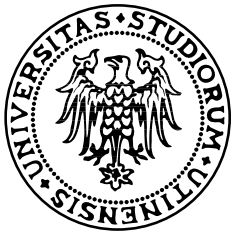
Peak-tracking for stepwise perturbed NMR spectra: development of a new analysis method.



Tommaso Banelli

Under the supervision of Dott. Alessandra Corazza
and co-supervised by Prof. Andrea Fusiello

University of Udine

A thesis submitted for the degree of

PhD in Biomedical Sciences and Biotechnology (Ciclo XXVIII)

academic year
2016/2017

Reviewer:

Prof. Daniel Oscar Cicero

Day of the defense:
20 April 2017

Signature from head of PhD committee:

Abstract

Stepwise perturbed NMR spectra analysis is a powerful tool capable of describing kinetic, thermodynamic, structural aspects of proteins at a residue level and of following the physical and chemical changes of the system. The analysis of an NMR spectrum still offers compelling challenges to the automatic identification of the chemical shift evolution.

We designed and developed a data-analysis method which allows automatic peak detection in every spectrum, peak tracking between spectra and peak reconstruction for BLUU-Tramp sessions, a stepwise isotopic exchange experiment producing few hundreds of 2D NMR spectra.

The method has been named TinT (Trace in Track), referring to the idea that a gaussian decomposition traces peaks within the tracks recognized through 3D mathematical morphology. TinT 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. The comparison with a state-of-the-art algorithm showed very good results for great numbers of spectra and low signal to noise ratios, when the graduality of the perturbation is appropriate.

In the thesis, in addition to the description of the current version of TinT, some observations and considerations that can allow future revisions or improvements on BLUU-Tramp protocol or its analysis are also described.

To all my family

Contents

List of Figures

LIST OF FIGURES

List of Tables

LIST OF TABLES

1

Introduction

In the last few years it seems that we are witnessing an epochal transition fostered by a rapid evolution of techniques, technology and tools that is modifying humankind habits and behaviors in every aspects of our life: communications, commerce, knowledge and research. Following the distinction that Dyson suggests in his article "Is science mostly driven by ideas or by tools?" [1] it seems that we are in between a Galisonian and a Kuhnian scientific revolution.

On one hand, Peter Galison highlighted in his 1997 book, "Image and Logic" [2], the key role of tools in driving science, that is what we are assisting [3]. In fact, increasing computing powers of CPU and GPU allow users to face increasingly complex calculations in less time; vastness of RAM and Hard-disk are moving the attention from "less but good" to the chance to explore more and more new horizons with detailed description; finally faster communications lines (ADSL, fiber, 4G, 5G) have created pressure to share all kind of data in web site, in data banks or in the Cloud. The cloud computing gathers all these positive aspects and it is arising as the newest powerful service available for researchers and no.

On the other hand, data explosion and the falling cost of computing, storing and communications are also affecting the scientific method. Shared data means that researchers can work not only with their own data collection but also manipulate a huge amount of data produced all over the world. In this way they can generate models and hypotheses combining and mining the pool of data already available. Data exploration becomes a valuable approach to generating new knowledge. It is rising a shift in the scientific thinking, a 'paradigm' shift in the Kuhnian sense [4], dubbed by Bell,

Hey and Szalay "the 4th paradigm" [5]. In fact, science was born empiric, as observational description of natural phenomena. Then, in parallel to the development of the mathematical sciences, a theoretical branch was grown for the generalization and modeling. These are classically considered the first two paradigms. Subsequently, in the last decades a computational science or simulation has emerged as the third pillar of the scientific process, completing the experimental-theoretical framework. Driven by inexpensive, opportunity and necessity, more and more data are acquired. In front of this bulk of available data, science is discovering the abundant vastness of data as a source of knowledge. This is the new era of data-intensive science [6].

In Jim Gray's last talk [6] to the Computer Science and Telecommunications Board on January 11, 2007, he said: "We have to do better at producing tools to support the whole research cycle – from data capture and data curation to data analysis and data visualization". All the scientific community is invited to consider the many opportunities and challenges for data-intensive science.

1.1 NMR spectroscopy opportunities

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 [7, 8, 9] 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 [10], SAR by NMR [11, 12], protein-protein interaction [13, 14, 15, 16, 17], chemical shift covariance analyses [18], isotopic exchange at single temperatures [19, 20, 21], BLUU-Tramp [22, 23] and titration in general [24].

Very often the NMR experiment of choice, for such studies on proteins, is the highly sensitive 2D HSQC of ^{15}N-enriched proteins [25, 26] or fast acquisition versions such as HMQC-SOFAST [27] and BEST-TROSY [28], because they allow one to obtain information at single residue resolution with one signal for each N-H pair. When no enriched product is available, 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 [29], 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 of peak picking and peak tracking because of noise, peak overlap, cross shifting and long distance correlation peaks.

1.2 Literature review

In the 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 [30] in 1990 with a naive point by point reconstruction and the software currently used are based on a variety of methods, such as peak properties [31, 32], machine learning algorithms [33, 34, 35, 36], spectral decomposition [37, 38, 39, 40], wavelet smoothing [41], Benjamini-Hochberg procedure [42], computer vision [43], Monte Carlo stochastic approxi-

mation and Bayesian statistics [44]. A comprehensive review has been given by Liu and coworkers in the introduction of the article describing their algorithm WaVPeak [41].

Fewer articles deal with the automation of the tracking procedure. The proposed methods in the literature are: APET/PROPET (in Felix-Autoscreen) [45], based on bipartite graph matching by systematic tree search methods and simulated annealing approach with heuristic simplification, Nvmap (NMRViewJ [32]) [46], based on search of the nearest pair with a greedy algorithm, GAPT [47] based on best-score-selection under constrain with heuristic simplification, PeakWalker [48] 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-completeness of the problem [48].

1.3 BLUU-Tramp

Our attempt in automation stems from the need to analyze data resulting from BLUU-Tramp experiment [22, 23]; therefore some methodological choices reflect features of the method.

An experimental session of BLUU-Tramp produces two sequences of around 200 2D HMQC spectra (Fig. 1.1), that are acquired at regular temperature increments (usually 0.1-0.2 K) 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 TinT, 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.

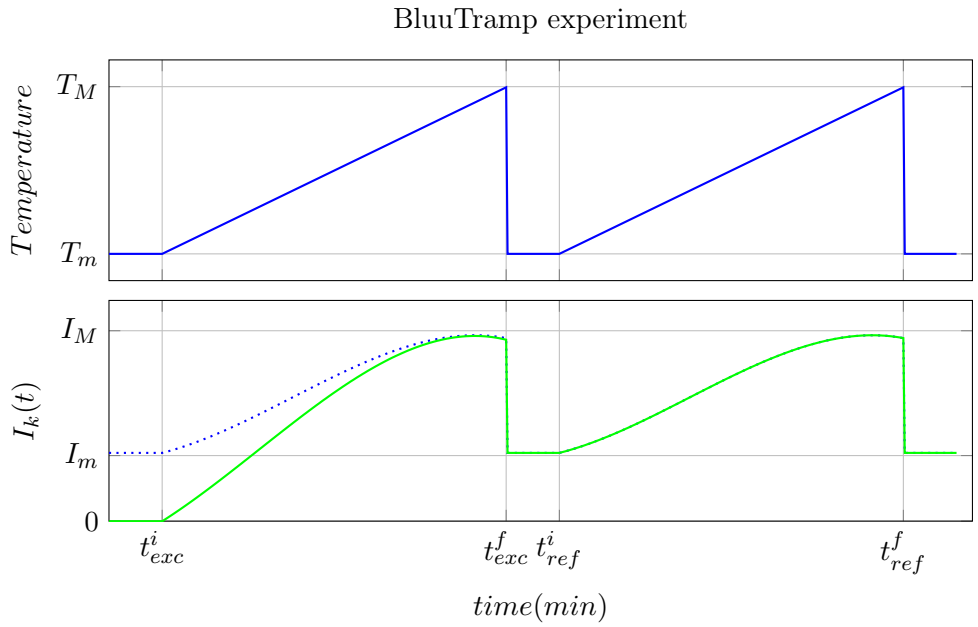


Figure 1.1: Design of BLUU-Tramp experiments: plots of temperature and intensity of one peak over time. In the second graph, the observed intensity of one peak is plotted in green. The dotted line represent the intensity of the same peak without the H-D exchange process.

For each peak, the derived quantity used to evaluate the isotopic exchange process during a BLUU-Tramp session, cleaned from the (unwanted) contribution of the thermal dependence of spin relaxation, is the following:

$$N[s] = \frac{I_{ref}[s] - I_{exc}[s]}{I_{ref}[s]} \quad \forall s \tag{1.1}$$

where $I_{exc}[s]$ and $I_{ref}[s]$ are arrays that collect the intensity evolution of one peak in each spectrum s during the two thermal ramps, with and without an exchange process, respectively (Fig. 1.2).

$N[s]$ becomes $N(t)$ once a regression model is defined (see chapter 6).

Theoretically [22, 23], the physical phenomenon of the isotopic exchange is described by the following formula:

$$N(t) = N(0) \exp\left\{ - \int_0^t k(T(t'))dt' \right\} \tag{1.2}$$

Solving equation (1.2) for k(T(t)), the following description is obtained:

$$k(T(t)) = -\frac{d}{dt} \ln N(t) \tag{1.3}$$

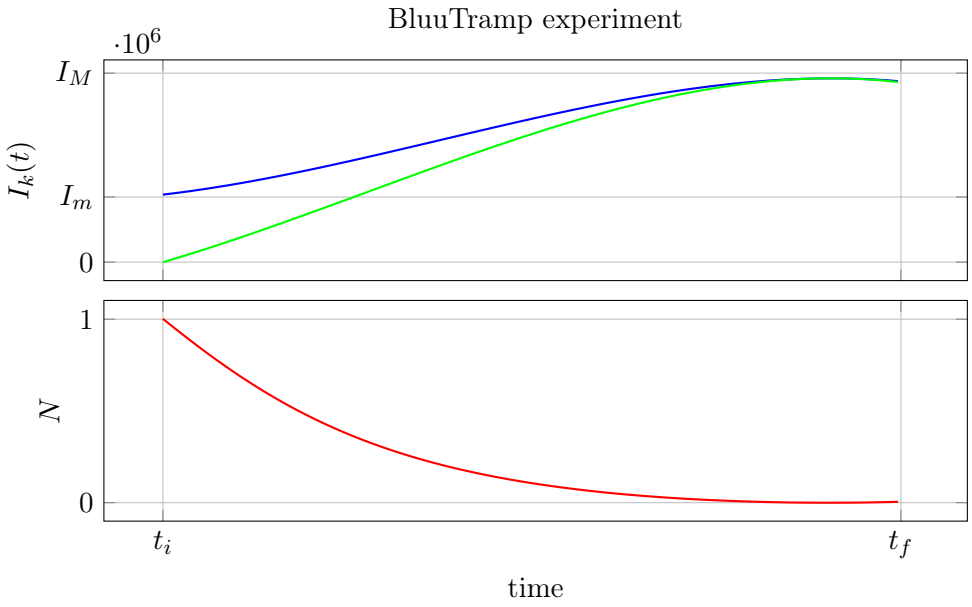


Figure 1.2: Main data from BLUU-Tramp experiment: plots over time of reference and exchange intensities and their derived quantity N(t) in blue, green and red respectively.

