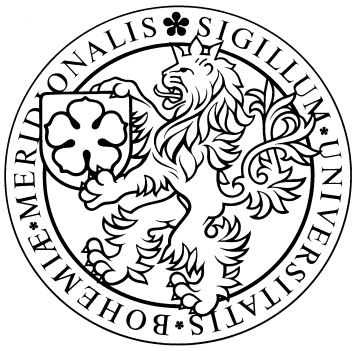


University of South Bohemia

in České Budějovice

Institute of Physical Biology

Department of Bioengineering

Ph.D. Thesis

in the field of Biophysics

LC-MS analysis based on probabilistic approach

Ing. Jan Urban

Supervisor: Doc. RNDr. Dalibor Štys, CSc.

Nové Hrady 2010

Declaration of Authenticity

I herewith declare that I autonomously carried out the PhD thesis. To the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference has been made.

In Nové Hrady 15.06.2010 Jan Urban

Acknowledgments

This work was partly supported by the Ministry of Education, Youth and Sports of the Czech Republic under the grant MSM 6007665808 and grant HCTFOOD A/CZ0046/1/0008 of EEA funds.

I also would like to thank to

- my supervisor Doc. RNDr. Dalibor Štys, CSc.
- my colleagues and coworkers Vítězslav Březina, Jana Cimlová, Anna Churilova, Jan Fesl, Pavel Hrouzek, Irina Kishko, Petr Kohout, Jiří Kopecký, Martin Lukeš, Tomáš Levitner, Lucie Marková, Harald Martens, Karel Matouš, Štěpán Papáček, Bård Rasmussen, Pavel Souček, Jiří Soukup, Petr Šimek, Radek Tesař, Jan Vaněk and Petr Zelík for relevant discussion, assistance with the experiments and software testing.
- my students Lukáš Machlica, Jiří Maršálek and Andrey Timoshenko.
- my mother and my wife as well as whole my family for support during my study.

Motto

'To see what is general in what is particular, and what is permanent in what is transitory, is the aim of scientific thought.'

Alfred North Whitehead, 1911.

Annotation

Liquid chromatography (LC) in tandem with mass spectrometry (MS) is a measurement tool for obtain information about the compounds in the investigated extracts. There were already developed methods for processing and analysis of measured data sets. However, only partial problems of processing/analysis task were handled independently.

Therefore, the first part describes existing methods and techniques commonly used in the LC-MS for the processing and analysis today.

In this thesis an approach based on the theory of systems is used for description of abstract model above the measured data. This model encapsulated all processing/analysis steps into appropriate and consistent mathematical space. The creation of this model via description of the measurement device and data outputs is introduced.

Abstract model of LC-MS data set is used to decompose the measurement into three partial contributions, the analyte signal, the random noise and the systemic noise. The separation process of the signal could be estimated using the probabilistic approach.

That probabilistic approach to the LC-MS analysis was implemented in the developed software, which was published in the Bioinformatics Journal.

Anotace

Kapalinová chromatografie (LC) ve spojení s hmotnostním spektrometrem (MS) představuje měřící techniku, která umožňuje získat informace o látkách ve zkoumaném extraktu. V minulosti již bylo vyvinuto mnoho metod, jak ke zpracování, tak k analýze naměřených dat. Nicméně byly řešena pouze část problematiky a to nezávisle na ostatních.

Tudíž první část práce se zabývá popisem existujících metod a technik používaných v současné době k analýze a zpracování měření z LC-MS.

Tato disertační práce představuje přístup založený na teorii sytému, jímž popisuje abstraktní model naměřených dat. Takto vytvořený model uzavírá veškeré kroky ke zpracování i k analýzy na vhodné a odpovídající matematické místo. Tvorba modelu je ukázána během popisu měřícího zařízení a jeho výstupů.

Abstraktní model LC-MS dat je použit k rozdělení naměřených dat na jeho tři částečné příspěvky, vlastní signál analytu, náhodný šum a systémový šum. Rozdělení na jednotlivé příspěvky lze odhadnout pomocí pravděpodobnostních metod.

Pravděpodobnostní přístup k analýze LC-MS měření je implementován v software, který byl vyvinut a publikován v Bioinformatics Journal.

List of abbreviations

ASCII	American standard code for information interchange
APCI	Atmospheric pressure photoionization
BBTA	Blank based time alignment
C12	The most abundant isotope of carbon (98.89%)
C13	Natural and stable isotope of carbon (1.1%)
CI	Chemical ionization
COW	Correlation optimized warping
CPU	Central processing unit
DCOW	Direct correlation optimized warping
DOL	Data object library
DTW	Dynamic time warping
ESI	Electrospray ionization
EOF	End of file
EOL	End of line
GC	Gas chromatography
GPU	Graphic processing unit
HDD	Hard disk drive
HPLC	High performance liquid chromatography
IEC	International electrotechnical commission
IS	Internal standard(s)
ISO	International organization for standardization
IUPAC	International union of pure and applied chemistry
JCAMP	Joint committee on atomic and molecular physical data
LC	Liquid chromatography
LOD	Limits of detection
LSERs	Linear solvation energy relationships
M	Set of mass values
MALDI	Matrix assisted laser desorption/ionization
MB	Mega bytes
MeOH	Methanol
MS	Mass spectrometry

MSD	Mass selective detector
MSI	Metabolomics standards initiative
MSC	Multiplicative signal correction
MVA	Multivariate data analysis
MzXML	Mass spectrometry eXtensible markup language data format
m/z	Mass to charge ratio
Netcdf	Network common data form
NOMIS	Normalization using optimal selection of multiple internal standards
ODE	Ordinary differential equation
PARAFAC	Parallel Factor analysis
PC	Personal computer
PCA	Principal component analysis
PIE	Peak integration error
PDA	Photodiode array
PDF	Probability density function
PLSR	Partial least squares regression
PTW	Parametric time warping
RT	Retention time
RSTD	Robust standard deviation
SIC	Single ion chromatogram
SNR	Signal to noise ratio
SNV	Standard normal variete
SP	Solvent (injection) peak
SPC	Shape peak count
Sp.	Species
T	Set of time values
TIC	Total ion current (or Total ion chromatogram)
Th	Threshold
UV	Ultraviolet
WOT	Wash-out tail
XCMS	R package for processing mass spectrometry
XML	eXtensible markup language

List of Figures

2.1 The relationship of chemometrics 5

3.1 Information channel in LC/MS. 19
3.2 Example of LC-MS measurement 26

4.1 Two examples of blank measurements 34
4.2 Example of blank markers selection 43
4.3 Example of analyte time marker selection 45
4.4 Comparison of all TICs 53
4.5 Detail of Nystatin part of TICs in DCOW and BBTA 55
4.6 Details of several TICs parts 58
4.7 Example of two mixture of standards 59

5.1 3D Example of blank measurements in decimal and logarithmic scale 63
5.2 Example of mass signal max values 64
5.3 Examples of filtered blank and analyte measurement TICs 67

6.1 Chemical structure of Nystatin molecule. 76
6.2 Chromatograms, spectra of Nystatin, probability factors 77

List of Tables

4.1 Values of blank and analytes time sets values. 52
4.2 Time values of blank and analytes set to the reference time set. . . 55
4.3 Comparison of COW, DCOW and BBTA parameters 56
4.4 Average computed distance between pairs of spectra 57

6.1 Spectra correlation of *Pure* and *Mix* with *Reference* 79

Contents

Declaration of Authenticity **I**

Acknowledgments **II**

Motto **III**

Annotation **IV**

Annotation CS V

List of abbreviations **VI**

List of figures **VIII**

List of tables **IX**

Contents **X**

1 Introduction **1**

2 Current state of LC-MS data processing and analysis **3**

2.1 Data filtering 6

2.2 Feature detection 8

2.3 Alignment 9

2.4 Comparison 12

2.5 Omics 13

2.6 File formats and arithmetics issues 16

2.6.1 Overflow 18

3 LC-MS according to the system theory 19
3.1 Liquid chromatography 20
3.2 Mass spectrometry 21
3.3 System theory 22
3.4 Abstract model 24
3.4.1 Mass resolution 27
3.4.2 Mass accuracy and mass precission 28

4 Blank based time alignment 30
4.1 Step 1.:Reduction of blank points 35
4.2 Step 2.:Markers selection 37
4.3 Step 3.:Transformation function(s) 46
4.4 Comparison of BBTA with COW 50

5 Adaptive filter for baseline thresholding 62
5.1 Theory and calculation 63
5.2 Results of baseline filtration 66

6 Noise filtration via probabilistic theory 68
6.1 Probabilistic approach 69
6.2 Estimation of random noise characteristics 72
6.3 Estimation of systematic noise characteristics 74
6.4 Advantages of probabilistic approach 75
6.5 Probabilistic filtration of Nystatin in the Nostoc sp. extract 76

7 Conclusion 81

Bibliography 83

Appendix A: Getting more information from LC-MS A-1

Appendix B: Cytotoxicity and Secondary Metabolites B-1

Appendix C: Concentration response dependence C-1

Appendix D: Expertomica Metabolite Profiling D-1

Appendix E: MetDB v2.5 **E-1**

Appendix F: Software for LC-MS **F-1**

1 Introduction

Contemporary paradigms of real systems assume that any natural or artificial process under study fulfills the general set of nature laws. Those laws are a priori stochastic (probabilistic) descriptions, where a deterministic case is just a special case of stochasticity (with the probabilities equal to one). The stochastic behavior is given by our inability to measure (observe) exact values of all system attributes with infinite accuracy. All individual objects of interest are ordered to the proper subtraction of general laws, usually just by parameterization. It is necessary to mention that any thought construction above the object behavior never works with the real object itself, only with the abstract object. Thus, the abstract objects in every single theory are created as more or less homomorphic models of the real objects [32].

Every model is created for a reason. One of the most justified purpose is to discover (literally emphasize) congruent description (and parameterization) of known general law in the given case of study. This process verifies or rejects both the generality of the law in nature observation and hypothesis of exact instance of this law in the object of interest. The situation that rejects generality demands modification of paradigms and it is not considered in this thesis. On the other hand, the exact instance should be verified statistically by huge amount of measurements (observations) and their fits to the model.

The states of the LC-MS dataset model are not stationary nor invariant in some attribute. Also, they are not periodic (however, state of the sub-system could be periodic, like pumping contribution on the baseline) or zero. Therefore, they are also not ergodic. There is not any average metric that could be the limit value for the state variables. The variables should be limited only in set of possible values. The states are just stochastic and this stochasticity is causal and

dependent (on injection peak, column history, etc.).

However, even the measured data were obtained by measurement device which was designed according to some model of physical (chemical, biological, mathematical) process of the measure and they are always quantized in the value domain. It is done by analog-digital converters on the input of (control, storage and processing) computers and at many other instances which reflects primarily our inability to measure with infinitesimal accuracy and precision. Therefore, all possible datasets are already models according to the theory of systems. Moreover, proper description of the mathematical space results in description of the abstract model. The initial hypothesis in this work is that model of data from Liquid Chromatography in tandem with Mass spectrometry (LC-MS) fulfills the contemporary paradigms.

The abstract model of dataset itself and its intuitive construction is presented in this thesis via the description of measurement process. A comprehensive abstract model of measured data is necessary for consequential processing or analysis. The reason of mathematically described data model is to encapsulate behavior hypothesis into appropriate mathematical space. Layout of the possible domain values ensures that created behavior models also fulfill the mathematical presumptions of data model. Furthermore, outputs of the processing and analysis are therefore also consistent with the theory of abstract systems.

Thus, a following hypothesis is assumed and tested:
(a) Raw measurement data output of LC-MS consist of three partial contributions, the analyte signal, the random noise and the systemic noise.
(b) Partitioning process of data into individual signal source contribution could be estimated using the probabilistic approach.
(c) All necessary information for the processing and analysis are already present in the data model and arise from the model structure according to the probabilistic theory. In another words, evaluation of individual probabilities is unsupervised.

The presumption for this hypothesis as well as current state of processing approaches are discussed. The estimated probabilities are then used for subsequent measurement decomposition and analytes spectra filtration.

2 Current state of LC-MS data processing and analysis

Liquid chromatography (LC) in tandem with Mass spectrometry (MS) is widely used in many chemical and biochemical analytical setups, especially in so-called omics science to analyze the content of measured samples ([68, 73]). Output of omic sciences is utilized as basis for systemic approach to organism analysis, the systems biology([74]). The omics technologies make the systems biology realistic and experiment-based science. They reveal hidden properties of the compounds present in biological samples. Metabolomics ([20], [60], [11],[63], [51], [52], [61]), proteomics ([77],[78]) or lipidomics ([15],[17],[16]) profiling lies in the heart of gene products profile identification ([4]). Inclusion of metabolomics and proteomics into systems biology often assumes certain relatively high degree of comprehensiveness and quantitative estimation. Complete experimentation to meet this requirement is seldom achieved. In this thesis the data reliability is examined, measured and utilized in practice.

LC-MS measurement is one of the key tool for the biochemical pathways analysis ([75]). Difficulties in finding the correspondence between the experiment and the biochemical-model-based predictions of systems biology lead towards the definition of integrative biology with 'greater emphasis on the process of developing such models ' ([5]). This leads to inclusion of biological experiments modeling on the basis of its description as an abstract system as legitimate way of description of biological experiment. The advantage of this approach is the naturalness with which the characteristics of the biological model ([6, 7, 8]) and that of the experiment itself may be integrated together. Nevertheless, it would be clearly missing the target of creation of biological models if the two models, model of technical

experiment and model of biological experiment, would not be dissected.

Various approaches were developed for LC-MS data processing and data analysis. Those two terms are often interchanged and definitions are unclear. However, Katajama ([127]) separates the data handling tasks into processing and analysis in reasonable way. Filtering, feature detection, alignment and normalization tasks belong to data processing group. The processing is necessary step before the analysis. It transforms the raw data into more transparent format for the analysis. The analysis is then interpretation of the processed data.

In this thesis, the creation of abstract model, the time alignment, evaluation of the probabilities and peak detection are considered as processing tasks. The separation of the measurement into analyte signal, random and systemic noise using the probabilities has character of both processing and analysis. It is filtration from the noise removal (or level tuning as will be explained later) point of view. It is also the analysis from the content identification (what is noise) point of view. Exactly the same question may be asked in peak detection problem. Determination of retention time, molecular ion or full width at half maximum are features, and therefore data processing. Integrative spectra, compounds identified as parallel peaks and distinguish between overlapping peaks are data analysis. Fortunately, it is clear that comparison tasks across the samples is definitely analysis, even using probability based processing. Therefore, both data processing and data analysis are described in this thesis. The terms processing and analysis may be misinterpreted or join sometimes, because of fuzzy border between definitions or points of view.

There is one important philosophical question that requires discussion. Common definition of data processing is that any processing changed the data content. This is obviously true in filtration and alignment case. However probabilistic approach adds the information about probabilities to every single point of the data. Use of this additional information in further computations changes the data content (removes low probability signal). Therefore, the basis of probabilistic approach should be considered as data analysis and the instances of data handling according to this probabilities are data processing.

Of course, both processing and analysis could be also together considered as chemometrics ([42, 43]). Chemometrics as a whole is a set of approaches for

extraction of information from chemical or biochemical systems using methods of applied mathematics. It consists of two main tasks a) modeling relationships and structure of the system and b) predicting new properties or behavior of the system. Chemometrics uses techniques as calibration, classification, pattern recognition, clustering, multivariate analysis, experimental design, signal processing, etc.

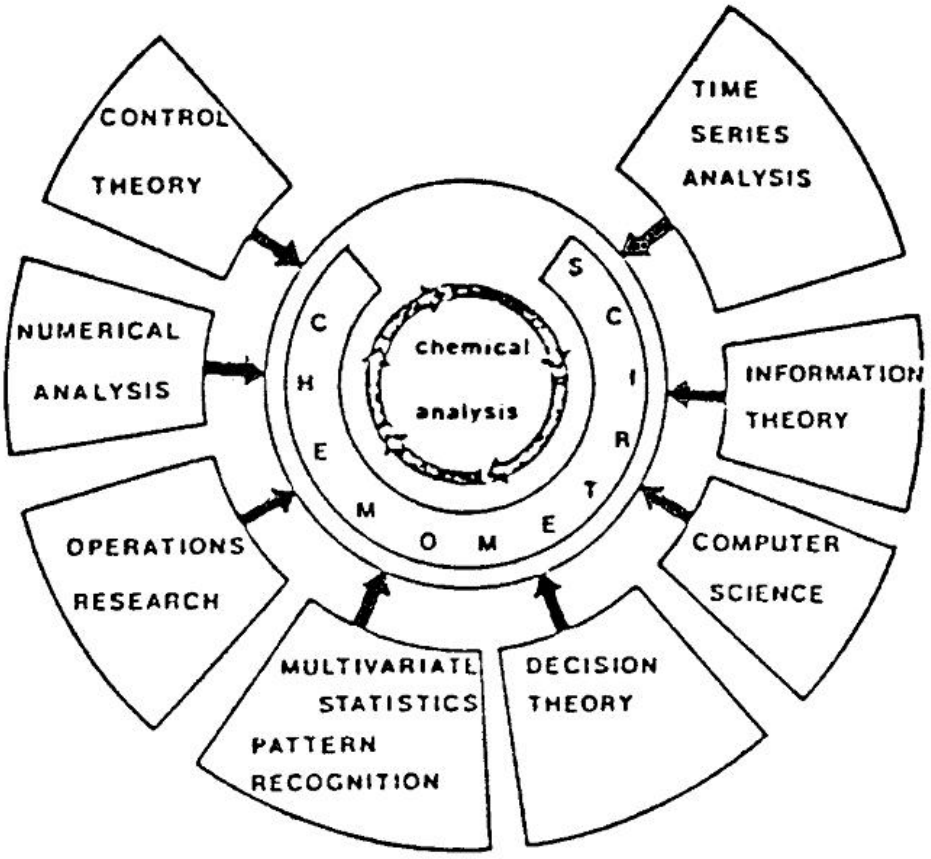


Figure 2.1: The relationship of chemometrics. Source: B. G. M. Vandeginste, Analytica Chimica Acta, 150 (1983) 199-206.

The processing steps could be simply interpreted as transformation of the raw data into more transparent format for the analysis, and includes modeling. The analysis itself is just the interpretation of the processed data sets, especially

comparison, clustering, decomposition and identification. The current state of individual steps of both data processing and data analysis are described in this section.

2.1 Data filtering

Liquid chromatography in tandem with mass spectrometry (LC-MS) produce terabytes of measurements daily around the world [45, 77, 78]. Systemic (instrumental and chemical) and random noise complicate the dataset. Correct interpretation of mass spectrometry (MS) is affected by presented noise across all kinds of MS techniques. The noise addition may produce fake peaks or hide small intensities in the measurements. Thus, LC-MS data are 'crowded' and have a uneven baseline [76]. It is a common subject in the chromatography produced by both mobile phase and column bleeding. This systemic noise causes extraneous peaks or rising baseline during gradient elution [46]. The interpretation of LC-MS is not trivial mainly because of the vast amount of noise especially in complex samples ([14]).

It is necessary to consider approaches for denoising and baseline subtraction [45]. Common algorithms based on thresholding or wavelet transformation (([44]),[76]) are not resistant to the losses of information from their principle. Thresholding methods, even in the adaptive form, still discard parts under threshold level(s) from the whole measurement. The wavelet transformations directly change the information content and are sensitive to the window length. Therefore, some information could not be used for for further analyzing process, including peak detection.

Omitting the presence of baseline (also called background, systemic noise or mobile phase) in Liquid chromatography - Mass spectrometry (LC-MS) impedes objective analysis. That contribution has to be removed from the signal response. Behavior of the baseline content is not constant in time axis. As is also often necessary for experiments with gradient changes. The results may be measured by increase of data mining output, both qualitatively and quantitatively.

The filtration is necessary processing step to emphasize features which are relevant for other steps, especially segmentation of the measurement into indi-

vidual eluted compounds. Sooner or later, the segmentation task leads to the feature detection problem. This features are related to the data peaks which represent compounds, ideally (see chapter 2.2). Strong peak candidates give the alignment additional flexibility [85, 87]. Robust peak detectors require advanced analysis like noise filtration, baseline subtraction, pattern recognition or curve fitting [87, 88, 84]. Noise additions are produced not only by random errors (random noise) but also by influence of baseline from the Liquid chromatography. Sum of the noise and the signal may produce false interpretation or hide the signal under reasonable level. Therefore, baseline in LC-MS negatively affects the measurement analysis and represents the systematic noise in nonlinear level on the time axis. However, the LC-MS data are complex and algorithms based only on filtering in chromatographic domain (Total Ion Current - TIC) or only in individual mass spectra can not have performance as good as algorithms which incorporate information from both domains together. Generally, any filtration which uses hard fixed threshold values are problematic. Its results are often inconsistent between runs, instrumentation and methods because the values from nearest threshold neighborhood may be easily misclassified.

The baseline contribution is related to the instrument and to the particular experiment. However, the baseline is always presented in any analyte measurement, even in the blank measurement. In the context of this thesis, the blank is considered as the chromatographic measurement without addition of the sample. So, it is usually just the mixture of solvents. Hence, the blank is easily obtained for every kind of experiment and it is often done without any further use. Thus, methods to baseline subtraction based on direct subtraction of the blank from the measurement are used. However, their results are not optimal because of random influences which add additional noise into the measurement. Moreover, the baseline characteristic in the blank is not chemically affected by analysed substances.

Chemical noise (e.g. sodium adducts) results from mobile phase impurities. It is more difficult to remove then random noise, because they have a pattern similar to the signal. Chemical noise can reduce mass accuracy by shifting peaks centroids. Denoising model for chemical noise was developed by Andreev et al. ([21]). Noise produced by random errors is caused by minor variation of the distribution surface. Systemic errors become more noticeable as they create

borders effects, that are systematically over or underestimated ([58]). The baseline removal and data smoothing are usually the basic preprocessing steps ([76]).

2.2 Feature detection

The comprehensive comparison of complex mixtures of similar compounds by HPLC-MS has been major issue in 1980s and 1990s ([69, 71, 72, 70]) and became again highly interesting with extension of so-called -omics approach from genomics to proteomics and metabolomics. There, LC-MS is one of the prime experimental tools.

There is a wide spread of features which are extracted from the LC-MS measurements. In time alignment tasks, there are markers to be aligned in warp techniques (2.3). The most investigated features are compounds or mass peaks. The peaks are signal rising in time above the baseline or noise signal with a priori unknown time, shape and characteristics.

Peak detectors are important processing steps in LC-MS. Their performance directly affects the subsequent process (alignment, identification). The major problem in detecting peaks of low amplitude is complexity of the signal and different noise sources ([76]). Peaks are assumed to have characteristic shapes and patterns determines by geometric construction. Unfortunately the most expected shape is Gaussian or its derivative (so-called Mexican hat). Even in the matching in the wavelet space. Common understanding of peak is defined by peak maximum and by the ratio of height and width of peak at half its maximum height (cite-GoldBook). However, the correct model of the peak shape has to consist of tail which expressed the left side of peak ([13]). The uncertainty in enrichment analysis is caused by the stochastic nature of the results obtained by high-throughput experimental techniques ([64]). The interesting smoothing algorithm based on m/z distances to the nearest neighbor was introduced by Stolt et al. ([14]);

Seemingly, it is correct and preferable approach to use of internal standards (IS), i.e. the addition of known substance(s) into the sample(s) ([73, 80, 81]). At the best, these samples should be isotopically labeled versions of the same compound. This approach may become extremely expensive, time and experimentally demanding. Often, the design of standards follows certain logic, i.e.

hydrophobicity index ([82]). There is no universal set of standards which would map the behavior of any solvent mixture on any column. As well, there is no idealized column which would separate compounds only according to one chemical parameter, often also idealized. Also dynamic parameters, rate of binding of a compound to the column and release from it and column capacity affect the retention time of all compounds which interact with the column at a time. From this point of view, some combination of standard compounds may even be misleading. In practice, IS are much less often applied than they would be needed. In some cases, they are not applicable due to lack of adequate standards on uniqueness of the sample.

Addition of known substance to the measured sample relieves to quality of measurement ([95]). However, the addition itself is not obviously easy, exact substances selection depends on the current measurement ([96]). It has to differ from analyte, which could be a priori unknown in study of chemical fingerprints of specific processes like metabolite profiles ([97]). Obtained data output still require computation to fit internal standards response from slightly different measurements together. This step can not be skipped and the addition helps only (but substantially) to locate the marker data points or statistical parameters ([80]) for the retention time alignment.

2.3 Alignment

The compounds of interest (analytes) are found as complex mixture in the sample an LC decrease the complexity by improving analyte separation. That produce the time element of the measurement, called retention time (RT). Separation process shows shifts and distortions in the RT when two or more measurements are compared. This fact makes the assignment of similar compounds difficult, since the mapping to each other is not known in advance. But it is crucial to correct for those warps. Otherwise, it is hard or even impossible to find the corresponding partners ([79]).

In many cases of complex samples, it is recognized as crucial, difficult and nontrivial task to compare two or more measurements obtained by LC-MS. Even the measurements of samples identical in content but differing in amounts

of applied quantity on the same chromatographic column with the same experiment settings are affected by nonlinear shifts in retention times. Therefore, the 'same' results do not fit together in the time axis. Comparison of samples requires transformation (normalization) function(s) to compare retention time values and other characteristics. Because of nonlinearity of the shift(s), also the normalization function has to be nonlinear.

Current philosophies for time normalization are divided into two major categories: Statistical models (MVA, DTW, Peak detection) and empirical rules based on internal standards. Actually, there is no restriction for the model to be based on internal standards (IS). Recently, there were developed methods for estimation of semi-optimal set of single or multiple IS, like NOMIS ([80]) or excellent idea of Linear solvation energy relationships (LSERs ([81])). The LSERs is based on selection of open windows in the chromatograms for prediction of IS candidates. This is time (and standards) saving approach which minimize the errors of samples and IS compounds mutual influence or competitions. However, both ways (NOMIS and LSERs) demands to think about it before the own measurements. Also a few of forgone experiments to choose the proper set of standards for given samples, column or method(s) are required. This increases the amount of necessary sample, experiment time and significantly the costs. Either of this, but typically mainly the experiment time, is often limiting.

On the other hand, the non-supervised models and derived algorithms are based on time warping approaches ([79]). It all started with the Dynamic time warping (DTW) in speech recognition tasks. The main idea is on partial shrinking and stretching of the time axis. Naturally, reference set or piecewise transformation differ in several warping techniques. Namely, the parameters for the transformation function are in Linear time warping, Fast dynamic time warping, Parametric Time Warping (PTW) and Correlation Optimized Warping (COW) determined by maximizing or minimizing the sum of coefficients between data segments in pairs of samples ([79, 83, 84, 85, 86]). Time warping algorithms separate the time dimension into segments but preserve the temporal order.

The segmentation task leads to the peak detection problem. All methods using peak detection for time alignment are error propagating. Any error from the peak detection process is propagated into the further processing, obscure initial

errors may emphasize errors in the output ([123]). This is called a power-law dependence. Let assume two peak features X and Y, where both values are uncertain.

- The square of the uncertainty in the sum or difference of two features is evaluated as sum of the squares of individual absolute errors (a priori unknown). Therefore, the uncertainty of the result (e.g. full-width at half-maximum) is larger than either individual uncertainty of the features X and Y.
- The relative uncertainty as a result of multiplication or division is evaluated as the square root of the sum of the squares of the relative uncertainties in the individual features.
- When an uncertain value is multiplied by a constant, the absolute uncertainty on the result is the constant times the uncertainty of the original feature.
- The relative uncertainty on value raised to an exponent is the exponent times the relative uncertainty of the original feature.
- Complicated functions applied on the original features demand application of the rules given above for each mathematical operation (so-called Chain rule).

While the peak dependent methods are effective for simple samples, they could be insufficient for more complex biological analytes ([87]), for the given reason.

The time alignment strictly depends on correct peak definition and detection. Incorrect peak definition and detection brings dangerous presumptions into account, if incorrect. For example, in XCMS ([124]) toolbox for R is used information from blank signals for time alignment. The ability of XCMS time alignment depends on initial matching of peaks into reasonable groups ([124]). XCMS approach of filtration also changes the shape of the peak according to the idealized model. Another example, a pre-processing tool for PARAFAC modeling ([125],[53]) slightly extend the COW algorithm by correct idea of using covariance instead of correlation. However, the PARAFAC modeling (([57], [56])) is more proper analytical tool for LC-MS/MS. It uses regression analysis, while MS-Resolver ([54]) uses pattern recognition and Mass Works ([55]) uses calibration

(internal) standards. The piecewise alignment similar to the COW was introduced by Pierce [126] with over-combined feature selection. However, warps might be easily confused by single metabolite, as it will be shown later. Exhausting overview of both, commercial and freely available softwares for metabolomic data processing as well as time alignment was done by [127]. Some level of peak detection or binning is assumed in most of the available products. For completeness sake, exhausting survey of possible alignment approaches was done by ([92], [93] and [127]).

An attempt to two dimensional semi-parametric warping as alignment of retention time as well as mass spectra was modeled by de Boer ([94]). Here, the series of bicubic splines poses a linear constraint on the warping coefficients limiting their variability.

2.4 Comparison

Despite the best efforts, there still remains the problem of comparison of metabolite profiles obtained, for example, under different instrumental setups, or simply enough, at different amount of sample loaded on the column and arriving simultaneously to the ion source. From that point of view, the data analysis standardisation ([3]) may be incomplete.

Nowadays extremely popular approach is principal Multivariate data analysis (MVA), especially its Principal component analysis (PCA) ([90, 91, 88, 12, 60, 61]). It is a method of classification based on correlation and linear combination. It finds a new coordinate system from the original variables. PCA advantages are mainly the reduction of dimensionality of the data sets and better visualization of major trends in the data. It has to be realized, that two principal components would be comparable only if they represent exactly the same linear combination. That is hardly fulfilled in completely different mass spectra (with possible exception only for noise contribution). However, PCA is powerful mathematical tool when it is used with wisdom. Linear scaling, normalization and transformations failed to recombine the groups properly in equivalent problem ([60], [62]).

PCA is not a classification technique, it is an unsupervised clusterring or data reduction. It determines an optimal linear transformation for a collection of

data to display the correlation or sample patterns or groups. PCA can help quantify information content. However it is sensitive to experimental noise. There are many other statistical techniques like partial least squares discriminant analysis, k-mean clusterring or soft independent modeling of class analogy. Models to correctly assign chemical compounds also uses heuristic topological or geometric subgraphs for classifying ([59]). large number of features increasely classify time and determination can be computationally expensive. Especially in the drug discovery ([59]).

Chemometrics softwares like ParLes ([44]) for multivariate modeling and prediction were developed to perform transformations a data pretreatments. Cross validation as well as bootstrap aggregation provides assessment statistics (PCA, partial least squares regression - PLSR). Methods as multiplicative signal correction (MSC), standard normal variete (SNV) transform or discrete wavelet transform might be used as preprocessing techniques to improve the robustness of the PCA and PLSR models ([44]).

Identification of mass spectra becomes important in omics science, Most tools employ only linear scans or databases. Shared peak count (SPC) does not account small peak shifts and calibration errors. The improvement was done by coarse filtering-fine ranking by reducing the matrix space dimensionality ([18]). Other metrics were used and will be explained later in this thesis (tab. 4.4).

2.5 Omics

Biologically relevant formalism is necessary for common understanding of biological function. Systems in biology are complicated and have many a priori unknown attributes to warrant ordinary differential equations (ODE) modeling (photosynthesis models, metabolomic pathways enumeration, etc.). The resulting sets of ODE are too complex to be analytically solvable. As an alternative, dynamic behavior can be approached by stochastic methods. Probabilistic models can be particularly useful ([1]) as they constitute theory that generate hypothesis. ([19])

Mass spectra represent valuable information used in biology and experimental medicine. Many fields deal with the problem of statistic evaluation of mass spectra in biology research ([13]). Nowadays, most biochemists and bioin-

formaticians are familiar with newly emerging 'omics' (proteomics, metabolomics, lipidomics, etc.) fields research (analyzing the interactions of biological information objects). Metabolomics is the study of the study of their small-molecule metabolite profiles in a biological cell, tissue, organ or organism, which are the end products of cellular processes. Proteomics is the large-scale study of proteins, particularly their structures and functions. Lipidomics may be defined as the large-scale study of pathways and networks of cellular lipids in biological systems.

Namely, when LC-MS is used in complex lipid mixtures in proteomics, experiments are poorly reproducible because of low data quality, systemic bias or stochastic sampling effects. 'The true roots of the phenomenon are presently incompletely understood' ([48],[49]). Generic problems are also in databases for correct characterization of proteins. The search engines cannot distinguish among different identifiers from the way of database construction. Algorithms to calculate molecular weight are variable ([49]).

Metabolomics identifies small molecules that participate in the metabolomic activity of the biological system ([20], [60]). Metabolites and metabolic pathways outlines the needs in databases, data standards and modeling. ([11],[63]). Metabolomics has two basic approaches. Metabolite profiling ([4]) aims on compounds identification and quantization. This is dependent on existence of databases of known compounds. In chemometrics are compounds not necessarily identified, only their features. Chemometrics holistic approach has strength in absence selective ignoration or selective inclusion of data in diagnosis. Quantitative metabolomics has to deal with biases that distort relationship between the original measured metabolite concentration ant measured peak area ([50]). Lot of of effort in metabolomics is invested in the annotation of unknown peaks ([51], [52]). 'Statistical analysis might be biased due to dependency between peaks currently considered as independent in the metabolomic profiles dataset' [50]. Many users are strictly focused only in compounds which are searchable in the databases.

In recent years, movement toward standardization of biological experiments description becomes definite and inevitable step in current data boom caused mainly by general accessibility of omics technologies ([2],[63]). However, in scientific papers significant deviations from standard experimentation are still to be expected as well as division in approach between applied omics and omics

for basic science. In the standardization of data representation themselves, one approaches substantial problems coming (a) from the fact that data are significantly instrument-dependent both in the actual technical setup of the instrument and on the technical setup of the chemical and physico-chemical experiment preceding the chromatographic analysis and (b) from data handling approaches which are not standardized either in mathematical principles or in actual provision. In this thesis is emphasized, and on the case of a metabolomic experiment demonstrated, that understanding of biological instrumental data as probabilistic problem, may lead to principally different approach to data transformation which is not conform with standard terms used in data transformation ([3]).

Real performance of a metabolomic experiment using LC-MS never meets the expectations of generality included in the term of metabolite profiling. Various steps in the metabolic profile measurements such as state of the sample, chemical procedures included in metabolite extraction and the setup of the LC-MS experiment are among the major factors that influence experimental results and reliability. Great effort was made by the metabolomics standards initiative (MSI) ([9]) to standardize the definition of the experiment including its chemical part ([10]). These efforts shall definitely lead to better comparable datasets from the biological and chemical point of view. This contributes to dissection of error and noise originating in biological or chemical part of the experiment and that originating from the physico-chemical events in the analytical instrument.

