BMC Bioinformatics

RESEARCH Open Access

A coherent mathematical characterization of isotope trace extraction, isotopic envelope extraction, and LC-MS correspondence

Rob Smith[1*], John T Prince[2], Dan Ventura[3]

From The 11th Annual Biotechnology and Bioinformatics Symposium (BIOT-2014)
Provo, UT USA. 11-12 December 2014

Abstract

Background: Liquid chromatography-mass spectrometry is a popular technique for high-throughput protein, lipid, and metabolite comparative analysis. Such statistical comparison of millions of data points requires the generation of an inter-run correspondence. Though many techniques for generating this correspondence exist, few if any, address certain well-known run-to-run LC-MS behaviors such as elution order swaps, unbounded retention time swaps, missing data, and significant differences in abundance. Moreover, not all extant correspondence methods leverage the rich discriminating information offered by isotope envelope extraction informed by isotope trace extraction. To date, no attempt has been made to create a formal generalization of extant algorithms for these problems.

Results: By enumerating extant objective functions for these problems, we elucidate discrepancies between known LC-MS data behavior and extant approaches. We propose novel objective functions that more closely model known LC-MS behavior.

Conclusions: Through instantiating the proposed objective functions in the form of novel algorithms, practitioners can more accurately capture the known behavior of isotope traces, isotopic envelopes, and replicate LC-MS data, ultimately providing for improved quantitative accuracy.

Background

Liquid chromatography-mass spectrometry (LC-MS) is a popular technique for elucidating the composition of liquid samples. Data processing considerations are essential to accurately determine the identity of molecules (analytes such as lipids or peptides) contained in the sample (a process called identification), as well as their quantity in sample (a process called quantification).

Information about sample quantity is captured directly in survey scans, or MS (aka MS1) data. Fragmentation spectra of one or more analytes constitute MS/MS (or MS2) data, and this information is typically used to corroborate or ascertain the identity of a molecule. Partitioning/clustering MS1 signal from complex samples and mapping the signal to other analyses (correspondence) is challenging. Some quantification strategies bypass these challenges by using information derived directly or indirectly from MS/MS data. These methods include spectral counting [1] and isobaric tags for relative and absolute quantification (iTRAQ) [2]. Though these methods have been successful, the amount of quantifiable signal embedded in MS1 data is estimated to far exceed what is currently available by MS/MS [3]; however, most MS1 data remains unused by current software. Hence, improving methods for partitioning and mapping MS1 signal stands to significantly (˜10 fold) increase the sensitivity of a typical label-free or isotope-labeling MS-omics experiment, both for experiments currently being run and for past experiments where raw data is still available.

Subdivision of raw mass spectrometer output data into smaller signal partitions attributed to specific analytes in the sample is critical prior to achieving analyte identification

* Correspondence: 2robsmith@gmail.com
[1]Department of Computer Science, University of Montana, 59812 Missoula, USA
Full list of author information is available at the end of the article

and quantification. The larger partition unit, called an isotopic envelope trace, is the signal pattern generated by each analyte/charge combination (see Figure 1). Because mass spectrometers can only detect charged analytes, the sample must be subjected to an ionization method, which imputes a charge on each detected analyte. Since multiple instances of each component exist in the sample, and since each instance is charged independently, there exist in each output the signals of multiple analytes, each with (potentially) multiple charge states. These create a distinct signal–the isotopic envelope trace–for the total signal detected for each analyte/charge state combination. Each isotopic envelope trace is composed of a series of isotope traces, which are manifestations of the fact that each analyte is composed of chemically similar compounds that differ in the weight of certain isotopes (such as ^{12}C vs ^{13}C). At each charge state, each molecular variant of the analyte is detected at a particular m/z offset, creating one isotope trace per molecular variant/charge-state/analyte combination.

Mass spectrometry data, in its raw form, is not ideal for isotope trace extraction or subsequent processing. After internally accumulating signal over discrete time slices, the mass spectrometer outputs raw data condensed into the form of many narrow profiles wherever signal is present. Conversion to centroid mode integrates the abundance of each of these profiles into a single tuple called a centroid. This is considered a routine conversion for which ample software is readily available. We adopt the typical convention of using centroid data.

Despite the ubiquity of LC-MS experiments, to the best of our knowledge, no concise, complete description of the LC-MS isotope trace and isotopic envelope extraction problems exists. Here, we describe constructs for isotope traces and isotopic envelopes, as well as formally describe the relationship of centroids, isotope traces and isotopic envelopes. In this context, we review extant objective functions for isotope trace extraction, isotopic envelope extraction, and correspondence. Finally, we propose novel objective functions for each of these tasks that address shortcomings in current approaches.

