Data and text mining

Expertomica metabolite profiling: getting more information from LC-MS using the stochastic systems approach

Jan Urban*, Jan Vaněk, Jiří Soukup and Dalibor Štys

Institute of Physical Biology, University of South Bohemia, Academic and University Center, Zámeček 136, 373 33 Nové Hrady, Czech Republic

Received on April 21, 2009; revised on June 10, 2009; accepted on July 7, 2009
Advance Access publication July 14, 2009

Associate Editor: John Quackenbush

ABSTRACT

Mass spectrometers are sophisticated, fine instruments which are essential in many applications. However, the data they produce are usually interpreted in a rather primitive way, without considering the accuracy of this data and the potential errors in identifying peaks. Our new approach corrects this situation by dividing the LC-MS output into three components: (i) signature of the analyte, (ii) random noise and (iii) systemic noise. The systemic noise is related to the instrument and to the particular experiment; its characteristics change in time and depend on the analyzed substance. Working with these components allows us to quantify the probability of peak errors and, at the same time, to retrieve some peaks which get lost in the noise when using the existing methods. Our software tool, Expertomica metabolite profiling, automatically evaluates the given instrument, detects compounds and calculates the probability of individual peaks. It does not need any artificial user-defined parameters or thresholds.

Availability: MATLAB scripts with a simple graphical user interface are free to download from http://sourceforge.net/projects/expertomica-edata/. The software reads data exported by most Thermo and Agilent spectrometers, and it can also read the more general JCAMP-DX ASCII format. Other formats will be supported on request, assuming that the user can provide representative data samples.

Contact: urban@greentech.cz

1 INTRODUCTION:

Mass spectrometers are sophisticated, fine instruments which are essential in many applications. However, the data they produce are usually interpreted in a rather primitive way, without paying attention to the accuracy of this data and to its effect on the final results we get. This has several reasons:

1. The data accuracy vary during the experiment, and depend on the analyzed substance.
2. The results have a discrete character. For any time and mass, we want to know whether we have a peak.
3. Manufacturers of mass spectrometers fiercely compete, and revealing details about the errors of their instruments could negatively impact on their business.

The spectrometer samples the intensity $y$ of its output in selected times $t$, and for selected values of mass to charge ratio, $m$. This 3D relation $y(t,m)$ is usually shown as two 2D graphs, $y(t)$ and $y(m)$, which are projections of $y(t,m)$ to planes $(y,t)$ and $(y,m)$.

Our software tool, Expertomica metabolite profiling (EMP), uses the original 3D representation, and is based on the new systems theory proposed by Pavel Žampa, see Žampa and Arnošt (2004). We assume that, at a given point $(t,m)$, the output intensity $y(t,m)$ represents only one of three possible components:

$$y(t,m)=x(t,m)\lor r(t,m)\lor q(t,m),$$

where $x(t)$ is the signature of the analyte, $r(t)$ is a random noise and $q(t)$ is a systemic noise. The characteristics of the systemic noise change in time and depend on the instrument and on the analyzed substance. In the first phase, our software determines the following probabilities:

$$p_r(t,m) = \text{probability that } y(t,m) \text{ is not a random noise;}$$

in other words, it is either a part of the analyte signal or systemic noise

$$p_q(t,m) = \text{probability that } y(t,m) \text{ is not a systemic noise;}$$

in other words, it is either a part of the analyte signal or random noise

The program converts these internal probabilities into

$$p(t,m) = p_r(t,m) p_q(t,m)$$

which gives the probability that $y(t,m)$ represents the signature of the analyte, $x(t,m)$.

The key idea is that the presence of potentially useful signal can be detected as a violation in the noise behavior. For each time slice ($t = \text{const.}$), we determine characteristics of the random noise, and for each mass/charge slice ($m = \text{const.}$), we determine the characteristics of the systemic noise. By comparing the characteristics of the two noises with the spectrometer output $y(t,m)$, we can find the probabilities we need. Isolated spikes are treated as a special case and are removed.

This splitting of noise along the $t$ and $m$ directions is based on our observation that the random noise changes more in time $t$, while the systemic noise changes more with the mass to charge ratio, $m$.

Note that this approach is based on a model of the physical process inside the instrument, and has a potential for significantly better...
results than blind regression or pattern matching. Our probabilities unlock a valuable information which has always been in the spectrometer outputs but, until now, it has been completely ignored.