Altogether, there are plenty of softwares, packages, toolboxes, applications, programs and recommendations for data processing and/or analysis of LC-MS data sets, as was shown in previous subsections. They vary in the used data source, data format as well as used methods, techniques or ideas in processing steps. Some of them are combined with the specific database or focused only on narrow part of the investigated molecules. The main producers of the measuring device (LC,MS) or its parts are also producers (or owners) of control and processing softwares. Short review is attached in the Appendix F. Exhausting overview of both, commercial and freely available softwares for metabolomic data processing was done by Katajama [127]. Summary of the metabolomic databases was done by Wishart ([11]). Browser to explore multidimensional information in omics data servers was introduced by Toyoda et al ([47]).

2.6 File formats and arithmetics issues

Usually every kind of data storage require data format according to some data logic. In principal, there are two main approaches, text file format and binary file format. Text file format contains only writable characters from the standard ascii table (character-encoding scheme based on the ordering of the English alphabet, ISO/IEC 646, ISO/IEC 8859) and control characters (end of line, end of file, etc.). Ascii files are easy to write, easy to read, however the usage of 8-bits to store each character leads to the large files.

Binary file format contains any type of data, encoded in binary form. For example, in ascii file format is each digit of a number encoded by one byte (8 bits) and in binary file format is one byte enable to encode any integer number from the range 0-255. There is not only one type of binary encoding ([41]).

As example of ASCII dataset, I discuss DataAnalysis for LC/MSD Trap from Bruker Daltonic GmbH allows to export measured data in text file format (ASCII) with intuitive structure of stored data. Each mass spectrum is written on one line, ended with EOL symbol. As separation marks are used comma and space. First number is the value of retention time in minutes. Then, there is type of ionization (sign + for positive and sign − for negative.) Source of ionization follows after second comma (for example ESI as Electrospray ionization). Next information describes the level fragmentation (msn, where n is natural number, excluding zero). Also the range of detectable mass to charge ratios is stored. Number after seventh comma represents amount of pairs mass and intensity. Then follows the designated amount of pairs, separated by comma. Mass to charge ratio and intensity value are separated by space.

Reading of ascii is trivial, however requires to reshape the data after import. Set M of mass values could be determined after union of all mass values in all mass spectra. Also whole matrix of Cartesian product $M \times T$ is determined when reading is finished.

Xcalibur from Thermo Scientific works with own raw format ([36]) and allows to convert stored data into text (ASCII) file with defined structure. It is started with $RunHeaderInfo$ with information of measuring process, settings and measurement device, amount of time scans, time duration, low and high detected

mass value and maximal integer intensity. $RunHeaderInfo$ is followed by individual time scans, each with its own ScanHeader. This file format is redundant, many information are repeated almost every line. Exactly the same data in low-res m/z could be stored in Bruker Daltonic ASCII with size of 3 MB or in Thermo Scientific ASCII with size of 70 MB.

ASCII based attempt for file format standardization was done by JCAMP (Joint Committee on Atomic and Molecular Physical Data). Their DX-JCAMP Spectroscopic Data Exchange Format is a Standard format for the exchange of ion mobility spectrometry data ([35]). Further development/refinement of standards is now under the auspices of IUPAC. However, the support of this format from the third party in the future is unclear.

Situation in binary file formats is quite a bit wilder. Each software and hardware producer prefers own file format. Access to the Thermo raw format is defined in the headings library shared to the customers which bought Thermo Xcalibur software package ([36]). Nowadays, MzXml data format ([39]) becomes as a standard, based on eXtensible markup language (XML), especially in the proteomics. MzXml is extremably redundant and huge (Gigabytes for the same information stored in few MB in Bruker Daltonic ASCII). It also takes long to load and save files. 'Xml is like political speeches: it requires a pointless prologue, establishes no dialog; by attempting to represent everything it ends up representing nothing; it rambles on about namespace and territory, uses a completely different language to define it; the precious little information it contains is buried in verbose noise... ' ([40]).

Some of the file formats should be converted between each other by applications like Waters Inspector([37]) or Vx_capture([38]). However not all file formats specification are available to the customers or even free. The management, storage and standardization is absolutely critical ([11]). Standardization lies in making data more uniformly and exchanged. The self-describing portable machine independent file protocol NetCDF was developed by the Unidata Program Center in Boulder, Colorado. However, MzXML is more preferred, for an unknown objective reason.

2.6.1 Overflow

LC-MS measurements stored in different (ASCII or binary) file formats are processed on binary computers with potentially dangerous arithmetics issues. It is often possible, that result of arithmetic operations (summing, difference, multiplication, division) cross the lower (underflow) or upper (overflow) bound of representable numbers. For example, multiplicity of very big positive numbers will result as negative number. Therefore, it is necessary to keep in mind, that real numbers are not encoded with zero error, but rounded to the closest quantized level. It is caused by inaccuracy during the transformation from decimal to binary system. Exempli gratia, decimal 0.1 is irrational number in binary with period 1100 ([41]).

Overflow may be also caused by improper order of arithmetic operations, because associative and distributive law is not valid in digital computations. Rounding to the closest quantized level causes completely different results during combination of multiplicity and division in improper order. Moreover, zero is dangerous number in computation. Expected zero results are not always equals to the zero value, again because of rounding and possible overflow. That demands carefully algorithmised operations, even if the analytical equation is correct. Underestimating of this issue may lead to the errors of computation, which makes result nonsensical.

3 LC-MS according to the system theory

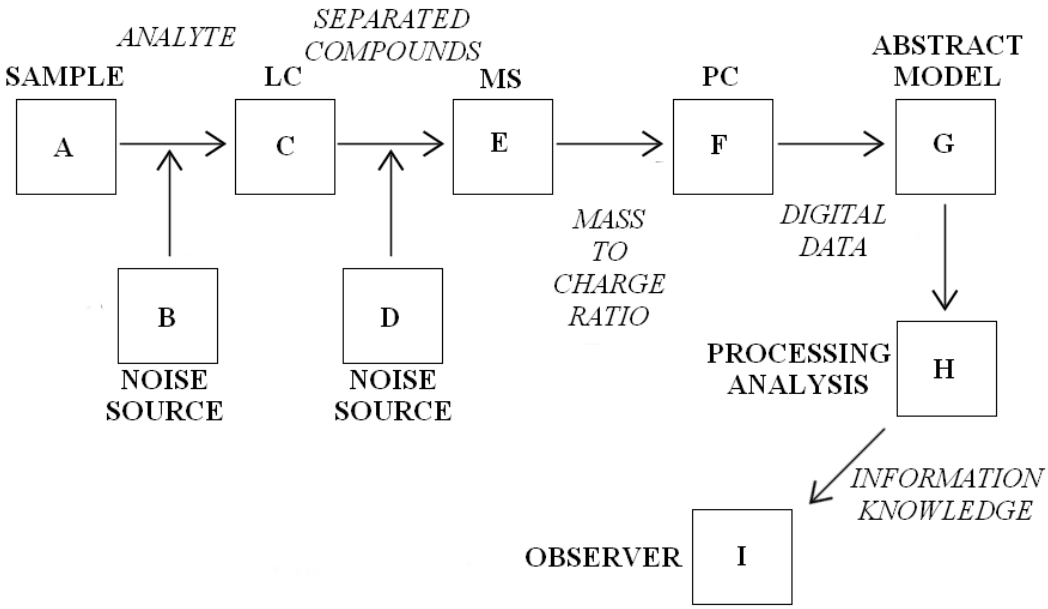


Figure 3.1: Information channel in LC/MS.

3.1 Liquid chromatography

The International Union of Pure and Applied Chemistry ([95]) defines chromatography as:
'A physical method of separation in which the components to be separated are distributed between two phases, one of which is stationary (stationary phase) while the other (the mobile phase) moves in a definite direction.'

Liquid chromatography is therefore described as:
'A separation technique in which the mobile phase is a liquid. Liquid chromatography can be carried out either in a column or on a plane. Present-day liquid chromatography generally utilizing very small particles and a relatively high inlet pressure is often characterized by the term high-performance (or high-pressure) liquid chromatography, and the acronym HPLC.'

A mobile phase in chromatography is described as:
'A fluid which percolates through or along the stationary bed, in a definite direction. It may be a liquid (liquid chromatography) or a gas (gas chromatography) or a supercritical fluid (supercritical-fluid chromatography). In gas chromatography the expression carrier gas may be used for the mobile phase. In elution chromatography the expression "eluent" is also used for the mobile phase.'

Finally, stationary phase in chromatography is described as:
'One of the two phases forming a chromatographic system. It may be a solid, a gel or a liquid. If a liquid, it may be distributed on a solid. This solid may or may not contribute to the separation process. The liquid may also be chemically bonded to the solid (bonded phase) or immobilized onto it (immobilized phase). The expression chromatographic bed or sorbent may be used as a general term to denote any of the different forms in which the stationary phase is used. Particularly in gas chromatography where the stationary phase is most often a liquid, the term liquid phase is used for it as compared to the gas phase, i.e. the mobile phase. However, particularly in the early development of liquid chromatography, the term 'liquid phase' had also been used to characterize the mobile phase as

compared to the 'solid phase' i.e. the stationary phase. Due to this ambiguity, the use of the term 'liquid phase' is discouraged. If the physical state of the stationary phase is to be expressed, the use of the adjective forms such as liquid stationary phase and solid stationary phase, bonded phase or immobilized phase is proposed.'

The comprehensive description of the liquid chromatography in varying amounts of detail was done by Meyer ([65]), Robards et al. ([66]), Lindsay ([67]), Ardrey ([73]) or McMaster ([68]). However, two aspects of the liquid chromatography are relevant enough to this thesis:

- LC-MS has not been guaranteed to provide the required analytical information.
- The main limitation is in inability to provide identification with any degree of certainty. ([73])

3.2 Mass spectrometry

Mass spectrometry (MS) identifies a simple molecule by its molecular ion mass. MS shows the mass of the molecule and the masses of pieces from it. Mass spectrum is the bar graph, where abscissa indicates the mass to charge ratio (m/z) and the ordinate indicates the intensity (relative or absolute). They are collected in sequence as the ratio increases, the ion current is amplified and it is then displayed by some means. With instruments of low resolution, peaks appear at unit mass numbers, but at high resolution the masses of individual ions can be measured with sufficient accuracy for the molecular formula of each to be determined. Molecules do not fragment in an arbitrary manner but tend to split at weaker bonds, such as those adjacent to specific functional groups. ([25, 15])

The analytical mass spectrometry was introduced in 1941. But mass spectrometry is in fact a much older technique. The basic principle to the separation of atomic masses have been demonstrated at the end of the 19th century.

The molecules need to be ionized before they reach the detector. There are several types of ionization (Electrospray ionization -ESI, Matrix assisted laser desorption/ionization - MALDI, chemical ionization, - CI, atmospheric pressure

photoionization - APCI, etc.) as well as devices (ion trap, quadrupole, Time of flight - TOF, Fourier transform/ion cyclotron resonance, Orbitrap) with different sensitivity ([36, 37, 25, 26, 95]).

The ions are separated according to their individual m/z values in a vacuum. If an ion collided with another one, its direction of travel could be changed and it may never reach the detector ([26]). The interpretation of the mass spectrum is based on the chemistry. Usually there is observed the molecular ion (single and/or multiply charged), its isotope(s), adducts and fragments. In complex spectra is hard to select the molecular ion explicitly. It does not have to be the most intensive ion. Application of chemical rules is also not unique and failed to be algorithmised so far. Usually the strategy for identifying an unknown compound is to compare its mass spectrum against a library of mass spectra. Mass spectrometry data analysis is a complicated subject that is very specific to the type of experiment producing the data. Results can also depend heavily on how the sample was prepared . Mass spectrometry cannot give evidence as to the stereochemistry or configuration of functional groups ([15]).

Limits of detection (LOD) in MS has direct relationship to the reproducibility of measurements. 'A common misconception is that the LOD is the smallest concentration taht can be measured. instead, it is the concentration at which we can decide whether an element is present or not - that is, the point where we can just distinguish a signal from the background' ([22]). Practical estimation requires the measurement of the background fluctuation ([23]).

3.3 System theory

System theory represents certain level of scientific though. During the history were gradually formed two basic approaches, analytical and synthetical. The analytical approach decomposes phenomenons into partial components mutually independent. Properties of the components are then used for deduction of the properties of original phenomenon using logical rules. The analytical approach started with Descartes ([27]) and resulted into Newton's laws of physics ([29]) as well as Linné's classification of biological species ([30]). However, Darwin's theory of natural selection ([28]) was based on the synthetical approach and on

the assumption of stochasticity. In the synthetical approach, the complicated phenomenon is more then just the sum of its components. The result of synthetical research is a model of appropriate (and full) set of events ([32]). Therefore, the synthetical approach refuses universality of the nature laws, it only assumes properties under certain conditions. The problem is, that both approaches had not gained corresponding position in the science:

- the system has not unique structure
- the system output may contain more information than the system state
- the system is not generally defined for complete set of input functions
- the system is usually described by equations which do not distinguish oriented and non-oriented causation
- the connection of two isolated theories may not be consistent theory

([33]).

'Inability to discern oriented and non-oriented dependencies between system variables does not cause essential inconsistences in each solved problem but usually in some limit cases only. In such situations a deep gap between theoretical and experimental conclusions is critical and results in an incredibility to the system theory as a whole...' ([32]).

Paradigms of the new system theory were introduced by prof. Ing. Pavel Žampa, CSc. in 1996 ([32, 33, 34]). He wanted to seek for a definition of an abstract system which would be sufficiently general as to cover any real problem and, at the same time, sufficiently specialized, as to enable to find an adequate physical realization to any theoretically given abstract system.

Thus, it is assumed that the real system is such a part of the real world for which holds that its properties are not affected by the other parts of the world. The processes of the system are governed by principle and law of causality. The system behavior is, in its nature, stochastic containing deterministic behavior as its special case. The system variables have to be primarily defined on finite sets and can be extended to infinite sets only when a continuity procedure is adopted.

There are such orientated connections among parts of the system which can be canceled or restored by a suitable external intervention ([32]).

On this base a new definition leads to the system structure, where:

- Inputs and outputs are elements of the system, even if they are not actually used.
- The variables which are neither the input nor the output variables are the inner variables of the system.
- The inner variables are the only system variables which are not accessible to measurement.
- The properties of any part of the system are conditional and are generally lost when the part of the system is transferred to another environment.

It offers formulation of a consistent system theory which can solve formerly non-treatable problems. The system is a unity unaffected by the outside of the system. From the principal reasons all the system variables were defined on finite sets only. However, under carefully chosen continuity hypothesis an extension to infinite sets is straightforward and brings good description of real phenomena ([33]).

3.4 Abstract model

Typical measurement output data from HPLC-MC is discrete set of points in discrete three dimensional space which is defined by discrete axis: retention time, molecular mass to charge ratio (mass) and intensity. Analytes elute at every time point from HPLC column, obviously because of delay caused by some physico-chemical interaction, and enter the MS ionization chamber. The delay is often related to a chemical property. Intensity for each detectable mass is measured inside the MS and this value represents amount of ionized molecules of individual mass in exact time point.

The time and mass are the attributes of the system. Let mark those attributes by sign a_k. Whole set of system attributes could be described as

$$A = \{a_k \mid k = 0, 1\}, \tag{3.1}$$

where a_k are names of corresponding attribute. Each attribute a_k takes on some value. In abstract systems the value of k^{-th} attribute $a_k \in A$ is represented by k^{-th} variable $v_k \in V_k, k = 0, 1$, where V_k is domain of definition of k^{-th} variable, it is set of all values that variable v_k can reach.

Attribute A_0 represent reference attribute, supposed to be time. For its variable v_0 can be used common sign $t \in T, T = t_0, t_1, t_2, \ldots, t_e$, where is true that $t_0 \prec t_1 \prec t_2 \prec \ldots \prec t_e$, and e is natural number. On the set T could (but do not have to) be defined a difference which represents the time period from time point t_i to time point t_j.

Value of the 2^{-nd} attribute a_1 is represent by variable $v_1 \in V_1$ and means mass to charge ratio. For its variable v_1 can be used common sign $m \in M, M = m_0, m_1, m_2, \ldots, m_n$, instead of *[m/z]* [1] (just for equations lucidity), where is true that $m_0 \prec m_1 \prec m_2 \prec \ldots \prec m_n$, and n is natural number. Now, are obtained two sets which describe the values of two axis: time and mass. Every individual measurement run generates intensity values for all possible pair time $t \in T$ and mass $m \in M$. Therefore, this generation process can be symbolically described as mapping

$$y : T \times M \rightarrow \cup_{t \in T, m \in M} \mid y(t, m) \in I, I = 0, 1, \ldots, i_{max}, \tag{3.2}$$

where I is set of natural numbers with zero and the value of mapping $y(t, m) \in I$ means intensity of mass $m \in M$ in time $t \in T$. Exact value of i_{max} is delimited by saturation level of MS detector and is called Mass limit. Thus, the abstract system ([32]) is then defined in domain by ordered pair (T, M) with image in I. Let mark the system by sign Ω. Notice, that Žampa's system is defined as a Cartesian product itself and in this sense it can be considered as more general than abstract system by Mesarović ([31]).

[1] In biochemistry, it is often used the dalton (Da) as a symbol for a mass unit which represents $1/12th$ the mass of C12 (the most abundant isotope of carbon). However, it has not been approved by the International System of Units. Therefore, dalton is not used in this thesis.

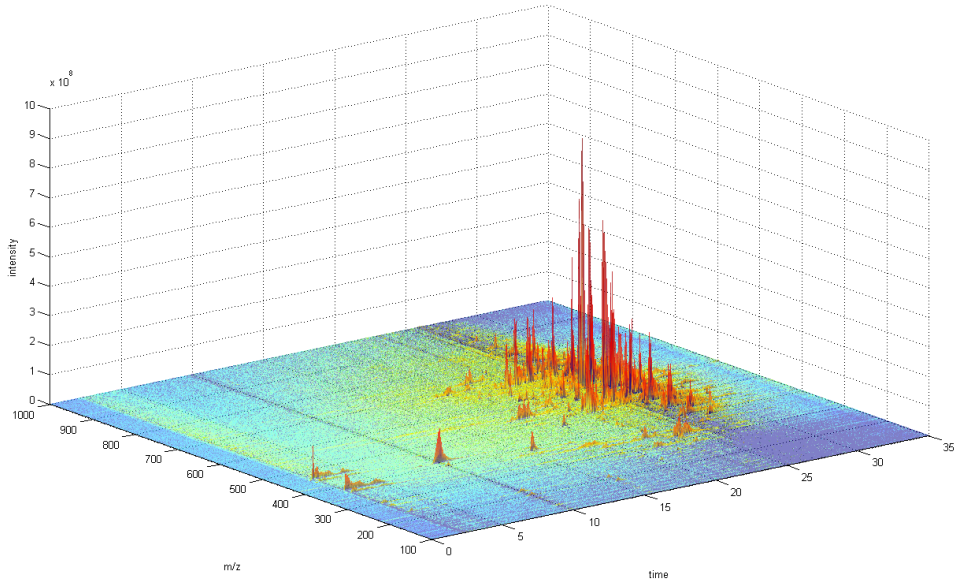


Figure 3.2: Example of LC-MS measurement, extract from alga Stigeoclonium in 70% MeOH.

Hence, could be also defined the state of the system Ω in time $t \in T$ as the exact mapping for each value $m \in M$ for given time $t \in T$:

$$y(t) = [y(t, m_0), y(t, m_1), y(t, m_2), \ldots, y(t, m_n)], \tag{3.3}$$

this n-tuple is usually called mass spectrum in time $t \in T$ and

$$\gamma(t) = ||y(t)|| = \sum_{k=0}^{n} y(t, m_k), \tag{3.4}$$

is value of total ion chromatogram (TIC) in $t \in T$. The symbol $||y(t)||$ means a metric of $y(t)$, which is the sum in TIC case.

3.4.1 Mass resolution

MS determines the attribute molecular mass of the molecule by measure of the mass to charge ratio (m/z) of its ion, marked by sign $m_\xi \mid \xi = 0, 1, 2, \ldots, n$ (i.e. $m \in M$) in the system Ω. In the mass analyzer are the ions resolved (separated) according to the m_ξ and they are counted into signal $y(t, m_\xi) \in I$ by ion detector. Therefore, in the single spectrum $y(t)$ (*eq.* 3.3) is obtained one intensity value $y_\xi(t) = y(t, m_\xi) \mid \xi = 0, 1, 2, \ldots, n$ for each m_ξ by mapping process (*eq.* 3.2) in any time point $t \in T$. But, the mapping (*eq.* 3.2) is dependent on the ions separating power (mass resolution R) and the measure of the m_ξ (mass accuracy) ([24]).

Simply, R of the MS analyzer is the ability of the device to distinguish between m_a and m_b in mass spectrum $y(t)$. Let mark the difference between two masses that can be separated as Δm:

$$\Delta m_b = \{m_b - m_a\} \mid m_0 \preceq m_a \prec m_b \preceq m_n \; \forall \{m_a, m_b \in M\}\,., \tag{3.5}$$

That definition has to be extended by condition:

$$if \; \exists m_c : m_a \prec m_c \prec m_b \Longrightarrow m_c \ni M \wedge \neg(a \prec c \prec b), \; \forall a, b, c < e \; natural, \tag{3.6}$$

to distinguish two closely adjacent mass values in the first instance (ability to distinguish any mass values from the measurable range does not require the extended condition). Then, according to the IUPAC definition ([95]):

$$R = m_\xi / \Delta m_\xi \mid m \in M, \; \Delta m_0 \equiv \Delta m_1, \tag{3.7}$$

In accordance to common meaning of the term resolution, R gives the count of measurable m_ξ in the range (m_0, m_e), the potency $\wp$ of set M. Its reverse $1/R * \eta$ represents resolution in some cases and denote relative proportion parts per η, typically $\eta = 10^6$ [ppm], a quantity with the dimensions of 1 (as R itself).

Let remind, that from (*eq.* 3.7) Δm is a function of m_ξ:

$$\Delta m_\xi = m_\xi / R. \tag{3.8}$$

Therefore, the set M of system Ω is determined (necessarily but not sufficiently)

by knowledge of MS resolution R, guaranteed by MS producer, and subsequently by function (3.8). Thus, the set M will fulfill the condition (3.6). As is clear from (*eq.* 3.8) the difference $\Delta m_b > \Delta m_a$, *for* $m_a \ll m_b$ for constant R across the range (m_0, m_e).

3.4.2 Mass accuracy and mass precission

Mass accuracy basically means the ability of MS to measure exact mass, it is the fidelity to give the best qualitative response to the true value of measured molecular mass. In terms of abstract system it transforms the actually measured m/z into the output data value m_ξ. The measured a priori unknown mass τ is affected by the resolution R:

$$\tau \in (m_\xi \pm \Delta m/2). \tag{3.9}$$

As was shown (in *eq.* 3.8) Δm is a function of m_ξ, so the mass accuracy also differs in $m \in M$ by m_ξ. It is also obvious, mass can be measured as accurately as the MS allows and mass accuracy (*eq.* 3.9) for output data makes sense only in case of calibrated measurement device. Modification of the (*eq.* 3.9) to include calibration inaccuracy is modest:

$$\tau_\xi \in (m_\xi - \hat{\varepsilon}_\xi \pm \Delta m/2) \mid \hat{\varepsilon}_\xi \to 0, \tag{3.10}$$

where $\hat{\varepsilon}_\xi$ is an average (in repetition, function marked by sign μ) shift between measured m/z (let mark it by sign $m_{\xi\prime}$) and true value τ_ξ:

$$\hat{\varepsilon}_\xi = \mu(m_{\xi\prime} - \tau_\xi), \tag{3.11}$$

and the calibration process minimize $\hat{\varepsilon}_\xi$, using known τ_ξ as MS input. With $\hat{E}$ (full set of ε_ξ, where $\xi = 0, 1, 2, \ldots, e$) of the potency E equal to the potency of M should be done the calibration process on the measurement output data:

$$m_\xi^{(eq.\ 3.9)} = m_\xi^{(eq.\ 3.10)} - \hat{\varepsilon}_\xi \mid \wp E = \wp M, \tag{3.12}$$

instead (but preferably) on MS device. Or some fitting transformation has to be estimated for $\wp E < \wp M$ (with injection from E to M known from the calibration process). Then only, ($eq.$ 3.9) is relevant for calibrated m_ξ.

Unfortunately, obtained value from the MS $m_{\xi\prime}$ (even after the calibration done) is still not equal to the 2^{-nd} system attribute value m_ξ, it is as close as possible because of the standard error of the reported value (imprecision):

$$||m_\xi - m_{\xi\prime}|| < \Delta m/2 \mid (eq.\ 3.9), . \tag{3.13}$$

On the interval $(m_\xi \pm \Delta m/2)$ it asserts how many decimal places are utile:

$$m_\xi = round(m_{\xi\prime}) - m_{\xi\prime}|| < \Delta m/2 \mid (eq.\ 3.9), . \tag{3.14}$$

Subsequently, MS device parameters defined by set of Δm_ξ and set of ε_ξ should be considered as extended attributes. However for the purpose of model of abstract system creation are relevant only for determination of the set M.

Given system is an abstract model of real observed object which does not always give the same output for given input. Origins of probabilistic behavior are multiple such as stochastic inputs, randomness in time delays, noise from different sources and of different characteristics etc. Biological systems, in these cases inputs, behave as stochastic systems. Different chemical and physico-chemical manipulation with the sample are other sources of stochasticity. Finally, the separation and detection system brings noise of different, generally uneven, characteristic. This approach is focus primarily on the influence of the noise produced by the instrument (LC-MS device) on analysis of measurement results. Such analysis may affect the biological interpretation of the data.

4 Blank based time alignment

Blank based time-alignment (BBTA) was developed as a strong analytical approach for treatment of non-linear shift in time occurring in HPLC-MS data. Need of such tool in recent large dataset produced by analytical instrumentation and amplified by requirements of so-called omics studies is caused mainly by the extensive number of datasets from widely variant samples. This overloads the operators capacity to handle optimal control samples and internal standards. The only common, and naturally accessible, overall standard is the blank sample. The approach is based on measurement and comparison of blank and analyzed sample evident features. In the first step of BBTA procedure, the number of compounds is reduced by max-to-mean ratio thresholding, which extensively reduce the computational time. Simple thresholding is followed by selection of time markers defined from blank inflex points which are then used for the transformation function, polynomial of second degree, in the example. BBTA approach is compared on real HPLC-MS measurement with Correlation Optimized Warping (COW) method. It was proved to have distinctively shorter computational time as well as lower level of mathematical presumptions. The BBTA is computationally much easier, quicker (more then 1000×) and accurate in comparison with warping. Moreover, markers selection works efficiently without any peak detection. It is sufficient to analyze only baseline contribution in the analyte measurement with sparse knowledge of blank behavior. Finally, BBTA does not required usage of extra internal standards and due to its simplicity it has a potential to be widespread tool in HPLC-MS data treatment. BBTA focus on measurements time alignment for comparison of multiple compounds in similar samples. For that, it is used the markers from selected spectra and the retention time values.

It is described in details, mathematically and experimentally justify

approach for time alignment of LC-MS spectra using blank measurement data as (inherent) internal standards (BBTA). BBTA utilizes solvent contaminants and other important events (inflex points) detectable both in blank run and the compared experiment for alignment of multiple 2D chromatograms. Addition of internal standards may increase number of data points available for calculation but is not necessary for general laboratory practice. Obvious advantage of BBTA is its readiness and essentially low expenditure level of its application. All mathematical descriptions are derived immediately from the system based description of the measurement data sets with respect to the common used definitions.

Naturally, the liquid phase interaction during the analyte measurement are sample dependent. Therefore, issues of those interactions are not necessarily represented in the blank. However, the processing is based on the opposite point of view. The compounds, presented in the blank are also still presented in the analyte measurement. The basis for this are trivial. Semi-similar samples (like in metabolomics) or concentration curves require sequences of analysis with the same settings, especially baseline contribution. Therefore, pertinent features pinpointed from the blank remain in the analyte measurements. They are, usually hidden in the noise contribution or peaks behavior in Total Ion Chromatogram (TIC), which is just the summary projection in one axis and therefore mathematically loss operation. However, in 3D data matrix space are still observable and detectable. Concisely, what is in the blank have to be also in the analyte measurement when the same liquid phase is used, out of the question. There should be also some shift of the retention time for certain elution according to the temperature. Small changes affect only the distance of the shifts, not the ordering and it is strictly recommended to keep the conditions constant for repetitive experiments. Therefore, temperature changes in comparable measurements are also similar from the principality (and occurred in corresponding parts of the measurement). Theoretically, ordering transpositions in retention time will be caused by large temperature changes between the samples. Thus, the presumption of samples similarity is not fulfilled. Therefore, it can be simply assumed that the temperature attribute is not important for the time alignment.

When corresponding retention time values are available, there may be compared peak positions by so-called Dynamic Time Warping (DTW). This

is a class of signal processing method to measure similarity and find optimal match between time axes. Warps produce highly reliable output across the different measurements. Namely, when the dataset is dominated by highly similar compounds (i. e. standards). However, the algorithms have heavy computational burden. DTW is based on re-calculation of main part of the original dataset. Crucial aspects of warps are discussed in detail in section 2.3.

In empty (or blank) run some relevant (inflex and marker) data points may be identified (not necessarily the peaks). Blank in the context of this thesis is the chromatographic measurement without addition of the sample. So, it is usually just the mixture of solvents, sometimes called baseline, mobile phase or systemic noise. Hence, the blank is easily obtained for every kind of experiment and is often recorded without any utilization for experiment evaluation. Such typical data points from blank are also present in datasets from real sample analysis performed technically under the same conditions. Instead of both DTW and Internal Standards (IS), information from the blank measurement is available for simple and immediate comparison of samples.

The key idea of the approach presented here is following: The common view of the LC-MS data considers that mobile phase complicates (negatively affects) the analysis of the measurement. It contributes to random noise and it is major cause of the systemic noise (ridges and interfering peaks) in nonlinear level on the time axis. Several works are focused on removal of baseline presence from the measured data ([89, 88]). The blank measurement should be considered as a permanent standard. The blank time axis has direct relation (homomorphism in fact) to all of the samples measurements obtained with the same settings, the same devices and the same mobile phase. Moreover, lower amount of relevant data points is needed to enter the computation process. These data points represent an inherent set of internal standards.

This section is focused on the study of the key idea to use the data from blank measurement directly for time-alignment, without any peak detection. It is done prior to any further and superfluous analysis and is of general character. The application of internal standards (IS) only adds additional information to it (mathematically just increase the amount of inflex points in the measurement). It is demonstrated on example that blank based approach is very robust, when only

few presumptions are fulfilled.

For the first step in the whole processing/analysis, the retention time alignment, was developed a method which is completely model-independent. This comparison is naturally more comprehensive than IS and does not require any compound identification. In some aspects, namely when abundant peaks are present, it preserves reliability.

With the knowledge of feature detections as well as time alignment issues, and without any other assumption, it can be put the following question: Where to look for internal standards fulfilling the condition to be 'friendly' (different, detectable, known properties, etc.) to given sample and experiment method? The most simple answer is usually neglected for no reason. Obviously, the baseline consist of substances with very relevant features: designated amount (rate, gradient) of solvents, known or predictable affection to the analyte(s), pertinence to the column, and therefore to the requested chemical separability and specific time of elution above all.

Mobile phase in LC-MS affect the measurement analysis, represent the systematic noise in nonlinear level on the time axis. As is shown here, the omitting presence of the baseline can be turn into the advantage considering it as the permanent standard addition measurable also alone in the blank. This section recommends it at the beginning of rough development of semi-optimal sets of internal standards or advanced comparison algorithms.

The reason, why the set of internal standards present in blank LC-MS measurement is so extensive, comes from measurement practice. The sample with solvent mixture is injected into a chromatographic column in LC-MS for the first separation and, due to the interaction with the column stationary phase, elutes at different retention times ([73, 79]). It is strictly recommended, but not always followed, to wash-out (clean up) the column for re-equilibration at the end of the measurements. The true wash-out takes as much as 24 hours ([68]), for this reason there are done only partial (short-time) wash-outs to remove the solvents and other impurities (rests of the sample, phthalate esters from preparation plastic dishes, etc.) at the end of every measurement. Therefore, it is obtained in most measurements at least one of these events, solvent (injection) peak (SP) and/or wash-out tail (WOT). If these part(s) of data were recorded, it is another question,

let assume the solicitous operator. It can not save the time of measurement to despise the beginning or end of the data. It is already done, so there is no reason for uncollect it.

In the blank measurement is SP or WOT (or both, in optimal case) the semi-dominant part of chromatogram (as is shown in Figure 4.1), even if the number of solvents in mixture is small. And, because of usage of the same settings, SP or WOT has to be also presented in the sample(s) measurement, perhaps less distinguishable. In given experiment series, due to incomplete wash-out of the column, some of the solvent contaminants may (and do) actually arise from samples (or blank) themselves. Thus, their use as effective internal standards is obvious.

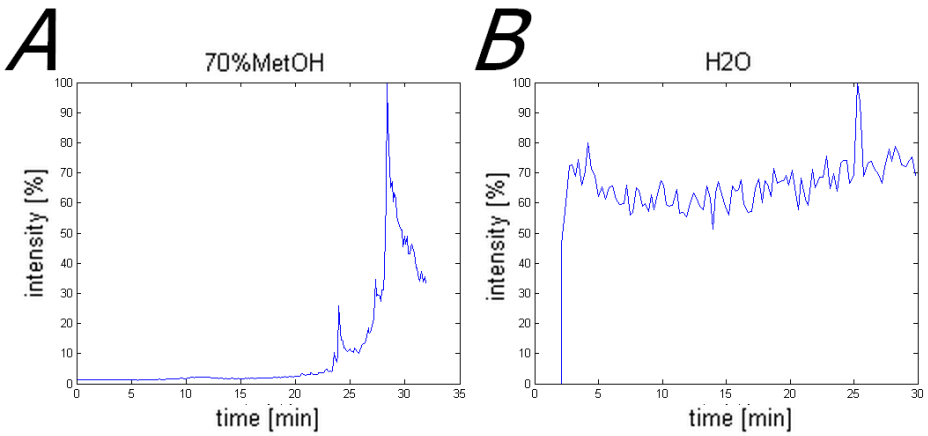


Figure 4.1: Two examples of blank measurements. Panel 1A shows the 70%MeOH mobile phase without solvent peak and with wash-out, panel 1B shows the H20 mobile phase with solvent peak far from ideality and without wash-out.

In this way, the time axis of the blank measurement is considered as reference time axis. It is congruent for all other sample measurements, which are done using the same settings and devices. The time-alignment consist of three main steps, each of them can be investigated by many different methods (already existed or developed in the future). In the following sections, all steps are extend in details. All relevant issues are precisely and mathematically described and justified.