Therefore, having a mathematical description of $N(t)$ and solving for that the equation (1.3), $k(T(t))$ can be obtained.

The first main point is to collect $I_{ref}[s]$, $I_{exc}[s]$ for every peak in each spectrum; this kind of analyses is related to peak tracking and reconstruction.

2

Aim of the project

The aim of this project was to design and develop an automatic data-analysis method which allows peak detection (picking) in every spectrum and peak tracking between spectra and their reconstruction.

The first part of my PhD course aimed at the conception and the implementation of the main part of the routine, developing the code with functions and procedures and defining the internal data structure. The software is named TinT (Trace in Track).

The second part of the PhD work was to optimize the implementation. I improved functions and routines to compute faster and to reduce the amount of memory used by the internal data-flow. To strengthen the automation, a starting subroutine was created minimizing the need to involve the user in the choice of the default starting parameters. The use of an input file, auto-generated the first time the routine is run, allow the user to edit these values. An effort for the simplification of the input and output procedures was made to allow the use of the software even by inexperienced users. Great attention was devoted to the visual and graphical inspection of the intermediate stages.

The last part of the work was intended to prove the validity of the software with both real and synthetic data. Three BLUU-Tramp sessions on three different proteins were analyzed by TinT and the results were compared with a state-of-the-art algorithm, PeakWalker. A procedure was implemented to generate synthetic BLUU-Tramp sessions from the results of the analysis on the real protein data. The results of this work were published in an article in high impact-factor journal [49].

An additional part of this project was the assessment of critical aspects in the analysis of BLUU-Tramp experiment that can allow future revisions or improvements

on the protocol or its analysis.

3

Methods and algorithms

3.1 TinT

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 [50, 51, 52] 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, modeled as gaussians. A statistical validation stage reduces the amount of artifacts.

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 by its application 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).

3.2 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 [51, 52] based on nonlinear geometric approach [50]. For an introduction to Mathematical morphology for image processing see e.g. a chapter by Glasbey [53]. 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. 3.1): 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 same way, but inverting the colors of the SE. These 2D-image operations can be easily extended

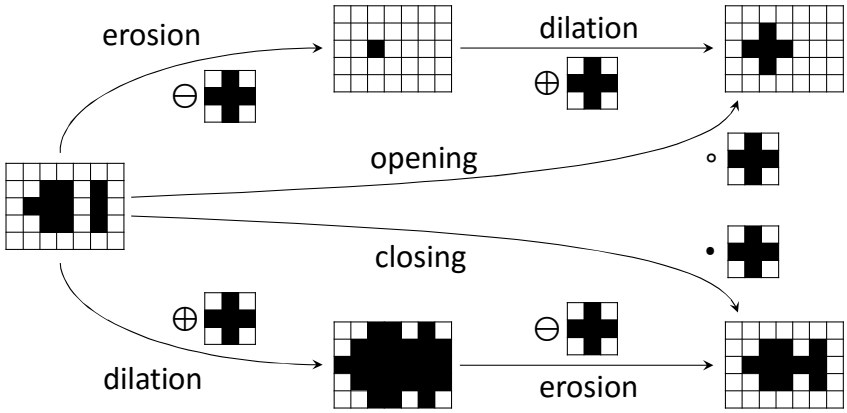


Figure 3.1: Behavior of the morphological operators. The arrows represent the specified algorithms that transform the input images in output images, with the given SE

to be applied on multidimensional binary matrices, defining:

$$\begin{aligned} &\textit{dilation} && A \oplus S = \bigcup_{z \in S} A_z \\ &\textit{erosion} && A \ominus S = \bigcap_{z \in S} A_z \end{aligned}$$

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. 3.1) such as:

$$\textit{opening} \qquad \circ = \ominus\oplus \tag{3.1}$$

$$\textit{closing} \qquad \bullet = \oplus\ominus \tag{3.2}$$

We took advantage from the potentiality of the opening operator to erase details smaller than a SE while maintaining unaltered the remainder (Fig. 3.5b). Then we exploit the capability of the closing operator to connect structures separated by volumes smaller than a SE (Fig. 3.5c).

3.3 TinT algorithm/structure

The TinT algorithm is coded in Octave (version 4.0.0) [54] with the following packages:

- image (version 2.4.1),
- optim (version 1.4.1)
- statistics (version 1.2.4)

The work-flow of the 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.
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 in every spectrum: intensity (I), ^{1}H and ^{15}N frequency positions (δH, δN) and linewidths (λH, λN).

In the following subsection, each step of the procedure will be described in some detail.

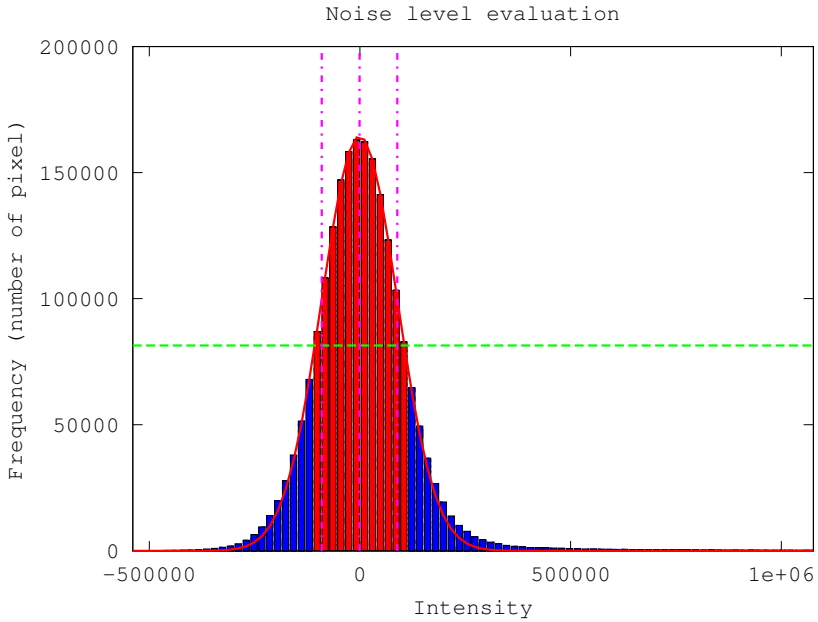


Figure 3.2: The MIDNE method applied to the first hAcP spectrum. The distribution of the intensity is represented with a histogram. The green dashed line was half of the maximum bin height. The bins higher than this value, shown in red, are used to reconstruct the gaussian distribution shown with the 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.

3.3.1 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 auto-generated 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 3.3.2), in the computation of σ for the weight function (see Eq. (3.5)) and as starting guess in the decomposition procedure (see section 3.3.4).

In order to let the user define the rough 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. 3.2. This estimation method has been named MIDNE (Modeling Intensity Distribution for Noise Estimation). TINT will propose setting the threshold at $B + 5N$ [55].

3.3.2 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. 3.3). 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. 3.4).

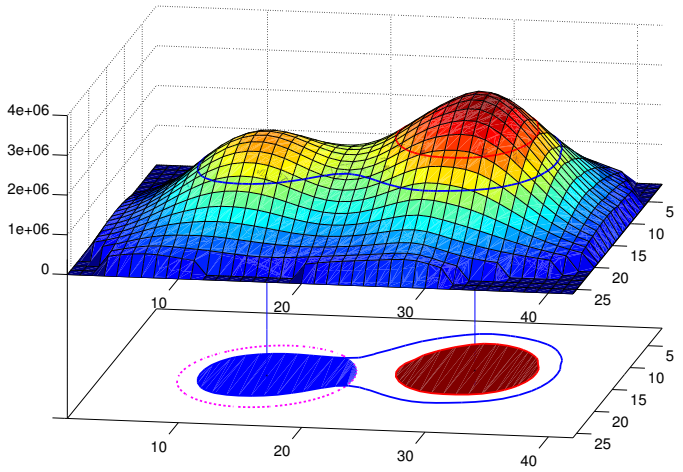


Figure 3.3: Example of spot evaluation for two overlapped peaks: PLTs are shown with contour lines in red and blue. The low intensity peak affects and overestimates the area around the higher peak. The magenta dotted line shows the circumference limiting the area of the lower peak to avoid the overestimation of areas. The filled areas are assigned to two maxima.

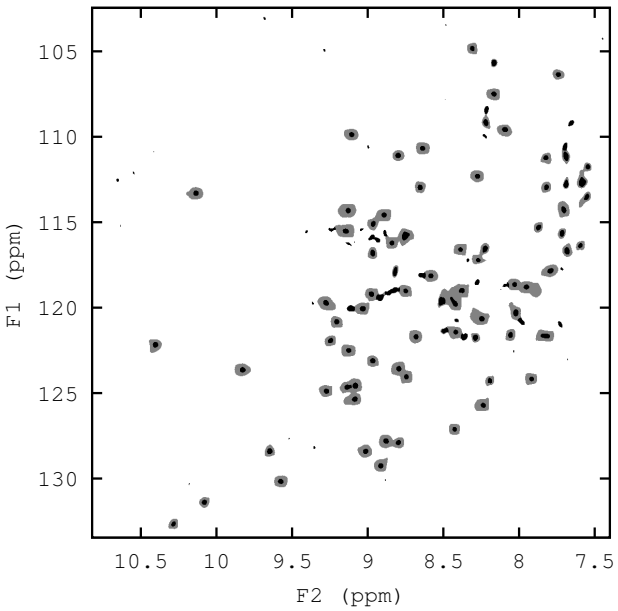


Figure 3.4: Comparison between global single threshold (in gray and black) and our PLT method (black only).

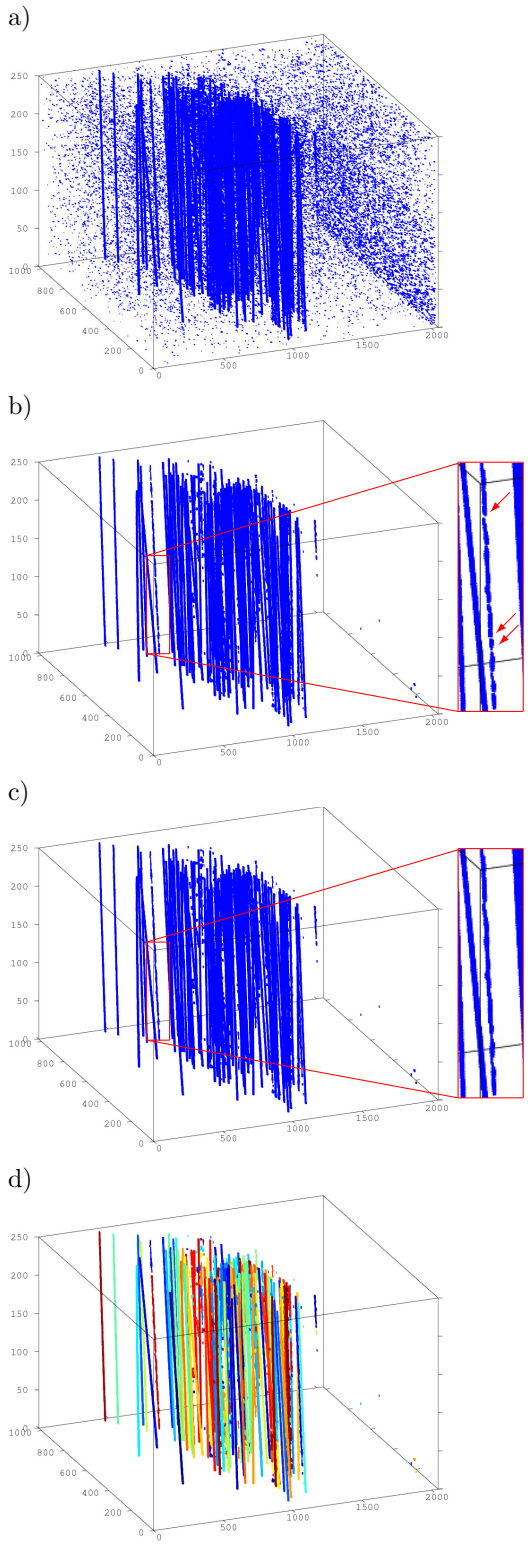


Figure 3.5: Elaboration of data during the morphological filter. The input 3D-matrix (a), the output of the opening algorithm (b), the output of the closing algorithm (c) and the output of the connected-components labeling algorithm (d) are shown. Gaps (filled by the closing operator) are indicated by the arrows.

