Computer assisted comparative analysis of large volume of mass spectral data originating from biological and medical samples.

Vegh Peter INCDTIM Cluj Napoca, Romania;
Vekey Karoly, Drahos Laszlo, Ozohanics Oliver KKKI Budapest, Hungary;

The analysis of various samples of natural provenience (i.e. biological, medical or environmental) is getting more and more widespread in Hungary and also in Romania since adjoins to the European Union. The analysis is mostly done with a liquid chromatograph coupled to mass spectrometer. The thorough and exact analysis of biological samples is hardened by the complexity of the sample and the presence of many components simultaneously. To be able to get good qualitative and quantitative data, we need to use both selective and sensitive techniques, like LC-MS. A huge quantity of data is acquired during these studies (several tens of thousands of spectra). In the past the evaluation of the data was done manually, but now days this is impossible. Even though, more and more international companies deal with development of informatics tools for data analysis (MatrixScience, Thermo Scientific), the quality of evaluated data is questionable in many cases.

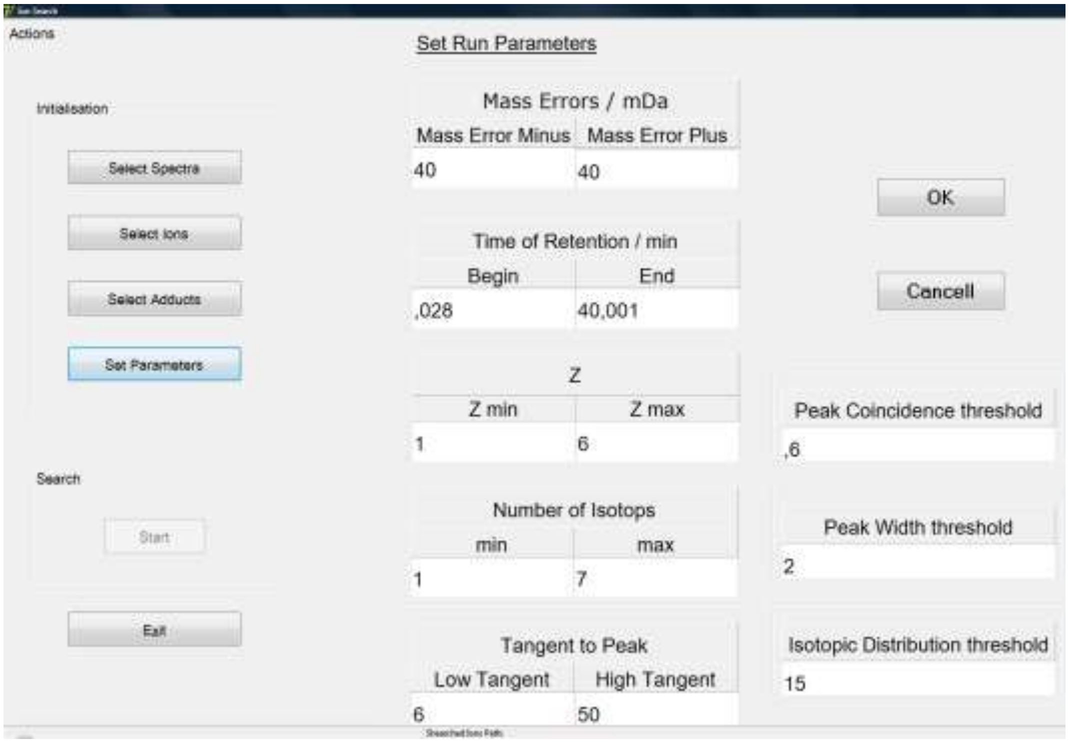
There are several situations, when there is need for comparative studies on samples from different locations (like different individuals). The acquired data must be processed and compare on the basis of a few selected criteria. Several hundred, sometimes several thousand measurements are done. This means there is need to evaluate large quantity of data, which is lengthy and the automation is not always solved.

