyses [6], isotopic exchange at single temperatures [15,16,41], BLUU-Tramp [40,39] and titration in general [48] .

Very often the NMR experiment of choice, for such studies on proteins, is the highly sensitive 2D HSQC of ^{15}N-enriched proteins [5,4] or fast acquisition versions such as HMQC-SOFAST [44] and BEST-TROSY [29], because they allow one to obtain information at single residue resolution with one signal for each N-H pair. However, homonuclear 2D TOCSY is also often used. Mapping and quantitative evaluation of peak evolution allows one not only to effectively gather experimental points, but also to accurately trace and reconstruct the function, modeling the process, to which thermodynamic formulae, statistical and clustering analyses can be applied.

Although algorithmic approaches have been developed, NMR spectroscopy has not yet raised an appropriate interest by the specialist programmers [19], that could address these issues with an automatic approach. In fact, though the human perceptual capability remains the source of inspiration for new methods, it is likely to fail providing the best results, especially regarding the performance precision, the timing and the analysis completeness, when dealing with massive data, i.e. circumstances in which computers outperform even an expert operator to give an unbiased result. Stepwise perturbed NMR spectra analysis and high throughput screening still offer compelling challenges to automation because of noise, peak overlap, cross shifting and long distance correlation peaks.

Our attempt in automation stems from the need to analyze data resulting from BLUU-Tramp experiment [40,39]; therefore some methodological choices reflect features of the method. An experimental session of BLUU-Tramp produces two sequences of around 200 2D HMQC spectra, that are acquired at regular temperature increments (usually 0.1-0.2K) and time intervals. The choice of a tiny temperature step provides a quasi-continuum evolution. During the first thermal ramp, the protein, previously deuterium-exchanged, undergoes a D-H isotopic exchange with the aqueous solvent. The second thermal ramp, analyzed by our routine, is used as reference in absence of isotopic exchange and it monitors the evolution of the NMR peaks as the temperature slowly changes: every peak shows a slow chemical shift drift (variation of the position in the ^{15}N and ^{1}H frequency space) along with a gradual modification of the intensity and linewidth.

The aim of this work is to implement an automatic data-analysis methodology which allows automatic peak detection (picking) in every spectrum and peak tracking between spectra, henceforth dubbed TinT (Trace in Track).

In literature peak picking and peak tracking analyses are two distinct concepts and are often considered separately.

As for the former procedure, the first proposal has been STELLA [25] in 1990 and the software currently used are based on a variety of methods, such as peak properties [20,24], machine learning algorithms [3,7,11, 43], spectral decomposition [2,27,28,34], wavelet smoothing [30], Benjamini-Hochberg procedure [1], Monte Carlo stochastic approximation and Bayesian statistics [8] and computer vision [26]. A comprehensive review has been given by Liu and coworkers in the introduction of the article describing their algorithm WaVPeak [30].

Fewer articles deal with the automation of the tracking procedure. The proposed methods in the literature are: APET/PROPET (in Felix-Autoscreen) [37], based on bipartite graph matching by systematic tree search methods and simulated annealing approach with heuristic simplification, Nvmap (NMRViewJ [24]) [18], based on search of the nearest pair with a greedy algorithm, GAPT [38] based on best-score-selection under constrain with heuristic simplification, PeakWalker [23] based on many-to-one mapping through maximum weighted k-dimensional matching of the graph. On one hand, all of them have implemented algorithms with a list-based approach considering matching among their own generated peak lists or given by one of the aforementioned peak pickers. Any error or artifact present in the peak list is not correctable since the main routine does not check the actual data. On the other hand, they differ in the score function used for the matching and in the level at which the best choice is made: peaks, pair of spectra or whole paths. Moving from local to global strategies brings all the approaches beyond computational possibilities for protein NMR spectra due to the NP-completess of the problem [23].

Our method is applicable to a series of spectra in which the features of the peaks undergo small modifications from one spectrum to the next due to quasi-continuum progressive sample perturbations. To deal with this issue we propose a novel approach based on morphological filtering [31,45,46] and decomposition. A selection of the region of interest is performed around local maxima collected over a threshold, roughly chosen low enough to maintain all the signal; subsequently signal-peaks are singled-out based on their persistence among the sequence of spectra, considered simultaneously. This is allowed by the application of 3D mathematical morphology which produces the removal of the fluctuating noise and the clustering by contact of the signal peaks. The result is a set of masks selecting each group of connected peaks. The estimated number of peaks of each selection is used by a later subroutine

that performs a decomposition to obtain the parameters of the peaks, modelled as gaussians.

The validity and the efficiency of TINT are demonstrated first on realistic synthetic data, where more than 90% of the tracks are correctly recognized, and then applying it to three BLUU-Tramp experimental sessions using ^{1}H-^{15}N HMQC spectra on three different human proteins: Acylphosphatase (hAcP), β2-microglobulin (β2-m) and Lysozyme (hLys).

2 Methods and algorithms

2.1 Mathematical morphology

The novelty of our approach is to find a solution of the tracking issue in the theoretical framework of mathematical morphology, a powerful theory for image processing [46,45] based on nonlinear geometric approach [31]. For an introduction to Mathematical morphology for image processing see e.g. a chapter by Glasbey [21]. The basic morphological tools are the dilation ($\oplus$) and the erosion ($\ominus$) algorithm that work on a black and white image by altering the distribution of the two colors in two opposite way (Fig. 1): dilation extends the white portion over the black one following a shape defined by the user through a "Structuring Element" (SE); erosion works in the opposite way.

These 2D-image operations can be easily extended to be applied on multidimensional binary matrices, defining:

$$\textit{dilation} \qquad A \oplus S = \bigcup_{z \in S} A_z$$

$$\textit{erosion} \qquad A \ominus S = \bigcap_{z \in S} A_z$$

where A_z is a translation of the image A and S is a SE.

Their combination defines new operators with more complex and sophisticated action (Fig. 1) such as:

$$\textit{opening} \qquad \circ = \ominus\oplus \tag{1}$$

$$\textit{closing} \qquad \bullet = \oplus\ominus \tag{2}$$

We took advantage from the potentiality of the opening operator to erase details smaller than a SE while maintaining unaltered the remainder (Fig. 5b). Then we exploit the capability of the closing operator to connect structures separated by volumes smaller than a SE (Fig. 5c).

2.2 Procedure

The work-flow of our routine is composed by five steps:

1. Default setting parameters definition: this procedure overcomes the need of manual parameter specification. An output file summarizes all the parameters and permits their modification.

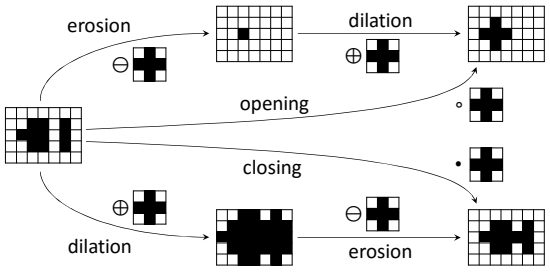

Fig. 1 Morphological operators behavior. The arrows represent the specified algorithms that transform the input images in output images, with the given SE

2. Selection of the region of interest: identification of all the local maxima over a rough intensity threshold in every spectrum and association of a corresponding area.
3. Morphological filter: sequence of morphological operators to track all peak representations along different spectra, while discriminating signal from noise. This stage allows us to group peaks that sooner or later overlap along their evolution, assembling them in a 3D structure that we shall refer to as 3D-blob.
4. Weighted decomposition: fitting all evolving peaks in each 3D-blob slice with 2D-gaussians.
5. Results validation: parameters statistical analysis to select internally coherent results.