Results and discussion

Isotope trace extraction

The most important data processing step in a typical quantitative LC-MS pipeline is isotope trace extraction [4]. Clustering centroids into isotope traces is a nontrivial problem due to the many sources of noise affecting centroid mass and abundance. Sources of noise affecting centroids include chemistry effects due to chromatography, abundance inaccuracy due to ionization efficiencies, m/z deviation due to machine calibration, occlusion/adulteration of low-abundance signal due to dynamic range limitations, and compounded inaccuracies in mass-to-charge ratio (m/z) and abundance due to centroid construction. Of course, these complications are propagated from the clustering of isotope traces to the clustering of isotopic envelopes to the identification of cross-experiment correspondence.

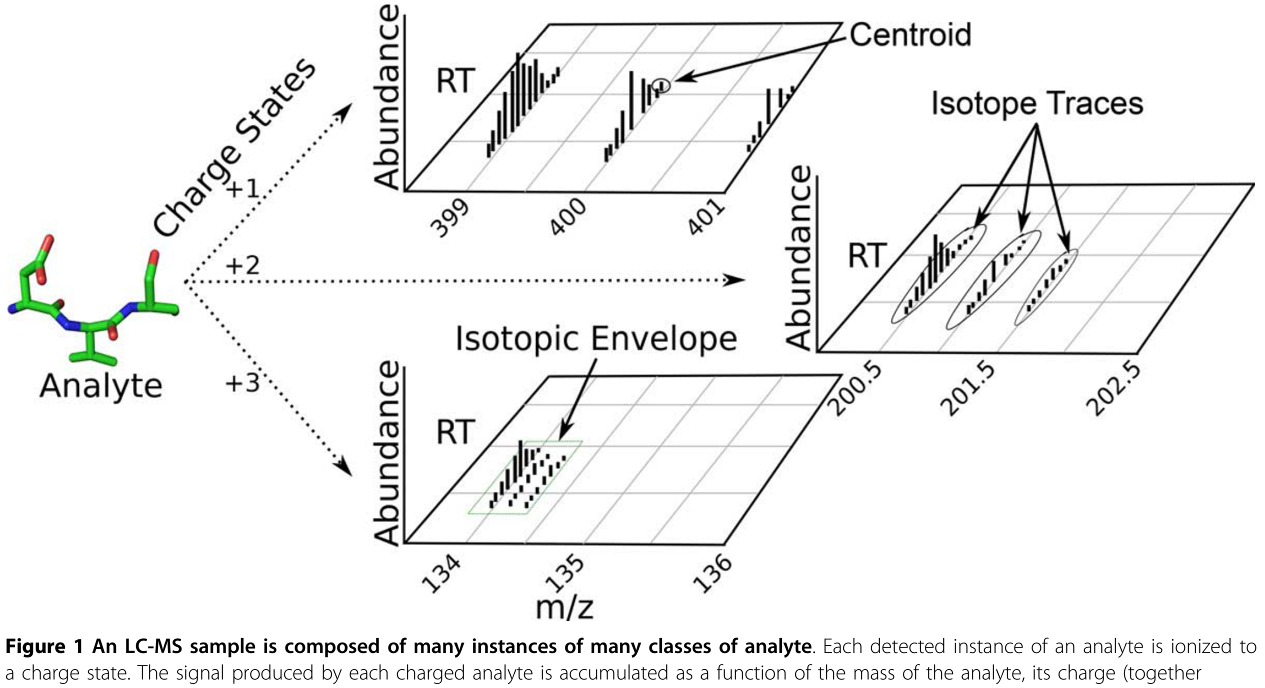

Figure 1 An LC-MS sample is composed of many instances of many classes of analyte. Each detected instance of an analyte is ionized to a charge state. The signal produced by each charged analyte is accumulated as a function of the mass of the analyte, its charge (together composing the mass-to-charge ratio (m/z)), and the time at which it is detected (dictated by the chromatographic system in use).

A centroid is denoted as $c = (\mu, \tau, \alpha)$ where μ, τ, α are values for m/z, retention time (RT), and abundance, respectively. A single MS run produces a set of centroids $C = \{c_i\}_{i=0}^{n}$, where n can readily reach into the millions.