Our aim was to develop computer software, which is able to detect and quantify many components simultaneously in a large number of samples. The task of the software is to search for spectral characteristics. This is hard to do manually, only based on component mass. To detect each component accurately one must take into account other features also (retention time, mass, intensity, adducts, isotope distribution, shape of the chromatographic peak). Another aim is to use the developed software for solving concrete scientific problems. The use of the software should considerably increase the evaluation of measurement data, thus reduce the time needed to obtain scientific information.

The current partnership increases the efficiency of research in the described field. Both research groups are able to gather experience in the research field of the cooperation partner. There is being possibility for use of modern mass spectrometers, which is highly important both scientifically and for industrial perspectives. We plan to publish future results in international journals and conferences.

During our work we used data recorded by a Q-ToF Premier mass spectrometer equipped with an elecrospray ion source (ESI). Sample inlet was achieved using a liquid chromatograph attached to the mass spectrometer. The ESI is one of the most widely used ionization technique in case of proteomic studies. The liquid sample solution is mixed with the eluents, and through a capillary, it reaches the ion source. Here, due to the high voltage applied the nebulising gases and the heating it will vaporize and reach the analyzer and the detector. During the ionization process several types of molecule-ion adducts can be formed. This considerably hardens the evaluation of the spectra.

In the previous faze of the project, we developed a computer software. This was our base for further development, during which we enhanced our software with a new graphical user interface (Figure 1.) This enables the easy usage of the program and a high degree of customization of input parameters. The software is able to identify chemical units from a chromatographic run. It is possible to search for 2000 chemical units simultaneously, and to identify and quantify these. A further possibility for development is the support for simultaneous search of several data files. To achieve the described properties, we developed several new algorithms. The algorithms designed for validating the evaluation results are described in the following.



Set Run Parameters

Mass Errors / mDa

Mass Error Minus	Mass Error Plus
40	40

Time of Retention / min

Begin	End
,028	40,001

Z

Z min	Z max
1	6

Number of Isotops

min	max
1	7

Tangent to Peak

Low Tangent	High Tangent
6	50

Peak Coincidence threshold
,6

Peak Width threshold
2

Isotopic Distribution threshold
15

Figure 1.: The graphical user interface of the developed software

The Align Algorithm

Our data was acquired using a liquid chromatograph coupled to a mass spectrometer. The advantage of the liquid chromatograph is that it is able to separate in time chemical entities with different properties, and the elution interval for a chemical unit is relatively short. This enables us to assign a retention time window to each chemical unit. The Align algorithm makes use of this property of chromatography. The time window has a width, which must be greater than a predefined value for the algorithm to accept it. This makes possible the elimination of some of the noise peaks.

One chemical unit is a sum of several mass to charge ratios (m/z) (due to the existence of isotopic peaks and different adducts). We wanted to determine which mass over charge values belong to the same chemical unit. For this a great help is the retention time. The retention time is the time (or place) of the gravity center of the chromatographic peak. In an ideal case all components of a chemical unit have the same retention time, i.e. for all mass to charge values belonging to the same chemical unit the center of their chromatographic peak are at the same point. Due to imprecision of experiment technique, a peak coincidence threshold is necessary. The Align algorithm searches for all possible m/z values between the given boundaries. First it estimates the retention time window for each m/z value. This is achieved by finding chromatographic peaks by monitoring the tangent to the current intensity values. After this, the gravity center of each peak is calculated, by fitting an ideal function (a Gaussian in most cases), and taking its middle. Comparing all these retention times, the algorithm selects those m/z values that have the same retention time as belonging to the same chemical unit. As an output we get a list of chemical units and retention times, found with high confidence factor. Most of the noise is filtered out, as are some of the coeluting chemical units.