The output of TINT is the evolution of the five fundamental descriptors of each peak over the time/ temperature: intensity (I), ^{1}H and ^{15}N frequency positions (δH, δN) and linewidths (λH, λN). In the following, each step of the procedure will be described in some detail.

2.2.1 Default setting parameters definition

A starting procedure was written to help the user to manage parameters needed by the algorithm, although the possibility of modification is maintained with an autogenerated output file. The automatic definition of all the default values through a fast analysis of the first spectrum avoids human bias.

The typical expected peak linewidths, $\lambda\bar{H}$ and $\lambda\bar{N}$, are statistically defined as the median of the set of evaluated linewidths since the signals of protein HSQC spectra can be guessed to exhibit similar shapes in principle. The peak linewidths are estimated, for the first twenty highest peaks, as half horizontal and vertical pixel dimension of the peak section with a threshold of 66% of the maximum. $\lambda\bar{H}$ and $\lambda\bar{N}$ will be involved in

the determination of the limit radius (see section 2.2.2), in the computation of σ of the weight function (see Eq. (5)) and as starting guess in the decomposition procedure (see section 2.2.4).

In order to let the user define the threshold used to initially filter the spectra, an estimation of baseline (B) and the noise level (N) are necessary. Following the white noise definition and the observation that the total signal area occupies a minority of the spectrum area, we compute the center (μ) and the width (γ) of the distribution of the intensities fitted by a gaussian, that estimate the baseline and noise level, respectively. The fitting is performed around the statistic mode value because the contribution of signals affects the gaussian shape only at very high intensities, as evidenced in Fig. 2. This estimation method has been named MIDNE (Modeling Intensity Distribution for Noise Estimation). TINT will propose setting the threshold at $B + 5N$ [42].

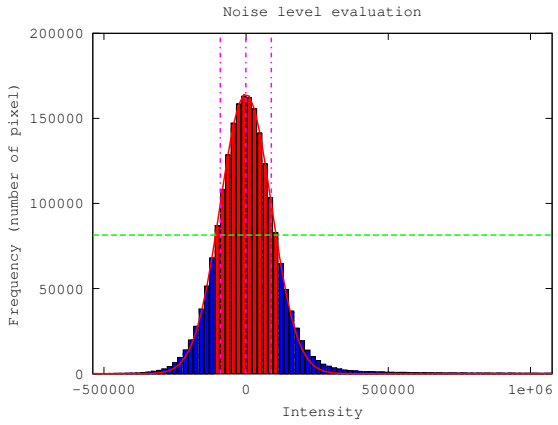

Fig. 2 The MIDNE method applied to the first hAcP spectrum. The distribution of the intensity is represented with a histogram. The green dashed line ws half of the maximum bin height. The bins higher than this value, shown in red, are used to reconstruct the gaussian distribution shown with the red line. The baseline (B) and the noise level (N) are estimated by the center (μ) and the width (γ) of the distribution. The three magenta dotted lines show $\mu - \gamma$, μ and $\mu + \gamma$ of the distribution.

2.2.2 Region of interest selection

All the spectra are uploaded with a zeroing of all points below the previously estimated intensity threshold. A selection of local maxima within a given window is implemented. To apply the subsequent morphological filter, an area, called spot, has to be assigned to each of the recognized maxima: we chose the base of the peak portion above a local threshold corresponding to a given percentage of the peak intensity (Percentage Local Threshold, PLT).

It was soon noted that low intensity local maxima can strongly affect the area of the selected spot when overlapping with higher intensity peaks (Fig. 3). The proposed solution is to limit the area of the lower spot by a disk of a radius equal to 1.5 times the maximum between $\lambda\bar{H}$ and $\lambda\bar{N}$. In this way, a black and white image is obtained from each spectrum containing all the signal spots (Fig. 4).

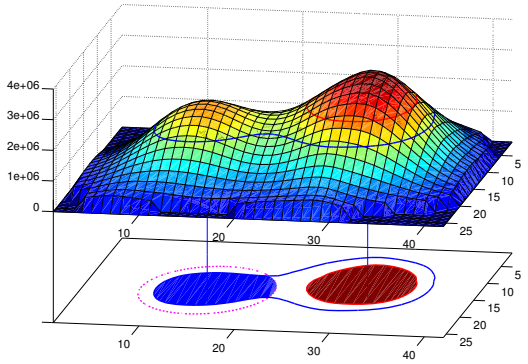

Fig. 3 Example of spot evaluation for two overlapped peaks: PLTs are shown with contour lines in red and blue. The low intensity peak affects and overestimates the area around the higher peak. The magenta dotted line shows the circumference limiting the area of the lower peak to avoid the overestimation of areas. The filled areas are assigned to two maxima.

2.2.3 Morphological filter

The input for the filter is a 3D-matrix composed by the obtained black and white images, stacked one over the other, in which the planar dimension are ^{1}H and ^{15}N frequencies of the 2D spectrum itself and the third dimension is the time/temperature (Fig. 5a). The signal spots are persistent from one spectrum to another or, at most, a slow shift is observed due to thermal drift. In this way, slanted columns, isolated or intersecting each other, will be assembled in the matrix along the third dimension. At variance, noise which is by definition uncorrelated between single spectra, tends to create shorter and unoriented 3D connected structures. Both described features are easily recognizable in Fig. 5a. We designed a filter that cleans the matrix exploiting these different behaviors. It works by erasing (with a morphological opening) black volumes smaller than a given SE (Fig. 5b) and then connecting (with a morphological closing) the surviving volumes closer

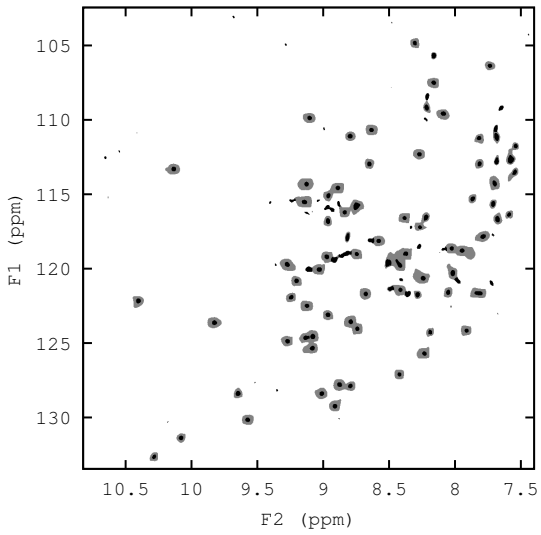

Fig. 4 Comparison between global single threshold (in gray and black) and our PLT method (black only).

than the proper SE (Fig. 5c). This SE has been chosen with a cylindrical shape with the height along the third dimension, to better resemble the silhouette of the column. Minimum radius and height are chosen to be sure to maintain the shape of the peaks however they can be adjusted by the user by modifying the parameters in the file autogenerated by the starting procedure. The connected-components labeling algorithm is subsequently used to uniquely identify subset of connected components, one for each surviving 3D structure (3D-blob) (Fig. 5d). The filtered 3D-matrix is used as a mask to select the signal data and each 3D-blob recalls one group of peaks at a time.