3.3.3 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. 3.5a). 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. 3.5a. 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. 3.5b) and then connecting (with a morphological closing) the surviving volumes closer than the proper SE (Fig. 3.5c). 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 auto-generated 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. 3.5d). 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.

3.3.4 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 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. 3.7). 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,

$$C(\bar{x}) = \sum_{\bar{p}} H^r\Big(w(\bar{p}) * \big[D(\bar{p}) - R(\bar{p}; \bar{x})\big]\Big) \tag{3.3}$$

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

$$H^r(v) = \begin{cases} \frac{1}{2}v^2, & |v| < r \\ rv - \frac{1}{2}r^2, & |v| \geq r \end{cases} \tag{3.4}$$

where r is a cutoff value equal to the initial threshold value (see section 3.3.1). The Huber function (shown in Fig. 3.6) 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

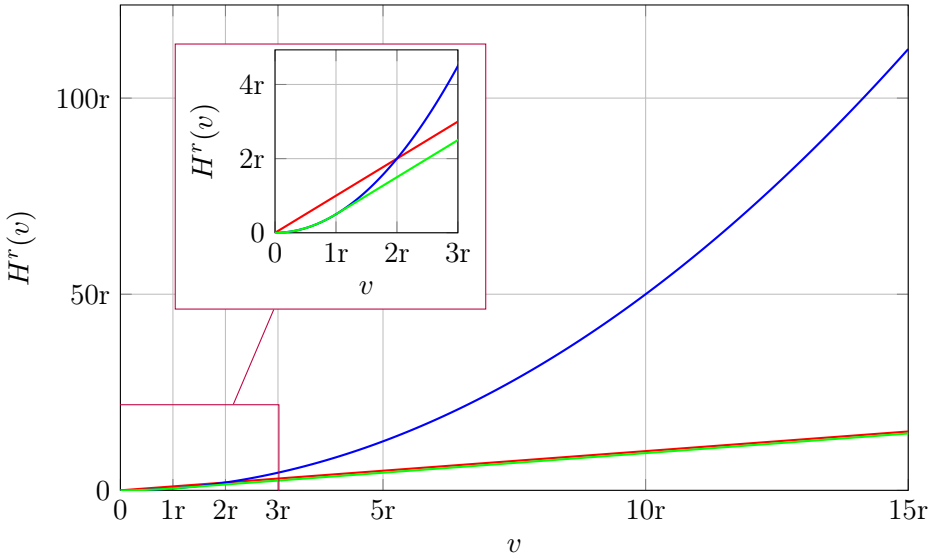


Figure 3.6: Comparison between Huber function (green), linear function (red) and square function (blue).

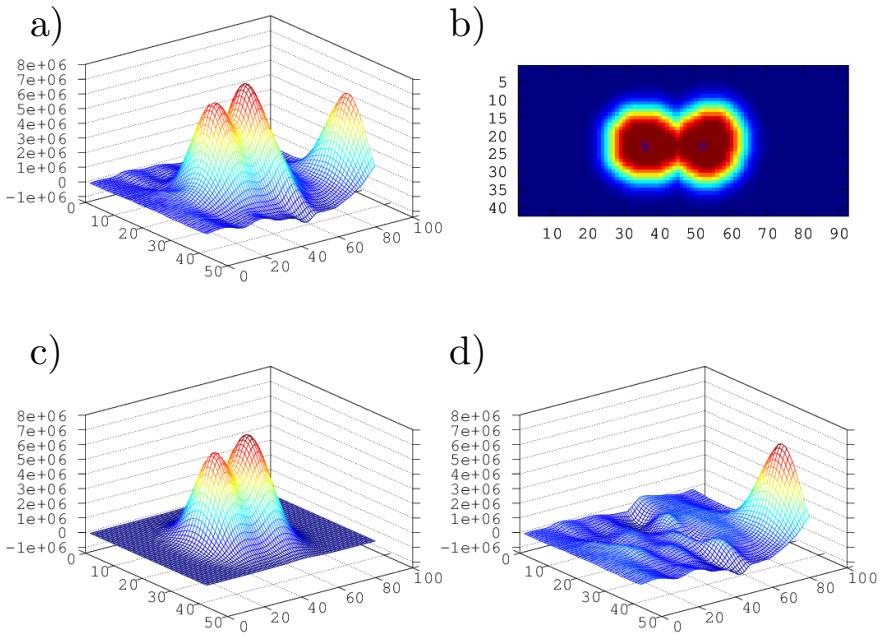


Figure 3.7: Weighted deconvolution: data (a), weight function (b), peak reconstruction (c) and remainder (d).

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. 3.7b):

$$w(\bar{p}) = \begin{cases} 1 & \bar{p} \in M \\ e^{-\left(\frac{d(\bar{p},M)}{\sigma}\right)^2} & \bar{p} \notin M \end{cases} \tag{3.5}$$

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 3.3.1).

The peak reconstruction (example in Fig. 3.7c) 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 3.3.2), while $\lambda\bar{H}$, $\lambda\bar{N}$ (section 3.3.1) 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.

3.3.5 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).

3.4 Experimental methods

3.4.1 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, relative to 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. 4.1). 210 spectra were generated. Noise was added by convolving gaussian white noise with a bidimensional gaussian with parameters optimized to reproduce experimental noise.

3.4.2 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 (130 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 [27] or ^{1}H-^{15}N BEST-TROSY [28]. 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 [22, 23].

3.4.3 Spectral data processing

The spectra were processed with NMRpipe [56] 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.

4

Results and Discussion

4.1 Results on synthetic data

Synthetic data, generated as described in 3.4.1, were analyzed 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 dataset. All results are reported in table 4.1 and table 4.2.

Table 4.1: 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 4.2: 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 %

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 4.2.

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 130 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. 4.1).

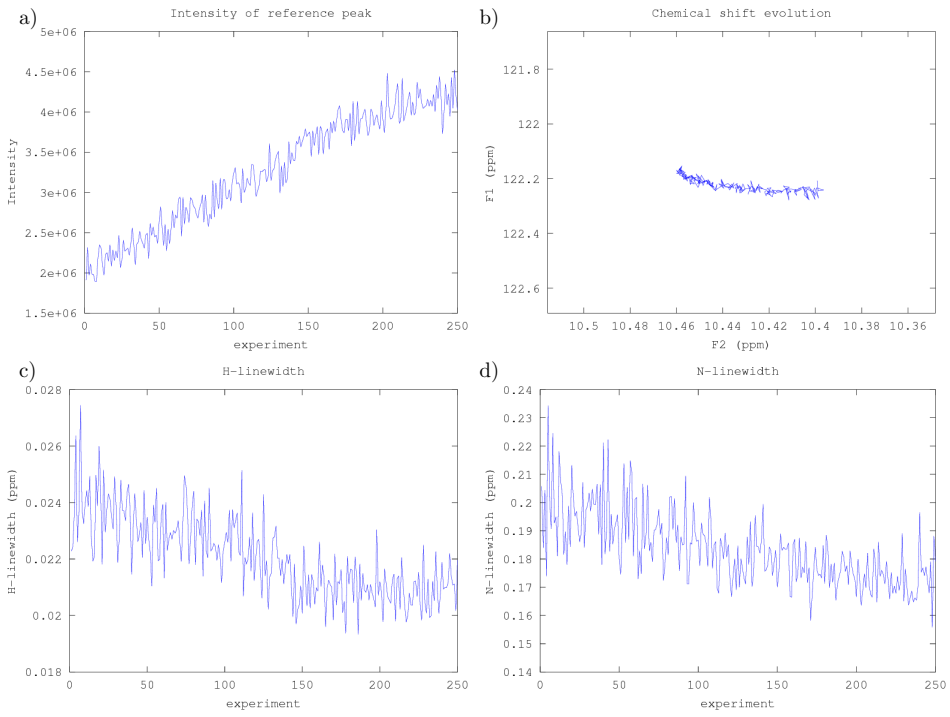


Figure 4.1: Evolution of the five fundamental parameters of one hAcP peak: $I(t)$ (a), $\delta H(t)$ and $\delta N(t)$ (b), $\lambda H(t)$ (c) and $\lambda N(t)$ (d).

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%, 94% for hAcP, β2-m and hLys, respectively (table 4.3).

Table 4.3: 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	130	123	94%

It must be noticed that our method allows us to recognize nonlinear paths, as shown in Fig. 4.2, similar to those seen in ligand binding studies [24].

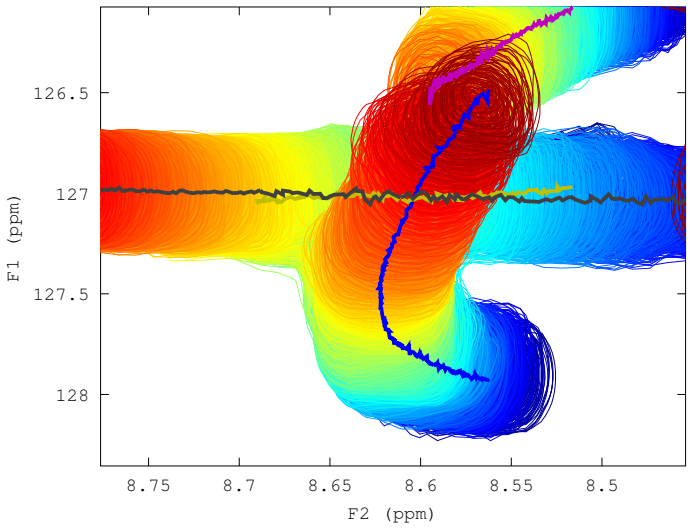


Figure 4.2: Recognition of a non linear path in hLys NMR spectrum. Lines show the evolution of the position of the recognized peaks. The flow of the time is represented with a chromatic scale.

4.3 Comparison with PeakWalker

To better evaluate the method we compared TINT output with the results of PeakWalker [48], 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 provides 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.