The software automatically evaluates the given instrument, detects peaks and calculates the probabilities of their existence. There are no artificial, user-defined parameters or thresholds. The results include the accuracy of the interpretation, and we often detect peaks which cannot be extracted from the noise, using the existing methods.

We believe that this statistical approach to metabolite profiling is new, and we are not aware of any publication or software based on this method. There are, however, several programs for parameter-free analysis of LC_MS spectra. For example, Parafac ([Arroyo et al., 2008]) uses regression analysis, while MS-Resolver (Pattern Recognition Software AS, 2004, http://www.prs.no/MS Resolver/Brochure.html) uses pattern recognition.

2 IMPLEMENTATION

2.1 Algorithm

Step 1: for each \( t_j \), determine probabilities \( p_{ij}(t_j,m) \), \( j=0,1,2, \ldots \) that we have a signal or systemic noise.

Step 2: for each \( m_j \), determine probabilities \( p_{ij}(t_j,m) \), \( i=0,1,2, \ldots \) that we have a signal, and not just pure noise. The shape of typical peaks, their tailings and the tail caused by the mobile phase are used in this calculation.

Step 3: multiply the two probabilities for each \( y(t,m) \): \( p(t,m) = p_{ij}(t_j,m) p_{ij}(t_j,m); \ 0 \leq p(t,m) \leq 1 \).

Step 4: set \( p(t,m) = 0 \) for all isolated spike locations, defined as \( p(t,m) = 0 \) for all \( i=0,1,2, \ldots \) except for \( t_j = t \).

Step 5: detect peaks by searching for local maxima and minima, and by searching for \( p(t,m) \) turning from 0 to positive values.

Compounds combine several substances (isotopes, fragments, adducts), each with a particular configuration of peaks. We assume that peaks with a similar time-shape belong to the same substance. Also, most substances can be identified by a single large peak, the mere subtraction does not eliminate the peak.

2.2 Using MATLAB script

Our software (EMP) is implemented with Matlab scripts. The input is currently restricted to the commonly used increment of 0.1 Da for the mass to charge ratio \( m \). High-resolution data cannot be processed because of the increased computational demands (two orders of magnitude). Also, some parts of our algorithm assume discrete data, but the high-resolution data display a continuous behavior. For example, in high-resolution data, isolated peaks may not be restricted to a single point \( t,m \). We are currently researching how to expand the capability of EMP to high-resolution data.

During the routine use of the software, the operators from our LC-MS lab observed that the quality of the results depends on the stability of the instrument. For some calculations, the existing algorithm rounds up all \( m \) values to integers, which makes it relatively insensitive to small drifts. We are working on a new methodology which, with the help of a special calibration, will allow our software to handle larger drifts.

3 EXAMPLE

In the automatic mode, our software converts the spectrometer output to a table of results, such as table in Figure 4. In the semi-automatic mode, the software extracts the peaks from the spectrometer output, reports them as traditional 2D chromatograms or as a 3D diagram such as Figure 2B, and lets the user to derive the substances and compounds manually. The description of this example considers both these possible uses.

Figure 1 shows an example of using EMP on highly difficult data, where only a few compounds can be found manually. Figure 1A shows the raw \( y(t,m) \) coming from the spectrometer; rainbow colors are used just to help the visualization. The graph is dominated by several solvent compounds which produce a rising systemic noise. Figure 1B shows the chromatogram after the EMP analysis. For the purpose of the display, the peaks with \( p(t,m) > 0.01 \) are not shown.

Figure 2 shows the breakdown of the spectrum to the signal, random noise and the systemic noise, all projected to the \((t,y)\) plane. Figure 2B shows the area under individual peaks, integrating by \( m \).

Figure 3 shows the second phase of the spectrum processing which includes compound assignment (Fig. 3B), intensities of most probable signals (Fig. 3C) and the integral spectrum (Fig. 3D) showing the area under individual peaks, integrating by \( t \).

Figure 4 shows the result table: the compound represented by the spectrum, time of the peak beginning, peak end, maximum, standard deviation and its probability. The probability of the peak and of the compound may be defined in several ways, and we have chosen the following formula: the peak probability is the highest probability of all signals assigned to the peak. The compound probability is the highest probability of all its peaks.

The program reads the spectrometer data and the data from the empty run. The user specifies the confidence level (probability) with which he/she wants to export the results. The results can be shown either as standard 2D chromatograms, as a table (compounds with peaks and their probabilities) or as the 3D diagram.