To exclude noise artifacts we analyze only blobs whose persistence is longer than an established cutoff (two times the length of the opening SE). This choice allows anyway to take in account peaks that disappear or emerge as the physical and chemical conditions are changing.

2.2.4 Weighted decomposition

The aim of this procedure is the reconstruction of synthetic spectra by decomposition and modeling of peaks as gaussians. This procedure allows one to obtain I, δH, δN, λH and λN of the 2D gaussian for each peak in each spectrum.

As a prerequisite, the procedure needs an estimation of the number of the involved peaks (n_G) in every 3D-blob. In fact, at this stage, the 3D-blobs that contain one recognized peak for each layer are already

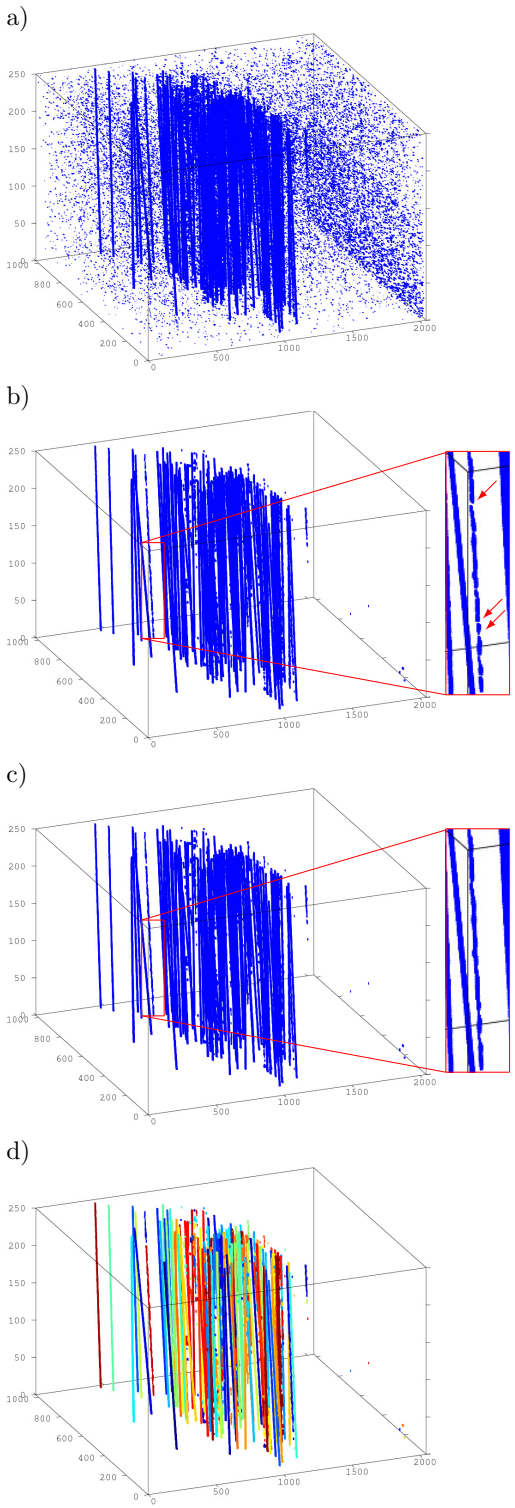

Fig. 5 Elaboration of data during the weighted decomposition. The input 3D-matrix (a), the output of the opening algorithm (b), the output of the closing algorithm (c) and the output of the connected-components labeling algorithm (d) are shown. Gaps (filled by the closing operator) are indicated by the arrows.

tracked in their evolution. If more than one peak is involved in a blob, the appropriate number of gaussians must be used for the decomposition to identify and isolate their contribution to the overall landscape.

A routine estimates n_G by the following steps:

- the number of recognized maxima for each layer in the 3D-blob is stored in an array, A;
- A is processed with a median filter to smooth sharp fluctuations;
- the first 2 modal values are calculated and the highest is chosen if its frequency is greater than a given percentage (5%) of the total number of spectra.

n_G also allows to identify the best layer in the 3D-matrix in which the decomposition procedure starts: within the longest interval of coincidence between n_G and A[i], the most distant layer from the interval borders is chosen, because in the corresponding spectrum the peaks are well distinguished.

The decomposition is implemented as a minimization algorithm with weighted data as target (Fig. 6). For each layer, the decomposition modifies the parameters

$$\bar{x} = \{I_k, \delta H_k, \delta N_k, \lambda H_k, \lambda N_k | k = 1, ..., n_G\}$$

to minimize a cost-function,

$$C(\bar{x}) = \sum_{\bar{p}} H^r \Big(w(\bar{p}) * \big[D(\bar{p}) - R(\bar{p}; \bar{x}) \big] \Big) \tag{3}$$

where w, D, and R are the weight function, the original data and the reconstructed spectrum in the spectral coordinate space ($\bar{p} = (F_1, F_2)$), respectively. H^r is the Huber function

$$H^r(v) = \begin{cases} \frac{1}{2}v^2, & |v| < r \\ rv - \frac{1}{2}r^2, & |v| \geq r \end{cases} \tag{4}$$

where r is a cutoff value equal to the initial threshold value (see section 2.2.1). The Huber function substitutes the square function, normally used for minimization, to reduce the contribution of original data outliers and to obtain a more robust evaluation of the reconstruction.

The weight function focuses the solution of the optimization on the neighborhood of the recognized maxima: we propose to use as weight function the section of the 3D-blob relative to the considered layer in which the sharp transition at the borders is smoothed by extending it with a gaussian shape having a linewidth comparable with the peaks (Fig. 6b):

$$w(\bar{p}) = \begin{cases} 1 & \bar{p} \in M \\ e^{-\left(\frac{d(\bar{p}, M)}{\sigma}\right)^2} & \bar{p} \notin M \end{cases} \tag{5}$$

where M is the morphological recognition set, $d(\bar{p}, M)$ is the minimum Euclidean distance between the $\bar{p}$ point and all the point in M and σ is a distance cutoff chosen as the maximum typical expected peak linewidth evaluated in the starting procedure (section 2.2.1).

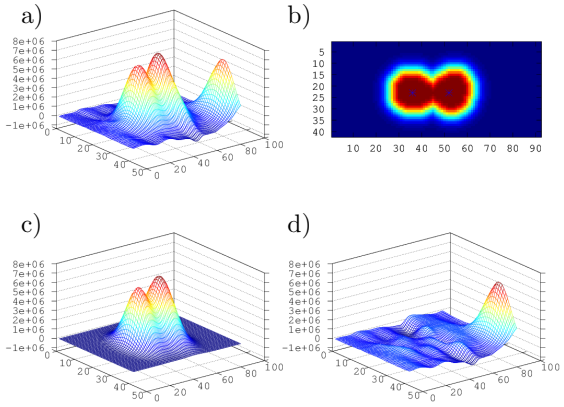

Fig. 6 Weighted deconvolution: data (a), weight function (b), peak reconstruction (c) and remainder (d).