An isotope trace $F \subset C$ is defined as a set of centroids: $F = \{c_i\}_{i=0}^{m}$, with each set F constrained so that all members of a given isotope trace F are within a distance threshold θ from other centroids in their neighborhood Υ (see Figure 2):

$$\max_{j \in \Upsilon_i} \quad \delta_F(c_i, c_j) < \theta^{\mu,\alpha,\tau} \tag{1}$$

where θ is a function of centroid m/z, RT, and abundance, δ_F is a distance function based on m/z, RT, and abundance, and Υ is a neighborhood demarcated by m/z, RT, and abundance. Additionally, the slope of a (abundance-weighted) linear regressor estimate for an isotopic trace is very nearly infinite (in the *m/z, RT*-plane). One way to formalize this is to use a weighted, inverse variant of the Theil-Sen estimator as follows (see Figure 3):

$$\frac{\sum_{c_i, c_j \in F} \dfrac{c_j^{\mu} - c_i^{\mu}}{c_j^{\tau} - c_i^{\tau}} c_j^{\alpha} c_i^{\alpha}}{\sum_{c_i, c_j \in F} c_j^{\alpha} c_i^{\alpha}} \approx 0 \tag{2}$$

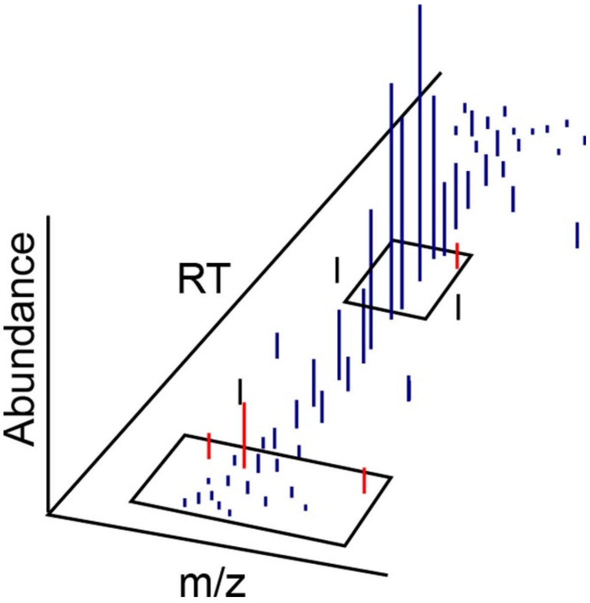

Figure 2 Each box illustrates an example candidate local neighborhood Υ defined by an algorithm-specific m/z and RT window. Blue centroids indicate the centroids pertaining to the isotope trace, while red centroids have been rejected due to differences in abundance, m/z, and/or RT compared to other centroids in Υ.

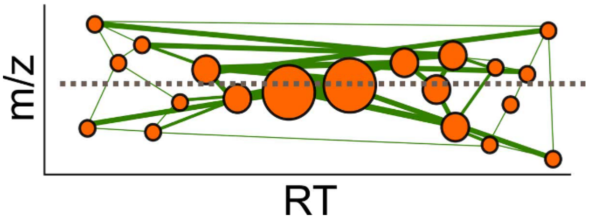

Figure 3 One way to characterize the relative lack of variance in m/z (compared to RT) of an isotopic trace is by using an inverse variant of the Theil-Sen estimator–a fully-connected graph is constructed with edges connecting each pair of centroids (circles whose radius indicates abundance), and weighted by the abundance of its connected centroids (represented by line thickness). An isotopic trace will have a weighted average (inverse) slope of approximately zero (not all connections shown).

where c^{α} is the abundance of centroid c and c^{μ} is the m/z of centroid c.

Note that the behavior of isotope traces are dependent on all three MS dimensions although many common approaches to isotope trace extraction ignore one or more of these dimensions. For example, most proprietary MS software uses hard m/z bins for isotope trace extraction.

Extant objective functions

The prominent algorithms for isotope trace extraction include centWave [5], MatchedFilter [5], centroidPicker [6], massifquant [7], and MaxQuant [8].

MatchedFilter operates on the simplifying assumptions that 1) isotope traces are completely contained within pre-processed hard m/z bins and 2) the shapes of all isotope traces in a run can be fit to the same shape. MatchedFilter minimizes the error of a Gaussian fit over prospective isotope traces, by attempting to find the set of isotope traces $\mathcal{F}$, a scaling factor b_F, and mean retention time F^t for each isotope trace that minimizes the summed abundance error over all isotope traces. Note the use of a single, global variance σ, an average RT width for all $F \in \mathcal{F}$:

$$\lambda_F = \sum_{F \in \mathcal{F}} \sum_{c \in \mathcal{F}} \left| bFe^{\dfrac{-(c^{\tau} - F^t)^2}{2\sigma^2}} - c^{\alpha} \right| \tag{3}$$

The centWave algorithm extracts isotope traces that fit a scaled and translated Ricker wavelet ζ (commonly called a Mexican hat function). The fit is calculated as a convolution between the shape function and the signal intensity (abundance), so the goal is to maximize the objective function:

$$\lambda_F = \sum_{F \in \mathcal{F}} \sum_{c \in \mathcal{F}} c^{\alpha} \zeta(c) \tag{4}$$

where

$$\zeta(c) = \left(\frac{1}{\sqrt{b_F}} \frac{2}{\sqrt{3}\pi^{\frac{1}{4}}} \left(1 - \left(\frac{c^\tau - t_F}{b_F} \right)^2 \right) e^{\frac{-\left(\frac{c^\tau - t_F}{b_F} \right)^2}{2}} \right) \tag{5}$$

with isotope trace-specific scaling parameter bF and translation parameter t_F chosen to maximize the convolutional fit over isotope trace F .