4.1 Step 1.:Reduction of blank points

The blank measurement as well as any LC-MS measurement (considering without msn or other extensions) produce data of three discrete axes: retention time, mass-to-charge ratio and intensity. In other words, there is obtained one intensity for each time and mass pair. This could be mathematically described as mapping from the set T of time values t and set M of mass values m into set Y of intensity values $y(t, m)$. It is more transparent when the sets T, M and Y are ordered, in the following text is considered that property and all sets are ordered increasingly. The LC-MS measurement is therefore defined by the sets (T, M, Y). Let mark the sets, that defined the blank measurement as (T_B, M_B, Y_B) to distinguish them in the following text from the experiment (analyte) measurements (T_{A1}, M_{A1}, Y_{A1}), (T_{A2}, M_{A2}, Y_{A2}) and so forth.

In the very first step, it is helpful to decrease the number of mass values in the blank. The reason is obvious, even the blank measurement is affected by the random noise and mass spikes. Only the true mobile phase compounds are required for the following computation. Furthermore, it is not a big pay to lose very small (in amount) compounds. They may not be present in the real sample measurement(s) for various reasons and contribute to useless increase of the computation time.

The basic way how to reduce amount of blank data points is to discard all intensity values under some thresholds value. This threshold could be general for whole blank or adaptive (different thresholds for different regions of blank), based only on the intensity value or computed via statistical parameters (PDF estimation, between-class variance, MVA) and other advanced techniques (entropy, space transformations, morphological segmentation). Four the purpose to show the usability of blank measurement for time alignment, is enough to compute general threshold from statistical moments. Actually, the precision of this step is not as important as in the next two steps (markers selection and estimation of transformation function). Decrease of data points for marker selection is more significant for computer memory (this could be overcome by HDD swapping which extents the total time) then for the total time of computation, using todays CPUs and/or GPUs.

Let analyze individual mass $m_b \in M_B$ in the time axis and compute the maximal intensity value $\mathcal{X}_Y$:

$$\mathcal{X}_Y(m_b) = max(y(t, m_b)),\ t \in T_B,\ y \in Y_B \tag{4.1}$$

and mean intensity value μ_Y:

$$\mu_Y(m_b) = mean(y(t, m_b)),\ t \in T_B,\ y \in Y_B. \tag{4.2}$$

As an input for thresholding process is used max-to-mean ratio R as standard method for automated data processing and observation ([128, 98, 99]):

$$R(m_b) = \mathcal{X}_Y(m_b)/\mu_Y(m_b). \tag{4.3}$$

Now, are computed two numbers from the max-to-mean ratio R (with a priori unknown distribution) using statistical moments. The number that separating the lower half of a sample from the higher half is the median, mathematically the value α that minimize

$$E(|R(m) - \alpha|), \tag{4.4}$$

where function $E(\xi)$ is considered as the average of its argument ξ (and in this case is $\xi = |R(m) - \alpha|$). Therefore, median α_{med} id defined as

$$\alpha_{med}\ :\ E(|R(m) - \alpha_{med}|) = min,\ \forall \alpha \in \mathbb{R}, \tag{4.5}$$

where $\mathbb{R}$ is set of real numbers. As a measure of the variability is used robust standard deviation ($RSTD$), because the max-to-mean ratio R has a priori unknown distribution:

$$RSTD = 1.25 * E(|R(m) - \alpha_{med}|). \tag{4.6}$$

The threshold value Θ for max-to-mean ratio R is set as

$$\Theta = \alpha_{med} - RSTD. \tag{4.7}$$

Consequently, all masses m_b with ratio $R(m_b)$ lower then threshold Θ are removed from blank in further computation. Let mark the new set of mass-to-charge ratio

with max-to-mean ratio R higher then threshold Θ as $\tilde{M}$:

$$\tilde{M_B} = M_B - \{m_b \ : \ R(m_b) < \Theta\}, m_b \in M_B \tag{4.8}$$

where M_B is ordered set of $[m/z]$ values in the blank measurement, $R(m_b)$ is max-to-mean ratio ([128, 98, 99]) and Θ is chosen threshold. Videlicet, $\tilde{M_B}$ is just a subset of M_B with property $R < \Theta$. However, the data reduction is not strictly necessary. Thresholding is not initial selection of alignment markers. It is just a simple random noise filtration.

Also could be the ratio set R separated only to lower and higher region by threshold equals to median value, whereas with threshold computed by equation (4.7) retain at least 2/3 of the blank measurement. In the blank with huge level of impurities may almost all data points pass through the thresholding, at least it still discards the low relevant of them (in meaning of capability for being markers in time-alignment).

4.2 Step 2.:Markers selection

The second step is the foot-stone for all comparison tasks and it is known as the selection of the markers ([106, 92]). In other words, the markers are point candidates for the alignment itself. The markers in this approach are defined only from the blank, instead of searching for similar values in compared sets. Without any hesitation, it is sure that they are present in the sample measurement(s) also. Therefore, the corresponding data points can be easily pinpointed from the sample, after finished definition.

As was described above, in every measurement (even in the blank) is presented at least one of SP or WOT event. Successfully, SP occurs on the first half (in time axis) of the measurement and WOT on the second half (not considering peculiar operator errors like two measurements in one data set, stored only middle of measurement or nothing, etc.). Therefore, one can split the blank in time into two subparts (time intervals), each possibly containing one expressive feature. Using gradient changes during measurement offers splitting into more subparts (not necessary equidistant) with simple selection of cutting times. Just

be sure, that the distinctive baseline inflex point (local minimum or maximum in intensity) is somewhere in the middle of selected interval (or leastwise not exactly on the interval borders). And it is known the exact time value of that inflex point from the settings of the experiment, as it was designed. Past question, maximal number of time intervals is equals to the number of measured time points in the discrete data set, i.e. equals to the cardinality ($\aleph$) of set T_B. The optimal number of subparts could be determined by statistically appropriate methods ([100, 101]), in case of equidistant intervals. Let assume that sets T_B, fulfill the sampling theorem ([102, 103, 104, 105]) and split the blank time axis (and therefore whole blank measurement) into n equidistant subparts, where $2 \leq n \leq \aleph(T_B)$. For the simple illuminating example, is n equals to 3. Now, they were obtain three time intervals $T1_B$, $T2_B$ and $T3_B$ (or $T\vartheta_B, \vartheta = 1, ..., n$ shortly) as the subsets of T_B:

$$(T1_B \subset T_B) \wedge (T2_B \subset T_B) \wedge (T3_B \subset T_B), \tag{4.9}$$

$$T1_B \wedge T2_B \wedge T3_B = T_B. \tag{4.10}$$

The intervals are defined with additional properties.

I.) The sets $T\vartheta_B$ are increasingly ordered sets.

II.) time interval $T1_B$ precede time interval T_B2 and time interval $T2_B$ precede time interval $T3_B$:

$$T1_B \prec T2_B \prec T3_B. \tag{4.11}$$

III.) The cardinalities of the subsets are equal or approximately equal:

$$\aleph(T1_B) \approx \aleph(T2_B) \approx \aleph(T3_B), \tag{4.12}$$

$$\aleph(T1_B) + \aleph(T2_B) + \aleph(T3_B) = \aleph(T_B), \tag{4.13}$$

because the time intervals $T\vartheta_B$ are equidistant or semi-equidistant (if cardinality of T_B is or is not divisible by $n = 3$ in natural numbers $\mathbb{N}$). In the worst case, cardinality of the shortest time interval differs to the others only by one.

The most common and understandable representations of LC-MS measurements are Total Ion Chromatogram (TIC) and mass spectrum. Mass spectrum is a measure of MS detector signal (intensities y) versus mass-to-charge ratio axis ($m \in M$ or $\tilde{m} \in \tilde{M}_B$ in the example now). One mass spectrum is just a slice of selected time in the whole measurement. The amount of all individual mass spectra in the measurement is equal to the cardinality of the set T. Therefore, is also the amount of mass spectra in each time intervals $T\vartheta_B$ equals to the cardinality of the related interval. TIC is a measure of detector signal versus time axis T_B. It is amount of all intensity values y in exact time point $t \in T_B$:

$$\gamma_B(t) = \sum_{\tilde{m}} y(t, \tilde{m}), y \in Y_B. \tag{4.14}$$

So, they are obtained three different sub-TICs $\gamma\vartheta_B$, after splitting the time axis T_B into $n = 3$ intervals:

$$\gamma\vartheta_B(t\vartheta) = \sum_{\tilde{m}} y(t\vartheta, \tilde{m}), t\vartheta \in T\vartheta_B, \vartheta = 1, ..., n, \tag{4.15}$$

one blank sub-TIC $\gamma\vartheta_B$ for each time interval $T\vartheta_B$.

The splitting of the time set T_B into n subparts (time intervals) $T\vartheta_B$ and therefore splitting of TIC γ_B into sub-TICs $\gamma\vartheta_B$ also define the amount of markers used for time-alignment. There is necessary only one point in each time interval and it is almost directly selected from the related blank sub-TIC. As a blank marker is considered the time value τ_B of the subset $T\vartheta_B$, where the sub-TIC value is the maximal value of that sub-TIC:

$$\tau_B(\vartheta) \mid \gamma\vartheta_B(\tau_B(\vartheta)) = max(\gamma\vartheta_B(t\vartheta)), \tau_B(\vartheta) \in T\vartheta_B. \tag{4.16}$$

In other words, is in time point $\tau_B(\vartheta)$ significant inflex point of blank sub-TIC $\gamma\vartheta_B$. Equation (4.16) produces the set $\{\tau_B\}$ of cardinality $\aleph = n$ as the set of blank markers for transformation function. Blank time axis T_B is in this approach considered as reference time axis for each time-alignment of measurement done with similar experiment conditions.

It is slightly trickier to identify corresponding markers in analyte measure-

ment time axis T_A. The minimal and maximal values of measurement TIC γ_A:

$$\gamma_A(t) = \sum_m y(t, m), y \in Y_A, \tag{4.17}$$

occurred in different parts of measurement, because of presence of the analyte. Cardinality of measurement mass-to-charge ratio set M_A is bigger then cardinality of blank mass-to-charge ratio set M_B. The reason is obvious, at least one m_A value of the measured analyte was added into the mobile phase to make the experiment meaningful. Usually, the amount of added mass values is higher than one. There is not only the analyte molecular ion, but ions of its isotopes, fragments, adducts and impurities. Therefore, cardinality of the intensities set Y_A has to be also bigger than cardinality of set Y_B. Bigger amount of molecules with bigger amount of possible mass-to-charge ratios in almost same measurement time length ($T_A \approx T_B$) produce wider dynamic range of intensity set Y_A:

$$\aleph(M_A) > \aleph(M_B) \wedge \aleph(Y_A) > \aleph(Y_B). \tag{4.18}$$

Surprisingly, the analyte measurement TIC γ_A is not relevant for selection of the analyte marker set $\{\tau_A\}$. The pinpointing process from sets (T_A, M_A, Y_A) differs from blank.

One more set of information is necessary to extract from blank measurement. With the knowledge of when ($in\ \tau_B(\vartheta)$) the maximal value of sub-TIC $\gamma\vartheta$ was obtained, is also profitable to ask where (in mass). Slice of selected time in the whole measurement (blank or analyte) represents the mass spectrum as tuple:

$$y(t) = [y(t, m_j)], m_j \in M, y \in Y. \tag{4.19}$$

Not every mass m_j was presented in detector in selected spectrum, i.e. some of the intensity values $y(t, m_j)$ are equal to zero in selected time. In mass spectrum, it is possible that two different and distinguishable mass values reach the exactly same intensity ($y(t, m_q) = y(t, m_w),\ q \neq w,\ t = const.$). Equality in non zero intensity values is not very frequent, however there is nothing bizarre on this fact. The probability is small, but it does not mean impossibility of the event, especially

in huge amount of different molecules detected by MS during the measurement. Hence, the mass spectrum is described as tuple and not as a set.

In time markers $\{\tau_B\}$ are corresponding n mass spectra tuples $y(\tau_B)$ of the blank. As a $\vartheta - th$ blank mass marker is considered the mass-to-charge value $\eta_B(\vartheta)$ of the set M_B, where in the mass spectrum $y(\tau_B(\vartheta))$ is the maximal value of intensity:

$$\eta_B(\vartheta) \mid y(\tau_B(\vartheta), \eta_B(\vartheta)) = max([y(\tau_B(\vartheta), \tilde{m_b})]), \tilde{m_b} \in \tilde{M_B}, y \in Y_B. \tag{4.20}$$

The cardinalities of blank time and mass markers are equal:

$$\aleph(\{\tau_B\}) = \aleph(\{\eta_B\}), \tag{4.21}$$

and time values $\tau_B(\vartheta)$ with mass values $\eta_B(\vartheta)$ make set of whole blank markers as n ordered pairs $\{(\tau_B, \eta_B)\}$.

Analyte measurement time axis T_A is also separated into n intervals $T\vartheta_A$, $\vartheta = 1..n$. Each analyte interval is approximate (means very similar) to blank interval ($T\vartheta_A \approx T\vartheta_B$) in equidistant case with approximately same start and end time point of the measurement ($T_A \approx T_B$). It is necessary to carefully choose the individual interval borders, when the time splitting was based on gradient inflex points. Corresponding gradient changes have to be situated in corresponding time intervals. Correct separation task could be simplified by proper timing of all measurements recording process and equipment synchronization.

Direction of analyte markers selection is opposite to the blank situation - from mass to time values. As analyte mass markers are considered blank mass-to-charge ratios $\{\eta_B\}$ that are present in the analyte mass set M_A:

$$\eta_A(\vartheta) \mid \eta_A(\vartheta) = \eta_B(\vartheta), \tag{4.22}$$

$$\eta_B(\vartheta) \in \tilde{M_B} \wedge \eta_B(\vartheta) \in M_A \Leftrightarrow \eta_A(\vartheta) \in M_A \wedge \eta_A(\vartheta) \in \tilde{M_B}. \tag{4.23}$$

Mass-to-charge ratios $\{\eta_B\}$ are supposed to be in the analyte measurements set M_A. Values $\eta\vartheta_B$ were taken from the blank set $\tilde{M_B}$ and belong to the molecules of mobile phase. Mobile phase is a part of the analyte measurement. This condition

is always fulfilled if whole blank markers selection was done on mass-to-charge subset $\tilde{\tilde{M}}_B$:

$$\tilde{\tilde{M}}_B \subset \tilde{M}_B \subset M_B \mid \tilde{\tilde{M}}_B = \tilde{M}_B \cap M_A. \tag{4.24}$$

In other words subset $\tilde{\tilde{M}}_B$ is defined as intersection of blank mass subset $\tilde{M}_B$ from Step1 and analyte mass set M_A. Therefore, values $\tilde{\tilde{m}}_b$ are present also in blank measurement and analyte measurement:

$$\tilde{\tilde{m}}_b \in \tilde{\tilde{M}}_B \Leftrightarrow \tilde{\tilde{m}}_b \in \tilde{M}_B \Leftrightarrow \tilde{\tilde{m}}_b \in M_A. \tag{4.25}$$

Instead of $\tilde{M}_B$ or $\tilde{m}_b$ is used $\tilde{\tilde{M}}_B$ or $\tilde{\tilde{m}}_b$ respectively in equations (4.14. .4.23). Thus, is redundant to distinguish signs η_B and η_A, because both tuples are equal. Let sign mass markers for further purpose only as η:

$$\eta(\vartheta) = \eta_B(\vartheta) = \eta_A(\vartheta) \mid \forall\vartheta = 1..n \Rightarrow \{\eta\} = \{\eta_B\} = \{\eta_A\}. \tag{4.26}$$

That is not as trivial as seems to be. Blank mass markers $\{\eta_B\}$ are values $\tilde{m}_b$ or $\tilde{\tilde{m}}_b$ from the subset $\tilde{M}_B$ or $\tilde{\tilde{M}}_B$ respectively. On the other hand, analyte mass markers $\{\eta_A\}$ are values from the set M_A. Therefore indexes b and a are not equal, even if the value m_b equals to the value m_a. Obviously, there is forbidden the exception of special case where set M_B or $\tilde{M}_B$ or $\tilde{\tilde{M}}_B$ respectively strictly equals to the set M_A, for two serious reasons. At first, set M_A contains additional mass values of the analyte itself, not presented in blank measurement. At second, some random noise is always presented. The probability is extremely low in our universe, that two measurements have exactly the same distribution of random noise occurrence which fits in values and positions. Sign simplification done by equation (4.26) is allowed just because blank mass subset $\tilde{\tilde{m}}_b$ is no more necessary in time-alignment process. However, b and a indexes inequality is important to consider in algorithm implementation (wrong index is one of the top common source code mistakes in programs development).

Only a part of analyte measurement is further investigated, once the mass markers $\{\eta\}$ were pinpointed. The behavior of single analyte mass value m_a in time could be described as mapping from that mass value $m_a \in M_A$ and the set T_A into the set Y_A of intensity values y. This mapping process produce Single Ion

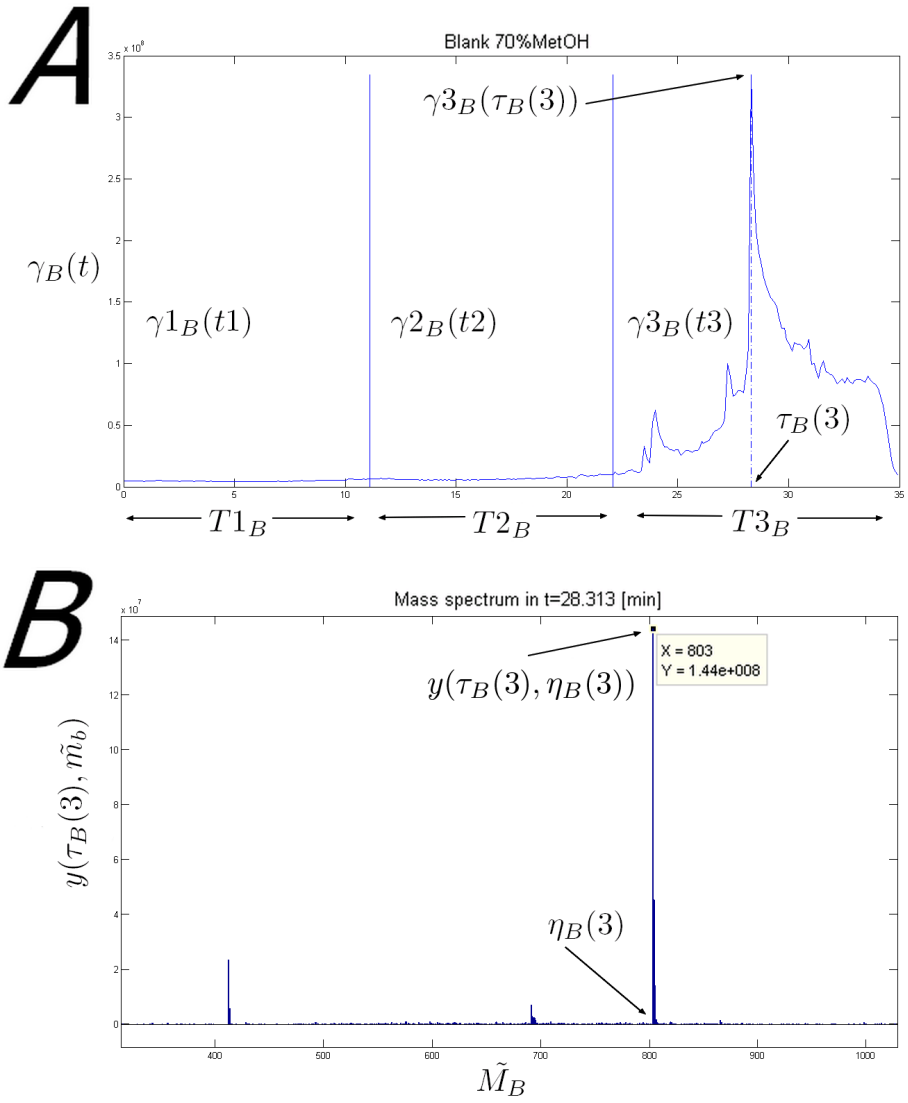


Figure 4.2: Example of blank markers selection. Panel 2A shows Total Ion Chromatogram (TIC) $\gamma_B(t)$ separated into $n = 3$ sub-TICs $\gamma 1_B$, $\gamma 2_B$ and $\gamma 3_B$ on time intervals $T1_B$, $T2_B$ and $T3_B$. Maximal intensity value $\gamma 3_B(\tau_B(3))$ is in time interval $T3_B$ located on time $\tau_B(3)$. Panel 2B shows mass spectrum in selected time $\tau_B(3)$. Maximal intensity $y(\tau_B(3), \eta_B(3))$ is obtained on mass $\eta_B(3) \in \tilde{M}_B$. Blank time marker value $\tau_B(3)$ is equals to 28.313 $[min]$ and blank mass marker value $\eta_B(3)$ is equals to 803 $[m/z]$ in this example. Apparently, there are no visible relevant features for markers selection. However, the range of intensity axis is 10^8, which dissable details in lower intensity values. That is exactly why observation only of TICs is not wisdom.

Chromatogram (SIC) as a function of time:

$$\gamma_{m_a}(t) = y(t, m_a),\ t \in T_A\ ,\ y \in Y_A. \tag{4.27}$$

Therefore, for each mass value m_a from set M_A exist one SIC ($\aleph(\{\gamma_{m_a}\}) = \aleph(M_A)$). Consequently, the analyte TIC $\gamma_A(t)$ is just a sum over $m_a \in M_A$ of all analyte SICs $\gamma_{m_a}(t)$:

$$\gamma_A(t) = \sum_{m_a} \gamma_{m_a}(t) = \sum_{m_a} y(t, m_a), m_a \in M_A,\ t \in T_A,\ y \in Y_A. \tag{4.28}$$

Note, that this seemingly means skipping the step of analyte measurement points reduction. In case of mass markers $\eta \in M_A$ is necessary only n number of analyte SICs, just $\gamma_{\eta(\vartheta)}$:

$$\gamma_{\eta(\vartheta)}(t) = y(t, \eta(\vartheta)), t \in T_A,\ y \in Y_A. \tag{4.29}$$

Therefore, decreasing of amount of points in analyte measurement is greater in contrast to the blank measurement reduction in Step 1. ($\aleph(\{\eta\}) \ll \aleph(M_A)$). Moreover, not whole SIC $\gamma_{\eta(\vartheta)}$ is required for selection of $\vartheta - th$ analyte time marker $\tau_A(\vartheta)$. The analyte measurement time axis T_A was separated into n intervals $T\vartheta_A$. It is quaranted to find the $\vartheta - th$ time value τ_A in time interval $T\vartheta_A$, when the time set separation was done correctly ($T\vartheta_A \approx T\vartheta_B$). Thus, analyte time markers pinpointing process works on n sub-SICs, instead of whole analyte measurement ((T_A, M_A, Y_A)). The $\vartheta - th$ sub-SIC is then defined as a part of mass marker $\eta(\vartheta)$ SIC $\gamma_{\eta(\vartheta)}(t)$ on time interval $T\vartheta_A$:

$$\gamma\vartheta_{\eta(\vartheta)}(t\vartheta) = y(t\vartheta, \eta(\vartheta)), t\vartheta \in T\vartheta_A,\ y \in Y_A, \vartheta = 1, ..., n. \tag{4.30}$$

As an analyte time marker is considered the time value τ_A of the subset $T\vartheta_A$, where the sub-SIC value $\gamma\vartheta_{\eta(\vartheta)}$ is the maximal value of that sub-SIC:

$$\tau_A(\vartheta) \mid \gamma\vartheta_{\eta(\vartheta)}(\tau_A(\vartheta)) = max(\gamma\vartheta_{\eta(\vartheta)}(t\vartheta)), \tau_A(\vartheta) \in T\vartheta_A. \tag{4.31}$$

The total space of values to be analyzed is rapidly decreased (from thou-

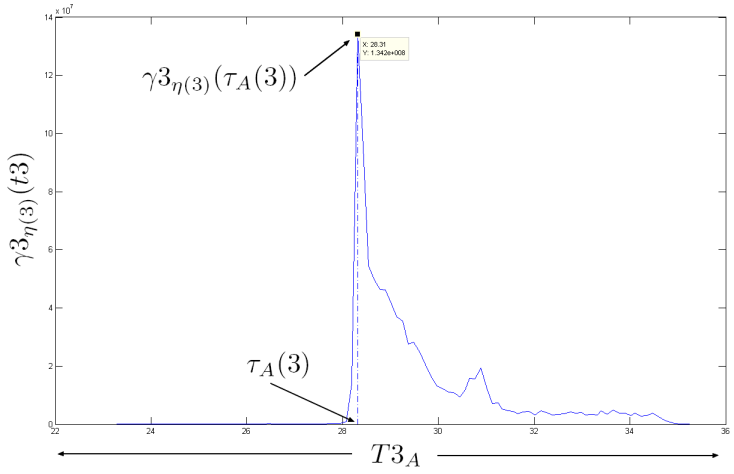


Figure 4.3: Example of analyte time marker selection. In the $3-rd$ sub-SIC $\gamma 3$ of the analyte mass $\eta(3)$ is maximal intensity obtained in the time value $\tau_A(3)$. Therefore, the $3-rd$ analyte time marker $\tau_A(3)$ value is equals to 28.31 $[min]$ in this example. There is no mass spectrum, because SIC consist (by its definition) of single $[m/z]$ value $=\ \eta(3)$.

sands to ones). Process of the selection of the markers is indicated on Figure 4.2. This is sufficiently robust approach because all blanks have discernible signals, even a watter (at least injection peak, however there are useful changes in span on the time axis). Once again, the determination of markers is enough to be done in blank processing and then pinpoint the corresponding markers in the analyte measurements.

Again, the cardinalities of analyte time and mass markers are equal:

$$\aleph(\{\tau_A\}) = \aleph(\{\eta\}). \tag{4.32}$$

and mass values $\eta(\vartheta)$ with time values $\tau_A(\vartheta)$ make set of whole analyte markers as n ordered pairs $\{(\tau_A, \eta)\}$. It follows from the equation (4.26) that mass markers η are the same for blank and analyte. Therefore, (using equations (4.32) and (4.21)) is also the amount of blank time markers equal to the amount of analyte time markers:

$$\aleph(\{\tau_A\}) = \aleph(\{\tau_B\}) = n. \tag{4.33}$$

This is exactly what is often demand (to have the same cardinality of two corresponding time sets) and makes the next step as easy as possible.

4.3 Step 3.:Transformation function(s)

Finally, the third step works with the time values of the selected markers from both sets (blank and sample), which are now of the same cardinality and in the same order. This last step actually produces the transformation function, it computes the description of the time-alignment. However, the procedure is not limited to the given algorithm. Nonlinear shifts in the retention time between measurements arise especially from stochastic changes in column chemistry over time and minor changes (also stochastic) in mobile phase composition ([88, 87, 93]). Considering this nonlinearity between time axes leads to the various normalization rules or shift corrections ([92, 106]). The blank measurement time axis T_B is considered as the reference time axis, in this approach. Generally, any analyte measurement time axis could be aligned onto blank time axis by a priori unknown non-linear transformation function $\mathcal{F}$:

$$t_b = \mathcal{F}(t_a, \beta), t_b \in T_B, t_a \in T_A, \{\beta\} \in \mathbb{R}, \tag{4.34}$$

where β denotes unknown parameter(s) of the function $\mathcal{F}$.

There is no strictly restriction for analyte time axis to be also considered as the reference one. Consequently, the blank measurement time axis could be aligned onto analyte time axis as by function $\breve{\mathcal{F}}$:

$$t_a = \breve{\mathcal{F}}(t_b, \breve{\beta}), t_a \in T_A, t_b \in T_B, \{\breve{\beta}\} \in \mathbb{R}, \tag{4.35}$$

and sign $\breve{\beta}$ denotes unknown parameter(s) of $\breve{\mathcal{F}}$, analogously. Function $\breve{\mathcal{F}}$ is in ideal case (in deterministic world without noise where all processes are purely equilibristic infinitesimal changes in non-fractal phase space) identical to the inverse function $\mathcal{F}^{-1}$ of $\mathcal{F}$. However, it may be misleading to select one of the analyte measurements time axis. There has to be very pertinent reason for using equation (4.35). Exempli gratia, using time axis of healthy patient blood sample as refer-

ence time axis for other 'sick' patients is just a wish for experiment purpose. The simplest standard is still represented by the blank for chosen setup of measurement device (LC column, solvents, gradient changes, MS ionization, detector focus, and so on). Once again, blank is general basic information independent on the experiment higher-level interpretation. Vice versa, the blank measurement depends only on the experiment setup and device properties. Therefore, correct and rigorous blank measurement (T_B, M_B, Y_B) describes the experiment. It is the knowledge ready to be used in time-alignment.

The transformation $\mathcal{F}$ is a description for adjustment of time axes relation. Time markers $\tau_B \in T_B$ and $\tau_A \in T_A$ are time values with superb property - the resemblance between $\tau_B(\vartheta)$ and $\tau_A(\vartheta)$ is congruent:

$$\tau_B(\vartheta) \cong \tau_A(\vartheta), \ \forall \vartheta = 1, ..., n. \tag{4.36}$$

In other words, time markers $\tau_B(\vartheta)$ and $\tau_A(\vartheta)$ match together. For the sake of completeness, relation between blank time axis T_B and analyte time axis T_A is homomorphism (structure-preserving mapping) and relation between time markers $\{\tau_B\}$ and $\{\tau_A\}$ is isomorphism (bijective homomorphism).

The most puzzling issue is the task of function $\mathcal{F}$ type specification ([107, 108, 109]), i.e searching for data analysis process for constructing mathematical mapping, that minimizes displacement of the data points (time values). Common approach is to create a class of possible models, but it is not always obvious what models should be used ([110]). Even with the understanding of underlying physical and chemical properties of the problem is difficult to choose the right model. Hence, both in linear and nonlinear modeling is used regression analysis ([111]) as investigation of the hypothesis about the relationship between the variables of interest. Specific cases are various iterative methods for value interpolation ([112, 113]), in which the function must go exactly through the time markers τ. The objective of regression analysis is to produce an estimate of the hidden parameters β ([114]). Unfortunately, any parameter analysis can only help in differentiating between hypothesis or models ([115]). Very strong results still do not prove that the correct function $\mathcal{F}$ was chosen ([116]).

Note, that the linear functions are just the evaluation of polynomial of first

degree. Consequently, the very first 'non-linearization' is the polynomial of higher degree. Insofar that, the most extremely primitive nonlinear function evaluate polynomial of second degree. The collection of eventual type of relations (models, mappings, hypothesis, functions, whatever) is huge. Harmonic analysis (wavelets, fast Fourier transformation, eigenvalues) and MVA are the famous and prevalent theories nowadays ([117, 118, 91, 90]).

Therefore, the task of the proper transformation function selection is always nontrivial. For instance, the mentioned simple function was chosen to illuminate the power of blank measurement. Accordingly, the relation between blank time set T_B and analyte time set T_A is considered as polynomial function of second degree:

$$\mathcal{F}(t_a, \beta): \; t_b = \beta_2 {t_a}^2 + \beta_1 t_a + \beta_0 + \varepsilon_a, t_b \in T_B, t_a \in T_A, \beta_\kappa \in \mathbb{R}, k = 0, ..., 2, \quad (4.37)$$

where $\varepsilon_a \in \mathbb{R}$ is an unobserved random variable, representing the errors in the data. Let define the parameters vector $[\beta]$, blank time markers vector $[\tau_B]$ and analyte time markers $[\tau_A]$ matrix:

$$[\beta] = \begin{pmatrix} \beta_p \\ \beta_{p-1} \\ \vdots \\ \beta_0 \end{pmatrix}, \; [\tau_B] = \begin{pmatrix} \tau_B(1) \\ \vdots \\ \tau_B(n) \end{pmatrix},$$

$$[\tau_A] = \begin{pmatrix} {\tau_A}^p(1) & {\tau_A}^{p-1}(1) & \cdots & \tau_A(1) & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ {\tau_A}^p(n) & {\tau_A}^{p-1}(n) & \cdots & \tau_A(n) & 1 \end{pmatrix}.$$

where p is degree of the polynomial (and therefore natural number, $p \in \mathbb{N}$) and n is cardinality $\aleph$ of time markers τ_A or τ_B ($\aleph\{\tau_A\} = \aleph\{\tau_B\}$). In the example are $p = 2$ and $n = 3$.

The unknown parameters β of polynomial transformation function $\mathcal{F}$ could be then estimated by regression analysis (using equation 4.36):

$$[\beta] \simeq [A] \backslash [B], \quad (4.38)$$

where sign \ is defined as matrix left division

$$[A]\backslash[B] = [A]^{-1} * [B], \tag{4.39}$$

because matrix multiplication is not commutative.

The problem is with the error ε_a, that causes only asymptotic equality in matrix equation (4.38) and leads to the inexactly specified system of simultaneous equations. The solutions is a particular estimation of the values of all parameters β that simultaneously satisfies all of the equations. Regression analysis offers numerous parameter estimation methods ([90, 91]), that differ in computational burdens and robustness depended on the distribution of unobserved error ε_a. Frequently used method to solving systems of equations is approach of least squares ([119, 120]). It is a technique that minimize the Euclidean length of a vector $[\varepsilon]$, defined as:

$$[\varepsilon] = [A] * [\beta] - [B], \tag{4.40}$$

This last step actually produces the parameters of transformation function, it computes the description $\mathcal{F}$ of the time-alignment:

$$\grave{t_a} = \beta_2 {t_a}^2 + \beta_1 t_a + \beta_0, \tag{4.41}$$

where time values $\grave{t_a} \in \grave{T_A}$ are analyte measurement time values $t_a \in T_A$ asymptotically aligned to the blank measurement time values:

$$t_b \simeq \grave{t_a}. \tag{4.42}$$

Furthermore, blank approach allows to align the time axes of all analyte measurements $(T_{A\lambda}, M_{A\lambda}, Y_{A\lambda})$, $\lambda \in \mathbb{N}$, done on the same chromatographic column under same experiment conditions. Simply, two given analyte time axis T_{A1} and T_{A2} are independently normalized to the blank time axis T_B:

$$\grave{t_{a1}} = \beta_{2(A1)} {t_{a1}}^2 + \beta_{1(A1)} t_{a1} + \beta_{0(A1)},\ t_{a1} \in T_{A1}, \tag{4.43}$$

$$\grave{t_{a2}} = \beta_{2(A2)} {t_{a2}}^2 + \beta_{1(A2)} t_{a2} + \beta_{0(A2)},\ t_{a2} \in T_{A2}, \tag{4.44}$$

where $\beta_{\kappa(A\lambda)}$ are the parameters of polynomial transformation function $\mathcal{F}_\lambda$ of each analyte time axis $T_{A\lambda}$. Normalized time values $t_{a\lambda}^{`}$ are asymptotically aligned to the time values t_b, by analogy of equation (4.42):

$$t_b \simeq t_{a1}^{`} \wedge t_b \simeq t_{a2}^{`}. \tag{4.45}$$

Therefore, also time values $t_{a1}^{`}$ are aligned to the time values $t_{a2}^{`}$.

$$t_{a1}^{`} \simeq t_{a2}^{`}. \tag{4.46}$$

However, equation (4.46) simplify any comparison of given analyte measurement $(T_{A\lambda}, M_{A\lambda}, Y_{A\lambda})$ using the knowledge of blank measurement (T_B, M_B, Y_B) and estimated parameters $\beta_{\kappa(A\lambda)}$ of functions $\mathcal{F}_\lambda$.