The peak reconstruction (example in Fig. 6c) in the starting layer needs a guess of the 5 parameters involved for each peak. I, δH and δN can be directly taken from the recognized maxima (section 2.2.2), while $\lambda\bar{H}$, $\lambda\bar{N}$ (section 2.2.1) are used to estimate λH, λN. After every estimation λH, λN will be forced to be positive.

For all the other layers, along the downstream direction the starting parameters are estimated using the evaluated ones of the previous spectrum, while the upstream starting parameters consider the next spectrum in the sequence.

2.2.5 Result validation

Statistical analysis is performed on position and linewidths of peaks resulting from the decomposition stage to select valid results. The following condition must be sequentially met:

1. δH and δN must be within the analysis window around the weight function;
2. δH and δN must not be within the areas of a different blob with respect to the analyzed one;
3. λH and λN must be limited by half of the spectrum dimension;
4. in every spectrum, λH and λN must be within the interval centered on the median of the linewidths of the isolated peaks at the same temperature and 6 MAD (Median Absolute Deviation) wide.

It must be noticed that filtering the results after the decomposition, instead of imposing the conditions

as a constrain to the optimizer, allows to use the same results as a discriminant for the quality of the peak identification. In fact results that do not meet the above criteria are symptoms of various kind of error such as:

- peaks that disappear or appear during physical evolution,
- excessive overlap between peaks,
- wrong estimation of involved peaks in a blob (n_G).

3 Experimental methods

3.1 Synthetic spectra

For testing purposes synthetic spectra were generated with realistic features. In particular, one hundred chemical shifts were randomly selected from the (N,HN) assignments of hLys. Consistently with most tracks, the chemical shift temperature dependence was approximated to be linear. The experimental reconstruction of complete tracks from the spectra of the three analyzed proteins, i.e. hAcP, β2-m, hLys, provided the temperature coefficients which were randomly assigned to the synthetic peaks. Intensities and linewidths were randomly assigned in a range of 0.4 to 1.5, and 0.8 to 1.2, respectively, of the averages observed on the three proteins. The time dependence of the intensity was assumed to be at most quadratic, whereas linewidths decrease exponentially towards 60% to %90 of their starting values, consistent with experimental observation (e.g. in Fig. 8). 210 spectra were generated. Noise was added by convolving gaussian white noise with a bidimensional gaussian with parameters optimized to reproduce experimental noise.

3.2 NMR Experiments

The BLUU-Tramp sessions of β2-m (100 amino acids, 279.2 K - 317.2 K), hAcP (99 amino acids, 290.9 K - 315.8 K) and hLys (140 amino acids, 283.0 K - 336.0 K) are used to demonstrate the effectiveness of our algorithm.

Chemical shift changes during the thermal ramp were monitored in ^{1}H-^{15}N HMQC-SOFAST [44] or ^{1}H-^{15}N BEST-TROSY [29]. The spectra of ^{15}N-labeled protein samples were acquired on a Bruker Avance operating at 500 MHz (^{1}H frequency) or a Bruker Avance III equipped with cryoprobe and operating at 600 MHz (^{1}H frequency), respectively at Biophysics laboratory of Udine University and Core Technology Platform of New York University Abu Dhabi. Data were collected over sweep widths of 14 ppm (^{1}H) and 32 ppm (^{15}N) with 768 and 80 points, respectively. All remaining conditions were set according to the protocol previously reported [39, 40].

3.3 Spectral data processing

The spectra were processed with NMRpipe [12] with a sinebell squared apodization function. 1K $\times$ 512 points real spectra were obtained after t1 linear prediction, apodization, zero-filling and finally Fourier transformation.

3.4 TinT

The TinT algorithm is coded in Octave (version 4.0.0) [14] with image (version 2.4.1), optim (version 1.4.1) and statistics (version 1.2.4) packages. The code is available from the authors upon request.

4 Results and Discussion

4.1 Results on synthetic data

Synthetic data, generated as described in 3.1, were analysed by TinT. In order to evaluate the accuracy of the method, for each spectrum the matrix of distances between reconstructed and original peaks was computed. All peaks which were closer than the original linewidths to an original peak, were considered compatible with the original peak. Compatibility at this stage is meant in a many-to-many relationship, due to overlaps or proximity within linewidths. Finally, the most persistent matches are used to produce the output one-to-one mapping between TinT and original tracks. Once a one-to-one mapping has been obtained, the accuracy of TinT was evaluated by two tests:

- the number of complete tracks reconstructed over the total number of original tracks;
- the number of correct peak assignments in all spectra over the product of number of original tracks times number of spectra, in order to account for both complete and partial track reconstructions.

The analysis was repeated for different noise level (with standard deviation from zero to 10% of the peak intensity mean, ranging up to 25% of the lowest intensity peak), and by progressive downsampling of the signal. All results are reported in table 1 and table 2.

It is seen from the tables that the performance of TinT in both complete tracks and detailed reconstruction is excellent for realistic noise levels (say up to 2%)

and it starts to deteriorate going to extreme noise levels. Although the effect of downsampling depends on the specific experimental conditions (e.g. the temperature interval between consecutive spectra) it seems important that a quasi-continuous variation of spectral features is met in the experiments, as it can be inferred from the results in table 2.

Table 1 TINT results for synthetic data with increasing noise level

N/S(mean)	N/S(min)	complete tracks	detailed reconstruction
0 %	0.0 %	97 %	99.96 %
1 %	2.6 %	98 %	99.8 %
2 %	5.2 %	96 %	99.7 %
3 %	7.7 %	87 %	97.2 %
4 %	10.3 %	89 %	97.8 %
5 %	12.9 %	83 %	96.8 %
6 %	15.5 %	73 %	96.9 %
7 %	18.1 %	65 %	90.4 %
8 %	20.6 %	64 %	86.2 %
9 %	23.2 %	52 %	80.4 %
10 %	25.8 %	42 %	75.0 %

Table 2 TINT results for downsampled synthetic data with a 2% N/S(mean) ratio.

downsampling	complete tracks	detailed reconstruction
1	96 %	99.7 %
2	93 %	99.1 %
3	90 %	97.4 %
4	76 %	90.2 %
5	69 %	85.4 %
6	56 %	79.8 %

4.2 Results on experimental data

We tested TINT on three BLUU-Tramp experimental sessions using ^{1}H-^{15}N HMQC spectra on three different human proteins hAcP, β2-m and hLys of 100, 99 and 140 residues, respectively.

For each recognized peak, $I(t)$, $\delta H(t)$, $\delta N(t)$, $\lambda H(t)$ and $\lambda N(t)$ were determined (an example is shown in Fig. 8).