The algorithm centroidPicker uses heuristic operations on a neighborhood graph to separated the data into connected components. It connects an undirected graph $G = (C, N)$ of centroids where the edges N are constrained such that:

$$N = \left\{ (c_i, c_j) \middle| \begin{array}{c} \delta_c (c_i, c_j) < \delta_c (c_i, c_k) \forall_{k \neq j} \\ c_i^\alpha > \theta \text{ and } c_i^\alpha > \theta \end{array} \right\} \tag{6}$$

for some intensity threshold θ and centroid distance function δ_c, resulting in G being composed of one or more connected components, each considered one isotope trace. Thus, $\mathcal{F} = \{F_i | \forall c_k \in F_i, \exists_{cl \in F_i} \{c_l \in \Upsilon(c_k)\}\}$, where the neighborhood function $\Upsilon\ (c)$ returns the set of nodes connected to c (and is symmetric because G is undirected).

The objective functions for massifquant and MaxQuant define $\mathcal{F}$ as the set of all F formed by iterating over values of time t, and adding c if $c^\tau = t$ and $\left| c^\mu - c_*^\mu \right| < \in$, where $c_* \in F$ and $c^\tau - c_*^\tau \leq c^\tau - c_j^\tau$ for all $c_j \in F$. For massifquant, $\in$ is prescribed by a Kalman filter induced from the variance in c^μ and c^α for all $c_j \in F$ such that $c_j^\tau < t$, with the added constraint that c^τ be unique in F . MaxQuant defines $\in$ simply as a distance threshold of 7 ppm m/z.

Proposed objective functions

We define F^μ, the m/z of isotope trace F, given by the weighted m/z of its component centroids:

$$F^\mu = \frac{\sum\limits_{c \in F} c^\alpha c^\mu}{\sum\limits_{c \in F} c^\alpha} \tag{7}$$

and using it propose an alternative objective function for isotope trace extraction:

$$\lambda_F = \sum_{F \in \mathcal{F}} \sum_{c \in F} \left| b_F(\tau) e^{\frac{-(c^\tau - F^t)^2}{2\sigma_F^2}} a_F(\alpha) e^{\frac{-(c^\mu - F^\mu)^2}{2h(\alpha)^2}} - c^\alpha \right| \tag{8}$$

where, again, centroid clustering $\mathcal{F}$ and retention time means F^t are chosen to minimize the Gaussian fit error; however, rather than using a single global variance in the RT dimension, each isotope trace F has a local variance σ_F; in addition, the scaling factors have become time-dependent scalar functions $b_F(\cdot)$. The second Gaussian factor, parameterized by mean F^μ and variance function $h(\cdot)$, models the m/z width of the isotope trace, which is a function of the abundance α. Isotope traces splay at low abundance and narrow at high abundance; thus, both the variance $h(\cdot)$ and the scaling factors $a_F(\cdot)$ are modeled as functions dependent on the abundance α. Note that while variance is trace-independent (depending only on abundance), each isotope trace has its own scaling function (which in turn is dependent on abundance).

Alleviating current limitations in isotopic trace extraction

Current objective functions for isotopic trace extraction fail to capture isotopic trace behavior formalized in this section: namely, a pattern of centroids forming a generally tight distribution through time around a specific m/z, with variation occurring as a factor of abundance, with normal abundance traces splaying at the beginning and end of elution, and lower abundance traces displaying high m/z variance in general. Moreover, isotope traces are skewed in time, with sharp onset of intensity followed by a post-peak long tail. The shape of traces is almost never strictly Gaussian (or even symmetric), as chromatography almost always deviates from the Gaussian in heading (which is more steep) and in tailing (which is less steep). Our objective functions account for each of these behaviors.