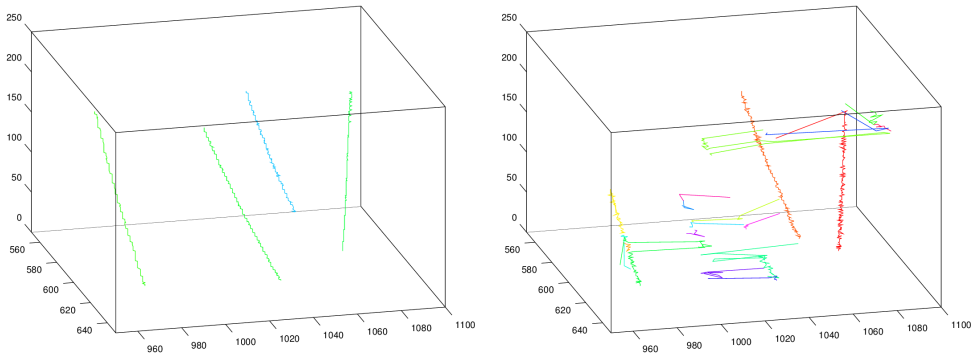


Figure 4.3: Comparison between TINT (left) and PeakWalker (right) resulting tracks

There were cases in which PeakWalker followed with the same track the evolution of more than one peak (Fig. 4.3). 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 [55], they were set as follows:

$$T_L = B + 4N \tag{4.1}$$

$$T_H = B + 8N \tag{4.2}$$

where B is the baseline and N is the noise level, both calculated with the starting procedure (see section 3.3.1) on the first spectrum, which is the one less affected by thermal noise. Tables 4.4, 4.5 and 4.6 report the number of signals at different thresholds, grouped by track length.

Table 4.4: 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

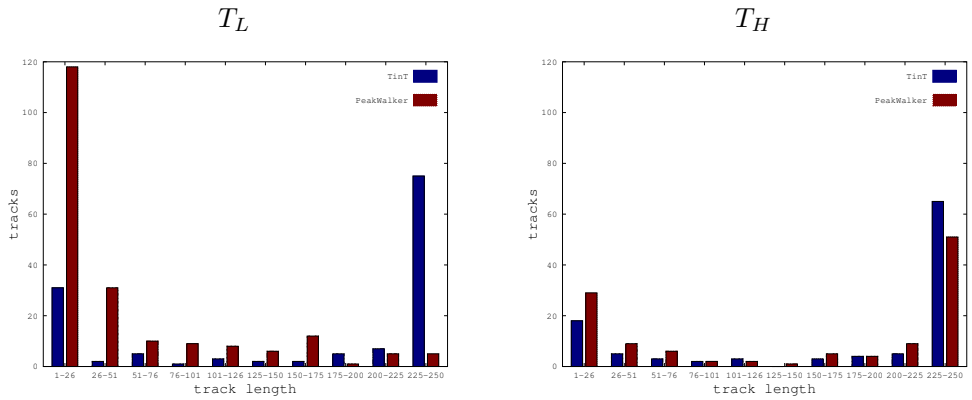


Figure 4.4: hAcP. Comparison between TINT (blue) and PeakWalker (red) results at T_L (a) and T_H (b): number of recognized tracks, grouped by track length.

Table 4.5: 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

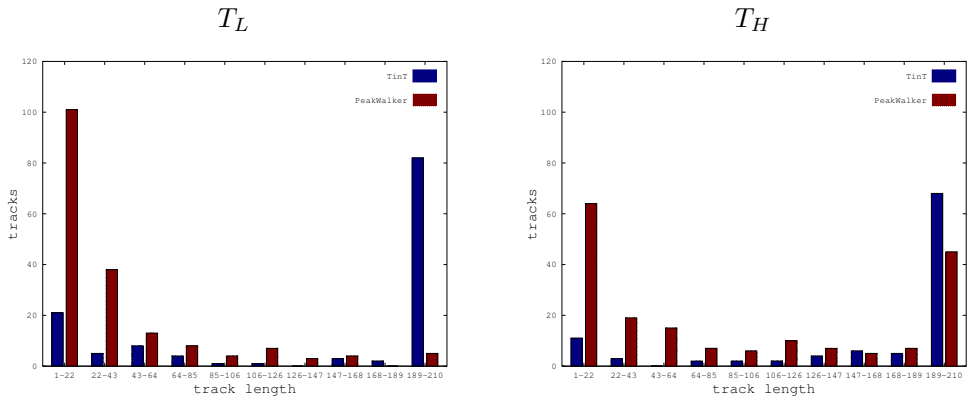


Figure 4.5: β2-m. Comparison between TINT (blue) and Peakwalker (red) results at T_L (a) and T_H (b): number of recognized tracks, grouped by track length.

Table 4.6: 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

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 4.4, 4.5, 4.6 and Figures 4.4, 4.5 and 4.6). Furthermore, on decreasing the threshold from T_H to T_L, the number of long tracks recognized by TINT increased, while decreasing for PeakWalker (tables 4.4, 4.5, 4.6).

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

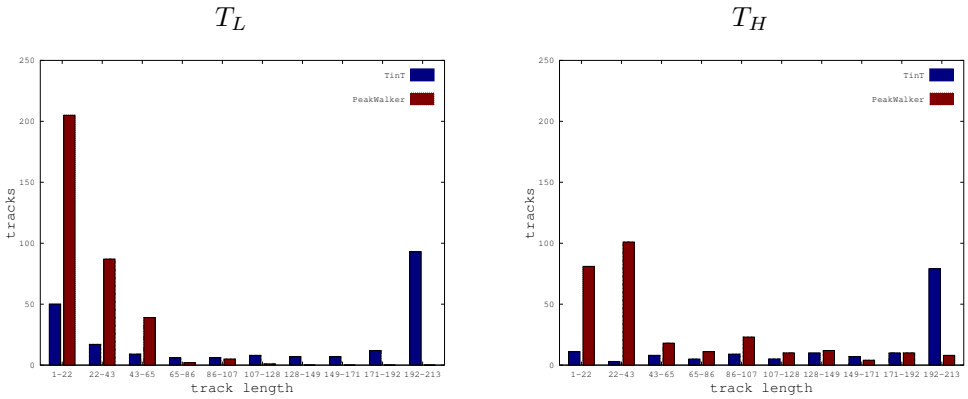


Figure 4.6: hLys. Comparison between TinT (blue) and Peakwalker (red) results at T_L (a) and T_H (b): number of recognized tracks, grouped by track length.

latter of recognizing much more long tracks than PeakWalker within the same pool of experimental data.

5

Model for loss correction

The TinT tracking routine, applied to BLUU-Tramp experiments, results in the detailed description of both the exchange and the reference subsession intensities, I_{exc} and I_{ref}, respectively. In this chapter, a simple model is described to understand how this output is affected by chemical and physical phenomena, such as conformational changes, aggregation and other unknown events that result in a permanent signal loss.

NMR theory states that the NMR signal is proportional to the concentration of the molecule in solution, therefore we can devise a simplified model in which each protein molecule can be represented like a sequence of signal emitters $a = \{a_k | k = 1 : \mathcal{K}\}$, with $\mathcal{K}$ the total number of H-N pairs; in this simple model, a_k represent the single molecule contribution of the k-th peak and it is an artificial function that hides all the dependence of each intensity peak with the relaxation time to isolate the dependence of I_k with the amount of protein in that conformation Q^0:

$$a_k(t) = \frac{I_k(t)}{Q^0}.$$

Every modification time- or temperature-driven of the amount of protein in the considered configuration affects also I_k^{obs} as following:

$$I_k^{obs}(t, T) \propto a_k(T) Q(t, T) \tag{5.1}$$

Chemical-physical phenomena that alter the quantity of the "emitters" can be expressed as a loss function P in the formula describing the evolution of Q with time:

$$Q(t) = Q^0 P(t) \tag{5.2}$$

where Q^0 is the initial quantity of molecules in the considered conformation.

(5.1) and (5.2) bring to the following:

$$I_k^{obs}(t,T) \propto a_k(T) Q^0 P(t,T) \tag{5.3}$$

It must be noticed that different phenomena can affect simultaneously intensities of all the peaks (e.g. aggregation) or have an effect only on a part of them (e.g. local conformational changes). Consequently even the loss function must be differentiated for each k-th peak: P_k.

In the following sections we will describe and graph the Ideal situation and three special cases: Time dependent loss, Thermal trigger loss, One-step loss. It must be stressed that none of these cases is meant to accurately represent a real case, but they can approximate basic behaviors.

The subscript k is omitted to simplify the writing.

5.1 Ideal situation

To work properly, a BLUU-Tramp experiment should be performed on a protein that can tolerate two thermal ramps. All the modifications occurring to the protein during the fist ramp must be reversible, so that the pause between two thermal ramps allows the protein to retrace the same intensity evolution. In this case the quantity of protein in the observed conformation can be described by a constant function, $Q(t) = Q^0$ (Fig. 5.1)

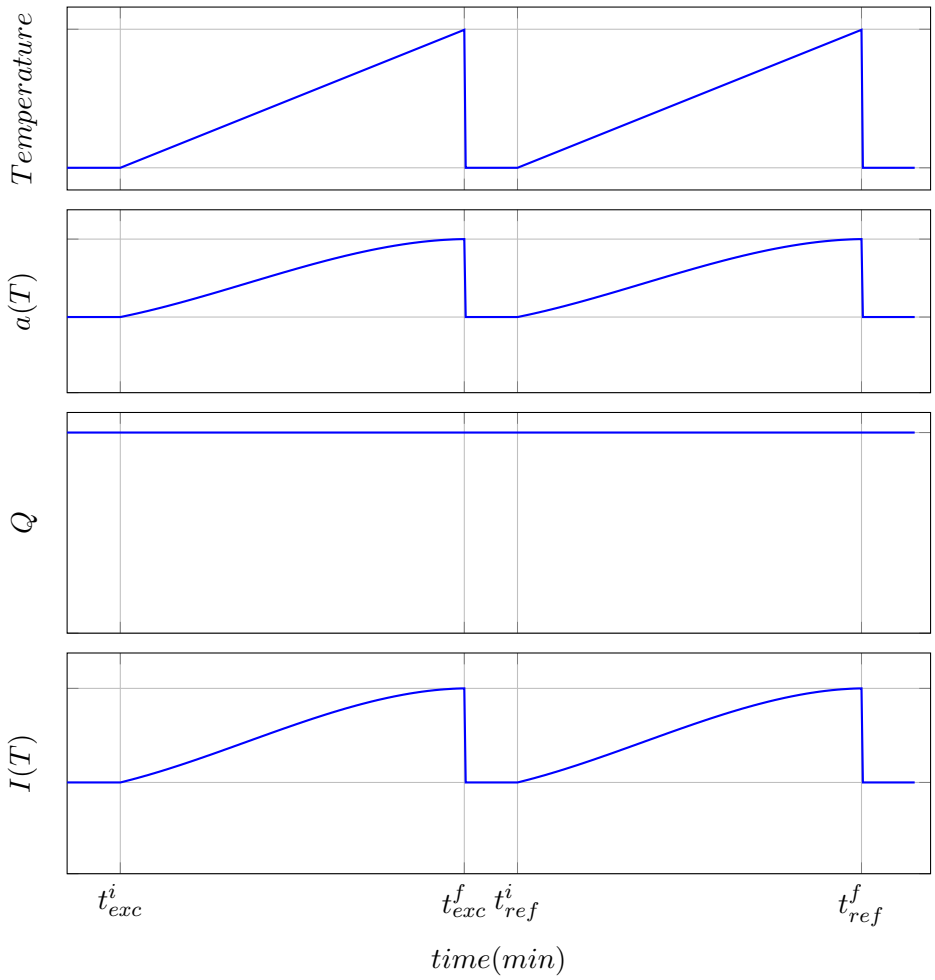


Figure 5.1: Plots of the k-th peak during a BLUU-Tramp experiment in the ideal situation: Temperature, single molecule contribution, amount of emitters and intensity of the peak.