The high number of spectra, peaks and complexities, such as the appearance and the disappearance of signals as well as extensive overlapping do not allow to establish a ground truth, i.e. exact number of peaks and their positions and shapes. Nevertheless in a single spectrum we expect to observe a number of peaks close to the number of N-H pairs. Thus we refer the number of the resulting tracks to the protein length. Considering a reasonable track the one spanning at least half of the number of experiments, the estimated percentages of TINT recognition are around 90%, 88%, 88% for hAcP, β2-m and hLys, respectively (table 3).

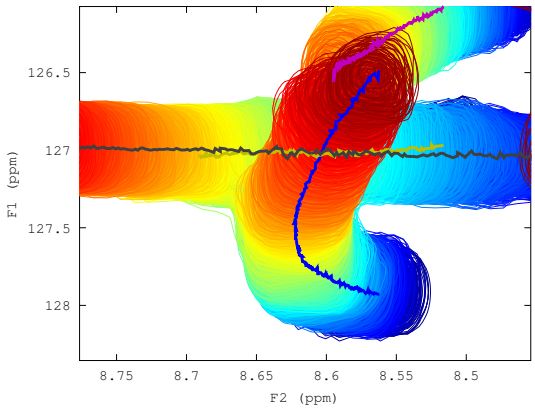

Fig. 7 Recognition of a non linear path in hLys NMR spectrum. Lines show the evolution of the position of the recognized peaks. The flow of the time is represented with a chromatic scale.

Table 3 TINT results for hAcP, β2-m and hLys considering tracks spanning at least half of the number of experiments

Protein	Residues	Tracks	Coverage
hAcP	100	90	90%
β2-m	99	87	88%
hLys	140	123	88%

It must be noticed that our method allows us to recognize nonlinear paths, as shown in Fig. 7, similar to those seen in ligand binding studies [48].

4.3 Comparison with PeakWalker

To better evaluate the method we compared TINT output with the results of PeakWalker [23], a state-of-the-art peak tracking algorithm. It must be noted that PeakWalker has been designed to analyze a smaller number of spectra with larger chemical shift changes than the one analyzed here. Moreover multiple runs of the program (which are not performed here) would result in better estimation of tracks, and finally it pro-

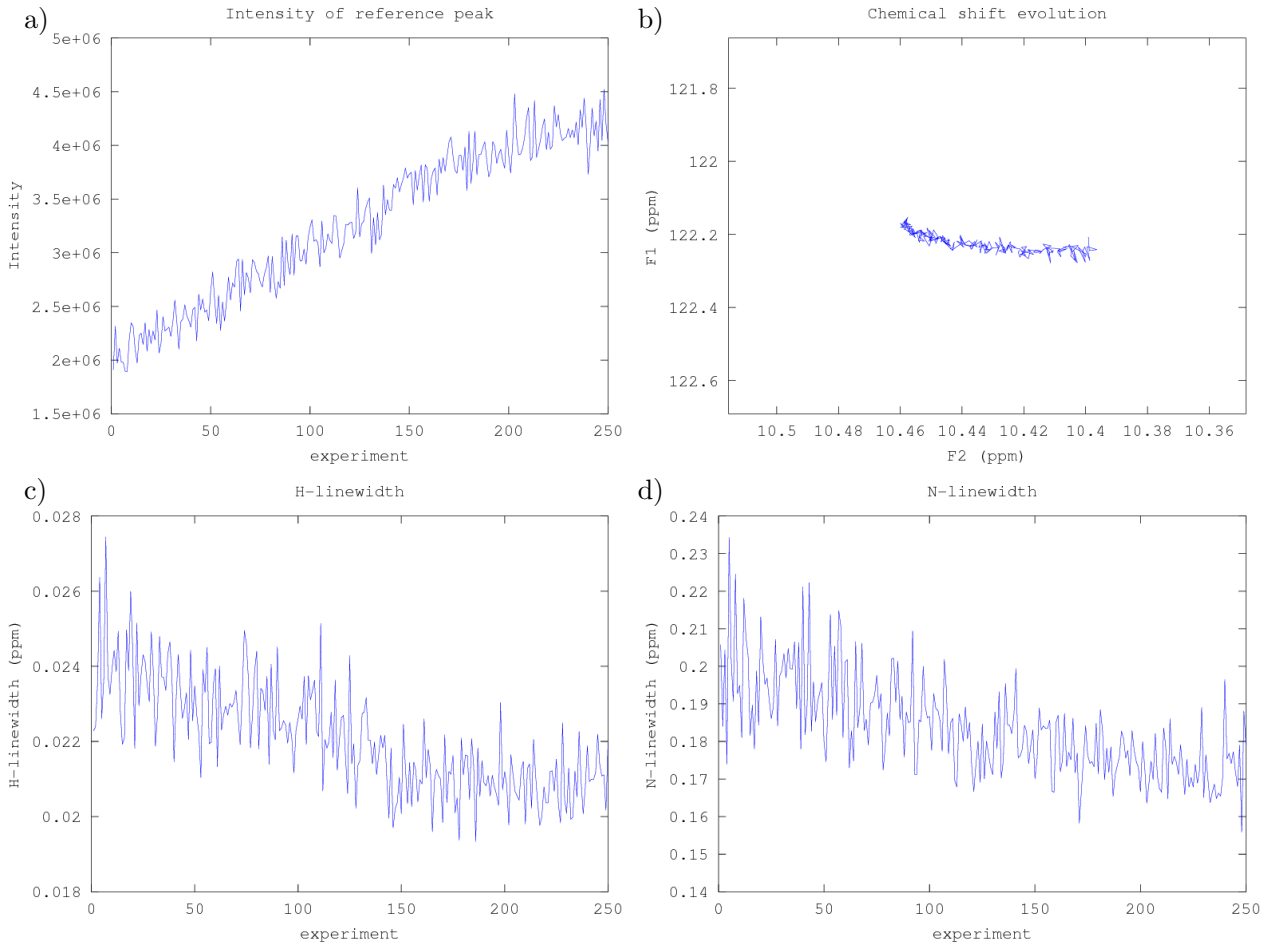

Fig. 8 Evolution of the 5 fundamental parameters of one hAcP peak: $I(t)$ (a), $\delta H(t)$ and $\delta N(t)$ (b), $\lambda H(t)$ (c) and $\lambda N(t)$ (d).

vides multiple choices for track end point mapping, whereas our focus is on track reconstructions. For this reasons the following comparisons should be regarded with some caution. In the following we describe our usage of the program to the best of our possibilities.

In order to reduce human biases in the comparison concerning thresholds, peak picking and validation, the following precautions were adopted:

- the same threshold was used for both methods;
- the peaks coordinates calculated by the region of interest selection were used to fill the peaklists needed for PeakWalker;
- no inferior limit was given to the tracks length.

There were cases in which PeakWalker followed with the same track the evolution of more than one peak (Fig. 9b). To better evaluate the performance of the algorithm, these cases should be filtered but a proper filter would need a prior knowledge of the position of each peak in all the spectra and reconstructing the whole tracks manually would have been impractical.

Each experimental set was analyzed with two different thresholds: a high threshold (T_H), that selects mainly the signals discarding low intensity peaks, and a low threshold (T_L), that keeps those peaks but allows also some noise to enter in the process. Following the Rose criterion [42], they were set as follows:

$$T_L = B + 4N \tag{6}$$

$$T_H = B + 8N \tag{7}$$

where B is the baseline and N is the noise level, both calculated with the starting procedure (see section 2.2.1) on the first spectrum, which is the one less affected by thermal noise. Tables 4, 5 and 6 report the number of signals at different thresholds, grouped by track length.