Isotopic envelope extraction

The LC-MS clustering problem is defined as a two-step partitioning problem. In the first step, isotope trace extraction, we require a partition φ of the set of all centroids C into the set of isotope traces $\mathcal{F}$, $\phi(C) = \{F_i\}_{i=1}^r = \mathcal{F}$ with the properties:

$$\bigcup_{i=1}^{r} F_i = C \quad \text{and} \quad F_i \cap F_j = \emptyset \quad \forall_{F_i \neq F_j \in \mathcal{F}} \tag{9}$$

In other words, 1) all centroids are assigned to an isotope trace; 2) isotope traces can’t share centroids. Because any sensor’s detection of a physical system will deviate somewhat from the true physical system, we can expect MS detections to contain extraneous centroids. However, all signal ought to be accounted for (even if some identified “traces” eventually are identified as noise) and, in a platonic model, ought to be assigned to an isotope trace.

In the second step, isotopic envelope extraction, we require a partition ψ of the set of isotope traces $\mathcal{F}$ into the set of isotopic envelopes ε, $\psi(\mathcal{F}) = \{E_i\}_{i=1}^p = \varepsilon$ with the property

$$\bigcup_{i=1}^{p} E_i = \mathcal{F} \tag{10}$$

The choice of partitions ϕ and ψ is guided by a set of distance functions Δ that define distances between centroids, isotope traces, isotopic envelopes, etc. and objective functions λ_F and λ_E that describe "good" isotope traces and isotopic envelopes, respectively. The choice of distance and objective functions, along with choice of optimization procedure, characterizes an algorithmic approach for solving this clustering problem. A defining general property of isotopic envelopes, however, is the regular spacing between component isotope traces. In addition, for virtually all molecules from biological sources we expect that if there is an isotope with index j and an isotope with index $j + 2$, then there exists an isotope with index $j + 1$.

An isotopic envelope E is the set of isotope traces F_i that are produced by a given analyte/charge state combination: $E = \{F_i\}_{i=0}^{q}$ subject to the constraint that the m/z difference between each consecutive (assuming an ordering of centroids from least mass to greatest mass) isotope trace in E must be equivalent to $\frac{k}{z_E} + \in$, where k is the mass of a neutron, z_E is the integer charge of E and $\in$ is a noise tolerance parameter. That is, assuming an indexing function $\iota^{\mu} : \varepsilon \times \mathcal{N} \rightarrow \mathcal{F}$ that returns the ith least massive isotope trace in an isotopic envelope:

$$l^{\mu}(F, i + 1) = l^{\mu}(F, i) = \frac{k}{z_E} + \in, \quad 1 \leq i \leq |E| - 1 \tag{11}$$

The m/z m of the jth isotope trace in E must be roughly equivalent to

$$m = \frac{\tilde{m} + jk}{z} \tag{12}$$

where $\tilde{m}$ is the uncharged molecular weight of the ion.

Every isotope trace consists of signal from at least one isotopic envelope, and, in the case of overlapping isotopic envelopes, an isotope trace may be composed of signal from more than one isotopic envelope.

Extant objective functions

FeatureFinder [9] is an isotopic envelope extraction algorithm in OpenMS that searches directly for E. Although the details are not completely clear, it appears that the algorithm attempts to minimize

$$\lambda_E = \sum_{E \in \varepsilon} \sum_{c \in E} G_E(c) \tag{13}$$

where the G_E compute a comparison between the (μ, τ, α) values for a centroid and the expected centroid values obtained from a heuristic isotopic envelope shape. Note that isotopic trace extraction is ignored.

MSInspect [10], another approach to isotopic envelope extraction, groups all coeluting signals and compares them to a simulated envelope calculated from a Poisson distribution parameterized by m/z, with the goal being to minimize the KL divergence between the Poisson distribution and the "distribution" of abundance in an instantaneous profile of the envelope at time τ :

$$\lambda_E = \sum_{F \in E, c \in^{\tau} F} \hat{P}(c^{\alpha}) \log \frac{\hat{P}(c^{\alpha})}{P_m(c^{\mu})} \tag{14}$$

where the notation $c \in^{\tau} F$ means that $c \in F$ at time τ, E is the maximal intensity (instantaneous) isotopic envelope (at time τ), $\hat{P}(\cdot)$ is the ratio of the intensity of isotope trace F (at time τ) to the total intensity of all isotope traces $F \in E$ (at time τ), and $P_m(\cdot)$ is the value of the Poisson distribution at c^{μ}.