The last two steps are very similar with DTW or IS. With standards addition, it is essential to locate their positions in the measurement data sets as input for time transformation function. The localization is algorithmically the comparison task, which is in principle time consuming and noise affected procedure. Some (or at least approximate) parameters of IS are known. This a priori information decreases slightly the complexity of comparison techniques. DTW is more difficult - the number of corresponding points in measurements is a priori unknown, data sets are large, impurities may be clear in signal but differ in order. Therefore, some filtration and preprocessing computation is optional. Of course, DTW could be also applied on IS to produce robust results, in case that IS are sufficiently dominant signals. Unfortunately, the strong and stable solutions are still far from quick and daily use in the rush lab during experiment tunning. As is shown in this thesis, BBTA has to deal only with minimal amount of selected points which are readily available.

4.4 Comparison of BBTA with COW

Two analyte measurements $A1$ and $A2$ are aligned using BBTA. This approach is compared with Correlation Optimized Warping ([83]), one of the well known warping algorithm ([84]). Both experimental samples were prepared by

mixing methanolic extract of the cyanobacterium Nostoc sp. with the antifungal drug Nystatin $C_{47}H_{75}NO_{17}$ (Duchefa Biochemie, cat. no.: 003042.03). Nystatin was added into measurement $A1$ in concentration = $0.5[mg/ml]$ as compound with known value of molecular ion = $926[m/z]$. Nystatin in different concentration = $0.05[mg/ml]$ was added into measurement $A2$.

The samples were analyzed on HPLC-MS (ESI) Agilent ([121]) 1100 Series LC/MSD Trap using C8 reverse phase column (Zorbax XBD C8, $4.6 \times 150[mm]$, $5[\mu m]$) eluted by MeOH / Water gradient with addition of 0.1% formic acid. The ion trap mass spectrometer was optimized for ions with [m/z] ratio 900 in positive mode. The data acquisition and exports were performed using ChemStation Software (Agilent) under WindowsNT operating system. The data analysis outputs were obtained by Expertomica metabolite profiling software ([89]) under Windows XP/Vista operating system.

The spray needle was at a potential of $4.5[kV]$, and a nitrogen sheet gas flow of 20 (arbitrary units) was used to stabilize the spray. The counter electrode was a heated ($200[^{\circ}C]$) stainless-steel capillary held at a potential of $10[V]$. The tube-lens offset was $20[V]$, and the electron multiplier voltage was $-800[V]$. Helium gas was introduced into the ion trap at a pressure of $1[mTorr]$ to improve the trapping efficiency of the sample ions introduced into the ion trap. The background helium gas also served as the collision gas during the collision activation dissociation (CAD).

Blank measurement B was obtained without presence of the analyte mixture (Nostoc extraction, Nystatin). Therefore, Nystatin addition is not considered as IS due to its absence in the blank measurement. Only the blank itself represents internal standards in the BBTA approach. The elements of time sets T_{A1}, T_{A1} and T_B differ to each other as is shown on 4.1. The cardinalities of analyte measurements are equal ($\aleph(\{T_{A1}\}) = \aleph(\{T_{A2}\}) = 322$), the cardinality of blank measurement is lower ($\aleph(\{T_B\} = 313)$).

The TICs of $A1$ (solid line), $A2$ (dotted line) and B (dash-dotted line) are shown on 4.4A. Blank measurement B is quite shorter by terminator of WOT decay beside to the analyte measurements $A1, A2$, as is clear from 4.1 and 4.4A. Analyte measurements time axes were artificially dis-aligned by basic replacement to emphasize time shifts. In principle, analyte time axes are replaced by blank time

	1	2	3	...	312	313	314
t_{a1}	0.0030	0.0963	0.1891	...	33.7272	33.8422	33.9575
t_{a2}	0.0042	0.1018	0.1952	...	33.8274	33.9436	34.0589
t_b	0.0042	0.1444	0.2265	...	31.8125	31.9277	$\emptyset$

Table 4.1: Values of blank and analytes time sets values.

axis. Let remind, that direct replacement has nothing to do with the alignment. Actually, it is the opposite process as is described further in this section.

Let denotes by sign ς the maximal amount of time elements in the given time sets:

$$\varsigma = max(\aleph(\{T_{A1}\}), \aleph(\{T_{A2}\}), \aleph(\{T_B\})), \tag{4.47}$$

and slightly extend the definition of the reference time axis:

$$T_R \mid T_B \subseteq T_R \ \wedge \ \aleph(\{T_R\}) = \varsigma. \tag{4.48}$$

The blank time T_B is a subset of reference time set T_R with cardinality equals to the ς:

$$t_r \equiv t_b \mid r = b, t_r \in T_R, t_b \in T_B, r, b \in \{1, ..., \aleph(\{T_B\})\}. \tag{4.49}$$

The missing time elements $\{t_{\aleph(\{T_B\})+1}, ..., t_\varsigma\} \in T_R$ could be set as equidistant continuation:

$$t_r = t_{\aleph(\{T_B\})} + \Delta t \times (r - \aleph(\{T_B\})), \tag{4.50}$$

where Δt is estimated as averaging of difference between two consecutive time elements in blank time set T_B:

$$\Delta t = \frac{1}{\aleph(\{T_B\}) - 1} \sum_{1}^{\aleph(\{T_B\})-1} (t_{\imath+1} - t_\imath), \ t_{\imath+1}, t_\imath \in T_B \tag{4.51}$$

Theoretically, there are more easy ways how to create the reference time set T_R. Maximal operator in equation (4.47) could be change into minimal and extension of equation (4.50) is no longer necessary. However, minimal reference set means time data reduction and that is not advisable as it was in mass case (Step1. in Methods). The pinpointing process of the time markers τ is crucial

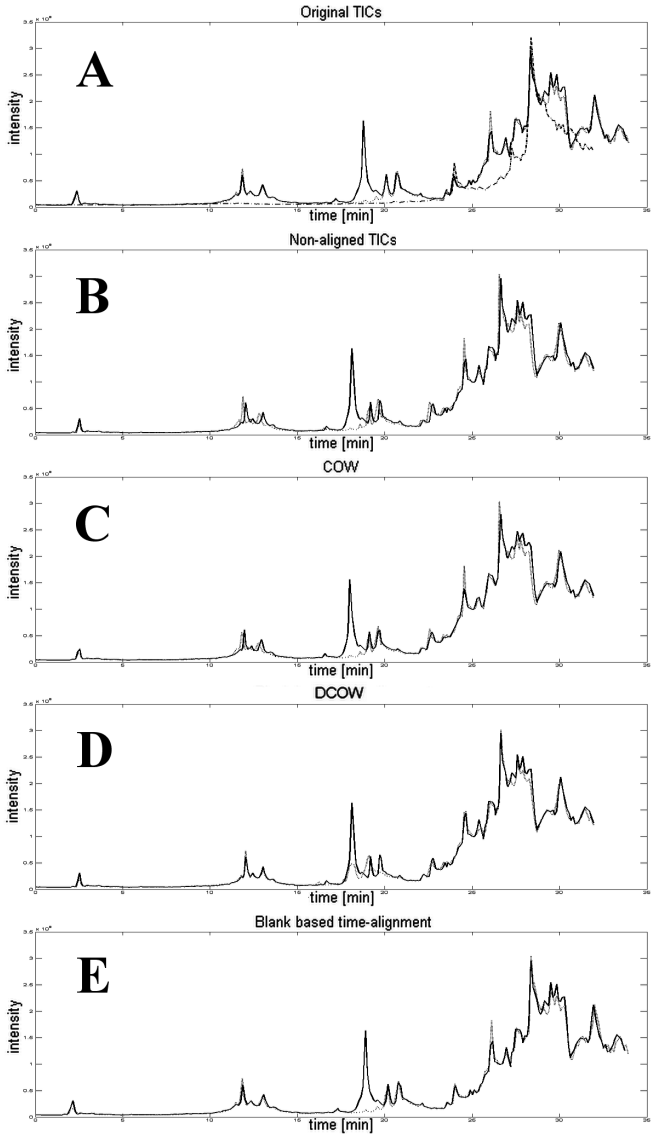


Figure 4.4: Comparison of all TICs. Panel A shows blank and analytes TICs $\gamma_B, \gamma_{A1}, \gamma_{A2}$ in original time axes T_B, T_{A1}, T_{A2}. Panel B shows artificially dis-aligned analyte TICs γ_{A1}, γ_{A2} in reference time axis T_R. Panel C shows results of analyte TICs γ_{A1}, γ_{A2} aligned to the blank TIC γ_B by COW algorithm in reference time axis T_R. Panel D shows results of analyt TIC γ_{A2} aligned directly to the anlyte TIC γ_{A1} by COW algorithm in reference time axis T_R. Panel E shows results of analyte TICs γ_{A1}, γ_{A2} aligned to the reference time axis T_R by Blank based time-alignment in aligned time axes $\grave{T_{A1}}, \grave{T_{A2}}$. Solid lines represents analyte TIC γ_{A1}, dotted lines represents analyte TIC γ_{A2}, dash-dotted line in panel A represents blank TIC γ_B.

part of time-alignment. Therefore, discarding time elements only for convenience reasons is dangerous way of thinking. No matter what the time elements values really are. Another option, the addition at the beginning of the reference set T_R is also possible, but complicated to no avail. The evaluation of missing time values and Δt has the same computational burden (as addition at the end). However, the indexes r has to be shifted and some of the added time elements may obtain negative values. The plots with negative time units on the reference time axis are not good exemplary candidates. The solution of setting all values added at the beginning to zero aims to the mismatch in TICs values. Therefore, is optional to follow the equations (4.47...4.51).

Apparently, in the definition (4.48) are missing some interval conditions. Time interval determined by minimal and maximal element of the reference time set T_R should be congruently inside the time intervals determined by minimal and maximal elements of any given time sets. The truth of the matter is that in this example were the blank time set T_B the set with minimal cardinality $\aleph(T_B) < \varsigma$ and cardinalities of analyte measurements are both equal to the ς. Furthermore, time interval congruent conditions are automatically fulfilled as is clear from the last row of 4.1.

Equations (4.47...4.51) as well as the reference time set T_R are necessary just for the comparison of BBTA with COW, into the bargain. The purpose is to made this example and comparison as illustrative as possible. Hence, all values of analyte time elements t_{a1} and t_{a2} with indexes $a1$ and $a2$ in the range $< 1..\varsigma >$ are replaced by the reference time values:

$$t_{a\lambda} := t_r \mid a\lambda = r, t_{a\lambda} \in T_{A\lambda}, t_r \in T_R, r \in \{1, ..., \varsigma\}, \lambda = \{1, 2\}. \tag{4.52}$$

Previous element values $t_{a\lambda}$ are forgotten. Description in equation (4.52) produces 4.2. All time sets T_{A1}, T_{A2}, T_B and T_R are now identical with also identical cardinality equals to ς. However, the TIC values $\gamma_{A1}(t_r)$ and $\gamma_{A2}(t_r)$ corresponding to the r-th time element t_r still differ to each other $(\gamma_{A1}(t_r) \neq \gamma_{A2}(t_r))$. The TICs did not change during time values replacing process:

$$\gamma_{A\lambda}(t_r) = \gamma_{A\lambda}(t_a) \mid r = a, t_r \in T_R, t_a \in T_{A\lambda}, \ \forall\ r, a \in \{1, ..., \varsigma\}, \lambda = \{1, 2\}. \tag{4.53}$$

Only the position of the TICs in the time axis has changed (4.4B.).

	1	2	3	...	312	313	314
t_{a1}	0.0042	0.1444	0.2265	...	31.8125	31.9277	32.030
t_{a2}	0.0042	0.1444	0.2265	...	31.8125	31.9277	32.030
t_b	0.0042	0.1444	0.2265	...	31.8125	31.9277	32.030

Table 4.2: Time values of blank and analytes set to the reference time set.

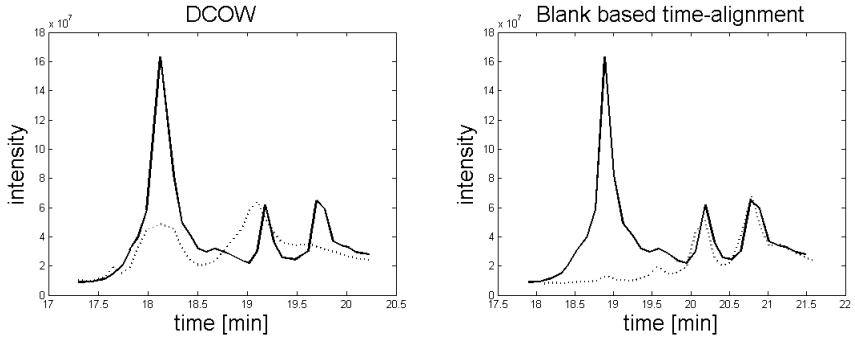


Figure 4.5: Detail of Nystatin part of TICs in DCOW and BBTA. Analyte measurement $A2$ TIC (dotted line) was aligned to the analyte measurement $A1$ TIC (solid line). First of the two peaks after the Nystatin elution in $A2$ is incorrectly aligned to the Nystatin in $A1$ in DCOW.

The COW algorithm aligns one or more data vector(s) onto reference vector via small changes in segments lengths on the data vector(s). Only the TICs values are considered as data vectors. For that reason, join reference time axis is required. Unfortunately, the time or mass sets are not taken into account in the available implementation ([83]). Theoretical possibility of COW for all SICs in the measurements collides with input file limitation. There are over 2000 individual SICs in each measurements $B, A1, A2$. Two tunable parameters are necessary for COW, the number of segments (borders) and maximal increase or decrease of segment length (so-called slack). Optimal values of both parameters are estimated during the computation. The outputs of COW algorithm are aligned TICs $\gamma_{A1}^{`}, \gamma_{A2}^{`}$. Two variants of the COW algorithm were tested. The analyte measurements TICs γ_{A1}, γ_{A2} were aligned to the blank TIC γ_B in the first one (signed simply as COW, Figure 4.4C.). In the second one (signed as DCOW), the analyte measurement TIC γ_{A2} was aligned directly to the analyte measurement TIC γ_{A1} (Figure 4.4D.).

The BBTA algorithm uses the three steps described in section Methods with default settings including automatic segmentation into three semi-equidistant segments and estimation of transformation function as polynomial function of second degree. Both analyte measurements TICs γ_{A1}, γ_{A2} were aligned to the blank TIC γ_B independently. The outputs are aligned time sets $\grave{T_{A1}}, \grave{T_{A2}}$ (Figure 4.4E.).

It is arduous to objectively evaluate the quality of any time-alignment. Comparison of the time values only is misguiding. The values in the Table S-2. are absolutely the same. Nevertheless, the corresponding TICs plots differ evidently (Figure 4.4B.). Another metric is so-called Peak integration error (Chae) defined as:

$$PIE = abs(\frac{area_{aligned} - area_{non-aligned}}{area_{non-aligned}}) \times 100\%, \tag{4.54}$$

where area is considered as integration of peak intensities. Therefore, area evaluation (and precision) is strictly dependent on used peak detection. Without any peak detector could be the area of whole measurement considered as input for equation (4.54), for instance (4.3.). Blank based time-alignment changed only the time sets of the analyte measurements. There are no changes of the TICs values, no changes of the peaks (whatever they are), and no changes of the areas. For these reasons, the PIE is nonsense in this case.

	COW	DCOW	BBTA
reference	γ_B	γ_{A1}	T_R
input data	$\gamma_B, \gamma_{A1}, \gamma_{A2}$	γ_{A1}, γ_{A2}	$B, A1, A2$
output data	$\grave{\gamma_{A1}}, \grave{\gamma_{A2}}$	$\grave{\gamma_{A2}}$	$\grave{T_{A1}}, \grave{T_{A2}}$
segments	84	30	3
slack	1	13	$\emptyset$
time of computation	$\sim$ 3 [min]	$\sim$ 3 [min]	$\sim$ 140 [msec]
PIE	0.32%	0.67%	0.00%

Table 4.3: Comparison of COW, DCOW and BBTA parameters. The main difference is in time of computation.

More objective metric of two similar LC-MS measurements is spectra comparison. A distance between a pair of spectra from two measurements in approximately same time has to be smaller in aligned case than in non-aligned

one. Also the average distance between all spectra pairs (in corresponding time values) has to be smaller for aligned measurements. The only remaining question is the choice of distance evaluation method. It is beyond the scope of this work, to discuss the properties and pertinences of known distance metrics. The results of most common used formulas are shown in 4.4. In all cases are the spectra of BBTA closer together then in the non-aligned measurements. Naturally, optimal distance is equals to zero. However, the presence of random noise excludes the optimality in principal always.

	eucl.	manh.	cos.	corr.	mink.	hamm.	cheb.
NA	5.1×10^6	3.8×10^7	0.17	0.18	5.1×10^6	0.382	3.7×10^6
BBTA	3.4×10^6	3.3×10^7	0.13	0.14	3.4×10^6	0.381	2.2×10^6

Table 4.4: Average computed distance between pairs of spectra in non-aligned (NA) data and blank based time-aligned (BBTA) data. Abbreviation: eucl. - Euclidean distance, manh - Manhattan distance (absolute difference), cos. - one minus angular cosine distance between spectra, corr. - one minus spectra linear correlation, mink. - Minkowski distance (generalization of both eucl. & manh. distance), hamm. - Hamming distance (% values in spectra that are not identical), cheb. - Chebychev distance (maximal difference of values in spectra).

Openly, the distinction between BBTA and COW alignment is quite unfair to the warping. The COW works only with the TICs, not with the whole measurements. However, full COW processing of all SICs exceeds the limits of available algorithm and may causes the mismatch in spectra. Obviously, the SICs can not be aligned to each other, the already pass together. The main problem with warps is more deeper and basic. Time warping is extremely powerful tool looking for parameters that minimize the distance between vectors. Therefore, it assumes that the alignment process is done for the same features that differ only in time duration and noise level. Thus, warp modification could be used as estimation for normalization function parameters as late as Step3, where the input warp features correspond to the time markers. Once again, using time warping directly on TICs confuses the algorithm unavoidably as it is shown on 4.6. On the $2 - nd$ column from the left, it is a part of TICs with Nystatin elution, which was described in Experimental section. The concetration of Nystatin addition differs between analyte measurements $A1$ and $A2$. In COW case, there are analyte TICs aligned

to the blank TIC. Therefore Nystatin can not affect the results in $3 - rd$ row from the top of 4.6. On the other hand, DCOW computes direct alignment of analyte measurement $A2$ TIC (dotted line) to the analyte measurement $A1$ TIC (solid line). As it is shown, one of the two peaks after the Nystatin elution in $A2$ is incorrectly aligned to the Nystatin in $A1$. That is not product of warping inefficiency, that is product of improper input.

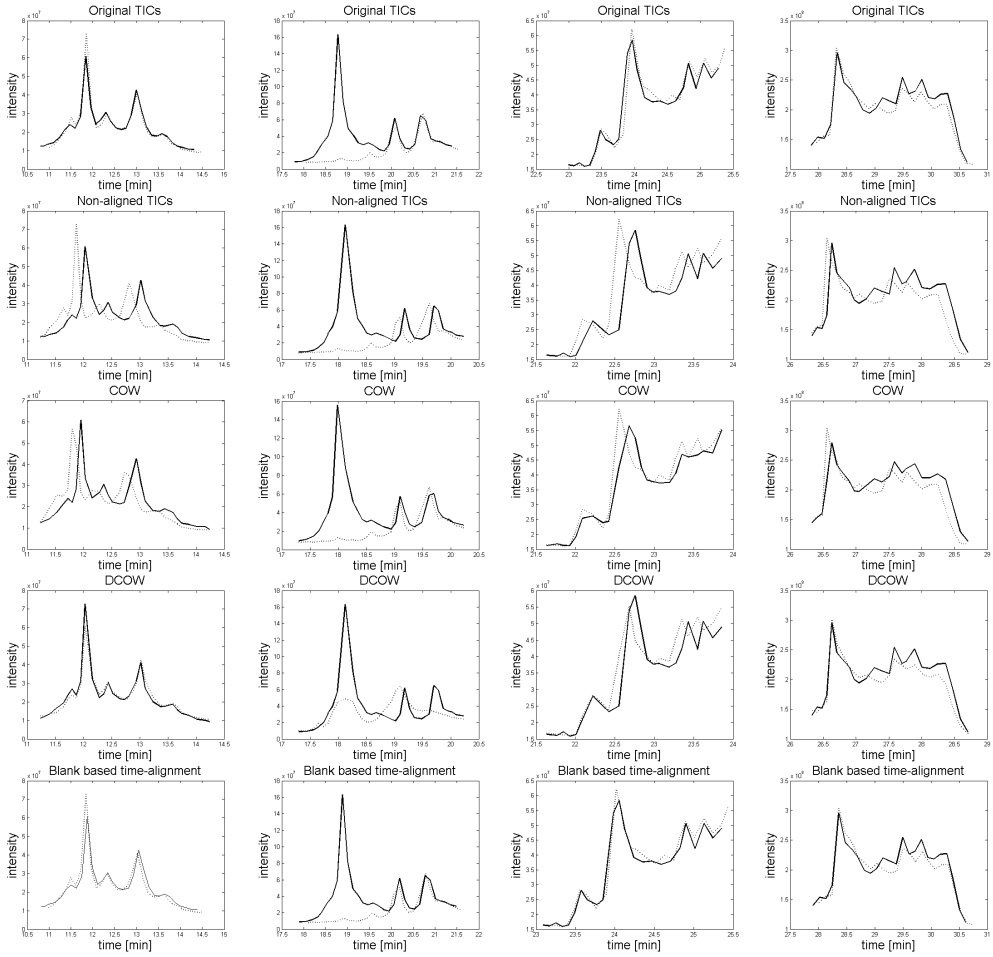


Figure 4.6: Details of several TICs parts (columns). Rows from top to down: original TICs, non-aligned TICs, COW alignment, DCOW alignment and BBTA approach. The results of time alignments were computed on whole measurements. There are visualized only several parts of final plots to enhance differences between approaches.

It is necessary to emphasize the information that the BBTA approach works

not only with the TICs. All markers selection process take into account whole measurement, therefore 3D matrix in time, mass and intensity space. It is also important, that markers selected from blank measurement are not usually significant in analyte measurement TIC, however they are still present in the matrix data. The BBTA approach is powerful enough to align data with simple blanks (with no patterns like peaks) even when the blank is just water (with some a priori unknown impurities) as is shown on 4.7.

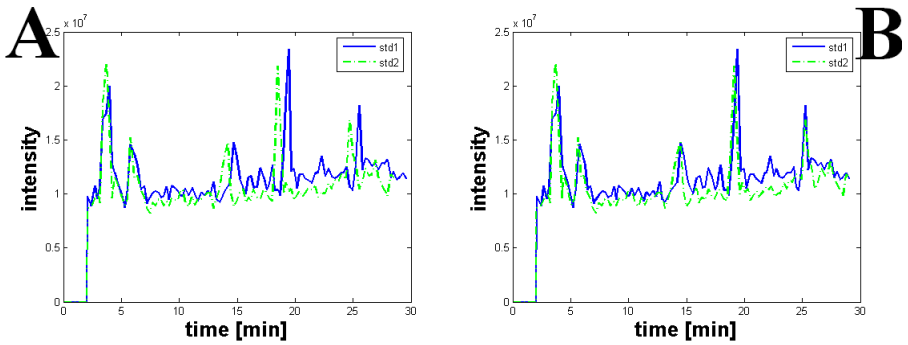


Figure 4.7: Example of two mixture of standards ($std1$ and $std2$). As blank was used the same H_2O as shown on Fig.4.1B without any standards addition. Panel A shows measurements before time alignment, panel B shows measurements after BBTA. Both measurements were aligned only to the blank, therefore there was no computation between $std1$ and $std2$.

In comparison to the advantages of known time alignment methods the BBTA is also opened for extensions. Using blank as internal standard set is not in violation of additional standards. The blank measurement (and therefore the analyte measurements) could easily include addition of compounds estimated by LSERs ([81]). The markers τ pinpointed as relevant inflex points from blank in Step2 are just an optional subset of all eventual markers. For example, robust point matching known as Amsrpm ([92]) is similar to the point of view to the systematic description of the measurements, used in this thesis. Finally, exact analytical and parametric model for transformation function is complicated to define. In the example, in Step3 it is used polynomial of second degree. This primitive function demonstrate the power of blank based time alignment approach in comparison of COW. However, mathematically expressed, the space of function is unlimited as well as criterion evaluation. One of the semi-supervised warps is implemented in

ChromA ([85]). Unfortunately, ChromA is mainly focused on last step of time alignment. The BBTA premise measurements obtained by the same settings and devices. The usage of geometric approach ([79]) is recommended for comparison of different measurements from different devices.

In summary, it is used one of the most primitive normalization function for Step3 in the simple example. Even then, the blank based time alignment results still prove blank usability. Step1 is not crucial for the BBTA approach, it is just for reduce of total time consumption. The main idea is presented in Step2. Selection of time markers with equal cardinalities solves problems with presumption fulfilling. Step3 is only regression analysis question and any algorithm belonging there could be improved. The idea of using blank measurements as internal standard is the main objective - the most simple and direct method for time alignment.

Over and above, IS in sufficient amount will also fulfill this approach. Additional standards in the blank measurement constitute highly significant markers, if they were distinguishable by the column. However, IS addition is just the extension of BBTA. Basically, it is not necessary for the time alignment itself. The common usage is the support for identification. And that is certainly different problem.

All analysis computations were performed in Matlab ([122]) 2008b on Intel CPU Centrino 2 P8600, 2.4 GHz, 4GB RAM.

BBTA is not general for comparison of any two or more measurements, but it is sufficient for measurements from the same chromatographic column with the same gradient settings. Nevertheless, these types of measurements represent everyday laboratory experiments in omics science, petroleum chemistry or pharmacology. One can directly afford the blank based approach, because of simple presumption. The mass values from the blank measurement are also presented in analyte measurement (or easily warranted). Moreover, the time behaviors of the blank mass values are preserved in analyte measurements by the utilized settings. Hypothetically, if some corresponding time inflex point in the measurement was caused by the analyte mass, then the experiment was designed wrongfully. This situation can happen only when the blank mixture contains a compound with identical mass value to the analyte (but with different elution time).

The aspect of transformation function selection requires more consistent

theory. However, it is a question of slightly different brand, especially nonlinear fits, regression analysis or genetic algorithms. This contribution still focused mainly on mechanism of simple, fast and reasonable markers definition from the blank measurement.

Theoretically, BBTA approach may also help to deal with the column aging. Mathematically, it is the problem of estimation of transformation between two or more blanks. When one of them is selected as the reference one, all other steps follow the described methods. Therefore, all analyte measurements could be aligned to the corresponding blank and hereupon aligned to the reference blank time axis. Unfortunately, data collection for column aging will take at least several months for everyday used column and years for rarely used column.

BBTA is a mathematically derived and algorithmically simple approach for time alignment of 2D LC-MS chromatograms which requires blank measurement data. The principle is more objective than many methods described shortly in the 2.3, inexpensive and readily available in any measurement series using the same procedure and devices. Moreover, all measurement spectra are preserved. Exemplificative transformation function could be easily supercede by any advanced estimation.

5 Adaptive filter for baseline thresholding

Exact time characteristics of the systemic noise vary for each mass. It is necessary to analyze the characteristics in every mass independently. The threshold value is the attribute of the measurement, not only an input parameter. Moreover, the baseline characteristic in the blank is not chemically affected by the analysed substances. In other words, the description of the baseline in the analyte has to be generalized description of the baseline in the blank. Therefore, an adaptive thresholding as unsupervised method for baseline removal from measurement data based on statistic moments is considered in this section. Behavior of the baseline content is not perfectly constant in time axis. As is also often necessary for experiments with gradient changes. However, the behavior could be parametrized using technique derived from statistical moments. Results on real analyte measurements are discussed to illustrate the efficiency.

In LC-MS measurement, the compounds elute from the chromatographic column, separated according to the column specific chemical properties in time axis. Output of the LC column enters into the ionization chamber in the mass spectrometry. Molecules are then separated according to the mass and charge (2nd axis) and detected by the MS detector. The thirds axis of measurement data represents intensity, amount of the molecules detected on certain position (section 3.3). Peaks created by compounds separated in time occur only in specific short time interval. On the other hand, the mobile phase, that carry the compounds through the column is therefore present in longer part of the measurement. In the blank measurement, presence of the baseline as dominant part of the measurements is expected. However, there are also random spikes (random errors, random noise)

and impurities peaks (chemical noise) as is shown on Figure 5.1A. The intensity scale range suppress contrast between m/z values of low and high occurrence. It is advisable to change the scale of intensity axis to non-linear (i.e. by logarithm) to increase the information visibility. On Figure 5.1B, it is observable that some masses are present during the whole measurement only with small changes in time axis. Log scale in 3D graph allows to illustrate the flows as emerging pattern with simple structure. The blank measurement gives there the opportunity to examine the description that separate the baseline from the peaks and random spikes. The baseline presence is similar in the analyte measurements as well as in blank. However, the characteristic is more hidden under analytes affection (presence of the analyte peak influent the systemic noise, or systemic noise influent the analyte, it is just the point of view). The threshold value that separates baseline signals from the analytes is derived from statistical parameters of the whole measurement.

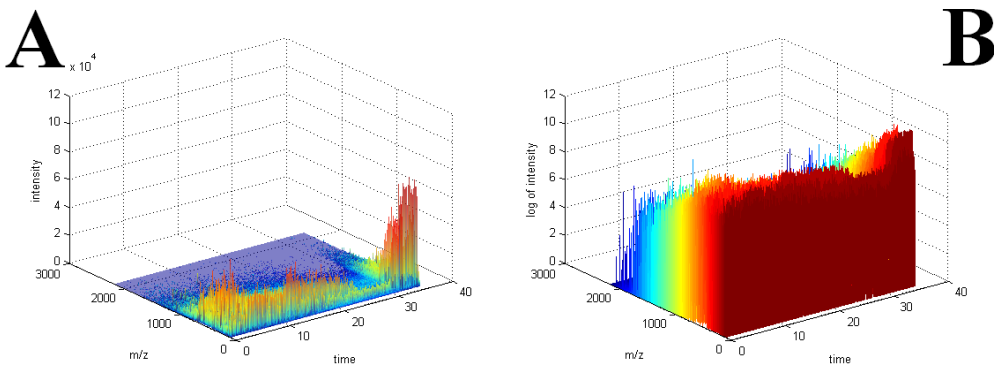


Figure 5.1: 3D Example of blank measurements in decimal (A) and logarithmic (B) scale.

5.1 Theory and calculation

The hypothesis arises from the following knowledge: analyte masses create peaks in time axis. Therefore, analytes have rapid increase of maximal signal above average signal. On the other hands, the baseline has just slow increase or decrease of signal. One of the standard methods for automated data processing

and observation is max-to-mean ratio R [99, 98, 128]:

$$R(m_i) = \mathcal{X}_Y(m_i)/\mu_Y(m_i), \tag{5.1}$$

where m_i $[m/z]$ is the $i-th$ mass of the measurement, $\mathcal{X}_Y(m_i)$ is maximal intensity of the $i-th$ mass and $\mu_Y(m_i)$ is mean intensity of the $i-th$ mass.

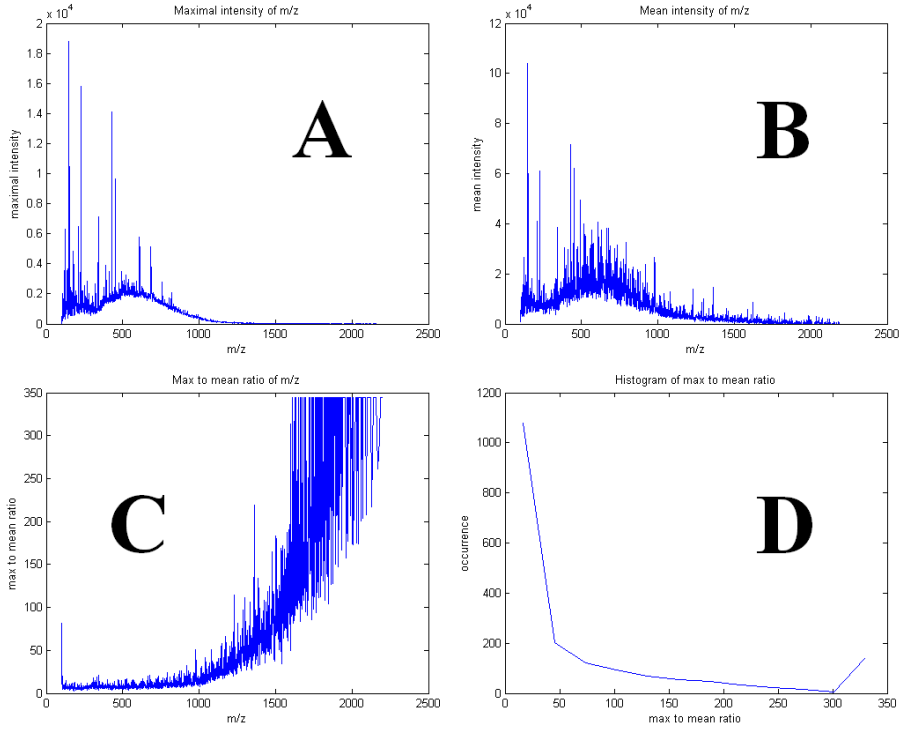


Figure 5.2: Example of mass signal max values (A), mass signal mean values (B), mass signal max-to-mean ratio (C) and max-to-mean ratio histogram in continuous representation (D) evokes heavy tailed distribution.

Max-to-mean ratio is still just a vector of values paired with the m_i. Any thresholding requires further pre-processing like smoothing or binning to reduce the effects of minor observation errors. The number of binning intervals bi is estimated via Sturges rule [101]:

$$bi = \lfloor 1.5 + 3.3 \times \log_{10}(\aleph(R)) \rfloor, \tag{5.2}$$

where $\aleph(R)$ is the amount of max-to-mean ratio values. Subsequently, the max-

to-mean ratio R is binned into histogram h according to the estimated number bi of intervals n. The most important value from the histogram is then the position p, where is the maximal occurrence in the histogram plot. This position p is equal to the average value of the max-to-mean ratio in normal distributions of R. However, the distribution in real samples is more or less shifted and skewed and also with heavy tails. The position of p is a priori unknown and identification of the distribution is nontrivial task. The suggestion is to approximate the standard deviation (positive square root of the second central moment) by weighted function of interval difference (Equation 5.4.).