The outcomes highlight that, at a given threshold, TINT was able to recognize a higher amount of longest tracks than PeakWalker and this is more evident when the T_L is used (Tables 4, 5, 6 and Figures 10, 11 and 12). Furthermore, on decreasing the threshold from T_H to T_L, the number of long tracks recognized by TINT

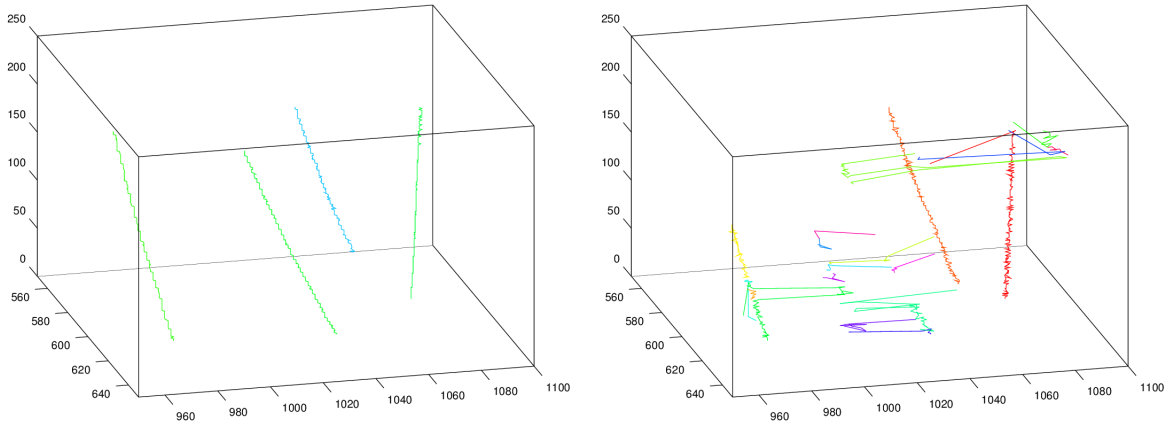

Fig. 9 Comparison between TinT (a) and PeakWalker (b) resulting tracks

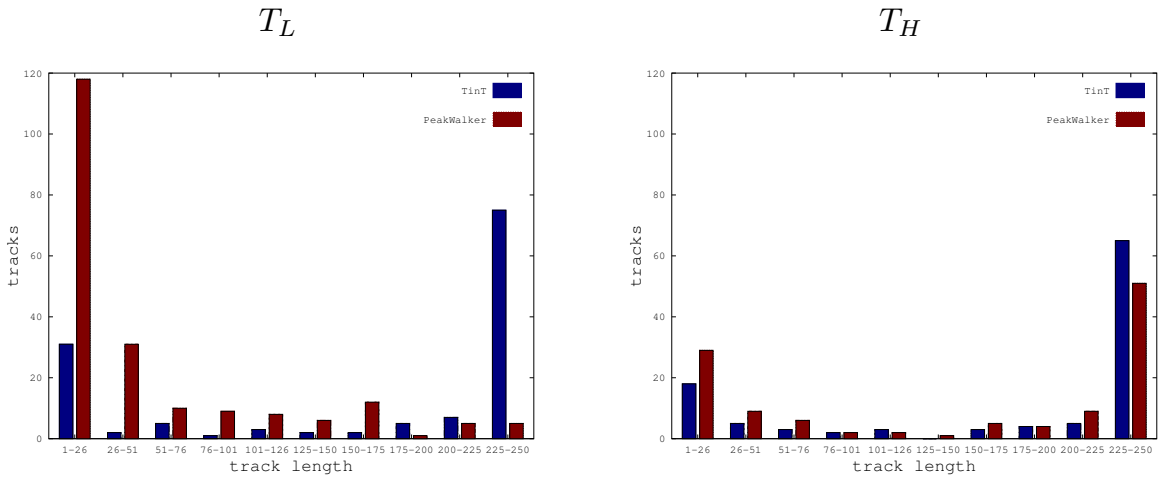

Fig. 10 hAcP. Comparison between TinT (blue) and PeakWalker (red) results at T_L (a) and T_H (b): number of recognized tracks, grouped by track length.

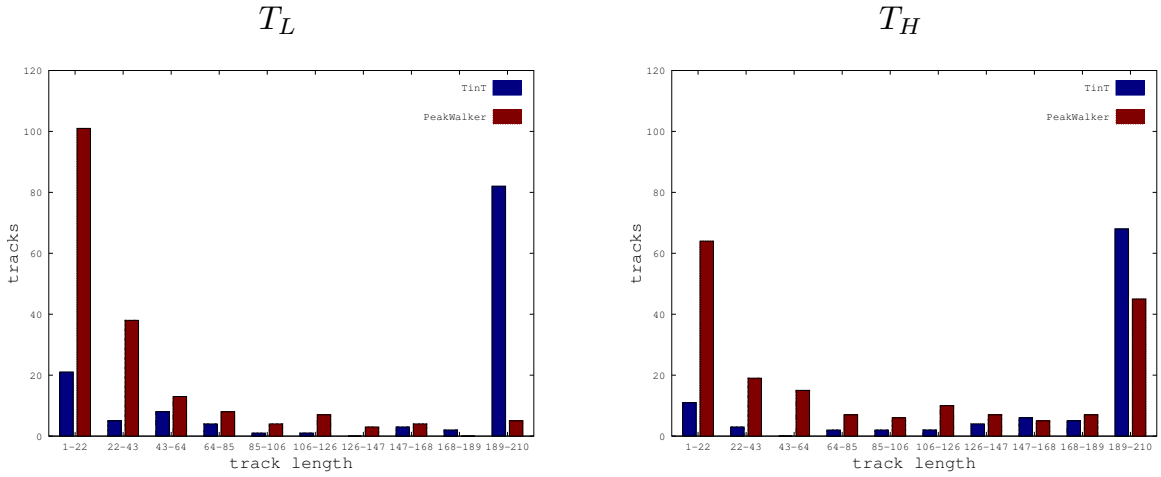

Fig. 11 β2-m. Comparison between TinT (blue) and Peakwalker (red) results at T_L (a) and T_H (b): number of recognized tracks, grouped by track length.