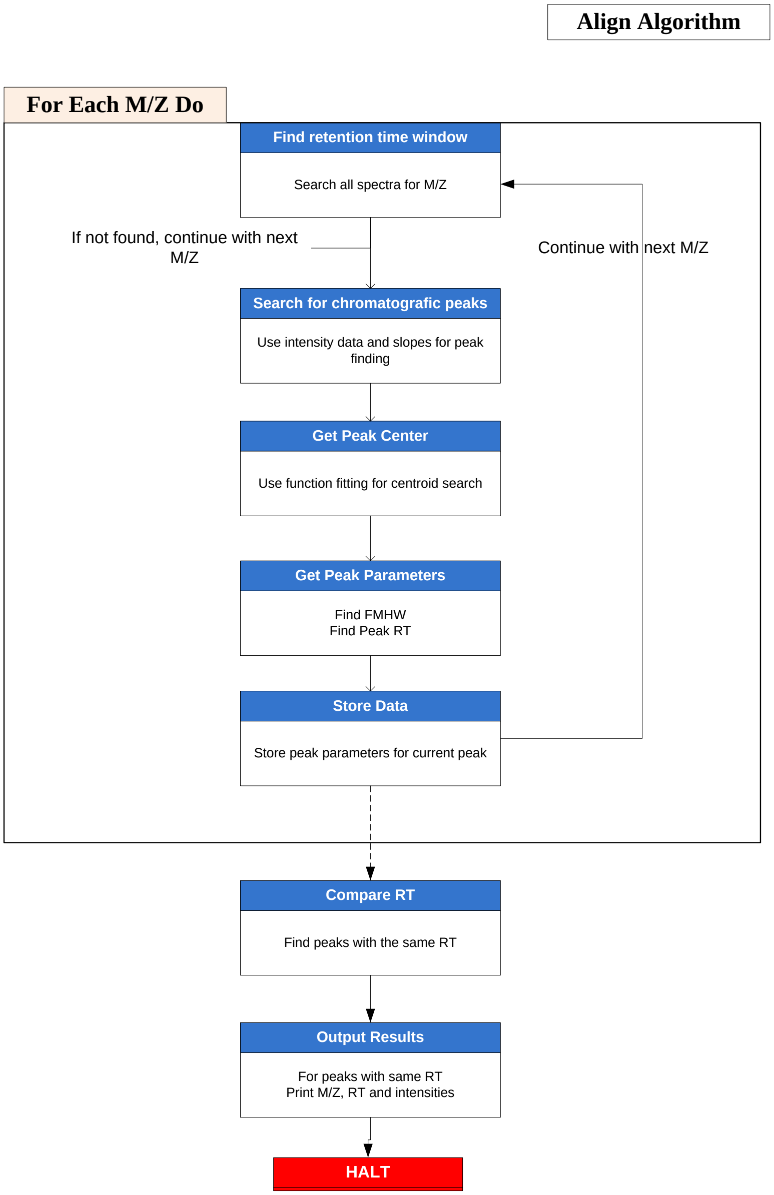

Align Algorithm
For Each M/Z Do
Find retention time window
Search all spectra for M/Z
If not found, continue with next M/Z
Continue with next M/Z
Search for chromatografic peaks
Use intensity data and slopes for peak finding
Get Peak Center
Use function fitting for centroid search
Get Peak Parameters
Find FMHW
Find Peak RT
Store Data
Store peak parameters for current peak
Compare RT
Find peaks with the same RT
Output Results
For peaks with same RT
Print M/Z, RT and intensities
HALT

Figure 2.: The Align Algorithm

The IsotopePattern algorithm

Most of the atoms on Earth have two or more stable isotopes.

Most of the atoms on Earth have two or more stable isotopes (with the same proton number but different number of neutrons in the nucleus). Consequently, the molecules formed by atoms may have different masses, although we speak the same molecules because the chemical properties are identical. All organic compounds characterized by the presence of ^{12}C and ^{13}C isotopes. These signs are visible in the mass spectra. This property is used the organic compounds identifying and the noise distinguishing, in the mass spectra. Algorithms were developed for the use of this feature. These algorithms compare the theoretical and measured isotope distribution and approve or disapprove the chemical unit.