The standard deviation for discrete variables x is defined as

$$\sqrt{\sum_{k=1}^{n}[x_k - E(x)]^2 \times p_k}, \tag{5.3}$$

where $E(x)$ is mean value of variable x and p_k is probability that variable x obtain value x_k.

Therefore, the histogram h represents the probability (non-normalized to unity) of the max-to-mean ratio R. Then, the position n is average value of interval defined by Sturges rule. The influence of binning simplification is approximated by the heuristic [129] constant shift:

$$s = \sqrt{\frac{1}{\aleph(R) - 1}\sum_{k=1}^{k=bi}[(n(k) - p)^2 \times h(k)]} + \frac{\pi}{2}. \tag{5.4}$$

Threshold value Th for max-to-mean ratio R depends on relation between the p value and approximated standard deviation as is shown in Table 5.1. It adapts to the actual position p via approximated standard deviation s as the compensation of the histogram h shift from the ideal normal distribution.

if p	$> 3s$	$> 2.576s$	$> 1.96s$	$> 1.645s$	$> 0.674s$	$\geqq \lvert p - s \rvert$	$< \lvert p - s \rvert$
Th	$p - 3s$	$> p - 2.576s$	$p - 1.96s$	$p - 1.645s$	$p - 0.674s$	$\lvert p - s \rvert$	p

Table 5.1: Description of adaptive threshold Th selection. Corresponding threshold Th is set to the p higher then expressions in the first row.

5.2 Results of baseline filtration

Computed threshold value is used for thresholding measurement mass values m/z in max-to-mean ratio R domain. In other words, the contribution of all m/z values with ratio R below threshold Th is classified as baseline. Those masses are then removed from the measurement. Adaptive thresholding is very fast analytical method. Total time of computation is about 0.03[sec] (CPU P8600, 2.4GHz, 4GB RAM).

Proposed adaptive thresholding removes baseline affection of individual m/z values according to their behavior. This method is independent on absolute intensity values. In other words, there is no fixed threshold on intensity levels. That is nontrivial property which causes two eminent features. At first, even the high intensity masses may be evaluated as baseline contribution, if present. And at second, some small peaks, hidden in the noise, will arise after filtration process as is shown in Figure 5.3C.

There is still open the question of max-to-mean ratio histogram fitting. Of course, proper identification of probability density function produces exact values of relevant central statistical moments instead of any approximation. However, the task of function type specification [107, 108, 109] it the most puzzling issue. Unfortunately, any analysis can only help in differentiating between hypothesis or models [115]. Very strong results still do not prove that the correct function was chosen [116]. Proposed approach of adaptive threshold method for baseline removal in LC-MS measurement is focused on statistical approximation of systemic noise contribution. On a real example, filtration of the m/z values belonging to the baseline was illustrated. However, mathematically stronger approach is intuitively developed in the next chapter.

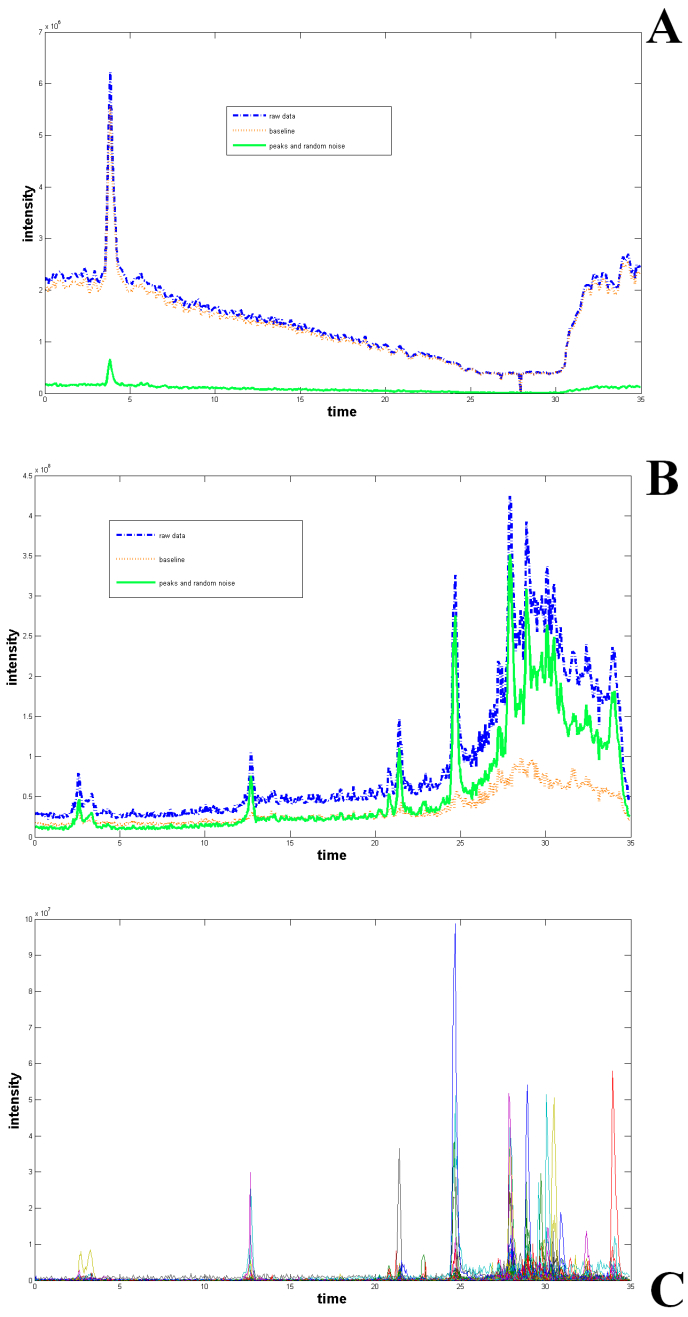


Figure 5.3: Examples of filtered blank (A) and analyte (B) measurement TICs. Dash dotted line represents original raw data, doted line is for removed baseline and solid line are the remaining peaks with random noise. Overlay of all peaks of analyte measurement after filtration are illustrated on panel C.

6 Noise filtration via probabilistic theory

Liquid chromatography with mass spectrometry (LC-MS) detection is one of the major tools in proteomics and metabolomics. Metabolite transformation by protein enzymes and protein- and lipid-mediated signal transduction are elements of the pathways responsible for the non-linear dynamics of living cells. The goal of experiments in metabolomics and proteomics is to identify the molecule (or its fragment) and quantify its amount, at the best inside the cell or in a representative sample of the culture.

The experiment on determination of protein or metabolite concentration in the sample consist in most cases of steps: 1) of sample collection, 2) set of physiochemical and mechanical operation such as filtration, extraction etc., 3) chemical or biochemical operations such as chemical modification or enzyme cleavage, 4) separation of chemical entities by chromatography, various electrochemical methods etc. and detection. In this section are discussed the information content of LC-MS measurement which combines chromatographic separation and detection of the compound identity and quantity.

In LC-MS the compounds are chromatographically separated on the basis of their physicochemical interactions with the chromatographic column material in particular solvent system. All separated compounds come to the ion source and where they are ionized. Resulting ions are resolved by (various types) of interactions with electromagnetic field. The signal is then detected by detectors of various constructions. Naturally, the various technical setups give rise to various detailed relations between the nature of the analysed compound - analyte, ultimately its chemical structure, and the signal measured at the detector. Yet, the

application of general systems theory enable to analyse this, in fact, whole class of experiments, in one natural generic way.

A novel method is proposed, based on system approach to LC-MS data analysis. Because the presented noises have to correspond to some probabilistic distributions, it is possible to approximate the distributions and identify the parameters. Using this parameters helps to describe measurement more accurately. This information could be used in addition for filtration or for further analysis process. In this section, LC-MS measurement data output is described according to the system approach point of view. Appropriate distributions of random noise as well as systemic noise are selected and parameters identification is described. Identified characteristics are used for probability factor evaluation. Two principles of using of this probability information are illustrated on real measurements.

In this section is therefore demonstrated mathematical definition of the problem of reliability of a LC-MS experiment, discussed the objective definition of the origin and character of noise. Probabilistic approach is to correctly describe the space created by chosen metabolite profile analysis (LC-MS) according to theory of systems, where the term system means mathematical description of measurement space. However, this approach is not confined to a particular experiment, it represents a rather more precise description of measurement results generally. Above all, this implement errors presence into the description. Thus, it put the question of data interpretation according to probability theory as well as it try to answer that question.

6.1 Probabilistic approach

Unfortunately, every measurement output data has its errors. Two basic principals of error occurrence in LC-MS have to be mentioned. The first one, is known as base line, sometimes called systemic noise, and in LC-MS is produced by presence of mobile phase, that carry analytes through the LC column into MS. The effect of the mobile phase is dependent on measurement device (LC), its setup and mobile phase composition. During the analysis of the measurement, it is necessary to keep in mind that in measurement output data the base line influence is always present. The other one, called random noise, includes all unwanted sources of

transient disturbances and it is always present too. Both of them affect the signal transparency and can be also described according to the theory of systems.

Random noise (r) can be described as mapping

$$r : T \times M \rightarrow \cup_{t \in T, m \in M} \mid r(t, m) \in I, I = 0, 1, \ldots, i_{max}, \tag{6.1}$$

and base line can be described formally in the same way marked by sign $b(t, m)$:

$$q : T \times M \rightarrow \cup_{t \in T, m \in M} | q(t, m) \in I, I = 0, 1, ..., i_{max}, \tag{6.2}$$

The conditions for random noise and base line mappings are equal to conditions for mapping of signal generating process. If one was able to define the generating process as mapping as well as noise then could be also defined mapping of real signal of pure analytes (signal) contribution $s(t, m)$. As was shown in *eq.* 3.3, the mapping in $t \in T$ defines the state. Consequently, the relation between mappings for random noise (r), base line (q) and signal (s) in $t \in T$ is

$$y(t) = s(t) + q(t) + r(t). \tag{6.3}$$

Common object of interest is description of $s(t)$ to reduce influence of presented noise, which can produce false peak or hide signal under reasonable level. Precise contributions of noise is unknown because of stochastic but the characteristics may be estimated via probabilistic analysis. Afterwords, with the knowledge of noise characteristic and measurement output data can be also estimated the signal state in time $t \in T$:

$$\tilde{s}(t) = y(t) - \tilde{q}(t) - \tilde{r}(t). \tag{6.4}$$

For values of estimation $\tilde{q}(t)$ and/or $\tilde{s}(t)$ are demanded to be the natural numbers or zero, because there is not measured negative amount of molecules in the MS detector. The signal is present or not and the contribution of base line is positive (mobile phase elute during whole measurement). Problematic is estimation of random noise $r(t)$. Because sometimes the sum of $||q(t)||$ and noise estimation $||r(t)||$ could be greater than measured $\gamma(t)$ and the result of estimation

$||\tilde{s}(t)||$ will be negative number. Simple solution of this problem is to set the negative value to zero. Basically, estimations of exact values of the noise may not be accurate even if the characteristic of noise could be estimated well.

Consequently, quantitative error and two kind of qualitative errors could be made. In quantitative case, this is the common error of the estimation solutions, the estimated value of analyte intensity $\tilde{s}(t, m)$ differs to real but unknown value $s(t, m)$. The qualitative errors of estimation are the same as in another detection tasks or in hypothesis testing. They are known as false reject (false negative in some literature) and false alarm (false positive, false accept) ([130]).

False reject happens where the analyte is present but $\tilde{s}(t, m)$ is equals to zero. On the other hand, false alarm means positive value of $\tilde{s}(t, m)$ when the analyte is not present. Generally, the quantity is not precise because of random noise and none processing of the data causes no false reject but remains all of false alarms. To reduce quantitative errors, it is advisable to replicate the measurement of the same sample many times. Therefore, it is not the key problem of analysis of single one measurement. But the estimation of exact values will produce the qualitative errors. It can decrease the false alarms but increase the false rejects. The most of filtration methods is designed to decrease the false alarms. The optimal rate between false alarms and false rejects is nontrivial question. In metabolomics research, several tasks vary in sensitivity to this qualitative errors. For example, false rejects in poison detection may cause wrong interpretation more frequently then false alarms of unknown analyte spike occurance. From this point of view, error ratio based directly on given task is more suitable.

The error ratio produced by filtration algorithms could be tuned via some parameters, but the relation between them is generally not evident. Especially when there are several steps in the filtration which can be tuned independently.

In this section another approach is proposed. Instead of value estimation of signal intensities $\tilde{s}(t, m)$ which is errorfull, it evaluate probability factor $p(t, m)$ that the measurement output data $y(t, m)$ is signal $s(t, m)$:

$$p(t, m) = p\left[y(t, m) = s(t, m)|\lambda_q, \lambda_r\right], \tag{6.5}$$

where λ_q and λ_r is estimated characteristic of mapping $q(t, m)$ and mapping $r(t, m)$

respectively. The *probability factor* $p(t, m)$ means probability that analyte with *molecular mass* m in *retention time* t has *intensity* $y(t, m)$. Probability $p(t, m)$ is multiplication of two independent probabilities ([89]). The first one is probability $p_r(t, m)$ that measurement data output $y(t, m)$ is not produced by random noise $r(t, m)$. The second one is the probability $p_q(t, m)$ that measurement data output $y(t, m)$ is not produced by systematic noise $q(t, m)$. And the final probability $p(t, m)$ is

$$p(t, m) = p_r(t, m) \times p_q(t, m). \tag{6.6}$$

With probability factor $p(t, m)$ which can be evaluated precisely with good noise characteristic, the error ratio can be tuned directly for any task. Subsequent filtration and/or analysing steps can propagate this probability to its outputs via probability theory formulas.

6.2 Estimation of random noise characteristics

In LC-MS measurement, it is considered as *random noise*, any unwanted influence during measuring process which causes imprecise equality of measured data output $y(t, m)$ to the analyte intensity $s(t, m)$ and it is not the contribution of mobile phase. Sources of *random noise* could be small substances eluted from stationary phase in LC column, impurities of the mobile phase, ionisation disturbances and short term variation in signal intensity on MS detector [73]. Increasing eluted amount from LC column also increases possibility to error occurrence. Therefore, characteristics of *random noise* vary in every *retention time*. It is necessary to analyse the characteristics in every *retention time* independently. Thus, it is assumed fixed *retention time* and as the region of interest becomes only actual mass spectrum.

This section is focused on random noise mapping characteristic estimation and its using to evaluate probability factor. Random noise during measuring process is considered as Gaussian probability distribution function (PDF) with differentiate in statistic order moments (median, variance, skewness, etc.). This is implicated directly from the common device feature known as sensitivity. From physical point of view, it is impossible to develop measurement device able to

measure precisely in whole range of values $w \in W = w_0, w_1, w_2, \ldots, w_{max}$. It produce changes not only in accuracy in values $w \in W$ close to the range borders w_0 and w_{max}, which is often considered, but also means that the possibility for values to be measured is not constant in the detectable range (3.4). There are several ways how to deal with the limitations. One of the suboptimal solutions is to separate the whole range W into higher number of intervals. Of course, for correct interpretation of the measurement data output is demand to know the PDF. If measuring of 'nothing', no given reasonable input signal, produce own PDF, then the signal, value of given mass in given time, is just a disturbance in noise PDF.

Range of detectable *molecular mass* is wide and a typical mass spectrum, produced by LC-MS, contains a few bars of 'high' *intensity* and a lot of 'small' ones. It is advisable to reduce intensity range by a compression function, like logarithm. Only positive *intensity* values are taken into account and logarithmic domain is assumed below

$$ly(m) = ln\left[y(m)\right]. \tag{6.7}$$

MS detector sensitivity is not strictly constant for various *molecular masses* (m/z) and should be normalized. Several methods to reach normalization function are possible. For example, sensitivity characteristics can be obtained by smoothing of mass spectrum with low-pass filter.

Probability $p_r(m)$ is evaluated as

$$p_r(m) = \frac{p\left[ly(m)|\lambda_{ls+lq}\right] p(s+q)}{p\left[ly(m)|\lambda_{lr}\right] p(r) + p\left[ly(m)|\lambda_{ls+lq}\right] p(s+q)}, \tag{6.8}$$

where $p(s+q)$ is a priory probability of analyte and *mobile phase* occurrence, $p(r)$ is a priory probability of *random noise* occurrence and sum of $p(s+q)$ and $p(r)$ is equals to one. λ_{ls+lq} is characteristic of sum of analyte *intensity* and *systematic noise* in logarithmic domain, λ_{lr} is characteristic of *random noise* in logarithmic domain. Those characteristics are a priory unknown but may be estimated from normalized logarithmed measurement data output. After analysis of data histograms, Normal distribution for *random noise* and shifted Rayleigh distribution for sum of analyte *intensities* and *mobile phase* as appropriate approx-

imations were chosen. Characteristic parameters of *random noise* distribution are mean value μ_r and standard deviation σ_r values of whole logarithm mass spectrum. Usually, there is a lot of small 'noisy' bars of impurities around main analyte bars present in mass spectrum. Therefore, analyte *intensities* are only disturbances in normal distribution and in logarithmic domain one can assume that

$$\mu_r \left[lr(m) + ls(m)\right] \cong \mu \left[lr(m)\right] \tag{6.9}$$

For λ_{ls+lq} distribution, its variance parameter is $4\sigma_r^2$ and offset is $\mu_r + \sigma_r$.

6.3 Estimation of systematic noise characteristics

For analyse characteristics of mobile phase contribution, helpful advantage is at disposal. Measurement with no analyte, but under same condition as measurement of analytes, called *blank* could be done. This measurement can not produce exact values of mobile phase *intensities* because of random noise and other disturbances (e.g. chemical influence, ionisation nonlinearities). But the *blank* measurement is a valuable information for analysis of its characteristics. It also can be estimated without the *blank* measurement but with higher error level. It gives the *molecular masses* presented in mobile phase, that are detectable by MS. Moreover, the run in *retention time* is available for every *molecular mass*. This time-run for every *molecular mass* which is present in the *blank* is analyzed independently. Thus, here is assumed fixed *molecular mass* and as the region of interest become only single one time-run below. Then, the approach of analyse is formally the same as in the *random noise* case. The logarithmic compression function

$$ly(t) = ln\left[y(t)\right], \tag{6.10}$$

is used to transform *blank* data output as well as further measurement with analytes under the same conditions. Mean value μ_q and standard deviation σ_q of *ly(t)* is computed as parameters of Normal distribution, which is used for esti-

mation of *systematic noise* probability $p_q(t)$:

$$p_q(t) = \frac{p\left[ly(t)|\lambda_{ls+lr}\right] p(s+r)}{p\left[ly(t)|\lambda_{lq}\right] p(q) + p\left[ly(t)|\lambda_{ls+lr}\right] p(s+r)}, \tag{6.11}$$

where $p(s+r)$ is a priory probability of analyte with *random noise* occurrence, $p(q)$ is a priory probability of *systematic noise* occurrence and sum of $p(s+r)$ and $p(q)$ is equals to one. λ_{ls+lr} is characteristic of sum of analyte *intensity* and *random noise* in logarithmic domain, λ_{lq} is characteristic of *systematic noise* in logarithmic domain. λ_{ls+lr} is shifted Rayleigh distribution, its parameters are evaluated in the same way as in previous section: $4\sigma_q^2$ and offset is $\mu_q + \sigma_q$.

Probability $p(t, m)$ is then evaluated with estimations of both noises via equation 6.6.

6.4 Advantages of probabilistic approach

Understanding to the measurement is more straightforward according to the estimated probability $p(t, m)$ because this information is available for all $y(t, m)$. Therefore, there is only one parameter which characterizes quality of the measurement data output during interpretation itself. No other parameters like SNR or intensity levels in *blank* need to be evaluated and tuned. In praxis, there are basically two principles how to deal with this probability information.

The first one, a fixed threshold value Th can be tuned for any step of further output analysis. For example, $Th = 0.5$ means that all intensity values $y(t, m)$ with probability $p(t, m)$ lower then 50% will be ignored. When the higher Th is set, the total number of credible data points decreases as well as number of *false alarms* but possibility of *false rejects* occurrence increases. When the lower Th is used the oposite situation happens, naturally.

The second principle is to use whole probabilistic information in further output analysis. This case is more advisable because no part of measurement data output and no probabilistic information are discarded. Of course, all analysis steps have to support processing of uncertain data characterized by probability values $p(t, m)$. Unfortunately, the most of common analysis algorithms assume exact data only although no real data are accurate and noise-free.

6.5 Probabilistic filtration of Nystatin in the Nostoc sp. extract

For clear illustration of the probabilistic approach and two principles how to deal with $p(t, m)$ information, a simple correlation between three measured spectra of known substance has been done. Well known antifungal drug Nystatin [131] was selected (Formula $C_{47}H_{75}NO_{17}$, mol. mass 926.09, structure is depicted in Figure 6.1). Every spectrum was selected from individual measurement. The first measurement was pure analyte in concentration $0.5mg/ml$ which was taken as a *Reference*. The second measurement (marked as *Pure*) was again pure analyte but in very low concentration $0.5\mu g/ml$. In both measurements, the noise level is similar but SNR in much more different because of different concentrations. The third measurement (marked as *Mix*) was mixture of Nystatin in concentration $0.05mg/ml$ and 70% MeOH extract from cyanobacteria Nostoc sp. This measurement simulated real conditions of unknown analyte detection. Analysed spectra were selected from retention time where the analyte intensity reached the highest value. All measurements were analysed by the probabilistic approach and corresponding $p(t, m)$ were computed. Examined spectra and their probabilities are shown in Figure 6.2.

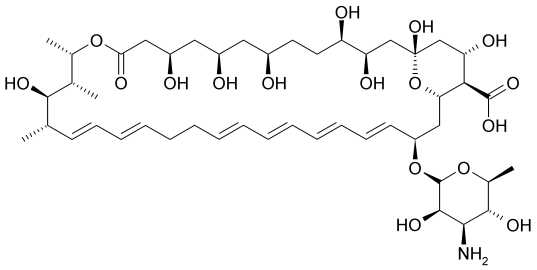


Figure 6.1: Chemical structure of Nystatin molecule.

Because there are two principles how to deal with probability information and some basic approaches also, the correlation criterion proposed by Vaněk ([132])

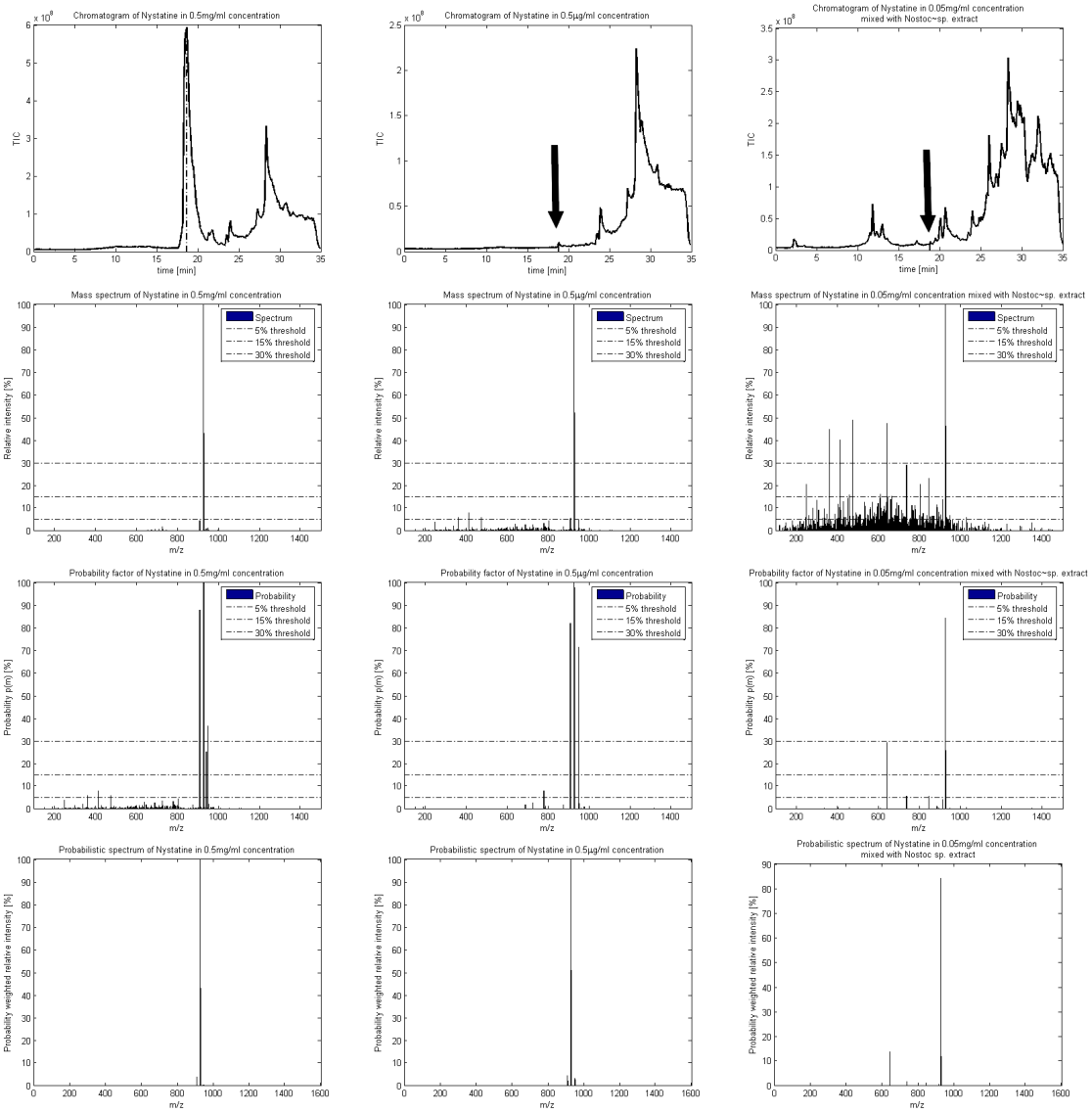


Figure 6.2: Chromatograms, spectra of Nystatin, probability factors and probability weighted spectra for all examined examples with thresholds illustration. The first row shows TIC chromatograms of source measurements (Nystatin peak is marked). The second row shows original Nystatin spectra in maximum peak time point. The third row shows evaluated probability factors of spectra from row two. The last row shows illustration of the spectra intensity according to the probability factors.

was used to show the differences:

$$R(Y_1, Y_2) = \frac{C(Y_1, Y_2)}{\sqrt{C(Y_1, Y_1)C(Y_2, Y_2)}}, \tag{6.12}$$

where Y_1 means reference spectrum and Y_2 means filtered spectrum. $C(Y_1, Y_2)$ is covariance defined as

$$C(Y_1, Y_2) = \frac{1}{M-1} \sum_{m=1}^{M} [y_1(m) - \mu_1][y_2(m) - \mu_2], \tag{6.13}$$

where $y_1(m)$ and $y_2(m)$ are intensity values of m-th molecular mass of spectra Y_1 and Y_2, respectively. μ_1 and μ_2 are average values of spectra intensities.

Four different measurement filtrations were examined using correlation criterion. The first one was original unfiltered data. In the second case, fixed relative intensity thresholds were used. For example, three various thresholds were evaluated as 5%, 15% and 30% of maximal intensity in spectrum. In the third case, the first principle of probabilistic approach was applied. For another example, the same three thresholds were evaluated but as limiting values of probabilty $p(t, m)$. The last case shows the second principle of probabilistic approach. The information about $p(t, m)$ is directly used during the correlation evaluation. The correlation weighted by $p(t, m)$ is defined as

$$R_p(Y_1, Y_2, P) = \frac{C_p(Y_1, Y_2, P)}{\sqrt{C_p(Y_1, Y_1, P)C_p(Y_2, Y_2, P)}}, \tag{6.14}$$

where P is $p(t, m)$ in retention time t and $C_p(Y_1, Y_2, P)$ is weighted covariance

$$C_p(Y_1, Y_2, P) = \frac{\sum_{m=1}^{M} p(m)[y_1(m) - \mu_1][y_2(m) - \mu_2]}{(M-1)\sum_{m=1}^{M} p(m)}. \tag{6.15}$$

All results of correlations of *Pure* and *Mix* with *Reference* are in Table 6.1. The correlation between reference spectrum and *Unfiltered Pure* spectrum is pretty high because in both measurements were presented only analyte alone. The SNR in *Pure* was lower then in *Reference* because of different concentration thus the correlation is 98.37%. In simulated real-like sample *Mix*, significantly higher noise

Table 6.1: Spectra correlation of *Pure* and *Mix* with *Reference*

$R(Y_1, Y_2)$	*Pure*	*Mix*
Unfiltered	98.37%	59.03%
Fix. Th. 5%	99.11%	60.37%
Fix. Th. 15%	99.54%	69.64%
Fix. Th. 30%	98.52%	76.23%
Fix. Prob. 5%	99.50%	84.45%
Fix. Prob. 15%	99.54%	90.75%
Fix. Prob. 30%	99.54%	90.68%
Prob.Corr.	99.67%	95.93%

level is presented and decreases the correlation to 59.03%. The results of various fixed relative intensity thresholds are unstable. A higher thresholds are better for *Mix* but not for *Pure*. The correlation is very sensitive to proper threshold value. In this case, the balance between *false rejects* and *false alarms* is difficult to tune because filtration results are unpredictable. In first principle of probabilistic approach, both results and balance between *false rejects* and *false alarms* are more stable. In addition, the correlation criterion is significantly higher especialy in *Mix* sample. It is produced by more objective noise characterisation. In the second principle of probabilistic approach, the correlation results are even better then in the first principle because no information was lost. Question about *false rejects* and *false alarms* is more complicated because no detected ions are discarded. Instead of it, probabilities $p(t, m)$ of true detections are still stored for further processing (see Figure 6.2).

In this section was proposed a probabilistic approach to analyse LC-MS measurement. This approach is focused on proper characterisation of presented noise. Noise produced by mobile phase is characterised separately to random noise contribution. Information about the both of noise characterisations were integrated into probability factor. Further, two principles of using the probability information were discussed. On a simple example was illustrated advantage of the probabilistic approach. Performance between the two principles and between classical fixed threshold approach was compared.

Recently, was published ([89]) this information-based approach for extrac-

tion of spectra of LC-MS data. There are reliable detect peaks, random and systematic noise (ridges) and store them and their statistical properties. Apart from electrical spikes, the whole spectra may be reconstructed from resulting dataset without loss of existent information. Certainly it rely on accepted model of LC-MS process, but are already introduced many amendments to it which can only make the model compatible with available data.

7 Conclusion

In this thesis was introduced the system based approach to the description and processing of LC-MS measurement data. The abstract model was constructed according to the system theory. Thus, definitions of attributes and their sets of variables are consistent and explicit for all processing/analysis steps as well as mapped Cartesian product(s).

In the introduction, hypothesis was assumed in which the raw measurement data output of LC-MS consist of three partial contributions, the analyte signal, the random noise and the systemic noise. In LC-MS there are also spike signals of several mass values in time axis (in SICs). They can be considered as random noise in the time or may represent the effect of Shannon-Nyquist-Kotelnikov aliasing. The determination of spikes origin requires construction of complex experiments which would be difficult to interpret, or even impossible.

The separation process of the data parts (signal and noises) could be estimated using the probabilistic approach. The verification process was described on example in the last chapter and in the Appendix A.

Also, the current state of data handling was studied. All the main types of processing and most popular methods were mentioned. However, the overview could not be exhausting and finite. Therefore, some literature was recommended for additional details.

The move toward practical part of this thesis was introduced by the construction steps of abstract model. Main attributes and properties of the system (mappings and conditional attributes) were described in a proper math-

ematical space. The rest of the thesis works with defined system of LC-MS measurements in certain level.

Blank based time alignment was introduced for LC-MS measurements obtained under the same conditions, as is usual in metabolomics. This approach was tested during several experiments and it is implemented in software which was developed (see Appendix D).

Tools for noise behavior estimation via probabilistic theory were described in the last section. Computation of probability that the obtained signal is a signal or one of the noises (random or systemic) allows to separate the measurement into parts proposed in the introduction. This approach is able to pinpoint the signal of low intensity hidden in the noise. Moreover, it can be tuned by the operator which level of probability worth for his consideration.

The probabilistic approach implementation as Matlab runtime application was published in the Bioinformatics Journal and it is now in use in the Department of the Phototropic microorganisms of the Institute of Microbiology of the Academy of Science of the Czech republic in Třeboň, where helps with analysis of measured extracts from green algae and cyanobacterias.

'It is also clear that there are still many opportunities for algorithmic development' ([11]). The resolwing power requires theory of direct determination, while IUPAC peak valley definition fails in low-res measurements. There is still no way how to fit exactly the function consist of two different PDFs, even the one is usually estimated from the small set of well known functions. General mathematical definition of the peak does not exist, it is usually aproximated by the Gaussian or its derivation, often ignoring fluctuations or tailings. Algorithmization of chemical rules for fragmentation and adducts has problems with huge order of possible permutations. Useful metric for spectra comparison is also missing.

Bibliography

[1] Ferrell, J. E. Jr., Question & Answer: Systems biology, Journal of Biology, vol. 8., Article 2., 2009.

[2] Field,D., Sansone,S.A. (2006) A Special Issue on Data Standards. OMICS: A Journal of Integrative Biology, Vol.10 No.2, 84-93.

[3] Goodacre, R., Baker, J. D., Beger, R., Broadhurst, D., Craig, G. C. A., Kell, D., Manetti, B. K. C., Newton, J., Paternostro, G., Sjöström, M., Smilde, A., Trygg, J. and Wulfert, F. (2007) Data analysis standards in metabolomics, Metabolomics, 3.

[4] Shulaev, V. (2006) Metabolomics technology and bioinformatics. Briefings in Bioinformatics, 7, 128139.

[5] Hodgman, C. (2007) Integrative biology–the way forward., Brief Bioinform. Jul;8(4):208-9. Epub 2007 Jul 18.

[6] Thieffry, D. (2007) Dynamical roles of biological regulatory circuits, Briefings in Bioinformatics 2007 8(4):220-225;

[7] Mjolsness, E. (2007) Towards a calculus of biomolecular complexes at equilibrium, Briefings in Bioinformatics 2007 8(4):226-233;

[8] Burrage, K., Hancock, J., Leier, A., Nicolau, D. V. Jr. (2007): Modelling and simulation techniques for Membrane Biology, Briefings in Bioinformatics, 8(4): 234-244, July 2007.

[9] Fiehn, O., Robertson, D., Griffin, J., van der Werf, M., Nikolau,B., Morrison, N., W. Sumner, L., Goodacre, R., Hardy, N. W., Taylor, Ch., Fostel, J.,

Kristal, B., Kaddurah-Daouk, R., Mendes, P., van Ommen, B., Lindon, J. C. and Sansone, S.-A. (2007) The metabolomics standards initiative (MSI),Metabolomics, Volume 3, Number 3.

[10] Sumner, LW., Urbanczyk-Wochniak E., Broeckling C.D. (2007) Metabolomics data analysis, visualization, and integration, Methods Mol Biol. 2007;406:409-36.

[11] Wishart, D. S., Current progress in computational metabolomics, Briefings in Bioinformatics, vol. 8. no.5., 279-293, 2007.