Table 4 hAcP. Comparison between TinT and PeakWalker results: number of recognized tracks at T_L and T_H, grouped by track length.

		track length									
		0-26	26-51	51-76	76-101	101-126	126-150	150-175	175-200	200-225	225-250
TinT	T_L	31	2	5	1	3	2	2	5	7	75
	T_H	18	5	3	2	3	0	3	4	5	65
PeakWalker	T_L	118	31	10	9	8	6	12	1	5	5
	T_H	29	9	6	2	2	1	5	4	9	51

Table 5 β2-m. Comparison between TinT and Peakwalker results: number of recognized tracks at T_L and T_H, grouped by track length.

		track length									
		1-22	22-43	43-64	64-85	85-106	106-126	126-147	147-168	168-189	189-210
TinT	T_L	21	5	8	4	1	1	0	3	2	82
	T_H	11	3	0	2	2	2	4	6	5	68
PeakWalker	T_L	101	38	13	8	4	7	3	4	0	5
	T_H	64	19	15	7	6	10	7	5	7	45

Table 6 hLys. Comparison between TinT and Peakwalker results: number of recognized tracks at T_L and T_H, grouped by track length.

		track length									
		0-22	23-44	45-66	67-88	89-110	111-132	133-154	155-176	177-198	199-213
TinT	T_L	50	17	9	6	6	8	7	7	12	93
	T_H	11	3	8	5	9	5	10	7	10	79
PeakWalker	T_L	205	87	39	2	5	1	0	0	0	0
	T_H	81	101	18	11	23	10	12	4	10	8

increased, while decreasing for PeakWalker (tables 4, 5, 6).

The high number of short tracks recognized by PeakWalker at low thresholds is not due to a better performance with respect to TinT, but rather to the capability of the latter of recognizing much more long tracks than PeakWalker within the same pool of experimental data.

5 Conclusions

We developed TinT, a novel method for peak picking and tracking based on mathematical morphology and decomposition, tailored for the analysis of stepwise perturbed spectra. TinT can be suitable for monitoring various kind of NMR and in general spectroscopic experiments and it is able to give a detailed description of all the fundamental parameters of each peak during a stepwise evolution. The method was tested on synthetic spectra showing excellent results on realistic noise levels an performing well even in extreme noise conditions. TinT was proved to be successful in tracking peaks in sets of hundreds of spectra resulting from BLUU-Tramp sessions performed on three different proteins amenable to NMR analysis. Quasi-continuous changes in spectral parameters between consecutive spectra are required for the method to work at best, as seen on downsampled synthetic data. In experiments where the latter condition is met the method is able to reconstruct complete peaks' evolution.

Acknowledgements TB would like to thank Richard Jang for providing PeakWalker and assisting in using it.
The work received financial support from PRIN project No. 2012A7LMS3.
TB was supported by the Social European Fund and sponsored by Bruker.

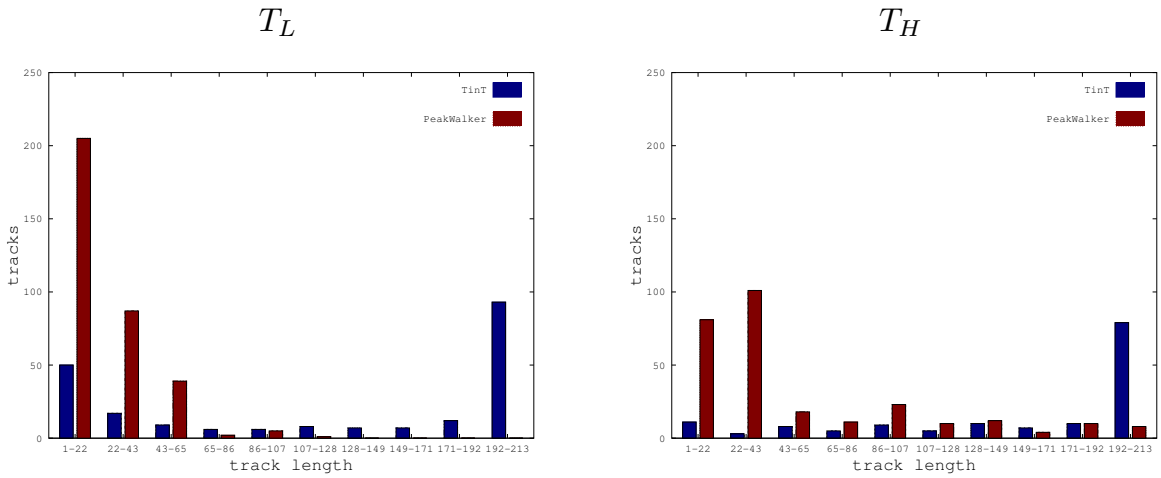

Fig. 12 hLys. Comparison between TinT (blue) and Peakwalker (red) results at T_L (a) and T_H (b): number of recognized tracks, grouped by track length.

References

1. Abbas, A., Xin-Bing, K., Zhi, L., Bing-Yi, J., Xin, G.: Automatic peak selection by a Benjamini-Hochberg-based algorithm. PLoS ONE **8**, 1–10 (2013)
2. Alipanahi, B., Gao, X., Karakoc, E., Donaldson, L., Li, M.: PICKY: a novel SVD-based NMR spectra peak picking method. Bioinformatics **25**, i268–i275 (2009)
3. Antz, C., Neidig, K.P., Kalbitzer, H.R.: A general Bayesian method for an automated signal class recognition in 2D NMR spectra combined with a multivariate discriminant analysis. J Biomol NMR **5**, 287–296 (1995)
4. Bax, A., Ikura, M., Kay, L.E., Torchia, D.A., Tschudin, R.: Comparison of different modes of two-dimensional reverse-correlation NMR for the study of proteins. J Magn Reson **86**, 304–318 (1990)
5. Bodenhausen, G., Ruben, D.J.: Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. Chem Phys Lett **69**, 185–189 (1980)
6. Boulton, S., Melacini, G.: Advances in nmr methods to map allosteric sites: From models to translation. Chemical Reviews **116**(11), 6267 – 6304 (2016)
7. Carrara, E.A., Pagliari, F., Nicolini, C.: Neural networks for the peak-picking of nuclear magnetic resonance spectra. Neural Networks **6**, 1023 – 1032 (1993)
8. Cheng, Y., Gao, X., Liang, F.: Bayesian Peak Picking for NMR Spectra. Genomics Proteomics Bioinf **12**, 39–47 (2014)
9. Christopher A. Lepre, Jonathan M. Moore, Jeffrey W. Peng: Theory and applications of NMR-based screening in pharmaceutical research. Chem Rev **104**, 3641–3676 (2004)
10. Clore, G.M., Schwieters, C.D.: Docking of protein-protein complexes on the basis of highly ambiguous intermolecular distance restraints derived from 1H/15N chemical shift mapping and backbone 15N-1H residual dipolar couplings using conjoined rigid body/torsion angle dynamics. J Am Chem Soc **125**, 2902–12 (2003)
11. Corne, S.A., Johnson, A.P., Fisher, J.: An artificial neural network for classifying cross peaks in two-dimensional NMR spectra. J Magn Reson **100**, 256–266 (1992)
12. Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J., Bax, A.: NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol NMR **6**, 277–93 (1995)
13. Dominguez, C., Boelens, R., Bonvin, A.M.J.J.: HADDOCK: a proteinprotein docking approach based on biochemical or biophysical information. J Am Chem Soc **125**, 1731–1737 (2003)
14. Eaton, J.W., Bateman, D., Hauberg, S., Wehbring, R.: GNU Octave version 4.0.0 manual: a high-level interactive language for numerical computations (2015). URL http://www.gnu.org/software/octave/doc/interpreter
15. Englander, S.W., Kallenbach, N.R.: Hydrogen exchange and structural dynamics of proteins and nucleic acids. Q Rev Biophys **16**, 521–655 (1983)
16. Englander, S.W., Mayne, L., Krishna, M.M.: Protein folding and misfolding: mechanism and principles. Q Rev Biophys **40**, 287–326 (2007)
17. Fahmy, A., Wagner, G.: TreeDock: a tool for protein docking based on minimizing van der Waals energies. J Am Chem Soc **124**, 1241–50 (2002)
18. Fukui, L., Chen, Y.: NvMap: automated analysis of NMR chemical shift perturbation data. Bioinformatics **23**, 378–380 (2007)
19. Gao, X.: Recent advances in computational methods for Nuclear Magnetic Resonance data processing. Genomics Proteomics Bioinf **11**, 29–33 (2013)
20. Garrett, D.S., Powers, R., Gronenborn, A.M., Clore, G.: A common sense approach to peak picking in two-, three- and four-dimensional spectra using automatic computer analysis of contour diagrams. J Magn Reson **95**, 214 – 220 (1991)
21. Glasbey, C., Horgan, G.: Image Analysis for the Biological Sciences. John Wiley and Sons (1995)
22. Hajduk, P.J., Dinges, J., Miknis, G.F., Merlock, M., Middleton, T., Kempf, D.J., Egan, D.A., Walter, K.A., Robins, T.S., Shuker, S.B., Holzman, T.F., Fesik, S.W.: NMR-based discovery of lead inhibitors that block DNA binding of the human papillomavirus E2 protein. J Med Chem **40**, 3144–3150 (1997)
23. Jang, R., Gao, X., Li, M.: Combining automated peak tracking in SAR by NMR with structure-based backbone assignment from 15N-NOESY. BMC Bioinformatics **13**, 1–15 (2012)