Proposed objective functions

We propose an alternative objective function for isotopic envelope extraction:

$$\lambda_E = \beta I(E) + (1 - \beta) J(E), 0 \leq \beta \leq 1 \tag{15}$$

where β is a relative importance weighting coefficient. The first term computes the deviation of member isotope traces from the expected charge-based m/z interval–we want the isotope traces in envelope E to fit expected m/z spacing:

$$I(E) = \sum_{\substack{F_i, F_j \in E \wedge \\ F_i^{\mu} < F_j^{\mu} \wedge \\ \forall_{F_k^{\mu} \in E} F_k^{\mu} > F_i^{\mu} \Rightarrow F_k^{\mu} > F_j^{\mu} \vee F_k = F_j}} |(F_i^{\mu} - F_j^{\mu}) - \frac{k}{z_E}| \tag{16}$$

The second term computes the deviation in elution time of member isotope traces–we want all the isotope traces in isotopic envelope E to co-elute within a small time window:

$$J(E) = \sum_{F_i, F_j \in E} F_i^{\tau} - F_j^{\tau} \tag{17}$$

where F^{τ} could be defined analogously to Equation 7, could be the maximum intensity for isotopic trace F or could be some other reasonable definition for isotopic trace elution time.

We want to optimize ε and the z_E so that λ_E is minimized; that is, we want to find charge-state/isotopic-envelope pairs such that the errors in expected m/z and co-elution time are minimized.

The isotopic envelope extraction segment of the MaxQuant [8] algorithm is one of the possible instantiations of this objective function, though many possibilities exist for how to set the allowable m/z and RT error and how to generate the prerequisite list of isotope traces.

Alleviating current limitations in isotopic envelope extraction

Isotopic envelopes are rich with data: the expectation of contiguous isotope traces with a uniform m/z charge gap, and similar maximal abundance across all isotope traces. Accounting for this behavior is not possible without adopting an isotope trace-centric approach to data extraction. Reliance upon maximal elution time alone–an approach that is susceptible to conflation with overlapping envelopes in complex samples–is not a sensitive approach in envelopes of lower abundance, where maximal elution times are not pronounced. Moreover, by first finding the isotope traces, the exact m/z of each isotope trace can be calculated using a weighted average, alleviating the need for larger than theoretically justified isotope trace gaps, which will not be sensitive in complex samples with overlapping isotopic envelopes. Instead, the proposed objective functions leverage a precise and reliable m/z charge gap and adjacency of isotope traces along with maximal elution times, using all the information in the data.

Correspondence

The final objective of almost every MS experiment is the differential analysis of more than one MS run. This comparison allows the identification of significant quantity and component differences, useful for applications such as drug design, disease treatment, biological processes research and chemical forensics. Correspondence yields a mapping between isotopic envelopes in different runs (see Figure 4), a prerequisite for differential analysis.

The combination of noise from within one run (enumerated above) and noise from run to run–most notable in retention time shifts, where an isotopic envelope appears at a different retention time or with a compressed or stretched RT length compared to another run–make LC-MS correspondence non-trivial.

The correspondence mapping should again optimize an objective function which, in turn, characterizes an algorithm choice for solving the correspondence problem.

Extant objective functions

According to a recent review on LC-MS correspondence algorithms [11], all extant approaches use either centroid data or a reduction of isotopic envelope traces into a single centroid. Of the almost sixty algorithms reviewed there, nearly all use the same objective function–finding a family of one-to-one partial functions χ_r : $\varepsilon_r \rightarrow \varepsilon_*$ (a different function for each experimental run *r*), where ε_* is the set of envelopes from a reference run, that minimizes global RT and m/z distance between isotopic envelopes (in any of their reduced forms, according to the authors):

$$\lambda_{corr} = \sum_{E \in \varepsilon_r} \delta(E, \chi_r(E))^{\tau,\mu} \tag{18}$$

where $\delta()^{\tau,\mu}$ is a distance function defined over RT and m/z.

The continuous profile model (CPM) [12] uses a different objective function, and thus is free from the reference requirement that most other algorithms have, allowing for a symmetric solution (one that is not dependent on the choice of a reference run). Additionally, the mapping is somewhat more localized than that of most correspondence algorithms. CPM minimizes the log likelihood of differences between a hidden Markov model $m\tau$ of the RT of a latent run and observed runs:

$$\lambda_{corr} = \log p(D|m^{\tau}) \tag{19}$$

where D is the set of observed runs.

Proposed objective functions

In contrast to existing LC-MS correspondence objective functions, the objective functions suggested here use the entire isotopic envelope. This allows greater discrimination by using isotope trace quantity and spacing to match isotopic envelopes from different runs. This extra discrimination is essential given the amount of RT variance and (to a lesser degree) m/z variance present in the data.