Firstly, the algorithm calculates the expected theoretical isotope distribution to the actual chemical unit. The theoretical isotope ratios are calculating in accordance with the binomial theorem. The intensity of each isotope peaks normalized to the total intensity.

$$(x+a)^n = \sum_{k=0}^{n} \binom{n}{k} x^k a^{n-k}$$

„x” isotope ratios, k index of isotope peak, n maximum of isotope number, a natural isotope ratios

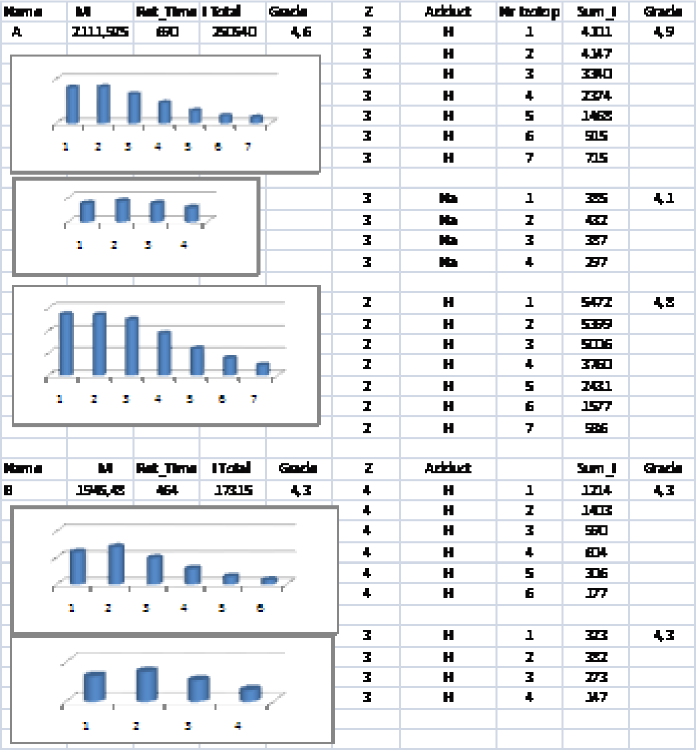
On the basis of measured data calculate the "measured isotope distribution", in this instance each intensity data was also normed on the total intensity. The "calculated isotope distribution" and the "measured isotope distribution" were compared. To do this, relative peak intensity of each isotopes used. The analogy to characterize, first calculated the square of the difference between the relative intensities, the amount thereof, then divided by the expected number of isotopes. The resulting number shows the similarity from measured and theoretical isotope distributions.

In order to be able to decide how much is the variation of the natural isotopic distribution, hundreds of spectra were examined. The standard deviation obtained from the study, used to give an acceptance criteria for the algorithm. This can be modified by the user (Isotopic distribution threshold, Figure 1.). If the similarity is within the specified threshold, the chemical unit is accepted.

Applications of the software. Achieved results

The Mass Spectrometry group of the Chemical Research Center of the Hungarian Academy of Sciences has been studying protein glycosylation for a long time. They use the alpha-1-acid glycoprotein (AGP) as a model compound. We used the data acquired of a tryptic digest of AGP to test our software, because it contains a large number of glycopeptides with different composition, all of which are well known to us. Running the software on afore mentioned data, it gave the expected results and no false identifications were present. This indicated that the software is working well, and we might it on try unknown mixture's data.

In the 3rd figure we present a part of the software's output. It found several m/z values belonging to the "A" molecule: doubly and triply charged species and Na adducts. The isotope distribution of these is visible in the picture. The quality of validation was designed with a grade, which must be between 3 and 5 to be acceptable, the higher the better.