24. Johnson, B.A., Blevins, R.A.: NMRView, a computer program for the visualization and analysis of NMR data. J Biomol NMR **4**, 603–614 (1994)
25. Kleywegt, G.J., Boelens, R., Kaptein, R.: A versatile approach toward the partially automatic recognition of cross peaks in 2D 1H NMR spectra. J Magn Reson **88**, 601–608 (1990)
26. Klukowski, P., Walczak, M.J., Gonczarek, A., Boudet, J., Wider, G.: Computer vision-based automated peak picking applied to protein NMR spectra. Bioinformatics **31**, 2981–2988 (2015)
27. Koradi, R., Billeter, M., Engeli, M., Guentert, P., Wuethrich, K.: Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. J Magn Reson **135**, 288 – 297 (1998)
28. Korzhnev, D.M., Ibraghimov, I.V., Billeter, M., Orekhov, V.Y.: MUNIN: application of three-way decomposition to the analysis of heteronuclear NMR relaxation data. J Biomol NMR **21**, 263–268 (2001)
29. Lescop, E., Schanda, P., Brutscher, B.: A set of best triple-resonance experiments for time-optimized protein resonance assignment. J Magn Reson **187**, 163 – 169 (2007)
30. Liu, Z., Abbas, A., Jing, B.Y., Gao, X.: WaVPeak: picking NMR peaks through wavelet-based smoothing and volume-based filtering. Bioinformatics **28**, 914–920 (2012)
31. Maragos, P.: Morphological filtering for image enhancement and feature detection. In: A.C. Bovik (ed.) Image & Video Processing Handbook (2nd ed.), chap. 3.3, pp. 135–156. Elsevier Academic Press, Amsterdam, The Netherlands (2005)
32. McCoy, M.A., Wyss, D.F.: Alignment of weakly interacting molecules to protein surfaces using simulations of chemical shift perturbations. J Biomol NMR **18**, 189–98 (2000)
33. Morelli, X.J., Palma, P.N., Guerlesquin, F., Rigby, A.C.: A novel approach for assessing macromolecular complexes combining soft-docking calculations with NMR data. Protein Sci. **10**, 2131–7 (2001)
34. Orekhov, V.Y., Ibraghimov, I.V., Billeter, M.: MUNIN: a new approach to multi-dimensional NMR spectra interpretation. J Biomol NMR **20**, 49–60 (2001)
35. Otting, G.: Experimental NMR techniques for studies of protein-ligand interactions. Curr Op Struct Biol **3**, 760–768 (1993)
36. Pellecchia, M., Montgomery, D.L., Stevens, S.Y., Vander Kooi, C.W., Feng, H.P., Gierasch, L.M., Zuiderweg, E.R.: Structural insights into substrate binding by the molecular chaperone DnaK. Nat Struct Mol Biol **7**, 298–303 (2000)
37. Peng, C., Unger, S.W., Filipp, F.V., Sattler, M., Szalma, S.: Automated evaluation of chemical shift perturbation spectra: New approaches to quantitative analysis of receptor-ligand interaction NMR spectra. J Biomol NMR **29**, 491–504 (2004)
38. Ravel, P., Kister, G., Malliavin, T.E., Delsuc, M.A.: A general algorithm for peak-tracking in multi-dimensional NMR experiments. J Biomol NMR **37**, 265–275 (2007)
39. Rennella, E., Corazza, A., Codutti, L., Bellotti, V., Stoppini, M., Viglino, P., Fogolari, F., Esposito, G.: Determining the energy landscape of proteins by a fast isotope exchange NMR approach. J Am Chem Soc **134**, 4457–60 (2012)
40. Rennella, E., Corazza, A., Codutti, L., Causero, A., Bellotti, V., Stoppini, M., Viglino, P., Fogolari, F., Esposito, G.: Single-shot NMR measurement of protein unfolding landscapes. Biochim Biophys Acta **1824**, 842–9 (2012)
41. Rennella, E., Corazza, A., Fogolari, F., Viglino, P., Giorgetti, S., Stoppini, M., Bellotti, V., Esposito, G.: Equilibrium unfolding thermodynamics of beta2-microglobulin analyzed through native-state H/D exchange. Biophys J **96**, 169–79 (2009)
42. Rose, A.: The sensitivity performance of the human eye on an absolute scale. J Opt Soc Am **38**, 196–208 (1948)
43. Rouh, A., Louis-Joseph, A., Lallemand, J.Y.: Bayesian signal extraction from noisy FT NMR spectra. J Biomol NMR **4**, 505–518 (1994)
44. Schanda, P., Brutscher, B.: Very fast two-dimensional NMR spectroscopy for real-time investigation of dynamic events in proteins on the time scale of seconds. J Am Chem Soc **127**, 8014–8015 (2005)
45. Serra, J.: Image analysis and mathematical morphology, vol. 1. Academic press, New York (1982)
46. Serra, J.: Image analysis and mathematical morphology: Theoretical Advances., vol. 2. Academic press, New York (1988)
47. Shuker, S.B., Hajduk, P.J., Meadows, R.P., Fesik, S.W.: Discovering high-affinity ligands for proteins: SAR by NMR. Science **274**, 1531–1534 (1996)
48. Williamson, M.P.: Using chemical shift perturbation to characterise ligand binding. Prog Nucl Magn Res Sp **73**, 1 – 16 (2013)
49. Zuiderweg, E.R.P.: Mapping protein–protein interactions in solution by NMR spectroscopy. Biochemistry **41**, 1–7 (2002)