5.2 Time dependent loss

Some chemical or physical events can affect the quantity of emitters as time progresses. In this case, the quantity of emitters decreases during the whole time frame in which the experiment takes place.

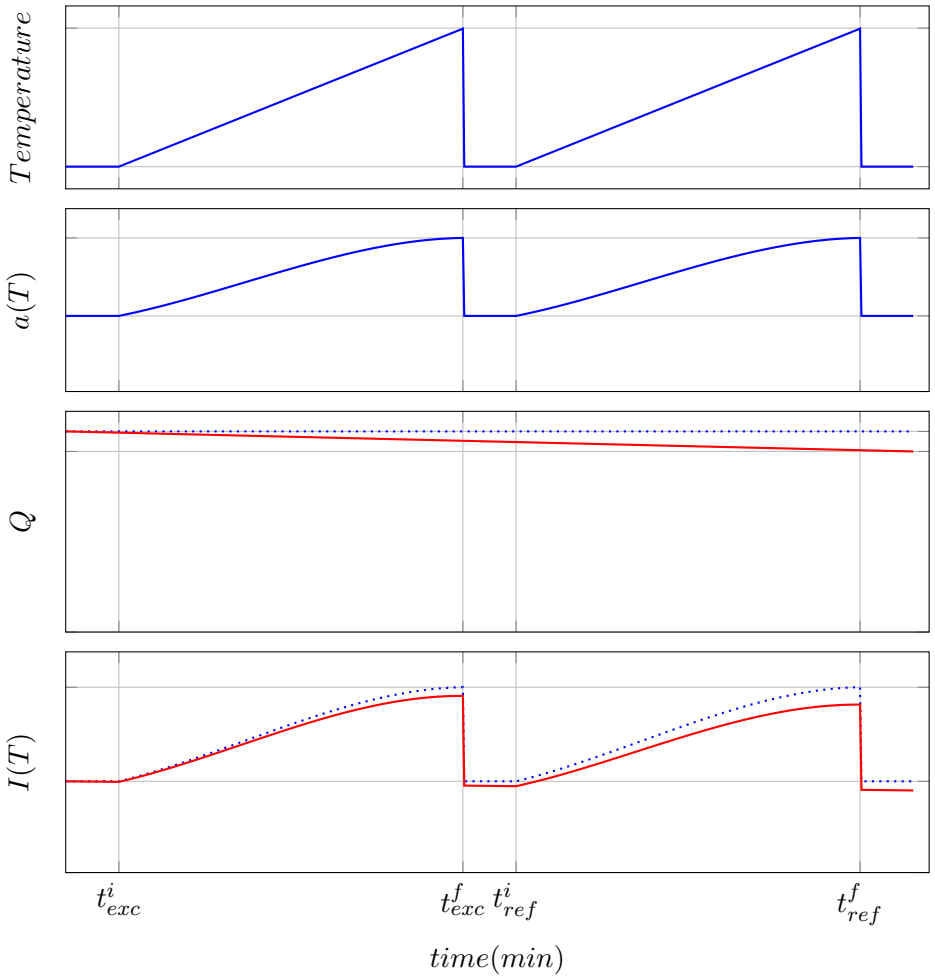


Figure 5.2: Plots of the k-th peak during a BLUU-Tramp experiment with time dependent loss: Temperature, single molecule contribution, amount of emitters and intensity of the peak. Dotted lines show the ideal situation.

5.3 Thermal trigger loss

In this case the quantity of the emitters decrease only when temperature is higher than a thermal threshold, T_t (Fig. 5.3).

Considering the theoretical linearity of the intensity evolution over the temperature, this model can potentially explain deviation from said linearity: as can be seen in Fig 5.3, the existence of a temperature threshold, that triggers the signal loss, lowers the intensity slope in the last part of the ramp. This effect can be reversible or not.

Therefore, the observation of a slope of the intensity curve which is lower than the expected one might be linked to events of conformational changes or aggregation.

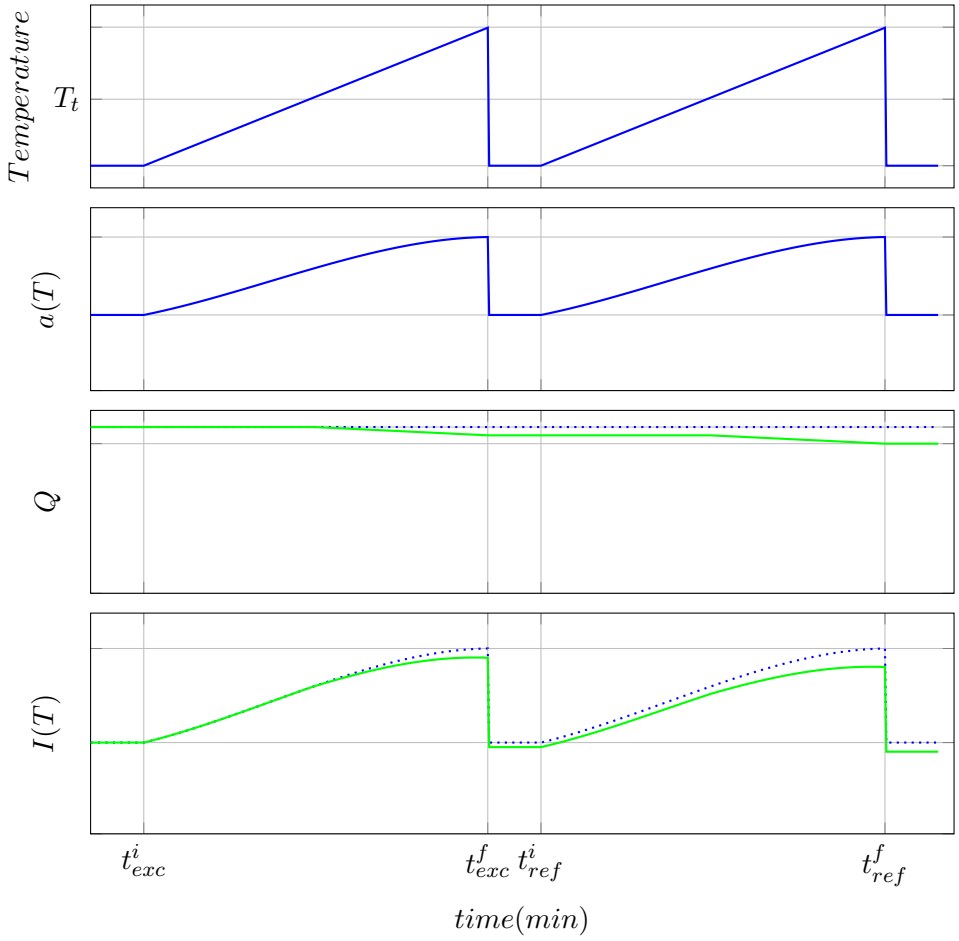


Figure 5.3: Plots of the k-th peak during a BLUU-Tramp experiment with thermal trigger loss: Temperature, single molecule contribution, amount of emitters and intensity of the peak. Dotted lines show the ideal situation.

5.4 One-step loss

In this model the whole loss of the emitters happens in the specific time frame between the two ramps.

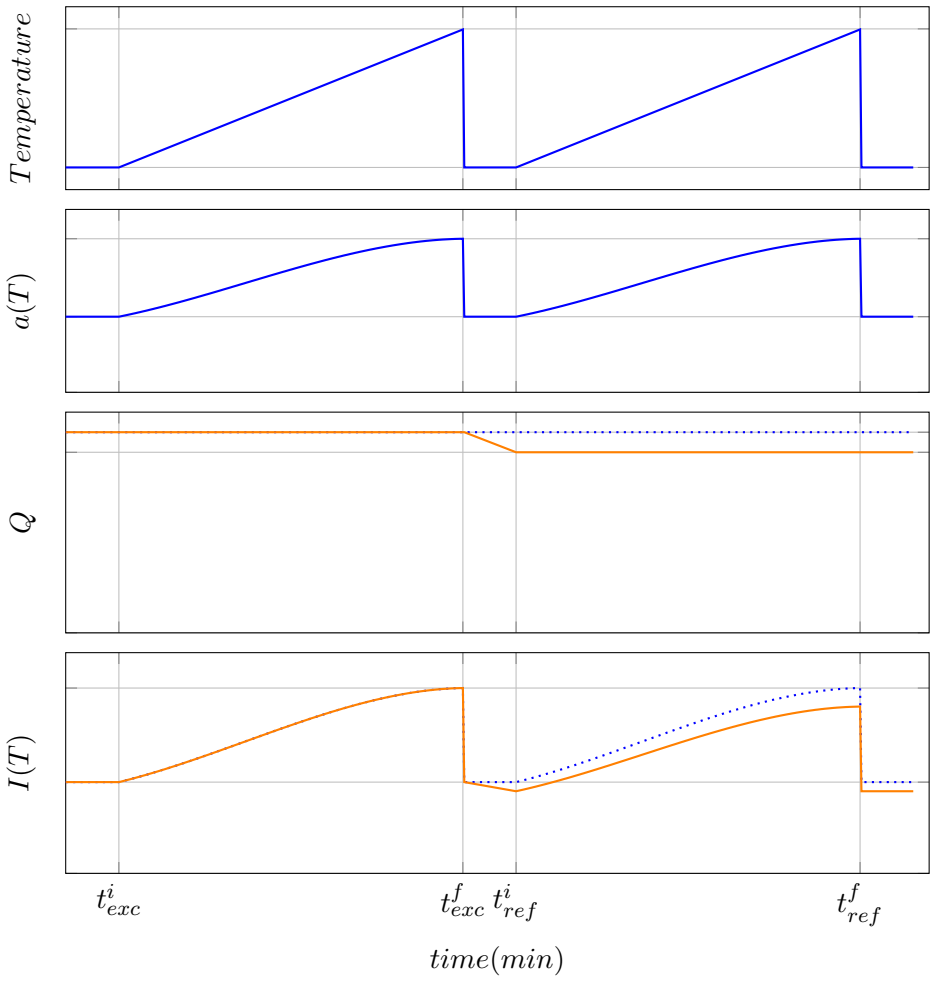


Figure 5.4: Plots of the k-th peak during a BLUU-Tramp experiment with one-step loss: Temperature, single molecule contribution, amount of emitters and intensity of the peak. Dotted lines show the ideal situation.

Among the models shown, this one is the most unrealistic and cannot be related to any physical and chemical process. However its loss can be easily corrected since BLUU-Tramp focuses its attention only on the intensities acquired during the ramps. During the second ramp, the quantity of emitters observed can be related to the one

before the loss by a multiplicative factor α defined as following:

$$\alpha = \frac{Q_i}{Q_f}$$

Consequently

$$I_{ref} = \alpha I_{ref}^{obs}$$

where I_{ref} is the reference intensity to be used in equation (1.1) and I_{ref}^{obs} is the observed intensity during the reference ramp.