The integral chromatogram is useful in both the automatic and semi-automatic modes of using the software. For each \( m \), the program detects the beginning \( t_0 \) and the end \( t_e \) of each peak, and then adds together all the values of \( y(t,m) \) for \( t = t_0, \ldots, t_e \). The resulting chromatogram shows the area rather than the height of the individual peaks. When deciding on substances and compounds, the area corresponds to the total mass of the peak, which is a better measure than its height.
Fig. 1. Example of the chromatogram analyzed by the EMP software. (A) The original chromatogram and (B) the chromatogram after the analysis.

Fig. 2. The three components of the chromatogram (signal, random and systematic noise) and their projection onto the \((t, m)\) plane. In (A), blue dots represent the sum of all \(y(t, m)\) at each \(t\). Green open circles represent the sum of all signals, here sum of all \(y(t, m)\) with \(p(t, m) > 0.01\). Red triangles represent the sum of random noise, here sum of all \(y(t, m)\) with \(p_{r}(t, m) \leq 0.01\). Black stars represent the sum of systemic noise, here \(y(t, m)\) with \(p_{q}(t, m) \leq 0.01\). (B) The sum of all the \(y(t, m)\) which, after the spike elimination, are still assigned to a peak.

4 PRACTICAL EXPERIENCE

The manual analysis of the LC-MS data is a time consuming, hard work and its results depend on the experience and patience of the operator. A typical operator cannot analyze the spectrum with the degree of precision and comprehensiveness provided by our software—see example in Figure 4. Single comprehensive analysis took between 3 and 7 h per sample. The operator, however, often considers not only the spectrometer data, but also uses chemical rules which are not included in our software. The computer run took 7.5 s for each sample, using the version implemented with Matlab (on 2.4 GHz CPU, 2 GB RAM).

As a test, we compared automatically produced results with the manual analysis on a series of 50 routine measurements from Agilent 1100 LC/MSD Ion Trap system. The samples contained biomass extracts of different cyanobacterial strains, with substances eluting at conditions similar to the oligopeptide nystatin which was added as a standard. Note that spectrum standards for most of the compounds we anticipated do not exist.

The manual analysis was done by a highly skilled operator familiar with this type of data. She detected the peaks from the 2D chromatogram showing the areas under the peaks, and reported all compounds which she could see. In 88% of the cases, the results were identical with the automatically generated results. In four cases (8% of the tests), the software identified one more compound which was subsequently approved by the operator. In two cases (4% of the tests), the operator found a compound which was not detected by the software. The operator’s clue was that 'in similar samples the compound was usually there'. There was no case of the software reporting a compound which would not be approved by the operator.

After about a year of experimental testing of this software in our lab, we made the following observations:

(a) When \(m\) is recorded with the increment of 0.1 Da, for a given time \(t\), the peak value of \(y(t, m)\) often repeats for several values of \(m\).

(b) This effect becomes more pronounced with the increase of the intensity \(y(t, m)\).

(c) This implies that \(y(t, m)\), \(p_{r}(t, m)\) and \(p_{q}(t, m)\) are not completely independent as our algorithm assumes.

(d) For different settings of the instrument, the relations among \(y(t, m)\), \(p_{r}(t, m)\) and \(p_{q}(t, m)\) vary, but these differences are insignificant for low-resolution data.

The software is being continuously upgraded to different types of data and instruments; a list of applications and instruments is available at http://sourceforge.net/projects/expertomica-eda/.
Fig. 3. Results for the time $t = 16.11$. (A) Raw spectrometer output. (B) Probability of the signal to be a peak $p(t,m)$. (C) The sum of $y(t,m)$ in $m$ positions of most probable signals. (D) Integral spectrum showing the area under the peaks.

Fig. 4. Example of the output table produced by the EMP software. It shows one compound and its five peaks with their masses (Mass), probabilities (CF), the relative content of the compound in sample, the relative content of the peaks in this compound and the statistical parameters of the peaks.

5 CONCLUSIONS

Compared with the existing methods based on heuristic rules, regression or pattern matching, the presented method has the advantage of using a statistical model of what happens within the LC-MS instrument. The software built on this idea can be used for multiple purposes: for expert data assessment, automated generation of compound databases, performance analysis of the instrument or for validity assessment of biological models.

ACKNOWLEDGEMENTS

Authors thank to Lucie Marková, Pavel Hrouzek, Pavel Souček and Jiří Kopecký for software testing and numerous suggestions.

Funding: 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 (in part).

Conflict of Interest: none declared.

REFERENCES