[12] Varmuza, K., Steiner, I., Glinsner, T., Klein, H., Chemometric evaluation of concentration profiles from compounds relevant in beer ageing, Eur Food Res Technol, 215:235-239, 2002.

[13] Beránek, L., Knížek, J., Pulpán, Z., Hubálek, M., Novák, V., Mathematical simulation of mass spectrum, Technical Computing Prague, 2005.

[14] Stolt, R., Torgrip, R. J. O., Lindberg, J., Csenki, L., Kolmert, J., Schuppe-Koistinen, I., Jacobsson, S. P., Second-order peak detection for multicomponent high-resolution LC/MS data, Anal. Chem. 78, 975-983, 2006.

[15] Christie, W. W., Gas chromatography and lipids, A Practical Guide, The Oily Press Ltd, Ayr, 1989.

[16] Rizov, I., Doulis, A., Separation of plant membrane lipids by multiple solid-phase extraction, Journal of Chromatography A, 992, p.347-354, 2001.

[17] Schwudke, D., Oegema, J., Burton, L., Entchev, E., Hannich, T., Ejsing, C. S., Kurzchalia, T., Shevchenko, A., Lipid profiling by multiple precursor and natural loss scanning driven by the data-dependent acquisition, Anal. Chem. 78, 585-595, 2006.

[18] Ramakrishnan, S. R., Mao, R., Nakorchevskiy, A. A., Prince, J. T., Willard, W. S., Xu, W., Marcotte, E. M., Miranker, D. P., A fast coarse filtering method for protein identification by mass spectrometry, Bioinformatics, doi:10.1093/bioinformatics/btl118 , 2006.

[19] Brent, R., Bruck, J., Can computers help to explain biology?, Nature, vol 440, 2006.

[20] Feist, A. M., Scholten, J. CM., Palsson, B. Ø., Brockman, F. J., Ideker, T., Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri, Molecular Systems Biology, 1744-4292, 2006.

[21] Andreev, V. P. Rejtar, T., Chen, H.-S., Moscovets, E. V., Ivanov, A. R., Karger, B. L., A new algorithm for minimizing chemical noise in LC-MS: Matched filtration with experimental noise determination (MEND), Proceedings of the 51st ASMS Conference on Mass Spectrometry and Allied Topics, Montreal, 2003.

[22] Thomson, V, Schatzlein, D., Mercuro, D., Limits of detection in spectroscopy, Spectroscopy 18(12), 2003.

[23] Prudnikov, E. D., Barnes, R. M., Estimation of detection limits in inductively coupled plasma mass spectrometry, Fresenius J Anal Chem, 362:465-468, 1998.

[24] Balogh, M. P., Debating resolution an mass accuracy in mass spectrometry, Spectroscopy 19(10), 2004.

[25] McLafferty, F. W., Tureček, F., Interpretation of mass spectra, University Science Books, Sausalito, 1993.

[26] Sparkman, O. D., Interpretation of mass spectra, Training course, 17th International mass spectrometry conference, Prague, 2006.

[27] Descartes, R.: Discourse de la method. In: Oeuvres de Descartes IV. sv. Adam et P. Tannery. Paris, 1908.

[28] Darwin, C.: On the Origin of Species by Means of Natural Selection, London, 1859.

[29] Newton, I.: Philosophiae Naturalis Principia Mathematica, London, 1687.

[30] Carl von Linné, Systema naturae per regna tria naturae, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis, Harderwijk, 1735

[31] Mesarovic, M.D., Takahara, Y.: Abstract Systems Theory. Springer-Verlag Berlin, Heidelberg, 1989.

[32] Žampa, P., On a New System Theory and Its New Paradigms, Cybernetics and Systems'96, Vienna, Austria

[33] Žampa P.: A new approach to (control) system theory, Preprints of the Fourth IFAC Symposium on Advances in Control Education, Istanbul, Turkey, 1997.

[34] Žampa P., Arnošt R. (2004): Alternative approach to continuous-time stochastic systems definition, Proc. of the 4th WSEAS conference, Wisconsin, USA ISBN:111-6789-99-3.

[35] IUPAC Recommendations 2001, Pure Appl. Chem., Vol. 73, No. 11, pp. 1765-1782, 2001

[36] www.thermo.com

[37] www.waters.com

[38] www.adronsystems.com

[39] tools.proteomecenter.org/mzXMLschema.php

[40] http://forums.thedailywtf.com

[41] Ježek, K., Klečková, J., Ledvina, J., Selected chapters of computers and programming, ZCU, 1998.

[42] Brown, S.D., Has the chemometrics revolution ended? Some views on the past, present and future of chemometrics. Chemometrics and inteligent laboratory systems 30 (1995) 49-58.

[43] Wold, S., Chemometrics; what do we mean with it, and what do we want from it? Chemometrics and inteligent laboratory systems 30 (1995) 109-115

[44] Viscarra Rossel, R.A., ParLes: Software for chemometric analysis of spectroscopic data, Chemometrics and Inteligent Laboratory System 90, 72-83, 2008.

[45] Andreev, V. P., Rejtar, T., Chen, H.-S., Moskovets, E. V., Ivanov, A. R., Karger, B. L., A universal denoising and peak picking algorithm for LC-MS based on matched filtration in the chromatographic time domain, Anal. Chem., 75, 6314-6326, 2003.

[46] Mihailova, A., Lundanes, E., Greibrokk, T., Determination and removal of impurities in 2-D LC-MS of peptides, J. Sep. Sci., 29, 576-581, 2006.

[47] Toyoda, T., Mochizuki, Y., Player, k., Heida, N., Kobayashi, N., Sakaka, Y., OmicBrowse: a browser multidimensional omics annottions, Bioinformatics, vol. 23 no. 4, p. 524-526, 2007/

[48] Aebersold, R., A stress test for mass spectrometry-based proteomics, Nature methods, vol.6 no.6, 2009.

[49] Bell, A. W., Deutsch, E. W., Au, C. E., Kearney, R. E., Beavis, R., Sechi, S., Nilsson, T., Bergeron, J. J. M. & HUPO test sample working group, A HUPO test sample study reveals common problems in mass spectrometry based proteomics, Nature methods, vol.6 no.6, 2009.

[50] Kanani, H.H., Klapa, M. I., Data correction strategy for metabolomics analysis using gas chromatography-mass spectrometry, Metabolic Engeneering 9, 39-51, 2007.

[51] Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, R.N., Willmitzer, L., Metabolite profiling for plant functional genomics. Nat. Biotechnol. 18, 11571168, 2000.

[52] Halket, J.M., Waterman, D., Przyborowska, A.M., Patel, R.K., Fraser, P.D., Bramley, P.M., Chemical derivatization and mass spectral libraries

in metabolic profiling by GC/MS and LC/MS/MS. J. Exp. Bot. 56, 219243, 2005.

[53] Arroyo, D., Ortiz, M. C., Sarabia, L. A., Palacios, F., Advantages of PARAFAC calibration in the determination of malachite green and ots metabolite in fish by liquid chromatography-tandem mass spectrometry, Journal of Chromatography A, 1187, 1-10, 2007.

[54] Pattern Recognition Software AS, 2004, http://www.prs.no/MS Resolver/Brochure.html

[55] Mass Works, A software for better MS, www.cernobioscience.com

[56] Tomasi G., Practical and computational aspects in chemometric data analysis, Ph.D. Dissertation, Frederiksberg, 2006.

[57] Skov T., Mathematical resolution of complex chromatographic measurements, Ph.D. Dissertation, Copenhagen, 2008.

[58] Makarenkov, V., Zentilli, P., Kevorkov, D., Gagarin, A. Malo, N., Nadon, R., An efficient method for the detection and elimination of systematic error in high-throughput screening, Bioinformatics, vol. 23 no. 13, p. 1648-1657, 2007.

[59] Deshpande, M., Kuramochi, M. Wale, N., Karypis, G., Frequent substructure-based approaches for classifying chemical compounds, IEEE Transactions on knowledge and data engineering, vol. 17, no. 8, 2005.

[60] Shurubor, Y. I., Paolucci, U., Krasnikov, B. F., Matson, W. R., Kristal, B. S., Analytical precision, biological variation, and mathematical normalization in high data density metabolomics, Metabolomics, vol. 1, no. 1, p. 75 -85, 2005.

[61] Duran, A. L., Yang, J., Wang, L., Sumner, L. W., Metabolomics spectral formatting, alignment and conversion tools (MSFACTs), Bioinformatics, vol. 19, no 17, 2283-2293, 2003.

[62] Paolucci, U., Vigneau-Callahan, K. E., Shi, H., Matson, W.R., kristal, B. S., Development of biomarkers based on diet-dependent metabolic serotypes:

Characteristics of component-based models of metabolic serotype, Omics J. integr. Biol., 8, p.221-238, 2004.

[63] The Standard Metabolic Reporting Structures (SRMS) working group, Summary recommendations for standardization and reporting of metabolic analysis, Nature Biotechnology, vol. 23, no. 7, p. 833-838, 2005.

[64] Vêncio, R. ZN., Shmulevich, I., ProbCD: enrichment analysis accounting for categorization uncertainty, BMC Bioinformatics, 8:383, 2007.

[65] Meyer, V. R., Practical High Performance Liquid Chromatography, Wiley, Chichester, UK, 1994.

[66] Robards, K., Haddad, P. R., Jackson, P. E., Principles and Practice of modern Chromatographic Methods, Academic Press, london, 1994.

[67] Lindsay, S., High Performance Liquid Chromatography, ACOL Series, Wiley, Chicester, UK, 1992.

[68] McMaster, M. C.; HPLC, a practical user's guide; Wiley, 2007.

[69] Hearn, M. T. W., Ed. HPLC of Proteins, Peptides and Polynucleotides. Contemporary Topics and Applications; Wiley: New York, 1991.

[70] Mant, C. T.; Hodges, R. S. In HPLC of Biological Macromolecules; Marcel Dekker: New York, pp 433-511, 2002.

[71] Snyder, L. R.; Glajch, J. L.; Kirkland, J. J. Practical HPLC Method Development; Wiley: New York, 1997.

[72] Snyder, L. R.; Dolan J. W. High-Performance Gradient Elution: The Practical Application of the Linear-Solvent-Strength Model; Wiley: New York, 2006.

[73] Ardrey, R. E., Liquid Chromatography Mass Spectrometry: An Introduction; Wiley, 2003.

[74] Nobel, D., The Music of Life: Biology beyond genes; Oxford University Press, 2006.

[75] Weckwerth, W. (ed.), Metabolomics: Methods and Protocols; Humana Press, Totowa NJ, 2007

[76] Du, P., Kibbe, W. A., Lin, S. M., Imroved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching, Bioinformatics, vol. 22 no. 17, p. 2059-2065, 2006.

[77] Kettman, J. R., Coleclough, C., Frey, J.R., Lefkovits, I., Clonal proteomics: one gene - family of proteins, Proteomics, 2(6):624-31, 2002.

[78] Corthals, G.L., Wasinger, V.C., Hochstrasser, D.F., Sanchez, J.C., The dynamic range of protein expression: a challenge for proteomic research, Electrophoresis, 21(6):1104-15, 2000.

[79] Lange, E.; Gröpl, C.; Schulz-Trieglaff, O.; Leinenbach, A.; Huber, Ch.; Reinert, K.; A geometric approach for the alignment of liquid chromatographymass spectrometry data, Vol. 23 ISMB/ECCB, pages i273i281, 2007.

[80] Sysi-Aho, M.; Katajamaa, M.; Yetukuri, L.; Orešič, M., Normalization method for metabolomics data using optimal selection of multiple internal standards; BMC Bioinformatics, 2007.

[81] Li, J., Prediction of internal standards in reversed-phase liquid chromatography: 1. Initial study on predicting internal standards for use with neutral samples based on linear solvation energy relationships; Journal of Chromatography A, 927, 1930, 2001.

[82] Krokhin O. V. and Spicer, V., Peptide Retention Standards and Hydrophobicity Indexes in Reversed-Phase High-Performance Liquid Chromatography of Peptides; Anal. Chem., 2009.

[83] Tomasi, G.; van den Berg, F.; Andersson, C., Correlation Optimized Warping and Dynamic Time Warping as Preprocessing Methods for Chromatographic Data; Journal of Chemometrics 18, 231-241, 2004.

[84] Chae, M.; Shmookler Reis, R. J.; Thaden, J. J., An iterative block-shifting approach to retention time alignment that preserves the shape and area of gas

chromatography-mass spectrometry peaks; BMC Bioinformatics, 9(Suppl 9):S15, 2008.

[85] Hoffman, N.; Stoye, J., ChromA: signal-based retention time alignment for chromatography-mass spectrometry data; Bioinformatics, Vol.25(16):2080-2081, 2009.

[86] Salvador, S; Chan, O., FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space, KDD Workshop on Mining Temporal and Sequential Data, pp. 70-80, 2004

[87] Norton, S. M., Methods for time-alignment of liquid chromatography-mass spectrometry data; Patent 6989100, 2006.

[88] Johnson, K. J.; Wright, B. W.; Jarman, K. H.; Synovec, R. E., High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis; Elsevier Science B.V., 2003.

[89] Urban, J., Vaněk, J., Soukup, J. Štys, D., Expertomica metabolite profiling: getting more information from LC-MS using the stochastic systems approach,; Bioinformatics, 25(20):2764-7, 2009.

[90] Martens, H.; Martens, M., Multivariate Analysis of Quality: An Introduction, Wiley 2000.

[91] Martens, H; Næs, T., Multivariate Calibration, Wiley 1992.

[92] Kirchner, M.; Saussen, B.; Steen, H.; Steen, J. A. J.; Hamprecht, F. A., amsrpm: Robust Point Matching for Retention Time Alignment of LC/MS Data with R; Journal of Statistical Software, Vol. 18, Issue 4, 2007.

[93] Podwojski, K., Fritsch, A., Chamrad, D. C. C.; Paul, W.; Sitek, B.; Mutzel, P.; Stephan, Ch.; Meyer, H. E. E.; Urfer, W.; Ickstadt, K.; Rahnenführer, J., Retention Time Alignment Algorithms for LC/MS Data must consider Nonlinear Shifts; Bioinformatics, 2009

[94] de Boer, W.P.H.; Horvatovich, P.; Lankelm, J.; Bischoff, R. Two-dimensional semi-parametric warping of LC-MS data, Conference of Systems Biology of Mammalian Cells, 2008, Dresden, Germany

[95] IUPAC. Compendium of Chemical Terminology, 2nd ed. (the Gold Book). Compiled by A. D. McNaught and A.Wilkinson. Blackwell Scientific Publications, Oxford (1997). XML on-line corrected version: http://goldbook.iupac.org (2006-) created by M. Nic, J. Jirat, B. Kosata; updates compiled by A. Jenkins. ISBN 0-9678550-9-8. doi:10.1351/goldbook.

[96] Bijlsma, S.; Bobeldijk, I.; Verheij, E.R.; Ramaker, R.; Kochhar, S.; Macdonald, I.A.; van Ommen, B.; Smilde, A.K., Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation; Anal Chem., 78(2):567-74, 2006.

[97] Daviss, B., Growing pains for metabolomics; The Scientist, Vol. 19, No. 8., pp. 25-28., 2005

[98] Bunting, C. F., Statistical characterization and the simulation of a reverberation chamber using finite-element techniques; Electromagnetic Compatibility, IEEE Transactions on Vol. 44, Issue 1, p:214 - 221, 2002.

[99] Kohl, P.; Medlar, S., Occurrence of Manganese in Drinking Water and Manganese Control; American Water Works Association, 2007.

[100] Perillo, G. M. E.; Marone, E., Determination of optimal numbers of class intervals using maximum entropy; Journal Mathematical Geology, Springer Vol. 18, N. 4, p401-407, 1986.

[101] Hyndman, R.J.; The problem with Sturgesŕule for constructing histograms; working papers, Monash University, Australia, 1995.

[102] Nyquist, H., Certain topics in telegraph transmission theory; AIEE Transactions, Vol. 47, pp. 617-644, 1928.

[103] Kotelnikov, V. A., On the carrying capacity of the ether and wire in telecommunications; Material for the First All-Union Conference on Questions of Communication, Izd. Red. Upr. Svyazi RKKA, Moscow, 1933.

[104] Shannon, C. E., A Mathematical Theory of Communication; The Bell System Technical Journal. vol 27, p 379, 1949.

[105] Shannon, C. E., Communication in the presence of noise; Proceedings IRE, Vol. 37, pp.10-21, 1949.

[106] Li, X. D., Retention time alignment in chromatography; European patent EP 1757929A1, 2007.

[107] Giatting, G.; Gletting, P.; Reske, S. N.; Hohl, K.; Ring, C. Choosing the optimal fit function: Comparison of the Akaike information criterion and the F-test; Med.Phys. 34 (11): 4285-92, 2007.

[108] Li, W.; Nyholt, D. R., Marker selection by Akaike information criterion and Bayesian information criterion; Genetic Epidemiology, 21(supp 1):S272-S277, 2001.

[109] Forni, S.; Piles, M.; Blasco, A.; Varona, L.; Oliveira, H. N.; Lôbo R. B.; Albuquerque L. G., Comparison of different nonlinear functions to describe Nelore cattle growth; Journal of Animal Science, 0845, 2008.

[110] Zçychaluk, K.; Foster, D. H., Nonparametric fitting of psychometric functions: How to choose the bandwidth?; Perception, volume 36, ECVP Supplement, 2007.

[111] Polettini, A. (ed.), Applications of LC-MS in Toxicology; Pharmaceutical Press, 2006.

[112] Prince, J. T.; Marcotte, E. M., Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping; Analytical Chemistry 78(17):6140-52, 2006.

[113] Cannataro, M.; Cuda, G.; Gaspari, M.; Greco, S.; Tradigo, G.; Veltri, P., The EIPeptiDi tool: enhancing peptide discovery in ICAT-based LC MS/MS experimentsl BMC Bioinformatics 8:255, 2007.

[114] Sykes, O. A., An Introduction to Regression Analysis; The Inaugural Coase Lecture, 1999.

[115] Ledvij, M, Curve Fitting Made Easy; The Industrial Physicist pp. 24-27,2003.

[116] Reed, J., Curve Fitting; Lessons on Introduction to Statistics and Probability, http://argyll.epsb.ca/jreed/, 2000.

[117] von zur Gathen, J.; Gerhard, J., Modern Computer Algebra; Cambridge University Press 2003.

[118] Childs, L. N., A Concrete Introduction to Higher Algebra (Undergraduate Texts in Mathematics); Springer 2008.

[119] Wolberg, J., Data Analysis Using the Method of Least Squares: Extracting the Most Information from Experiments; Springer, 12/2005.

[120] Moler, C. B., Numerical Computing with Matlab; Society for Industrial Mathematics, 2004.

[121] http://www.chem.agilent.com/en-US/PRODUCTS/SOFTWARE/ DATA-SYSTEMS/CHEMSTATION/

[122] MATLAB software, www.mathworks.com, The Mathworks, Natick, Massachusetts, USA.

[123] Lindberg, V., Uncertainties, Graphing, and the Vernier Caliper; Rochester Institute of Technology, 2003.

[124] Smith, C.A.; Want, E. J.; OMaille, G.; Abagyan, R.; Siuzdak, G., XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification; Analytical Chemistry, 78, p.779-787, 2006.

[125] Bylund, D.; Danielsson, R.; Malmquist, G.; Markides, K. E., Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography-mass spectrometry data; Journal of Chromatography A, 961, p.237-244, 2002.

[126] Pierce, K. M.; Hope, J. L.; Johnson, K. J.; Wright, B. W.; Synovec, R. E., Classification of gasoline data obtained by gas chromatography using a piecewise alignment combined with feature selection and principal component analysis; Journal of Chromatography A, 1096, p.101-110, 2005.

[127] Katajama, M.; Orešič, M., Data processing for mass spectrometry-based metabolomics, Journal of Chromatografhy A, 1158, p. 318-328, 2007.

[128] Chang L.; Yonghong Z.; Attallah, S., Max-To-Mean Ratio Detection for Cognitive Radio; Vehicular Technology Conference, p1959 - 1963, 2008.

[129] Polya, G., How To Solve It: A New Aspect of Mathematical Method, Princeton, NJ: Princeton University Press, 1945.

[130] Faber, N. M., Boqué, R., On the calculation of decision limits in doping control, Accreditation and Quality Assurance: Journal for Quality, Comparability and Reliability in Chemical Measurement, Volume 11, Number 10, p.536-538, 2006

[131] Akaike, N., Harata, N. (1994) Nystatin perforated patch recording and its applications to analyses of intracellular mechanisms, Jpn. J. Physiol. 44 (5): 43373. PMID 7534361

[132] Urban, J., Vaněk, J.; Štys, D., Mass spectrometry: system based analysis, The 11th Computer Applications in Biotechnology IFAC symposium, Leuven, Belgium, 2010.

[133] http://www.codefarms.com/dol

Appendix A

Jan Urban is the first author of followed Applications note in Bioinformatics Journal. He designed and performed the data processing, analysis as well as evaluation of the results. He is also the author of all figures, key idea of the article and had major contribution in writing of the paper.

doc. RNDr. Dalibor Štys, CSc.
Institute of Physical Biology
University of South Bohemia

Appendix B

The manuscript Cytotoxicity and Secondary Metabolites Production in Terrestrial Nostoc Strains, Originating From Different Climatic/Geographic Regions and Habitats: Is Their Cytotoxicity Environmentally Dependent? was accepted for publication in Environmental Toxicology.

Jan Urban performed the chromatogram analysis of the measurements in order to reveal molecular ions of low intensity, and for filtration of noise in the chromatograms. He also participated on the creation of figures 2., 3., and 6. and wrote the part of Chromatogram analysis.

doc. RNDr. Dalibor Štys, CSc.
Institute of Physical Biology
University of South Bohemia

Appendix C

Replicates of different concentration of pure cyanobacterial hepatotoxin microcystin-LR and its mixtures in extracts of the food additives were measured in the Nofima Mat, Ås, Norway. The first draft of the manuscript based on that measurements is on the following pages.

Jan Urban was collaborated with the preparation of Stigeoclonium extract as well as with the MCYST-LR dilutions. He also performed both, manual and automatic analysis of measured data. Jan Urban evaluated calibration curves fits from obtained reports and he is the author of all figures and tables. Together with Pavel Hrouzek he wrote the draft.

doc. RNDr. Dalibor Štys, CSc.
Institute of Physical Biology
University of South Bohemia

Appendix D

It was developed a software tool with simple User Graphical Interface in Matlab 7.7.0 (R2008b) for loading the measurements and/or blank (if available for discarding peaks presented in blank) and estimate the noise PDFs. Since there are no user-defined parameters or controls to play, the program runs completely automatically. It reads the results of the measurement and the results of the empty run. You specify the confidence level (probability) with which you want to detect peaks, and the program derives the peaks. The results are shown either as standard graphs, table of compounds, peaks and their probabilities possibly as a 3D diagram. These results are highly encouraging, exceeding the ability of the operator who performed the manual interpretation. The software is still being continuously adapted to different types of data and instruments. The presented method is based on a physical model of what happens within the LC-MS instrument, and is therefore superior to other existing methods usually based on general heuristic rules. The software may be used for multiple purposes: for expert data assessment, automated generation of the compound databases, performance analysis of the instrument, for validity assessment of biological models etc. Additional details are described in the Expertomica Metabolite Profiling manual.

Expertomica Metabolite Profiling

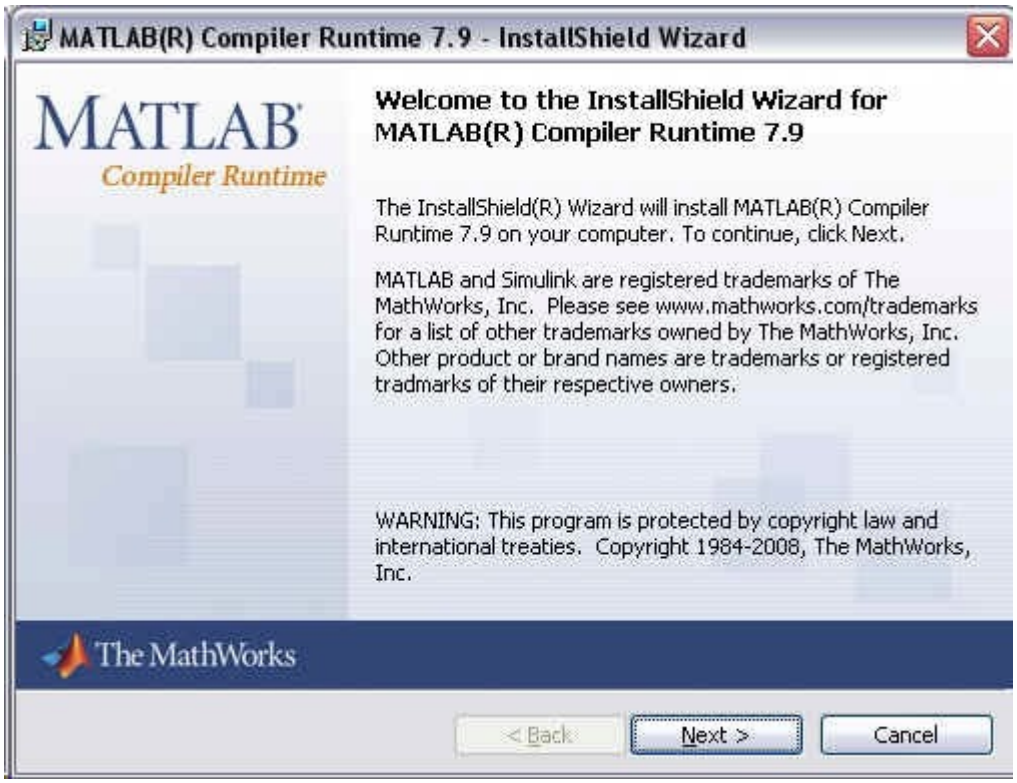
1. Introduction

EDA is Graphical User Interface and Matlab compiled application for filtration of LC-MS or GC-MS, based on probalistic methods. Including peaks and compounds segmentation with various visualizations.

Optimal RAM size installed on computer is 1GB or higher, speed of computations decrease rapidly with less amount of memory (caussed by swapping).

2. Install

a) Install `Matlab(R) Compiler Runtime` 7.9 running `MCRInstaller.exe`.



and follow the instructions

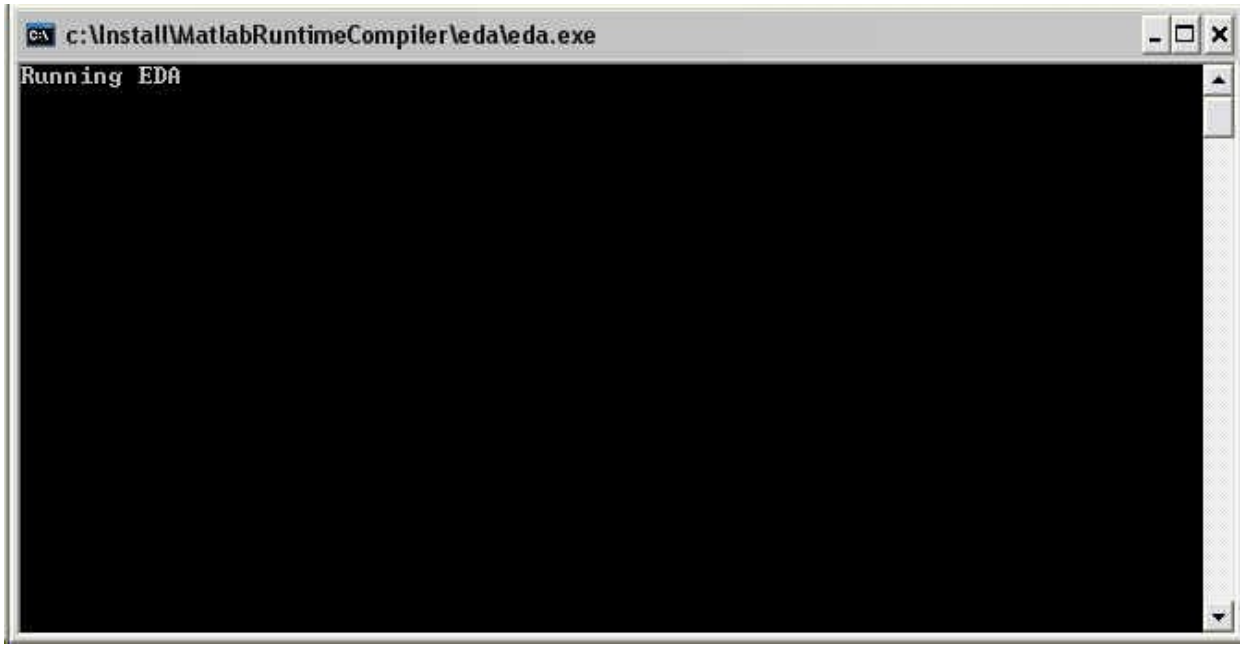
b) Reboot your computer

c) Unpack all files from `eda.zip` into your folder.

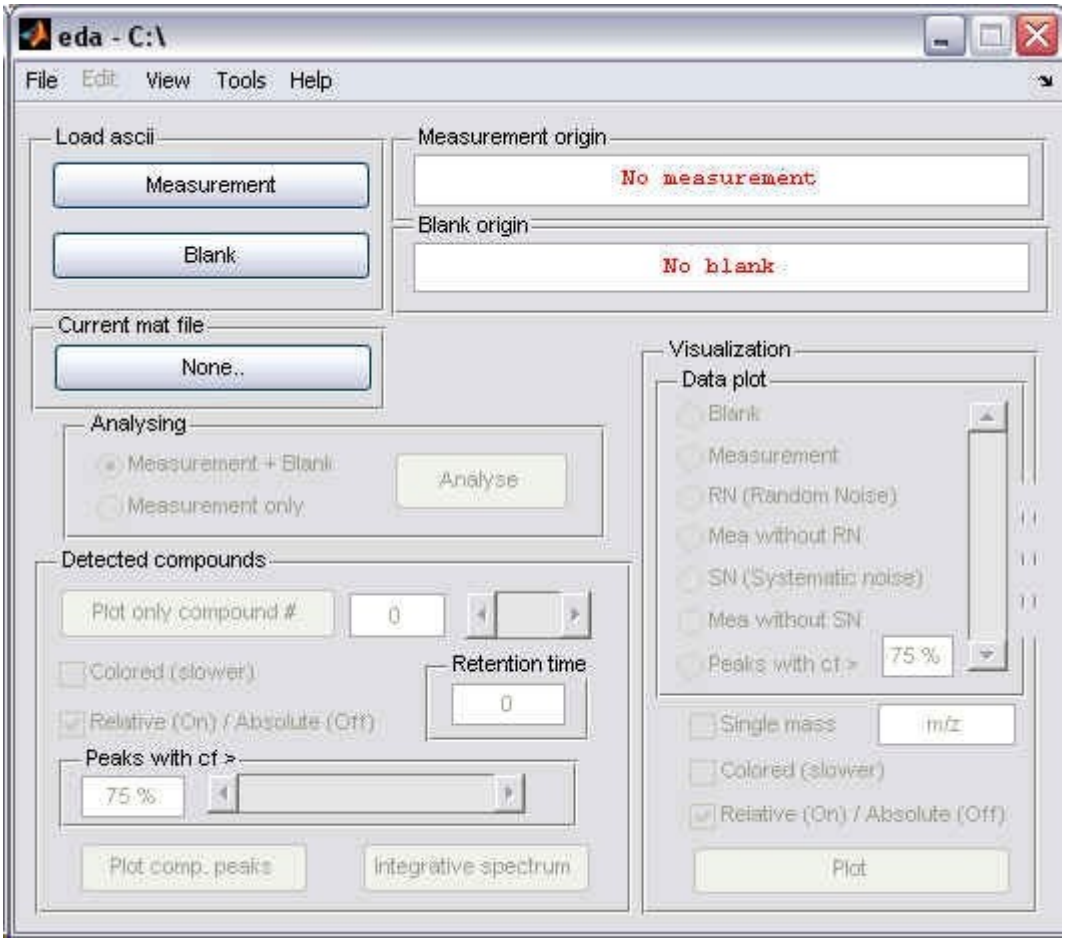
3. Start application

Run `eda.exe` and press Enter, You will see a command line window

(do not close - closing command line window will terminated whole application!)



and main window of EDA.

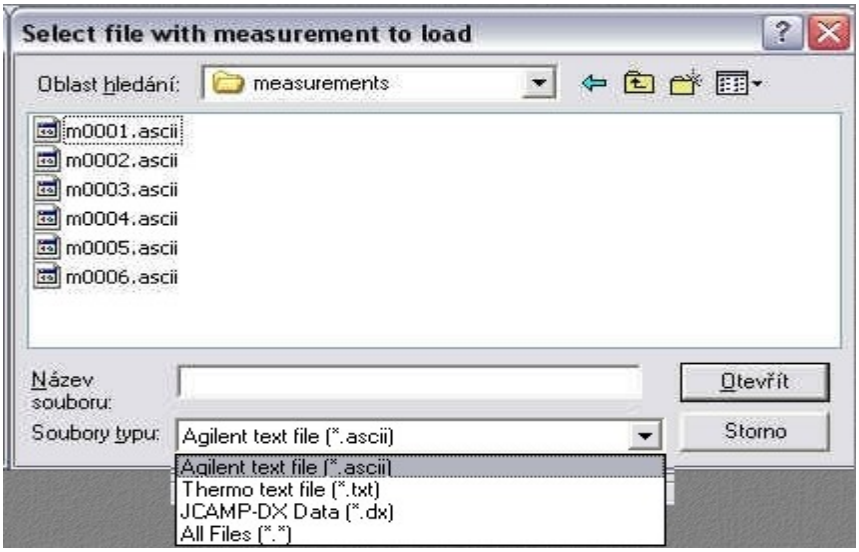


Initialisation may take a while.

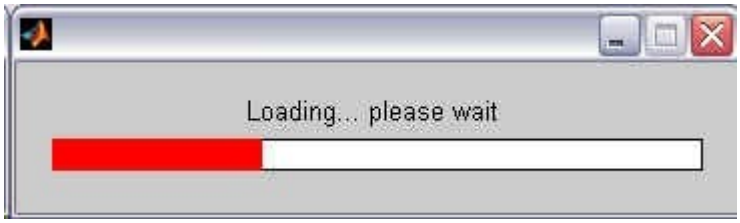
4. Insert measurement into application

EDA can currently read Agillent text files, Thermo text files and JCAMP-DX data files (ascii formats, see authors web page for new version or directly contact them for implementation of Your own file format).