Let R be a set of runs, each of which has an associated set of isotopic envelopes $\varepsilon_r = \{E_i^r\}_{i=1}^{pr}, 1 \leq r \leq |R|$ and let $\tilde{\varepsilon} = \cup_r \varepsilon_r$. We seek to find a binary equivalence relation ρ that induces a set of *correspondence classes* over $\tilde{\varepsilon}$ that is reflexive (an envelope corresponds with itself), symmetric (if envelope E_1 from run 1 corresponds with envelop E_2 from run 2, then E_2 also corresponds with E_1) and transitive (if envelope E_1 from run 1 corresponds with envelope E_2 from run 2 and envelope E_2 corresponds with envelope E_3 from run 3, then E_1 corresponds with E_3); and if $\rho(E_i^r, E_j^s) = \text{TRUE}$, then for $k \neq i$, $\rho(E_k^r, E_j^s) = \text{FALSE}$ and for $k \neq j$, $\rho(E_i^r, E_k^s) = \text{FALSE}$ (an envelope from one run may have 0 or 1 matches from any other run; note that due to reflexivity, this also means that two non-identical envelopes from the same run never correspond).

This relation should minimize

- The difference in charge state between corresponding isotopic envelopes, δ_{charge}.
- The difference in m/z between isotope traces in corresponding isotopic envelopes, $\delta_{mz_{it}}$.
- The difference in elution duration between isotope traces in corresponding isotopic envelopes, δ_{dur}.
- The difference in isotope abundance ratios between corresponding isotopic envelopes, δ_{ratio}.
- The difference in m/z between corresponding isotopic envelopes, $\delta_{mz_{ie}}$.
- The number of singleton correspondence classes, δ_{orphan}.
- The difference in retention time between corresponding isotopic envelopes, δ_{rt}.

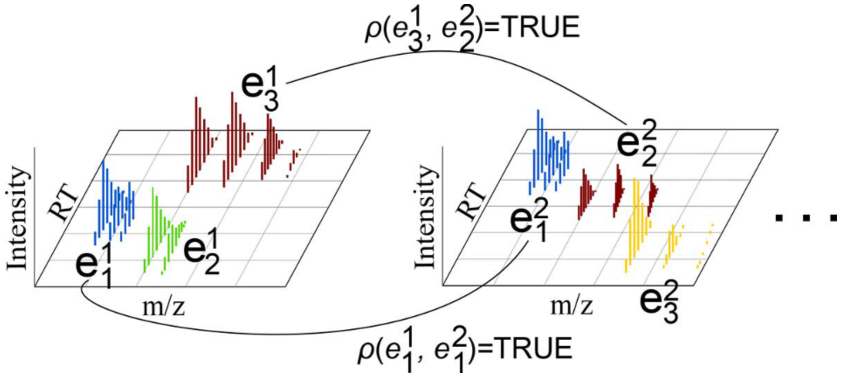

Figure 4 Objective functions for correspondence must allow a mapping from an isotopic envelope in one run to an envelope in another, or to none, if there is no corresponding isotopic envelope. Here, the unillustrated relations would yield FALSE.

An objective function incorporating all of these variables can take many forms, with perhaps the simplest generalization being a weighted linear combination, with weighting coefficients ω allowing relative prioritization:

$$\begin{aligned}\lambda_{corr} = \sum_{\rho(E_1,E_2)} & \omega_{charge}\delta_{charge}(E_1, E_2) + \omega_{mz_{it}}\delta_{mz_{it}}(E_1, E_2) \\ & + \omega_{dur}\delta_{dur}(E_1, E_2) + \omega_{ratio}\delta_{ratio}(E_1, E_2) \\ & + \omega_{mz_{ie}}\delta_{mz_{ie}}(E_1, E_2) + \omega_{orphan}\delta_{orphan}(E_1, E_2) \\ & + \omega_{rt}\delta_{rt}(E_1, E_2)\end{aligned} \tag{20}$$

with the summation over $\rho(E_1, E_2)$ meaning a summation taken over all pairs of envelopes $E_1, E_2 \in \tilde{\varepsilon}$ for which $\rho(E_1, E_2)$ = TRUE. Given the weighting coefficients ω, the most desirable correspondence would be that induced by the relation ρ^* that minimizes λ_{corr} (see Figure 4),

$$\rho* = \arg\min_{\rho} \lambda_{corr}$$

Alleviating current limitations in correspondence

Recently, several ubiquitous shortcomings were identified in a review of over 50 LCMS correspondence algorithms [11]. The most significant of these shortcomings was the fact that all current LC-MS correspondence algorithms make model assumptions that fail to capture common behavior. In other words, each algorithm is constructed in such a way that the algorithm is guaranteed to get the wrong answer under certain conditions that are common to real LC-MS data. The behaviors discussed included the ideas that:

- Not all analytes appear in all replicates.
- Elution order can swap.
- Shifts occur in m/z as well as in RT.