5.5 Comparison

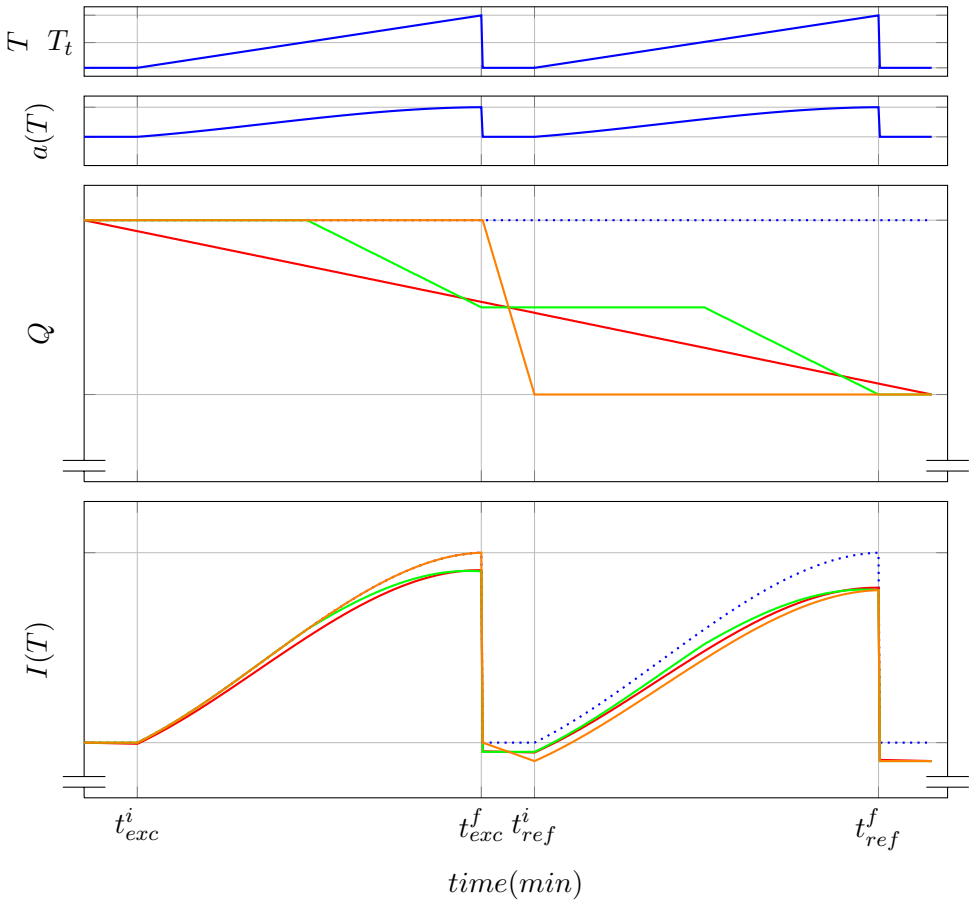


Figure 5.5: Comparison of the BLUU-Tramp experiment loss models: the ideal situation (dotted blu), time dipendent loss (red), thermal trigger loss (green) and one-step loss (orange).

For comparison purposes the final protein loss is equal in all aforementioned models.

It can be immediately noticed that the intensity curves for the different loss models Time dependent loss and Thermal trigger loss lead to very similar profiles during both the exchange and the reference ramp. Similarly, One-step loss adequately retraces their plots during the reference phase. In the exchange phase, this model does not correct the observed curve, however it sufficiently approximates the others during the first half of the ramp that is he relevant part since most of decays have already reached their inflection point in the first half allowing the evaluation of the exchange rate. These considerations along with the simplicity of One-step loss correction make it a good candidate for the emitter loss modeling.

6

Regression Model

The aim of TinT applied to BLUU-Tramp analysis is to obtain a reliable evaluation of the intensity evolution for each peak during the exchange and reference phases. As anticipated in the section 1.3,we can compute the quantity that describes the isotopic exchange process $N[s]$ point by point using the eq. (1.1):

$$N[s] = \frac{I_{ref}[s] - I_{exc}[s]}{I_{ref}[s]} \quad \forall s \tag{6.1}$$

At this stage, $N[s]$ is a sequence of data, but to apply the equation (1.3) and obtain $k(T(t))$ a regression function must be used.

It must be highlighted that, since there is no theoretical reasons behind the choice of the model, the regression can be valid in a limited range centered on the inflection point.

6.1 Regression functions

Due to the experimental BLUU-Tramp protocol (see section 1.3), the curve describing $N(t)$ must:

- be a monotonic decreasing function of the temperature,
- have an right horizontal asymptote at zero,
- have $N(T(0)) \leq 1$.

It is also experimentally observed that $N(t)$ has a **decay** trend (an example is shown in Fig. 6.1) or a **sigmoidal** trend (shown in Fig. 6.2).

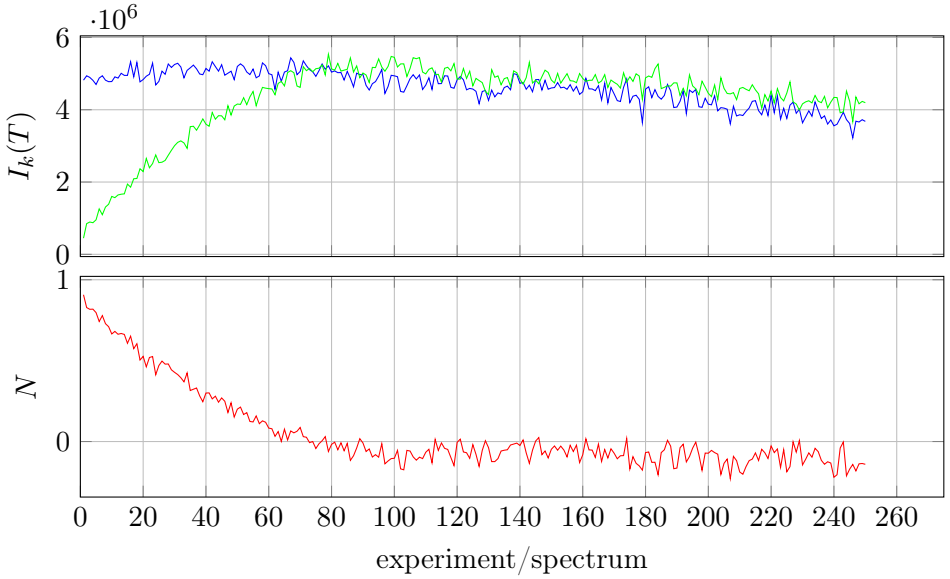


Figure 6.1: Main data from BLUU-Tramp with a decay trend: plots over time of reference and exchange intensities and their derived, N(t), for the 58th AcP track in blue, green and red respectively.

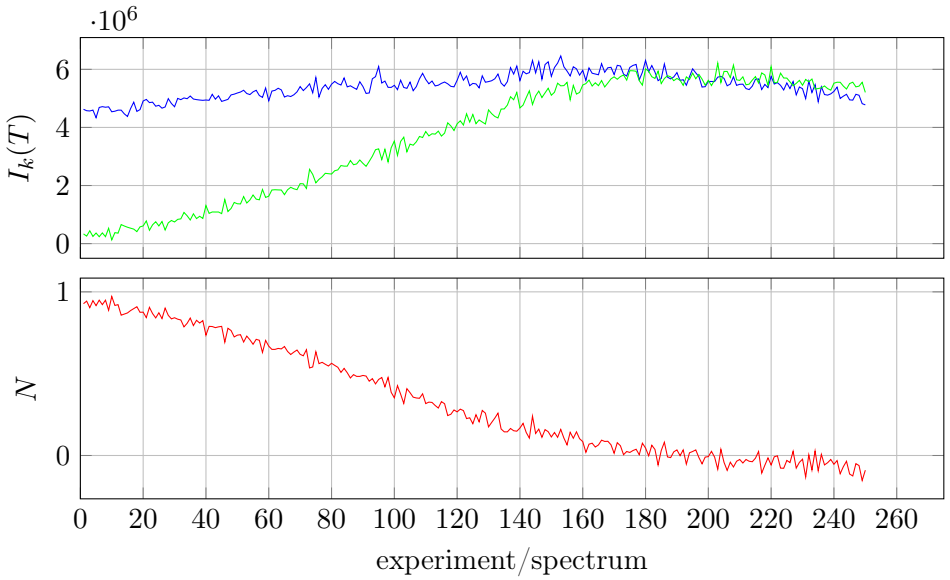


Figure 6.2: Main data from BLUU-Tramp with a sigmoidal trend: plots over time of reference and exchange intensities and their derived, N(t), for the 65th AcP track in blue, green and red respectively.

6.1.1 decay function

A decay plot like the one shows in Fig. 6.3 can be fitted with a decreasing exponential function:

$$N(t) = N(0)e^{-\lambda t}$$

It is mathematically described by only two parameters that define the initial quantity ($N(0)$) and the decay rate (λ).

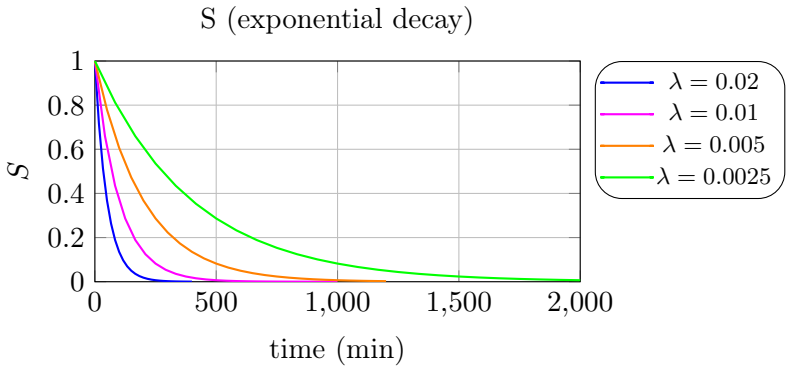


Figure 6.3: Examples of four exponential decays with $N(0) = 1$ and varying λ parameter.

Since two parameters are not enough to describe a curve with two curvatures like the one showed in fig 6.2 the exponential function can be ruled out as a suitable regression function. A sigmoidal function is instead a good candidate to describe the highest number of experimental curves being able to approximate an exponential decays when the decreasing part of the curve is around the initial values of the horizontal axis (Example in Fig. 6.4 with $b = 0.01$ and $a = -0.1$ or $a = -1$).

6.1.2 Gompertz function

Among all the sigmoid functions, the Gompertz Curve [57] (shown in Fig. 6.4) was historically chosen [22, 23]

$$N(t) = N(0)e^{ae^{bt}} \tag{6.2}$$

where parameters must be $a < 0$ and $b > 0$ to obtain a function that satisfy the first two theoretical constraints, while the third condition on $N(0)$ must be checked after the fitting or forced in cases in which data do not show the initial plateau because without it the fitting routine has not any data to find a realistic value for the left-hand

asymptote. In Fig. 6.4 the effect of varying parameters a and b are shown: a affects the displacement and b defines the slope of the function. Furthermore, in the figure it can be observed that the left-hand value (asymptote) is approached much more gradually by the curve than the right-hand asymptote.