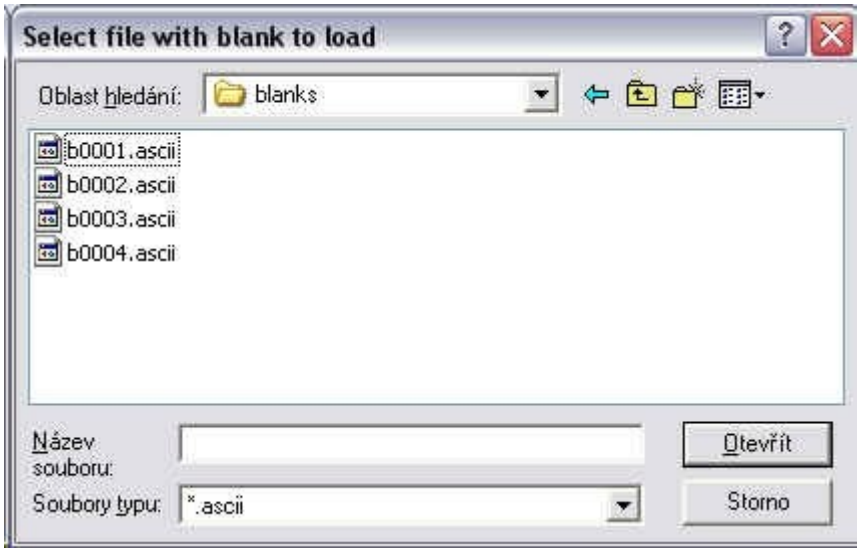
From menu `File \ Load Measurement` (or use button `Measurement` from panel `Load ascii` in EDA main window or hotkey **ctrl+m** instead) select ascii file with Your measurement in one of the supported fortmats and confirm.



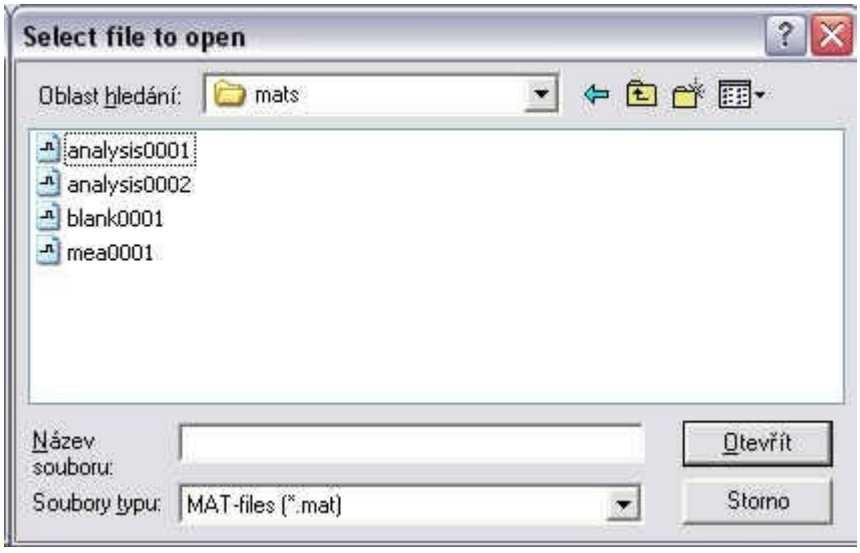
Loading ascii into data matrix takes a while, You will see a progress bar during process.



For loading blank (measurement without sample) use menu `File \ Load Blank` (or button `Blank` from panel `Load ascii` in EDA main window or hotkey **ctrl+b**), select Your ascii file and confirm. Again, You will see a progress bar.

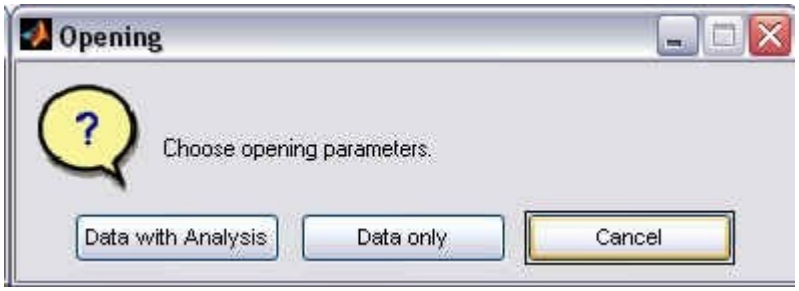


Once You have saved the data in `mat` file (Matlab format, see **5. Save dataset** for details), You can also load it. From menu `File \ Open mat` (or button in panel `Current mat file` in EDA main window or hotkey **ctrl+o**) select mat file with Your dataset and confirm.

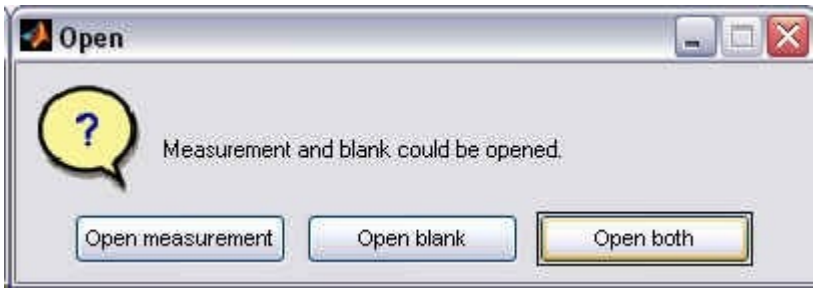


Application will check data integrity and content. Depending on Your decision during previous saving, can be be opened one of this items:

- `Data with Analysis` - measurement (and blank, if available) with computed probabilities, list of segmented compounds and its peaks



- `Data only` - blank or measurement or both (if available).

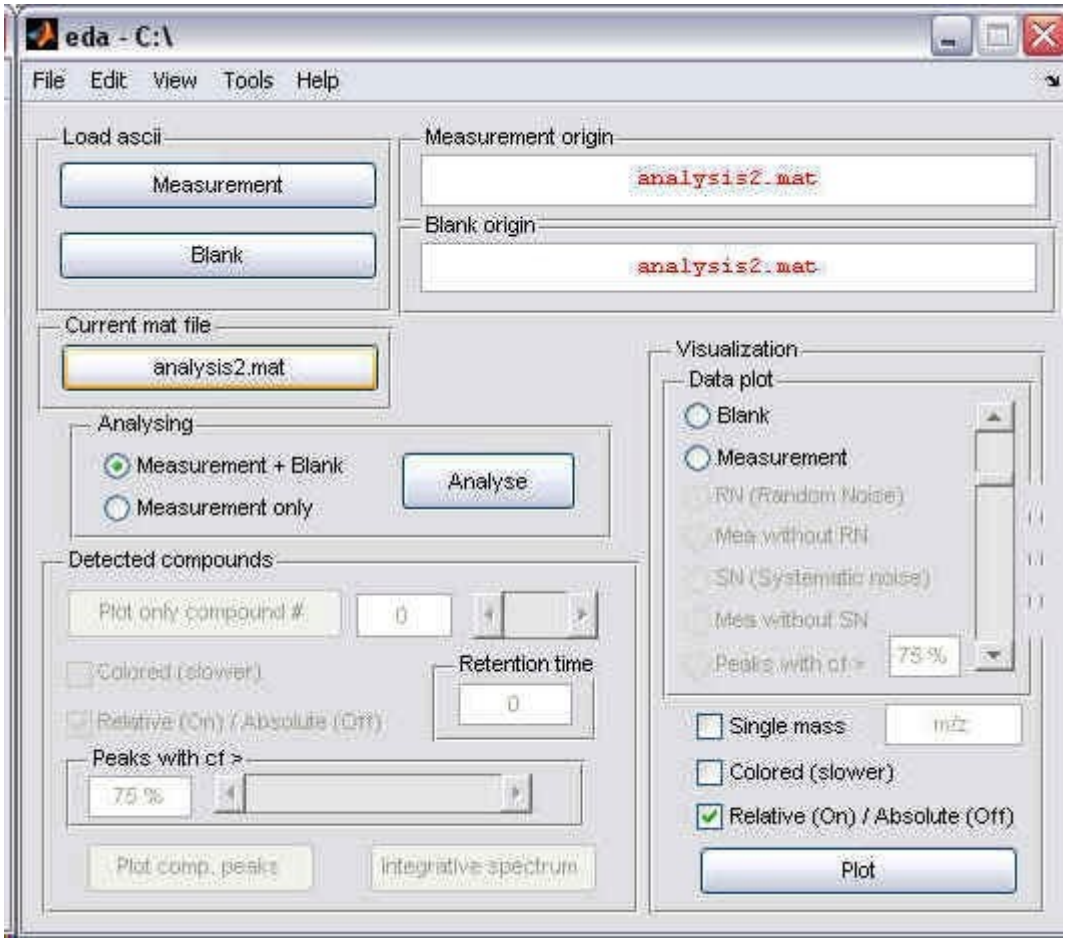


EDA will automatically ask You according to dataset presented in mat file.

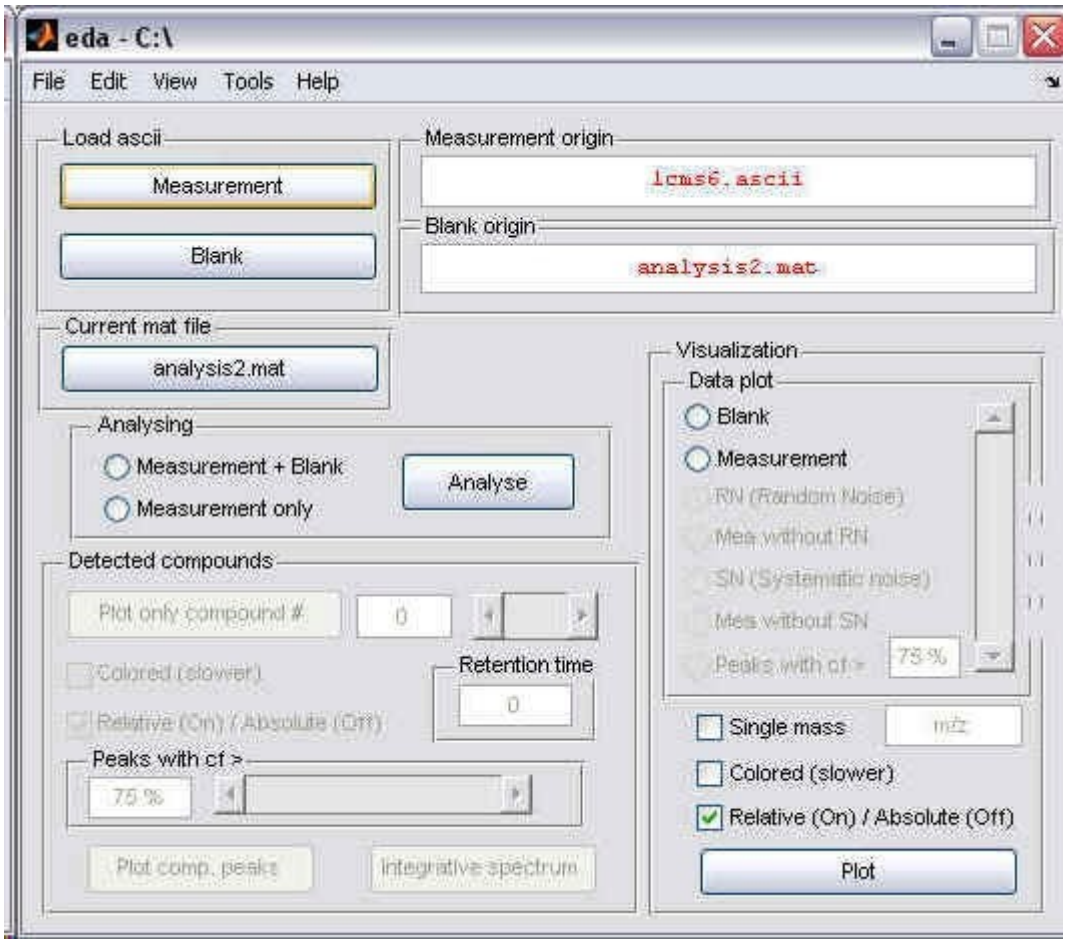
In panel `Measurement origin` You will see name of file with currently opened/loaded measurement (`No measurement` for none). Measurement can be loaded from ascii file or opened from mat file.

In panel `Blank origin` You will see name of file with currently opened/loaded blank (`No blank` for none). Again, blank can be loaded from ascii file or opened from mat file.

On button in `Current mat file` panel You will see name of currently opened mat file (`None...` for none)



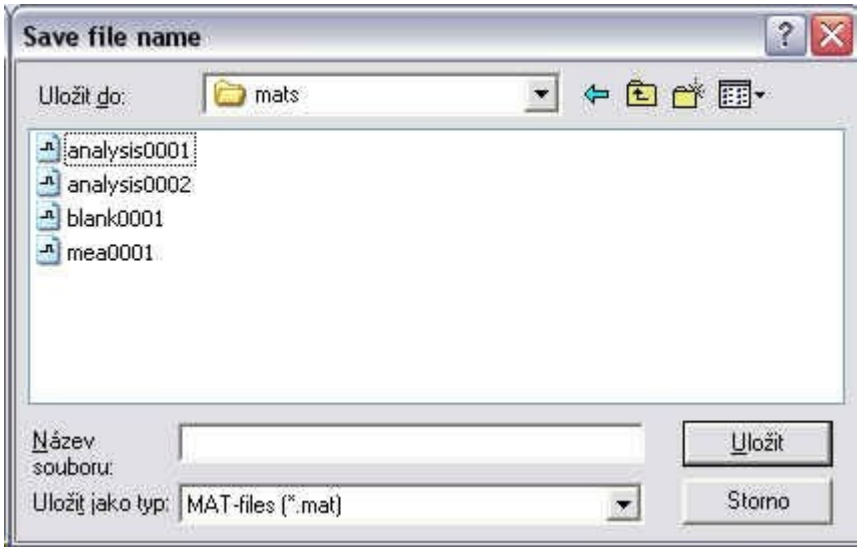
You can easily combine data from loading ascii and opening mat, for example: load measurement from ascii file and open only blank from mat file (mat file may include different measurement).



5. Save dataset

Loaded measurement or blank or both from ascii file can be simply saved in menu `File \ Save Data (M or B)` (or hotkey **ctrl+s**). Select name (default name is same as loaded ascii file, only extension is replaced by `mat`), where You want to have dataset stored and confirm.

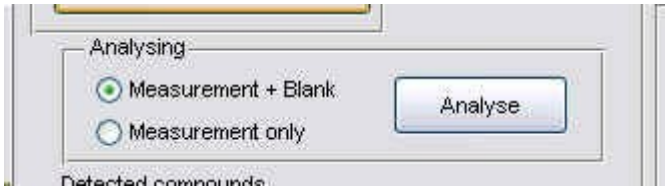
Computed analysis (see **6. Analysing**) of measurement (or measurement with blank) can be also saved as mat file. Measurement and blank (if available) used for analysing are stored together with analysis. From menu `File \ Save analysis (+Data) as` and select name for mat file to storage Your dataset and confirm.



EDA may stop responding to other actions during saving process.

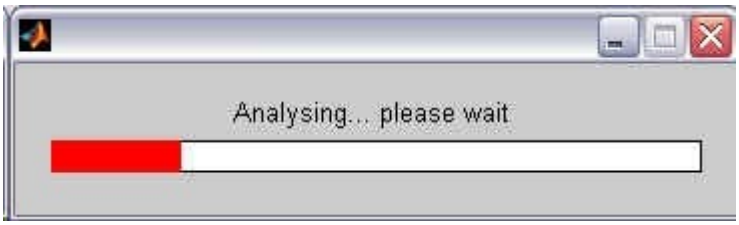
6. Analysing

After loading or opening measurement or measurement with blank or analysis dataset, EDA enables panel `Analysing`



where You can choose by radiobuttons analysis of `Measurement + Blank` or analysis of

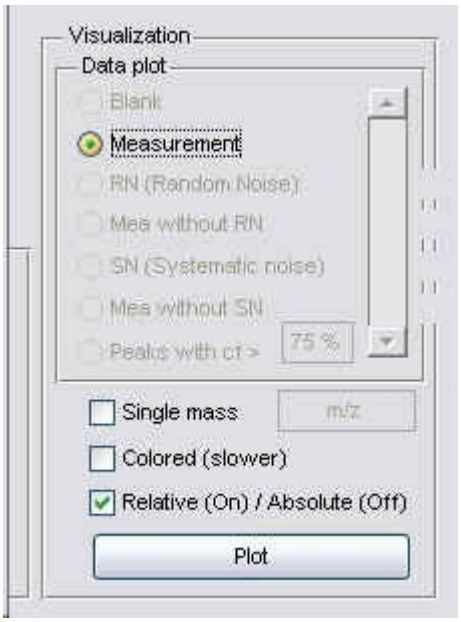
`Measurement only` (depends on opened/loaded dataset). Analysis itself is started by clicking button `Analyse` in the panel. Computation of all analysing methods, filtration, probabilities computation, peaks and compound segmentation is indicated by progress bar. Speed depends on computer memory and processor frequency.



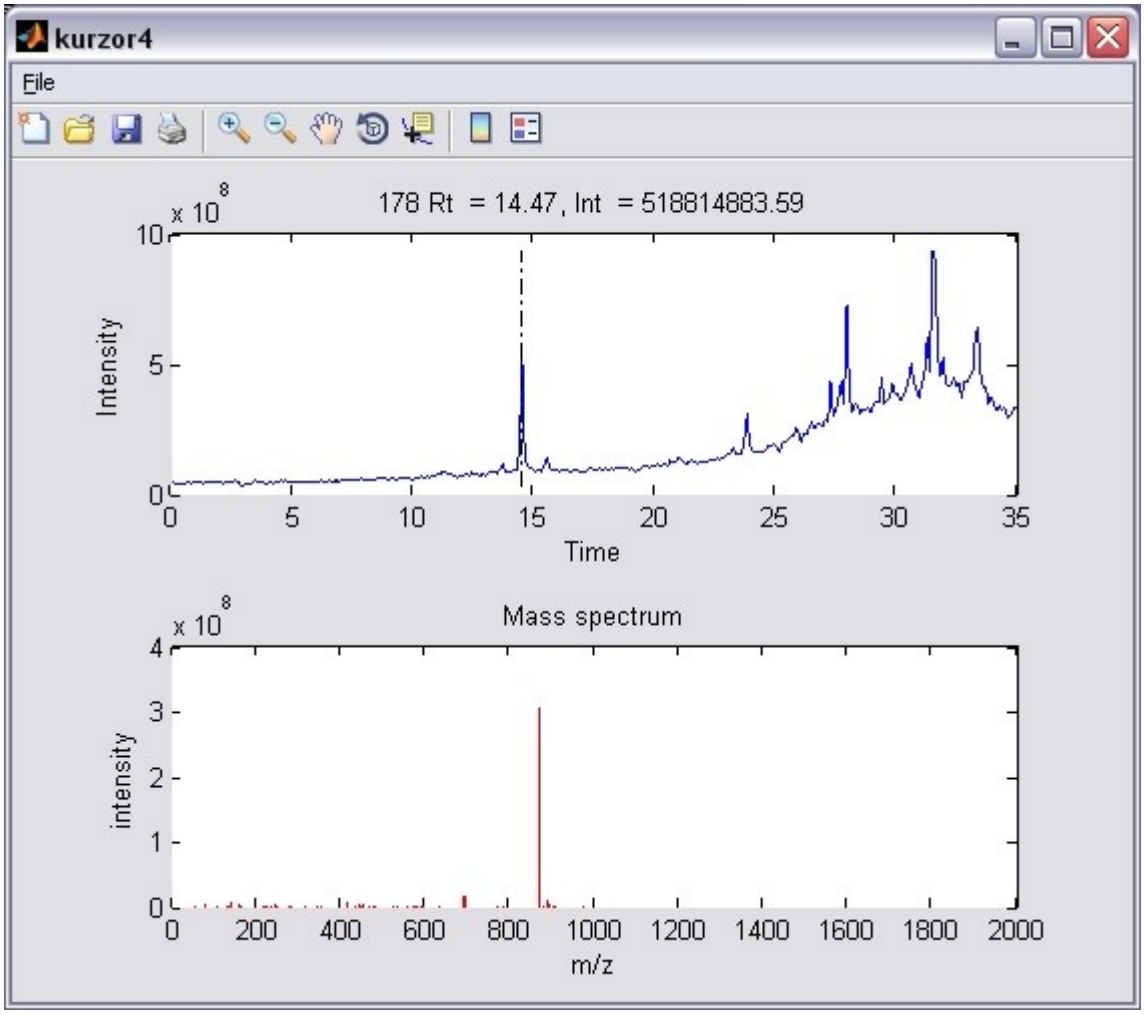
When analysing process finished, EDA enables two result panels - `Detected compounds` and `Visualization` (For detail information see 7. Visualization or 8.Peaks and Compounds).

7. Visualization

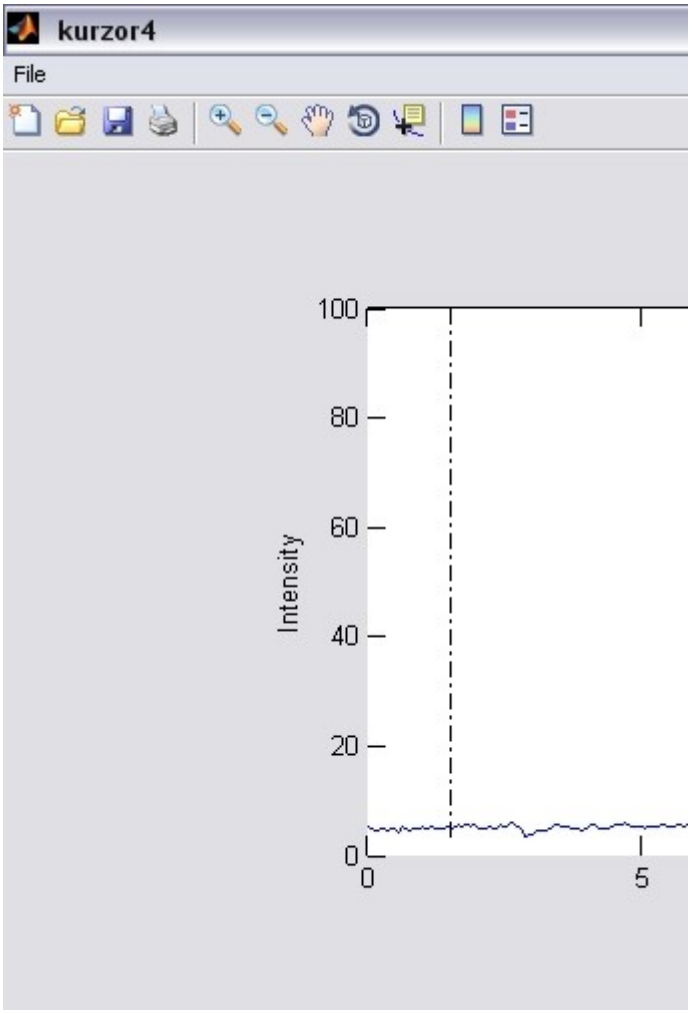
Loaded measurement or blank can be ploted using radio buttons in subpanel `Data plot` in panel `Visualization`, where you can also choose the type of data visualised from the measurement or blank.



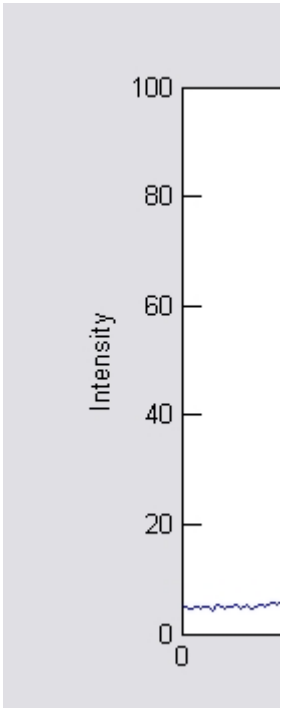
All plots are figured after clicking the button `Plot`. You will see a new figure with two subplots, upper one for TIC (Total Ion Chromatogram) and lower one for mass spectrum.

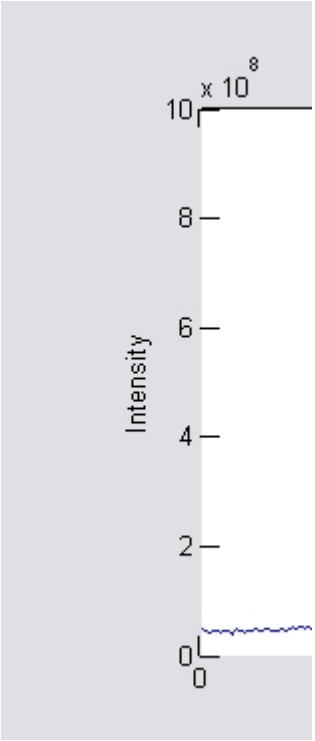


Use mouse right button in TIC to select exact time in upper figure for plottting related mass spectrum in lower figure.

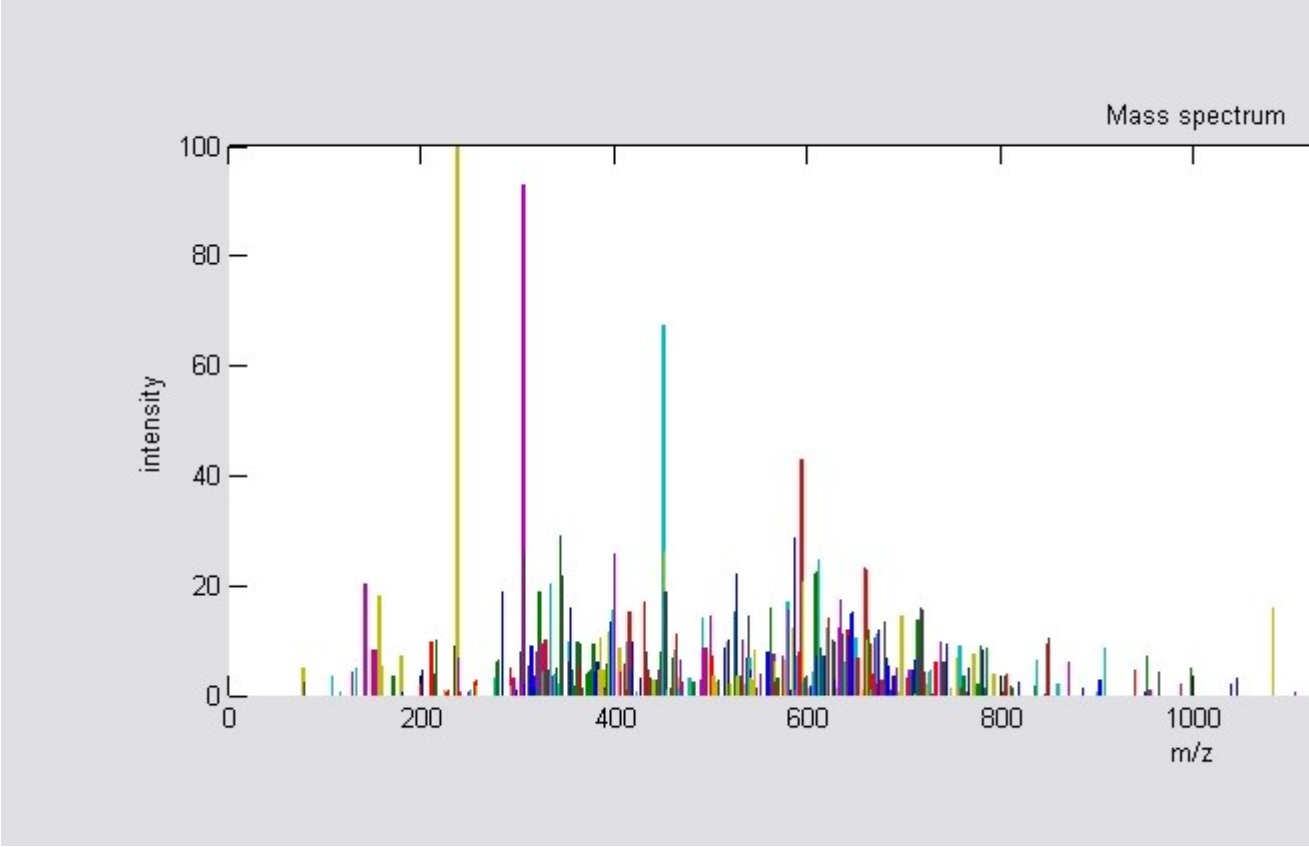


a) Check box `Relative (On) / Absolute (Off)` determine scale on y axis.



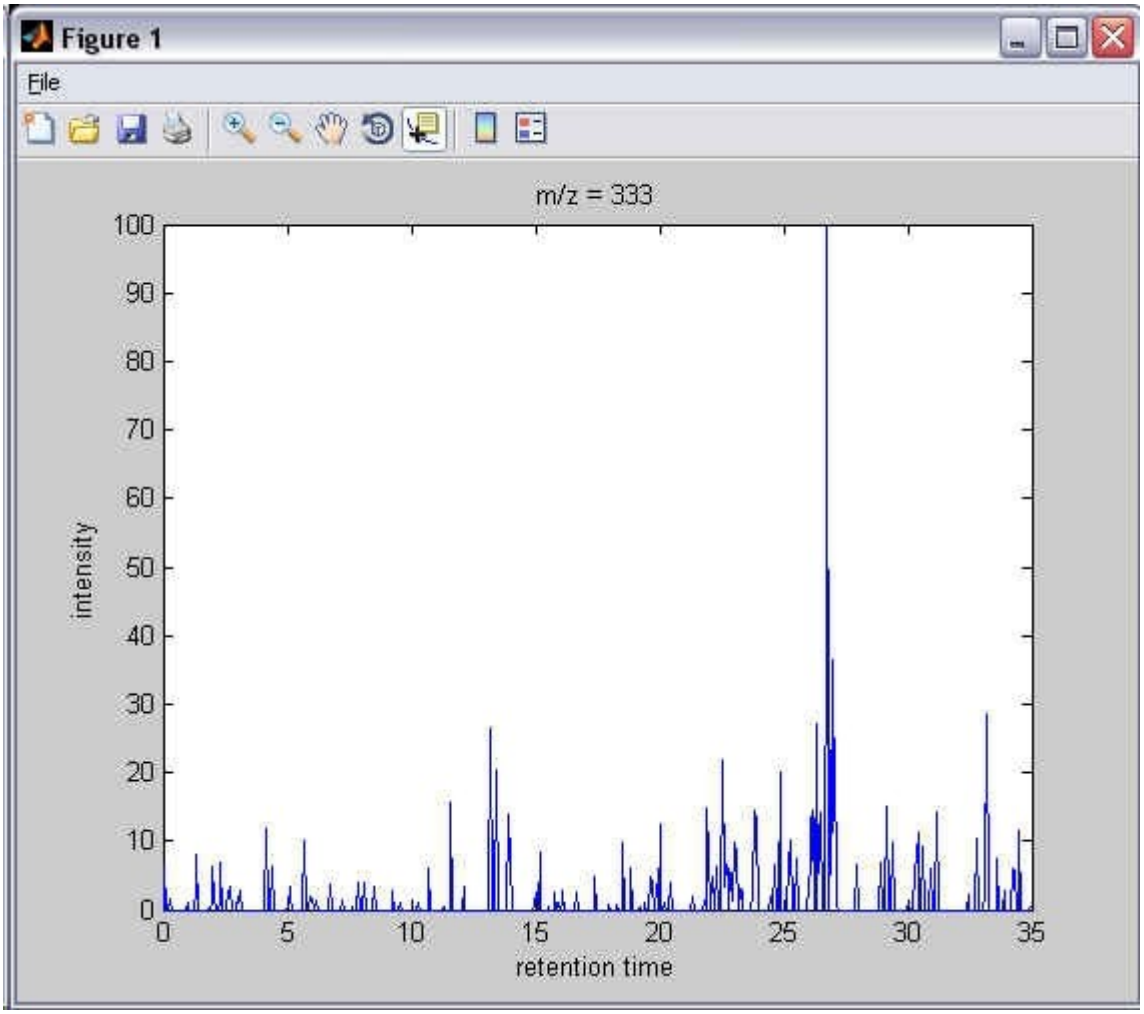


b) Check box `Colored (Slower)` determine using different colours for bars in mass spectrum. This option require little bit more computer memory to proceed.

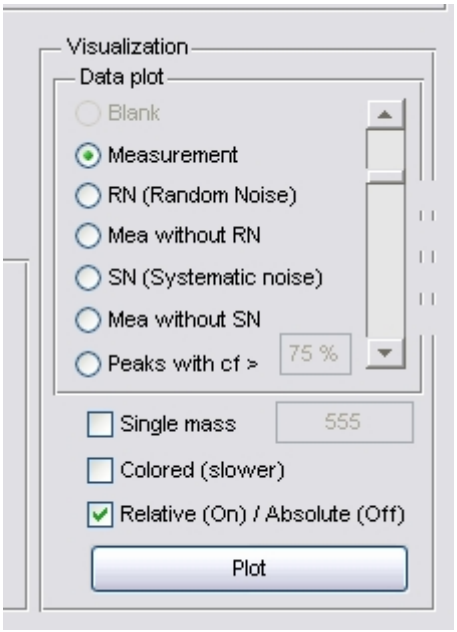


c) Check box `Single mass` allow to plot selected m/z value in time.

No mass spectrum will be ploted for single mass (it is nonsense).



After finishing analysis of Your data, more radio buttons will be enabled in `Data plot` subpanel.



All `check boxes` are the same as in previous case. Using different `radio buttons` You can plot contribution of `Random noise` or `Systematic noise` (base line) in Your measurement as well as `Measurement without random noise` or `Measurement`

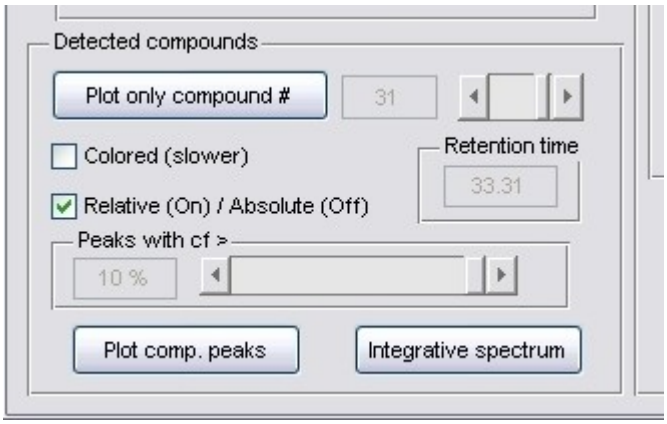
`without systematic noise` (means without both noises, random noise was removed before the algorithms for systematic noise were applied).

You can also plot only masses, that have peak behaviour in time with certain level of probability (`cf`). For selecting the minimal `cf` value of ploted peaks use the slider in subpanel `Data plot`. Default value is 75% and maximal value is equal to maximal probability in the whole measurement.

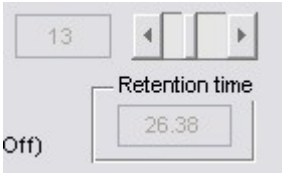
All plots are figured after clicking the button `Plot`.

8. Peaks and Compounds

After analysing process finished, EDA enables `Detected compounds` panel.