Some correspondence methods reduce isotopic envelopes to a single point representation. This deprives the method of a rich source of distinguishing data found in full isotopic envelopes–the expectation of contiguous isotope traces with a uniform m/z charge gap, number of isotope traces, and relative abundance ratio of isotope traces. Similarly, most correspondence algorithms conduct an initial RT alignment, where signals (almost always much-reduced from the full isotopic envelope, and rarely built up from isotope traces to isotopic envelopes) are shifted up or down in RT (preserving original order) in order to most closely match a reference run. This is invariably followed by direct matching. The problem is that the initial warping is a lossy procedure that adulterates the original RT time, which would be useful to probabilistically ascertaining the closest corresponding isotopic envelope.

The proposed objective function does not force matches between runs, as it is very common for species to either not be present or fall below the signal-to-noise ratio in differential studies. Instead, the proposed objective function leverages the full breadth of isotope envelope information, allowing a rigorous direct comparison of candidate correspondences based on all available data to select the most likely correspondence (in the sense of minimizing error), or no correspondence at all if that is the most likely case given the data.

Conclusions

We present a concise attempt to formalize LC-MS data clustering problems, describing the constructs of isotope traces and isotopic envelopes and their relational structure. We provide a review of current approaches to isotope trace extraction and LC-MS correspondence, and propose novel objective functions for both tasks that address shortcomings in current methods.

Competing interests and declarations

The authors declare that they have no competing interests. The publication costs for this article were funded

by the University of Montana Office of Research and Sponsored Programs.

This article has been published as part of *BMC Bioinformatics* Volume 16 Supplement 7, 2015: Selected articles from The 11th Annual Biotechnology and Bioinformatics Symposium (BIOT-2014): Bioinformatics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcbioinformatics/supplements/16/S7.

Authors' contributions

RS, JTP and DV all contributed in writing this manuscript.

Authors' details

[1]Department of Computer Science, University of Montana, 59812 Missoula, USA. [2]Department of Chemistry, Brigham Young University, 84606 Provo, USA. [3]Department of Computer Science, Brigham Young University, 84606 Provo, USA.

Published: 23 April 2015

References

1. Choi H, Fermin D, Nesvizhskii AI: **Significance analysis of spectral count data in label-free shotgun proteomics.** *Mol Cell Proteomics* 2008, **7(12)**:2373-2385.
2. Wiese S, Reidegeld KA, Meyer HE, Warscheid B: **Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research.** *Proteomics* 2007, **7(3)**:340-350.
3. Michalski A, Cox J, Mann M: **More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Marjority is Inaccessible to Data-Dependent LC-MS/MS.** *Journal of Proteome Research* 2011, **10**:1785-1793.
4. Cappadona S, Baker PR, Cutillas PR, Heck AJ, van Breukelen B: **Current challenges in software solutions for mass spectrometry-based quantitative proteomics.** *Amino Acids* 2012, **43(3)**:1087-1108.
5. Tautenhahn R, Bottcher C, Neumann S: **Highly sensitive feature detection for high resolution LC/MS.** *BMC Bioinformatics* 2008, **9(1)**:504.
6. Pluskal T, Castillo S, Villar-Briones A, Oresic M: **MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data.** *BMC Bioinformatics* 2010, **11(1)**:395.
7. Conley CJ, Smith R, Torgrip RJ, Taylor RM, Tautenhahn R, Prince JT: **Massifquant: open-source Kalman filter based XC-MS isotope trace feature detection.** *Bioinformatics* 2014, 359.
8. Cox J, Mann M: **MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification.** *Nature Biotechnology* 2008, **26(12)**:1367-1372.
9. Weisser H, Nahnsen S, Grossmann J, Nilse L, Quandt A, Brauer H, Sturm M, Kenar E, Kohlbacher O, Aebersold R, *et al*: **An automated pipeline for high-throughput label-free quantitative proteomics.** *Journal of Proteome Research* 2013.
10. Bellew M, Coram M, Fitzgibbon M, Igra M, Randolph T, Wang P, May D, Eng J, Fang R, Lin C, *et al*: **A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS.** *Bioinformatics* 2006, **22(15)**:1902-1909.
11. Smith R, Ventura D, Prince JT: **LC-MS Alignment in Theory and Practice: A Comprehensive Algorithmic Review.** *Briefings in Bioinformatics* 2013.
12. Listgarten J, Neal RM, Roweis ST, Wong P, Emili A: **Difference detection in LC-MS data for protein biomarker discovery.** *Bioinformatics* 2007, **23(2)**:198-204.

doi:10.1186/1471-2105-16-S7-S1
Cite this article as: Smith *et al.*: **A coherent mathematical characterization of isotope trace extraction, isotopic envelope extraction, and LC-MS correspondence.** *BMC Bioinformatics* 2015 **16**(Suppl 7):S1.