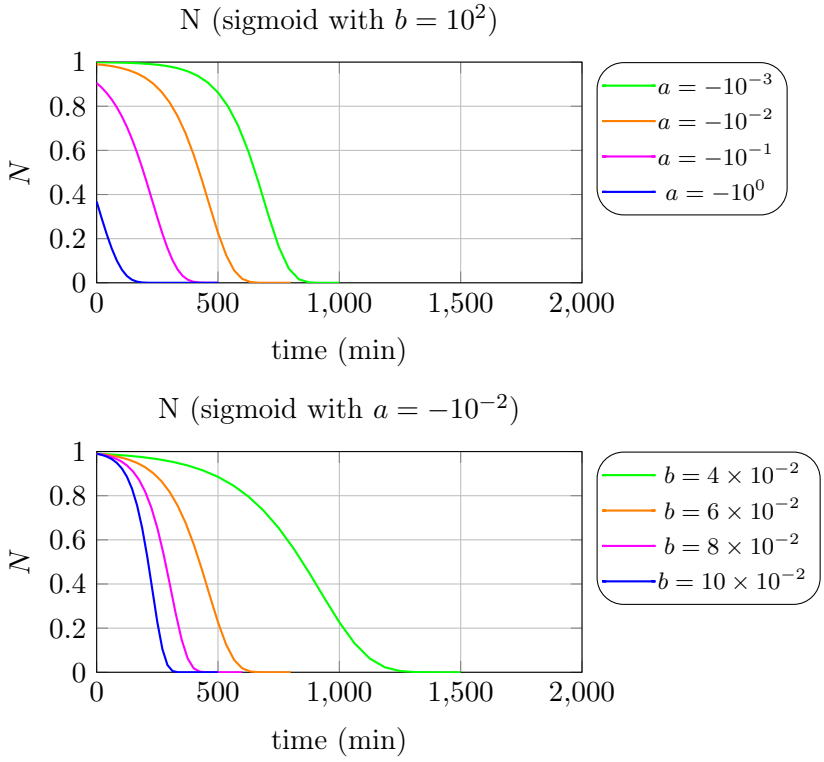


Figure 6.4: Graphs of Gompertz function varying separately a or b.

Applying 1.3 to 6.2, the following equation is obtained:

$$\begin{aligned} k(T(t)) &= -\frac{d}{dt}\left(\ln\left(N(0)\, e^{a\, e^{bt}}\right)\right) \\ &= -b \cdot a e^{bt} \end{aligned} \tag{6.3}$$

The latter equation (6.3) shows that the choice of Gompertz function (as the fitting function) forces the description of the target unknown function, $K(T(t))$, to have a pure exponential behavior. Using only three parameters the function 6.4 can't take into account data with a non-zeros left asymptote. Furthermore it can't shape differently the gradient of the slope or modify the intrinsic asymmetry.

6.1.3 Gompertz function with baseline

When data shows a right horizontal asymptote different from zero, the function (6.2) can be generalized in the following:

$$N(t) = N(0)\, e^{a\, e^{bt}} + d \tag{6.4}$$

Experimental reasons of this behavior could be found in a non complete final back exchange or due to an unknown intensity loss between the exchange and the reference stages (described in section 5).

Applying 1.3 to 6.4, the following equation is obtained:

$$\begin{aligned} k(T(t')) &= -\frac{d}{dt}\left(\ln\left(N(0)\, e^{a\, e^{bt}} + d\right)\right) \\ &= -\frac{b \cdot a\, e^{bt} \cdot N(0)\, e^{a\, e^{bt}}}{N(0)\, e^{a\, e^{bt}} + d} \\ &= -b \cdot a\, e^{bt} \cdot \frac{1}{1 + \frac{d}{N(0)}\, e^{-a\, e^{bt}}} \end{aligned} \tag{6.5}$$

6.1.4 Generalised logistic function

To overcome the limitation of the asymmetric description of the function 6.2 the following generalized logistic function [58] could be used:

$$N(t) = A + \frac{K - A}{\left(C + e^{-B(t-t_0)}\right)^{\frac{1}{\nu}}} \tag{6.6}$$

A and K define left and right asymptote value respectively. As it is shown in Fig. 6.5, t_0 and B define the displacement and the slope of the curve while ν affects the symmetry. Variability of C is not useful in our case and it should be fixed as $C = 1$.

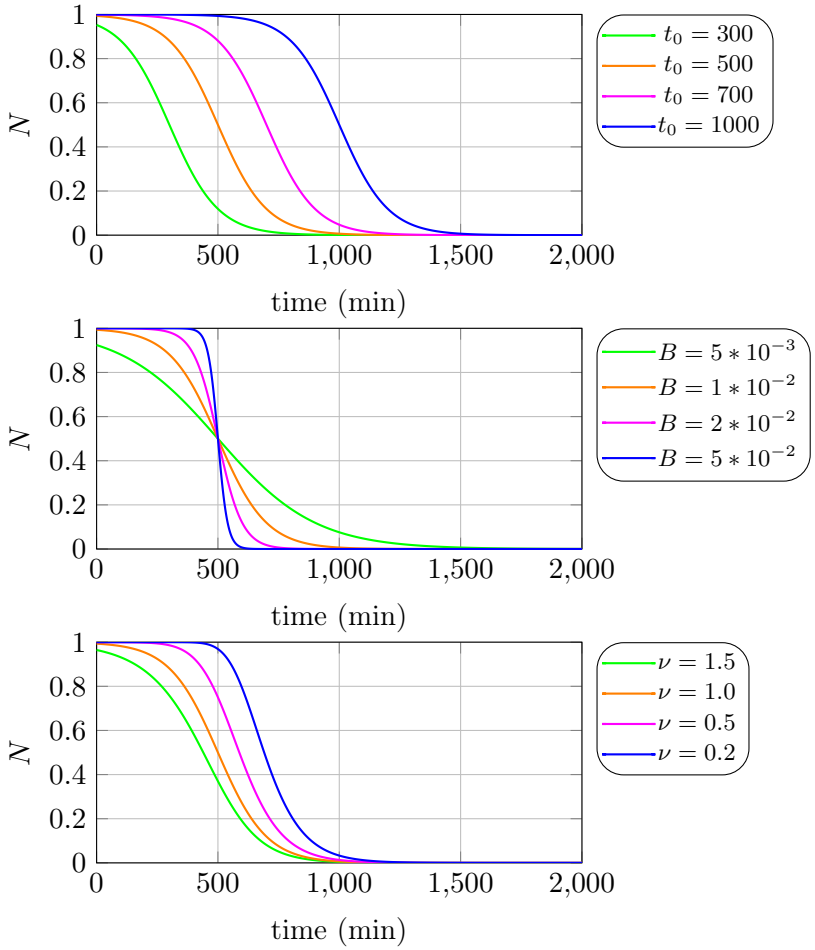


Figure 6.5: Plots of Generalized logistic function varying separately t_0, B or ν. Other parameters are fixed as: $A = 1$, $K = 0$, $C = 1$

Applying 1.3 to 6.6, the following equation is obtained:

$$
\begin{aligned}
k(T(t)) &= -\frac{d}{dt}\log N(t) \\
&= -\frac{d}{dt}\log\left(A + \frac{K-A}{\left(C + e^{-B(t-t_0)}\right)^{\frac{1}{\nu}}}\right) \\
&= -\frac{1}{A + \frac{K-A}{\left(C+e^{-B(t-t_0)}\right)^{\frac{1}{\nu}}}} \cdot \frac{(K-A)\cdot\left(-\frac{1}{\nu}\right)}{\left(C + e^{-B(t-t_0)}\right)^{\frac{1}{\nu}-1}} \cdot (-B)e^{-B(t-t_0)} \\
&= \frac{(A-K)\cdot\left(\frac{1}{\nu}\right)\cdot Be^{-B(t-t_0)}}{A + \frac{K-A}{\left(C+e^{-B(t-t_0)}\right)^{\frac{1}{\nu}}}} \cdot \frac{1}{\left(C + e^{-B(t-t_0)}\right)^{\frac{1}{\nu}-1}} \\
&= \left(\frac{(A-K)B}{\nu}\right) e^{-B(t-t_0)} \frac{1}{A\cdot\left(C + e^{-B(t-t_0)}\right)^{\frac{1}{\nu}-1} + \frac{(K-A)}{\left(C+e^{-B(t-t_0)}\right)}}
\end{aligned}
\tag{6.7}
$$

6.2 Considerations

k(T(t)) was derived for all the introduced models, showing their viability as regression models for $N[s]$.

Due to its greater versatility, the Generalised logistic function is consequently a promising regression in that, unlike the Gompertz function, it allows separate control on the slope and the asymmetry.

7

Conclusion

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. The results of this work were published in an article in high impact-factor journal [49]. Along with the development of TinT we further evaluated some improvements on critical aspects of the subsequent BLUU-Tramp data analysis, namely the loss correction and the fitting of the derived quantity N[s].

References

[1] Freeman J Dyson. **Is science mostly driven by ideas or by tools?** *Science*, **338**(6113):1426–1427, 2012. 1

[2] Peter Galison. *Image and logic: A material culture of microphysics*. University of Chicago Press, 1997. 1

[3] P Balaram. **Tools as Drivers of Science**. *Current Science*, **103**(12):1383–1384, dec 2012. 1

[4] Thomas S. Kuhn. *The structure of scientific revolutions*. University of Chicago Press, Chicago, 1970. 1

[5] Gordon Bell, Tony Hey, and Alex Szalay. **Beyond the Data Deluge**. *Science*, **323**(5919):1297–1298, 2009. 2

[6] Tony Hey, Stewart Tansley, and Kristin Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Redmond, Washington, 2009. 2

[7] Gottfried Otting. **Experimental NMR techniques for studies of protein-ligand interactions**. *Curr Op Struct Biol*, **3**:760–768, 1993. 2

[8] Maurizio Pellecchia, Diana L. Montgomery, Shawn Y. Stevens, Craig W. Vander Kooi, Hwa-Ping Feng, Lila M. Gierasch, and Erik R.P. Zuiderweg. **Structural insights into substrate binding by the molecular chaperone DnaK**. *Nat Struct Mol Biol*, **7**:298–303, 2000. 2

[9] Erik R. P. Zuiderweg. **Mapping protein–protein interactions in solution by NMR spectroscopy**. *Biochemistry*, **41**(1):1–7, 2002. 2

[10] Christopher A. Lepre, Jonathan M. Moore, and Jeffrey W. Peng. **Theory and applications of NMR-based screening in pharmaceutical research**. *Chem Rev*, **104**:3641–3676, 2004. 3

[11] Philip J. Hajduk, Jürgen Dinges, Gregory F. Miknis, Megan Merlock, Tim Middleton, Dale J. Kempf, David A. Egan, Karl A. Walter, Terry S. Robins,

Suzy B. Shuker, Thomas F. Holzman, and Stephen W. Fesik. **NMR-based discovery of lead inhibitors that block DNA binding of the human papillomavirus E2 protein**. *J Med Chem*, **40**(20):3144–3150, 1997. 3

[12] Suzanne B. Shuker, Philip J. Hajduk, Robert P. Meadows, and Stephen W. Fesik. **Discovering high-affinity ligands for proteins: SAR by NMR**. *Science*, **274**:1531–1534, 1996. 3

[13] Mark A. McCoy and Daniel F. Wyss. **Alignment of weakly interacting molecules to protein surfaces using simulations of chemical shift perturbations.** *Journal of Biomolecular NMR*, **18**:189–98, November 2000. 3

[14] Xavier J. Morelli, P. Nuno Palma, Françoise Guerlesquin, and Alan C. Rigby. **A novel approach for assessing macromolecular complexes combining soft-docking calculations with NMR data.** *Protein Sci.*, **10**:2131–2137, 2001. 3

[15] Amr Fahmy and Gerhard Wagner. **TreeDock: a tool for protein docking based on minimizing van der Waals energies.** *J Am Chem Soc*, **124**:1241–50, 2002. 3