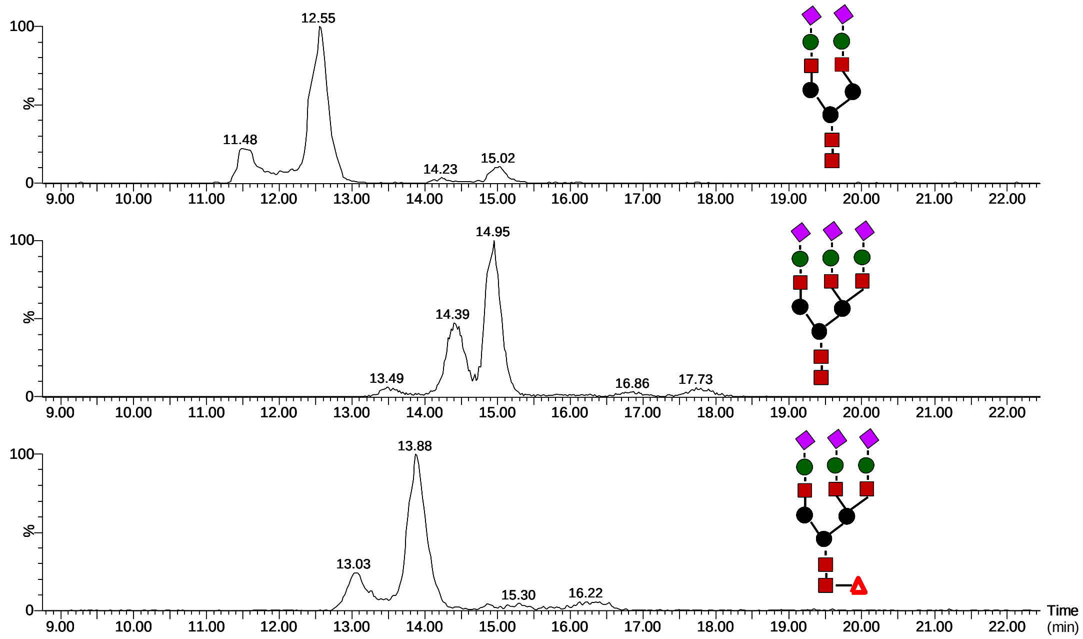
Name	M	Ret_Time	I Total	Grade	z	Adduct	Nr Isotop	Sum_I	Grade
A	2111,505	690	250640	4,6	3	H	1	4301	4,9
					3	H	2	4147	
					3	H	3	3340	
					3	H	4	2374	
					3	H	5	1468	
					3	H	6	915	
					3	H	7	715	
					3	Na	1	335	4,1
					3	Na	2	432	
					3	Na	3	387	
					3	Na	4	297	
					2	H	1	5472	4,8
					2	H	2	5399	
					2	H	3	5006	
					2	H	4	3760	
					2	H	5	2431	
					2	H	6	1577	
					2	H	7	586	
Name	M	Ret_Time	I Total	Grade	z	Adduct		Sum_I	Grade
B	1946,45	464	17315	4,3	4	H	1	1214	4,3
					4	H	2	1403	
					4	H	3	580	
					4	H	4	604	
					4	H	5	306	
					4	H	6	177	
					3	H	1	323	4,3
					3	H	2	352	
					3	H	3	273	
					3	H	4	147	

Figure 3.: A part of the software's output list

We applied our software to an unknown mixture of glycans generated by digesting AGP with PNGase-F enzyme. Since we didn't know what molecules to look for, a list of most frequent glycans was generated. This list is included in the supplementary data. The software is able to accept these masses as input. I will generate a list of possible m/z values and then search for these values in the mass spectrometric data file.

The found chemical units are listed the presented format. The findings of the software were verified manually in some cases. We searched for three m/z values belonging to three sugar compositions, using the MassLynx software. The results are presented in figure 4. For each m/z we get more than one chromatographic peak, which each must be manually analyzed, and the proper one accepted. This takes a lot of time and considerable knowledge. Our software is able to automatically find the correct retention time windows and to search for many m/z values simultaneously. Another advantage of our software is its considerable speed over manual search. Manual search could take hours or days, whereas the developed software search finishes in minutes or tens of minutes (2200 m/z values searched for in less than 40 minutes in a 70 minute chromatographic run).

Using our software it is possible to rapidly search many measurement data files, and for example make fast comparison of glycosylation profile of different individuals (like healthy and diseased).


Time
(min)