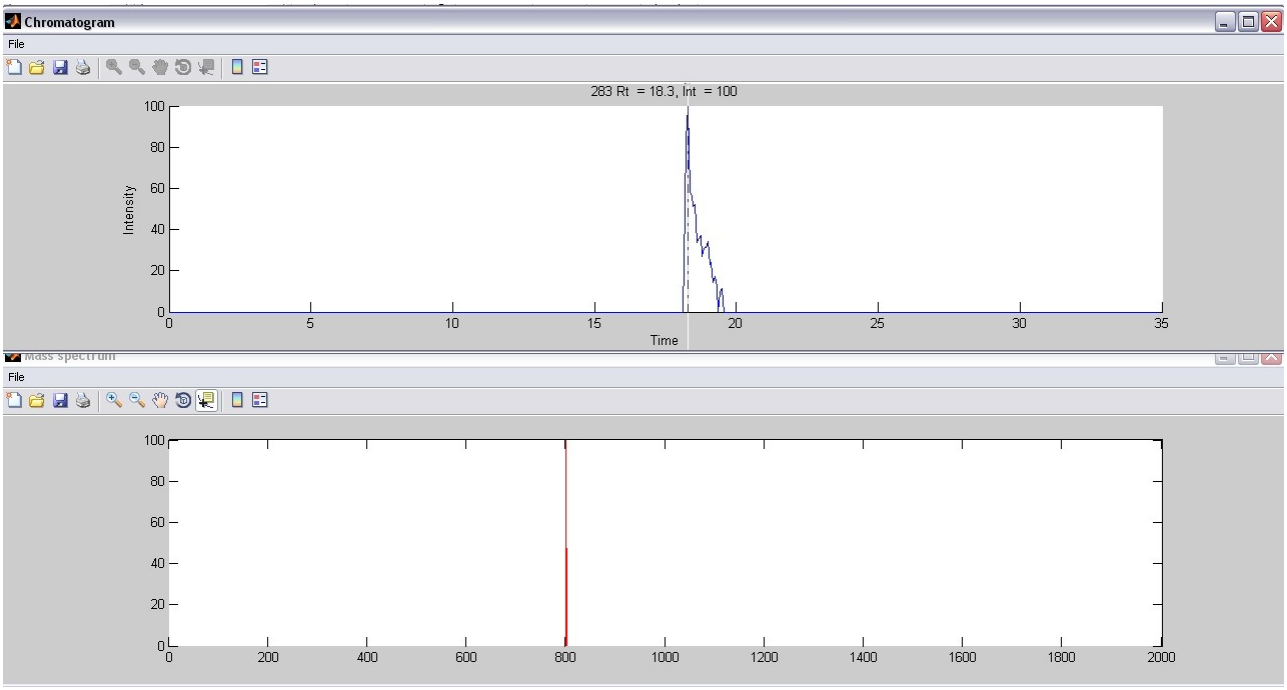


Between button `Plot only compound #` then You will see number of detected compounds. Using slider on the right-top will select which one of detected compounds You want to see in detail. Also the `Retention time` of compound intensity maximum will be displayed in window bellow



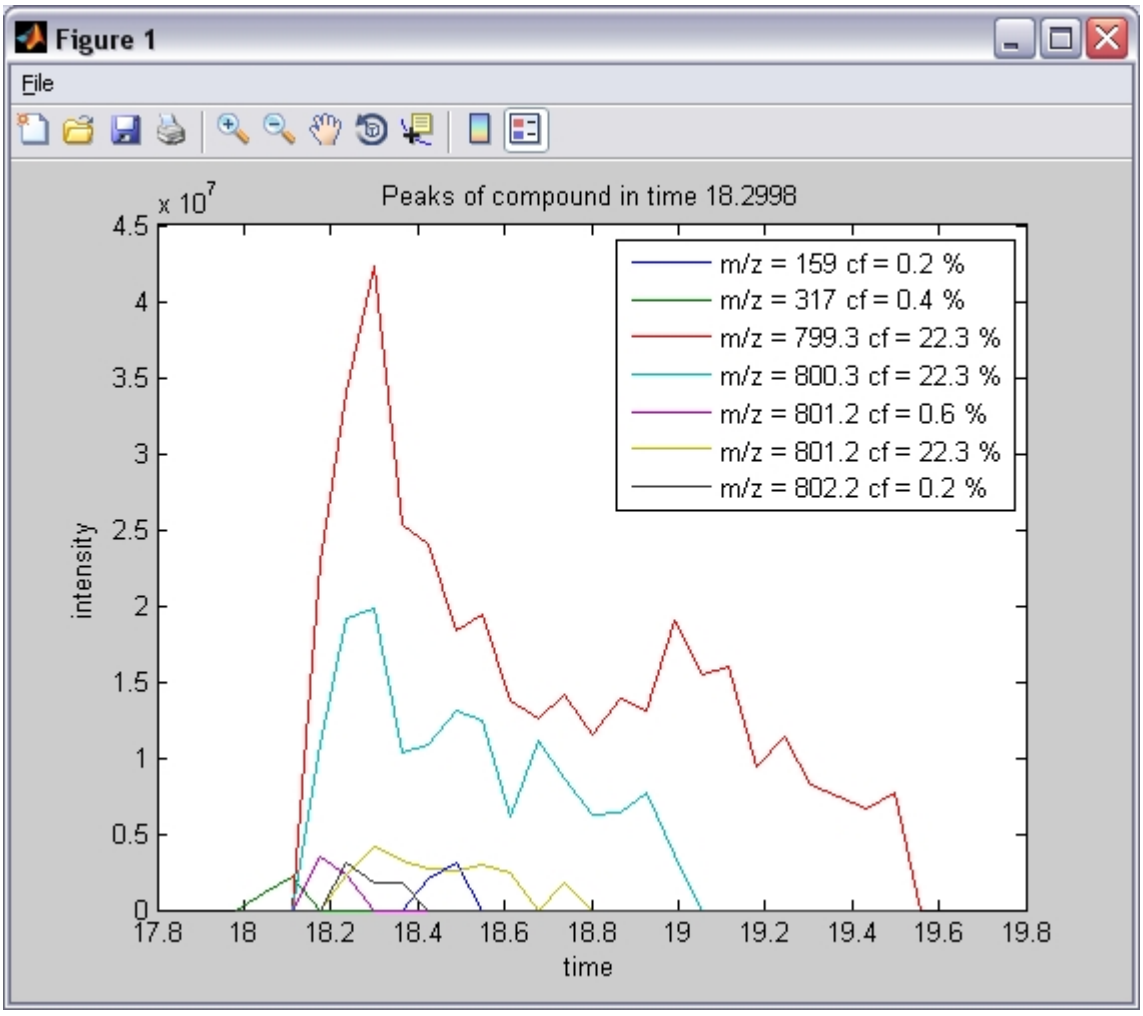
Check boxes `Colored (Slower)` and `Relative (On) / Absolute (Off)` have exactly the same function as in previous case (7. Visualization).

Button `Plot only compound #` will plot two figures, upper one for TIC (Total Ion Chromatogram) and lower one for mass spectrum. Both only for selected compound.

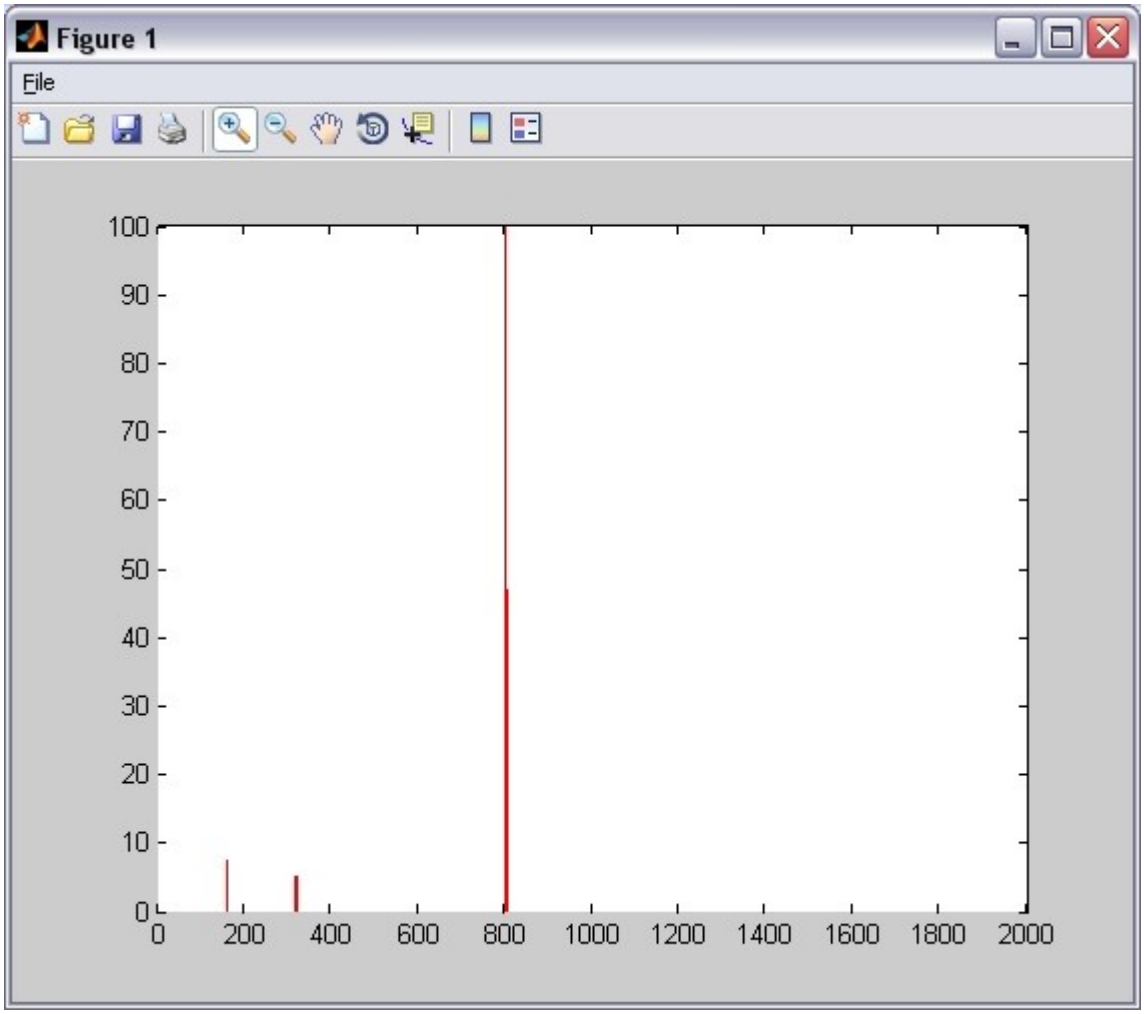


Slider for `Peaks with cf>` determine value of prabability for ploting TIC and mass spectrum.

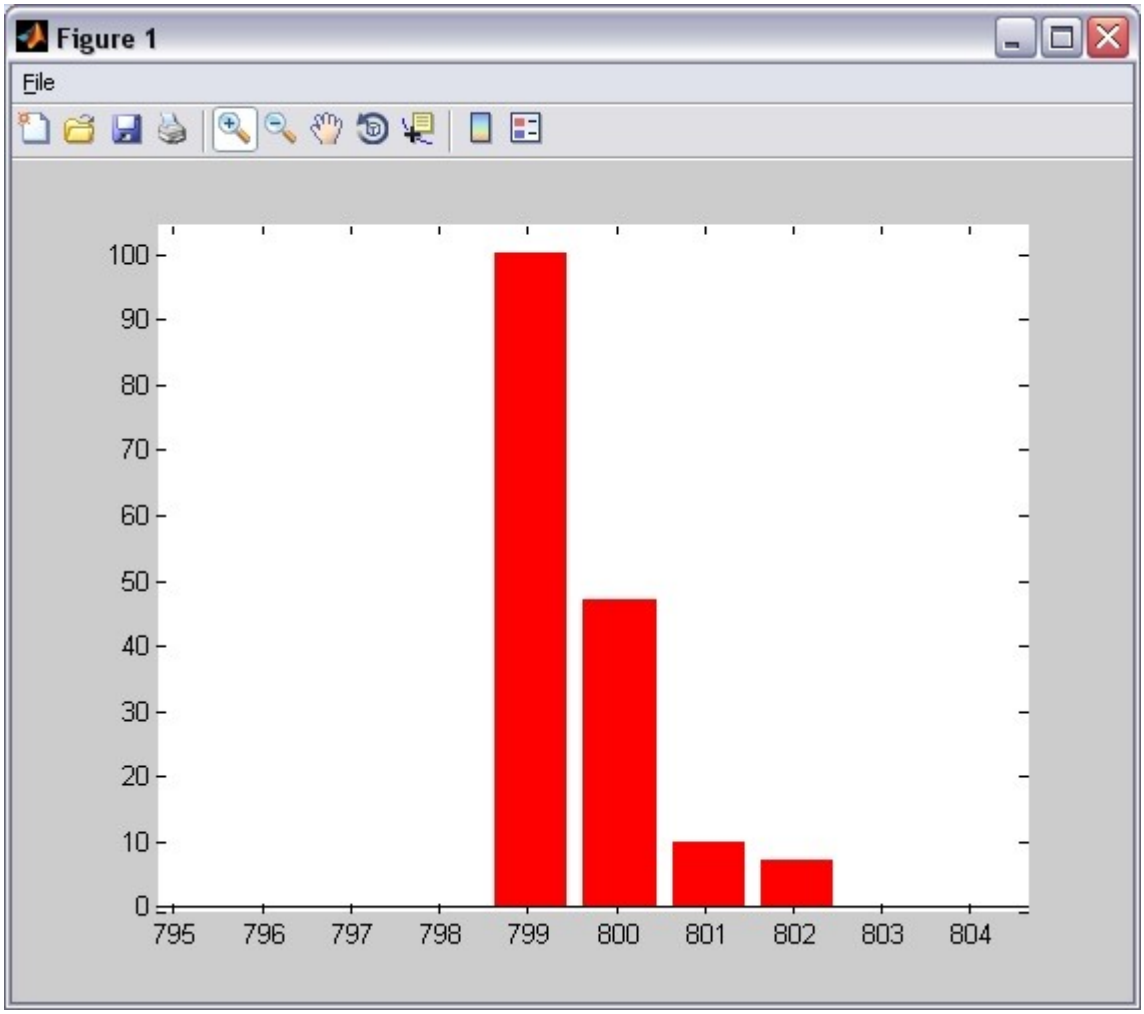
Button `Plot comp. peaks` will plot time behaviour of **all** masses in selected compound.



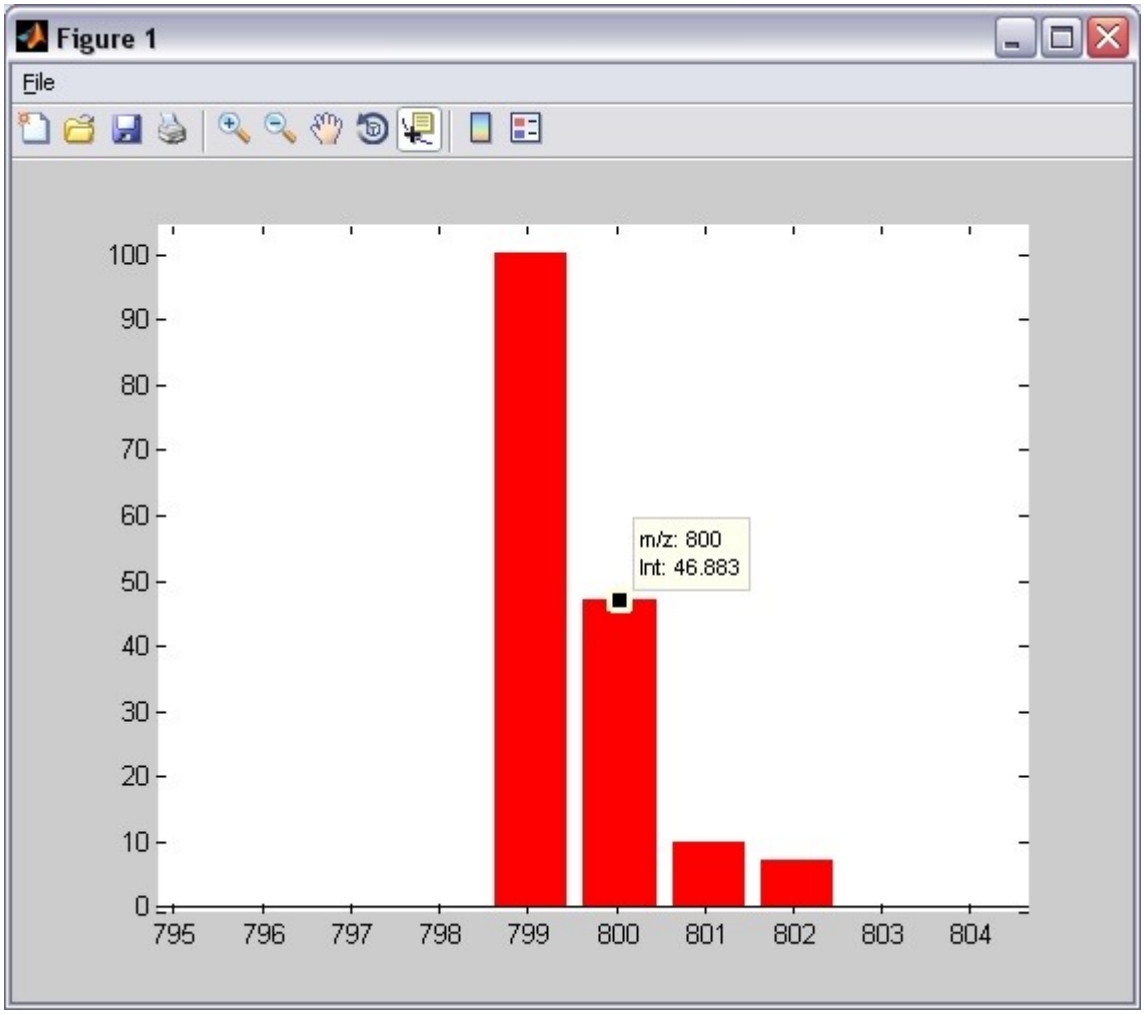
Button `Integrative spectrum` will plot mass spectrum across whole compound.



All mass spectrum plots may be zoomed using icon of macnifying glass and selection of region of interest. For zoom reset use double click of left mouse button.

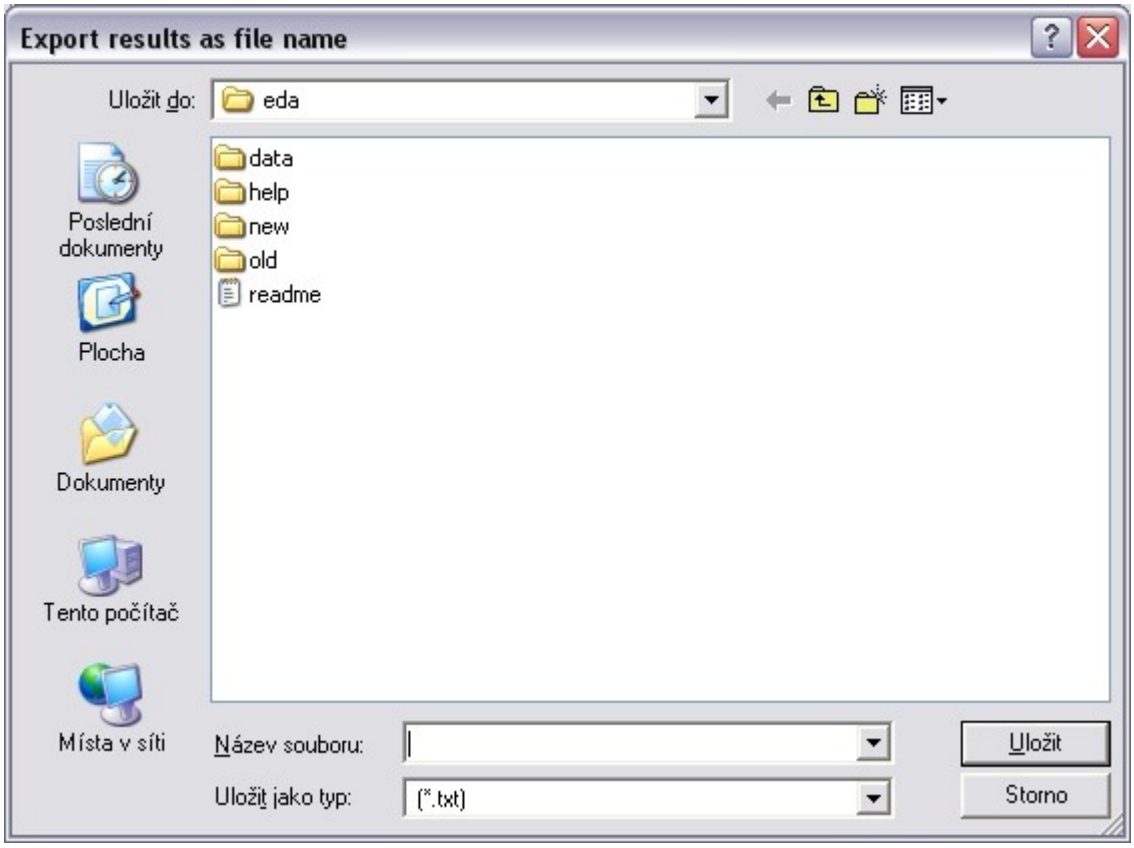


Another icon `Data cursor` allow to describe plot points. Hold `alt` on Your keybord to allow multiple data cursors.



9. Export Data

From menu `File/Export Data` create a txt file with the results of Analysis (table of compounds, peaks and their properities)



10. Main menu

In left upper corner You can find menu toolbar

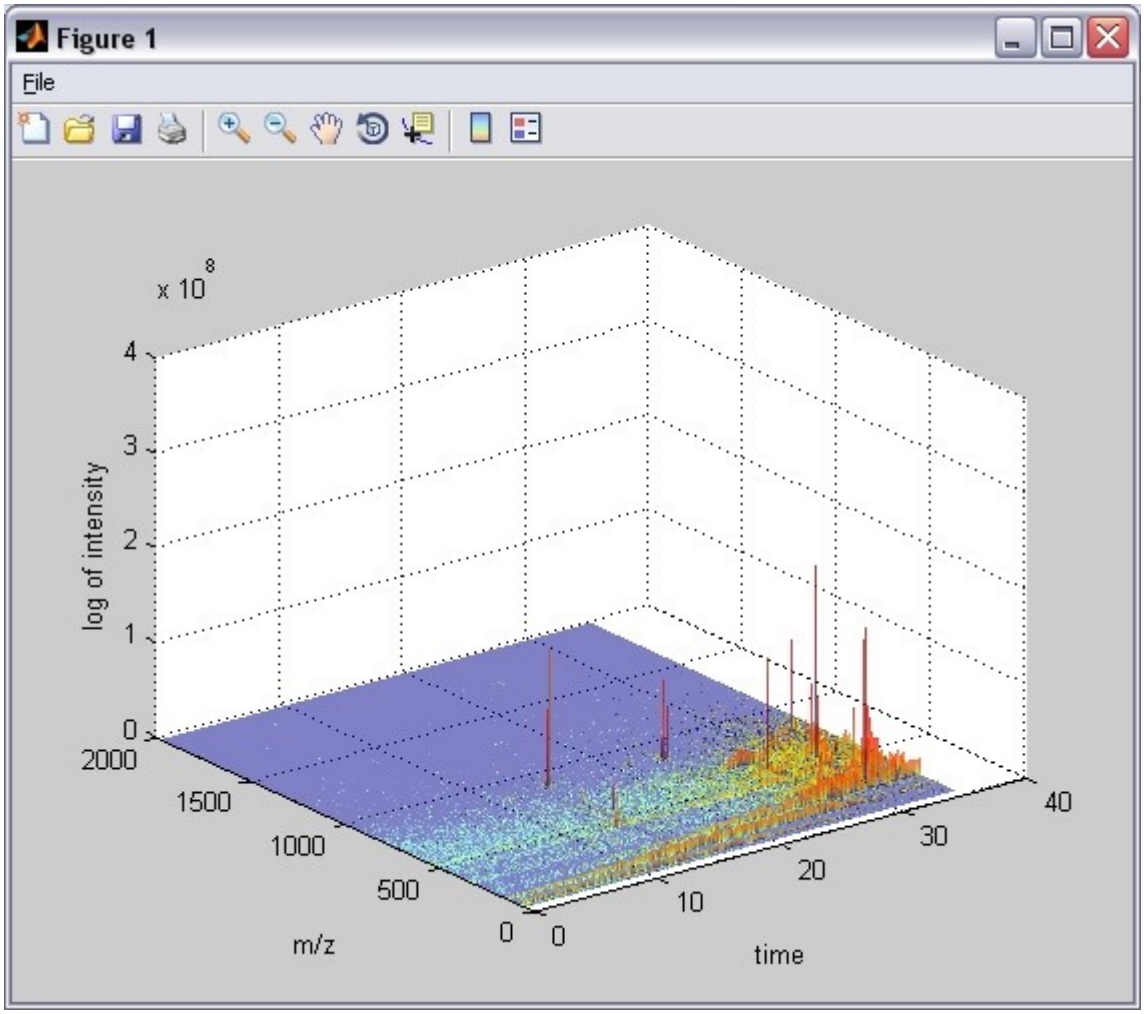
where

- `File` contains items for Open mat, Load Blank, Load Measurement, Save Data, Save Analysis, Export Data and Quit (see 4. Insert measurement into application and 5. Save dataset for details)

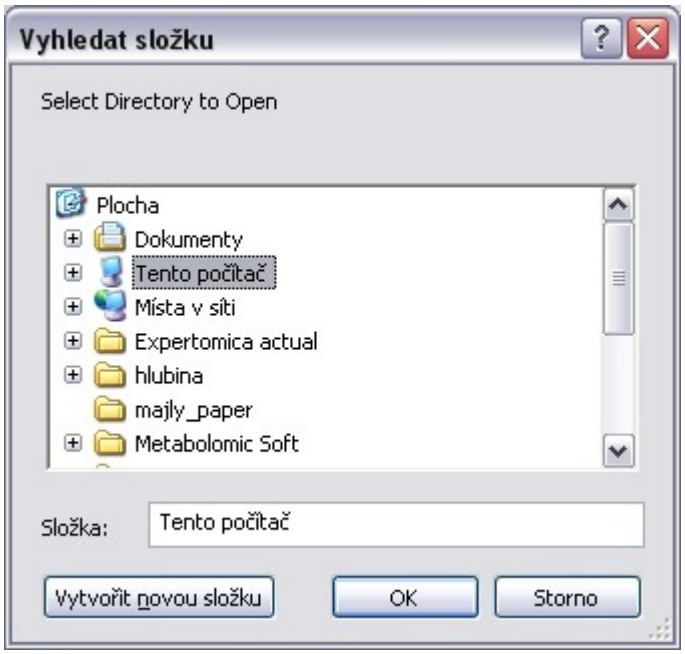
Before `Quit` You willbe asked for saving data.

- `Edit` contains items for Clear blank, Clear measurement, Clear analysis or Clear all from the computer memory.

- `View` items allows to plot 3D mesh of Blank, Measurement, or Analysis in decadic or log scale.



- `Tools` contains items for Analysis (see 6. Analysing for details) and Set Current directory



- `Help` allows to read this help and some basic informations about the program.

MATLAB(R) is registered trademark of The MathWorks, Inc http://www.mathworks.com

Appendix E

Non-optimized version of the processing/analysis steps was also implemented in C/C++ stand alone application called MetDB. The MetDB in version 2.5 includes database, using C++ Data Object Library 7.1. (DOL, [133]). Database uses two main classes according to the system based description of the measurement.

```
class Slice {
public:
 int numion;
 float time;
 float *mass;
 int *intensity;
 char Countms;
 float precursor;

 Slice(); //konstruktor
 ~Slice(); //destruktor
};

class Measurement {
protected:
 int numslices;
 Slice **slices;
```

```
public:
     int loadThermoTxt(char *name);
     int loadAscii(char *name);
     int loadJCAMP(char *name);
     int load(char *name);
     int save(char *name);
     Measurement(); //konstructor
     ~Measurement(); //destructor

};
```

Every measurement is composed of series of slices - mass spectra in one time point. Each slice has series of pairs mass and intensity. Integer variable numion is for amount of pairs.

The MetDB v2.5 is a console application without any GUI (Graphical user interface) as it is described in the attached user's guide. The results of the processing/analysis are PRT ascii reports with the some structure as in Expertomica metabolite profiling. Several functions of this application are still developed.

MetDB v2.5
User's Manual

MetDB is an object oriented database for storing and filtering metabolomic measurements from GC/MS or LC/MC. This version run under MS Dos or Command promt under Windows.

For start the run of the program please type „`metdb stdin`“ in your Metdb directory.

1.1 Opening New Database

Type „`openNewDB` *`dbPath adminID password`*“, where

- *`dbPath`* is path to your **existing** directory for new database files and name for database.
- *`adminID`* is your identification, it is case sensitive
- *`password`* is chain of numbers and letters, it is case sensitive

and press enter. Your new database was created.

1.2 Openig Existing Database

Type „`openDB` *`dbPath`*“, where

- *`dbPath`* is path to your **existing** directory with some database files.

Press enter and log in „*`M adminID password`*“, where *`M`* is mode you want to use for acces the database. Sellect from control user „`C`“ which represents database administrator; active user „`A`“ - can make changes in databese; and passive user „`P`“ - can only watch existing data.

2.1 Add measurement into database

As input it use Agilent ascii format or Thermo Xcalibur text format. Once you have opened your database, type „`import` *`data\file.ext F`*“.

`data\file.ext` is path to the file with the measurement you want to add into database and file's name and extension.

`F` is a symbol for file format, „`A`“ for Agilent or „`T`“ for Thermo. Usually Agilent use `ascii` as extension and Thermo `txt`, but it is **not always** true.

2.2 Process filtering and peak segmentation

Type „`PeaksH` *`measName`*“ where

`measName` is a name of measurement you want to process, it is same as the file name, but without extension.

2.3 Export results

Type „`export` *`measName`*“.

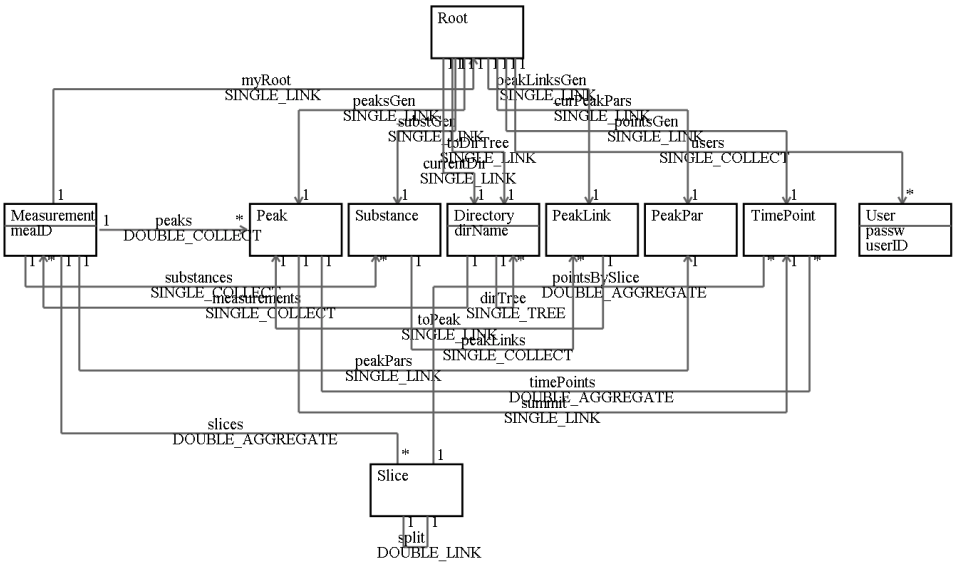
In your database directory you wil se two new files „*`measName`*`.svg`“ and „*`measName`*`.prt`“. `Svg` is image of 3D represantation of filtered data. Open it in your Internet browser. `Prt` is text file with list of detected substances and their fragments with mass, retention time, intensity and statistical characteristics.

3.1 Save

Type „`save`“ for saving all changes you made in your database.

3.2 Ending work

Type „`exit`“ for close your database. Be sure you **saved** your work before using this command.



UML class diagram of the structure of a MetDB system's classes and their relationships.

Appendix F

Software for processing/analysis of LC-MS measurements by main producers:

Thermo Scientific (www.thermo.com)

- Xcalibur - measurement device controll, basic processing, reports, integration of external modules.
- Quick Quan drug discovery tool, automatic procedures, high-throughput, LC-MS/MS.
- BioWorks identification of proteins, SEQUEST searching algorithm.
- Mass Frontier - analysis of measured datasets, especially MS/MS, prediction of fragmentation, compounds identification, database of fragmentation.
- MetWorks metabolites, structure identification, integration of Xcaluburu and Mass Frontier properties into one, molecules identification by combination of precise mass measurement and MS/MS measurement.

Waters Corporation: (www.waters.com)

- Mass Lynx measurement device controll, basic processing of measurements with rounded mass values, high precision mass values, MS/MS fragmentation. Consist of peak detector, quantitative estimation, molecular formula estimation fom high precision mass and isotopes.

- Inspector - converting tool of several data formats.

Applied Biosystems: (www.appliedbiosystems.com)

- Analyst measurement device controll, basic processing and analysis.
- BioAnalyst identification and characterization of peptides and proteins.
- LightSight searching and identification of metabolites, automatic peak detection, MS/MS spectra analysis.
- MarkerView tool for experiments analysis, biomarkers searching, statistical evaluation, changes of metabolites groups.

Bruker Daltonics: (www.bdal.de)

- MetaboliteTools prediction and detection of metabolites changes across samples, automatic peak detection.
- ProfileAnalyst analysis and statistical comparison of measurement series, searching for biomarkers.

Agilent Technologies: (www.agilent.com)

- ChemStation visualization of measured datasets, standard application for reports generator and quantitative analysis.
- Mass Hunter data analysis, metabolite identification, automatic peak detection, high precission mass values, MS/MS.

Shimadzu: (www.ssi.shimadzu.com)

- LCMSsolution measuremnt device controll and basic processing. Peak integration for MS, UV and PDA.

ACD Labs: (www.acdlabs.com)

- MS Processor basic processing of LC-MS/MS measurement.
- IntelliXtract analysis of LC-MS measurements, automatic or manual peak detection, spectra deconvolution, compounds assembling considering isotopes as well as fragmentations. Confidential estimation from C12/C13 ratio. Ability to read all main file formats.
- MS Manager overlay of MS Processor and IntelliXtract.

BioAnalyte: (www.bioanalyte.com)

- ProTrawler quick processing, ability to read main file formats, automatic peak detection, spectra deconvolution.
- Regatta overlay for results from ProTrawler, analysis of measurement differences, detected peaks comparison.