[16] Cyril Dominguez, Rolf Boelens, and Alexandre M. J. J. Bonvin. **HADDOCK: a protein-protein docking approach based on biochemical or biophysical information**. *J Am Chem Soc*, **125**(7):1731–1737, 2003. 3

[17] G Marius Clore and Charles D Schwieters. **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**(10):2902–12, 2003. 3

[18] Stephen Boulton and Giuseppe Melacini. **Advances in NMR Methods To Map Allosteric Sites: From Models to Translation**. *Chemical Reviews*, **116**(11):6267–6304, 2016. 3

[19] S Walter Englander and Neville R Kallenbach. **Hydrogen exchange and structural dynamics of proteins and nucleic acids.** *Q Rev Biophys*, **16**(4):521–655, 1983. 3

[20] S. Walter Englander, Leland Mayne, and Mallela M.G. Krishna. **Protein folding and misfolding: mechanism and principles**. *Q Rev Biophys*, **40**(4):287–326, 2007. 3

[21] Enrico Rennella, Alessandra Corazza, Federico Fogolari, Paolo Viglino, Sofia Giorgetti, Monica Stoppini, Vittorio Bellotti, and Gennaro Esposito. **Equilibrium unfolding thermodynamics of beta2-microglobulin analyzed through native-state H/D exchange.** *Biophys J*, **96**(1):169–79, 2009. 3

[22] Enrico Rennella, Alessandra Corazza, Luca Codutti, Araldo Causero, Vittorio Bellotti, Monica Stoppini, Paolo Viglino, Federico Fogolari, and Gennaro Esposito. **Single-shot NMR measurement of protein unfolding landscapes.** *Biochim Biophys Acta*, **1824**(6):842–9, 2012. 3, 4, 5, 21, 41

[23] Enrico Rennella, Alessandra Corazza, Luca Codutti, Vittorio Bellotti, Monica Stoppini, Paolo Viglino, Federico Fogolari, and Gennaro Esposito. **Determining the energy landscape of proteins by a fast isotope exchange NMR approach.** *J Am Chem Soc*, **134**(10):4457–60, 2012. 3, 4, 5, 21, 41

[24] Mike P. Williamson. **Using chemical shift perturbation to characterise ligand binding**. *Prog Nucl Magn Res Sp*, **73**:1–16, 2013. 3, 26

[25] Geoffrey Bodenhausen and David J. Ruben. **Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy**. *Chem Phys Lett*, **69**(1):185–189, 1980. 3

[26] Ad Bax, Mitsuhiko Ikura, Lewis E Kay, Dennis A Torchia, and Rolf Tschudin. **Comparison of different modes of two-dimensional reverse-correlation NMR for the study of proteins**. *J Magn Reson*, **86**(2):304–318, 1990. 3

[27] Paul Schanda and Bernhard Brutscher. **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**(22):8014–8015, 2005. 3, 20

[28] Ewen Lescop, Paul Schanda, and Bernhard Brutscher. **A set of BEST triple-resonance experiments for time-optimized protein resonance assignment**. *J Magn Reson*, **187**(1):163–169, 2007. 3, 20

[29] Xin Gao. **Recent advances in computational methods for Nuclear Magnetic Resonance data processing**. *Genomics Proteomics Bioinf*, **11**(1):29–33, 2013. 3

[30] Gerard J Kleywegt, Rolf Boelens, and Robert Kaptein. **A versatile approach toward the partially automatic recognition of cross peaks in 2D 1H NMR spectra**. *J Magn Reson*, **88**(3):601–608, 1990. 3

[31] Daniel S Garrett, Robert Powers, Angela M Gronenborn, and G.Marius Clore. **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**(1):214–220, 1991. 3

[32] Bruce A. Johnson and Richard A. Blevins. **NMRView, a computer program for the visualization and analysis of NMR data**. *J Biomol NMR*, **4**(5):603–614, 1994. 3, 4

[33] Christoph Antz, Klaus-Peter Neidig, and Hans Robert Kalbitzer. **A general Bayesian method for an automated signal class recognition in 2D NMR spectra combined with a multivariate discriminant analysis**. *J Biomol NMR*, **5**(3):287–296, 1995. 3

[34] Enrico A. Carrara, Franco Pagliari, and Claudio Nicolini. **Neural networks for the peak-picking of nuclear magnetic resonance spectra**. *Neural Networks*, **6**(7):1023–1032, 1993. 3

[35] Simon A Corne, A Peter Johnson, and Julie Fisher. **An artificial neural network for classifying cross peaks in two-dimensional NMR spectra**. *J Magn Reson*, **100**(2):256–266, 1992. 3

[36] Alain Rouh, Alain Louis-Joseph, and Jean-Yves Lallemand. **Bayesian signal extraction from noisy FT NMR spectra**. *J Biomol NMR*, **4**(4):505–518, 1994. 3

[37] Babak Alipanahi, Xin Gao, Emre Karakoc, Logan Donaldson, and Ming Li. **PICKY: a novel SVD-based NMR spectra peak picking method**. *Bioinformatics*, **25**(12):i268–i275, 2009. 3

[38] Reto Koradi, Martin Billeter, Max Engeli, Peter Guentert, and Kurt Wuethrich. **Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY**. *J Magn Reson*, **135**:288–297, 1998. 3

[39] Dmitry M Korzhnev, Ilghiz V Ibraghimov, Martin Billeter, and Vladislav Yu Orekhov. **MUNIN: application of three-way decomposition to the analysis of heteronuclear NMR relaxation data**. *J Biomol NMR*, **21**(3):263–268, 2001. 3

[40] Vladislav Yu Orekhov, Ilghiz V Ibraghimov, and Martin Billeter. **MUNIN: a new approach to multi-dimensional NMR spectra interpretation**. *J Biomol NMR*, **20**(1):49–60, 2001. 3

[41] Zhi Liu, Ahmed Abbas, Bing-Yi Jing, and Xin Gao. **WaVPeak: picking NMR peaks through wavelet-based smoothing and volume-based filtering**. *Bioinformatics*, **28**(7):914–920, 2012. 3, 4

[42] Ahmed Abbas, Kong Xin-Bing, Liu Zhi, Jing Bing-Yi, and Gao Xin. **Automatic peak selection by a Benjamini-Hochberg-based algorithm**. *PLoS ONE*, **8**(1):1–10, 2013. 3

[43] Piotr Klukowski, Michal J. Walczak, Adam Gonczarek, Julien Boudet, and Gerhard Wider. **Computer vision-based automated peak picking applied to protein NMR spectra**. *Bioinformatics*, **31**(18):2981–2988, 2015. 3

[44] Yichen Cheng, Xin Gao, and Faming Liang. **Bayesian Peak Picking for NMR Spectra**. *Genomics Proteomics Bioinf*, **12**(1):39–47, 2014. 4

[45] Chen Peng, Stephen W. Unger, Fabian V. Filipp, Michael Sattler, and Sándor Szalma. **Automated evaluation of chemical shift perturbation spectra: New approaches to quantitative analysis of receptor-ligand interaction NMR spectra**. *J Biomol NMR*, **29**(4):491–504, 2004. 4

[46] Lisa Fukui and Yuan Chen. **NvMap: automated analysis of NMR chemical shift perturbation data**. *Bioinformatics*, **23**(3):378–380, 2007. 4

[47] P. Ravel, G. Kister, T. E. Malliavin, and M. A. Delsuc. **A general algorithm for peak-tracking in multi-dimensional NMR experiments**. *J Biomol NMR*, **37**:265–275, 2007. 4

[48] Richard Jang, Xin Gao, and Ming Li. **Combining automated peak tracking in SAR by NMR with structure-based backbone assignment from 15N-NOESY**. *BMC Bioinformatics*, **13**:1–15, 2012. 4, 27

[49] Tommaso Banelli, Marco Vuano, Federico Fogolari, Andrea Fusiello, Gennaro Esposito, and Alessandra Corazza. **Automation of peak-tracking analysis of stepwise perturbed NMR spectra**. *Journal of Biomolecular NMR*, pages 1–14, 2017. 7, 47

[50] Petros Maragos. **Morphological filtering for image enhancement and feature detection**. In *Image & Video Processing Handbook*, chapter 3.3, pages 135–156. Elsevier Academic Press, Amsterdam, 2nd edition, 2005. 9, 10

[51] Jean Serra. *Image analysis and mathematical morphology*, **1**. Academic press, New York, 1982. 9, 10

[52] Jean Serra. *Image analysis and mathematical morphology: Theoretical Advances.*, **2**. Academic press, New York, 1988. 9, 10

[53] Chris A. Glasbey and Graham W. Horgan. *Image Analysis for the Biological Sciences*. John Wiley and Sons, 1995. 10

[54] John W. Eaton, David Bateman, Søren Hauberg, and Rik Wehbring. *GNU Octave version 4.0.0 manual: a high-level interactive language for numerical computations*. 2015. 11

[55] Albert Rose. **The Sensitivity Performance of the Human Eye on an Absolute Scale**. *J Opt Soc Am*, **38**(2):196–208, 1948. 13, 28

[56] Frank Delaglio, Stephan Grzesiek, Geerten W. Vuister, Guang Zhu, John Pfeifer, and Ad Bax. **NMRPipe: a multidimensional spectral processing system based on UNIX pipes.** *J Biomol NMR*, **6**(3):277–93, 1995. 21

[57] Dragan Jukić, Gordana Kralik, and Rudolf Scitovski. **Least-squares fitting Gompertz curve**. *Journal of Computational and Applied Mathematics*, **169**(2):359–375, 2004. 41

[58] F. J. Richards. **A Flexible Growth Function for Empirical Use**. *Journal of Experimental Botany*, **10**(2):290, 1959. 43

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor Dott Alessandra Corazza and my external reviewer Prof Andrea Fusiello, that soon became co-supervisor of the PhD project, for their guidance, motivation, patience and faith in giving me all the space I needed to learn, grow and work. It was a pleasure working with them.

I think my colleague and friend Dott Marco Vuano deserves a special mention for all concrete help and moral support and for his enthusiastic participation in all debates, discussions but also in foolishnesses that arise after so much time rubbing shoulders.

I am grateful to Prof Gennaro Esposito and Prof Federico Fogolari for their insightful comments, expertise and encouragement.

I must thank Jennifer Kay who has been the first external user of TinT and pointed out some bugs in the software.

I would like to thank Richard Jang for providing PeakWalker and assisting in using it.

I must also thank all my colleagues and friends Dott Haritha Haridas, Dott Maurizio Polano, Dott Raffaella Picco, Dott Fabrizio Serra, Dott Maurizio Ballico, Dott Chiara Mimmi, Boris Cvrtnjak, Cedrix Dongmo, Dott Aristeo Madera, Cristina Cantarutti, Dott Alex Stopar and all the students who have passed through the computational lab.

I would like to thank my wife Roberta and all my family for providing me with unfailing support and continuous encouragement throughout this project and for injecting enthusiasm into my life in general. Their support was essential to my success here!

I am thankful to all my friends for all the fun we have had in the last four years that reduce my frustrations and enhance my passion.

I would like to acknowledge the thousands of individuals who have coded for GNU Octave project and LaTeXproject for free. It is due to their efforts that we can generate professionally scientific software and typeset PDFs now.

Last but not least, I would like to express my gratitude to all my teachers who put their faith in me and urged me to do better.