

- 1 -

ANALYSING LIQUID CHROMATOGRAPHY ELECTROSPRAY IONISATION MASS SPECTROMETRY (LC-ESI-MS) DATA

Technical Field

The present invention relates, in general terms, to a system, and method implemented by that system, for analysing liquid chromatography electrospray ionisation mass spectrometry (LC-ESI-MS) data.

Background

High-resolution tandem mass spectrometry is a powerful data-rich analytical technique that provides high-dimensional data often complicated by background noise. Analysis of these complex datasets presents a challenging, time-consuming task that often bottlenecks project progress (e.g. in untargeted metabolomic discovery studies). Current commercial LC-ESI-MS analytical software solutions act directly on the raw data which (1) takes up a large amount of storage space at ~200MB per file and (2) is extremely time-consuming (e.g. can take 6-8 hours to perform a single query) due to the general nature of the program. Peak integration is also complicated by background noise and poor resolution (e.g. peak shoulders) resulting in erroneous fold change comparisons between samples. This bottleneck in data analysis prevents efficient extraction of scientific insights that could serve as building blocks for other applications.

It would be desirable to overcome at least one of the above-described problems by providing an automated data analysis workflow for rapid processing of LC-ESI-MS data, or at least to provide a useful alternative.

Summary

In view of the above problems, the solution presented herein provides a high

- 2 -

throughput data analysis workflow for automated processing and analysis of LC-ESI-MS data. The workflow utilizes a series of processes for (1) adjustable rule-based peak picking, (2) extraction of key features which, for some embodiments, reduces the file size by 10,000x, (3) automated annotation of chemical fingerprints and (4) creation of a dynamic summary for easy data exploration. In some embodiments, a 50x acceleration in data analysis turnover (from hours to minutes) is achieved, enabling faster query execution and easy data exploration.

The present invention provides a computer system for analysing liquid chromatography electrospray ionisation mass spectrometry (LC-ESI-MS) data, comprising:

- memory; and

- at least one processor, the memory storing instructions that, when executed by the at least one processor, cause the computer system to:

- implement a raw data processor that:

- obtains raw LC-ESI-MS data from a source and extracts a base peak chromatogram (BPC) from the raw data;

- detects, from the BPC, a peak for each signal intensity above a first predetermined threshold and, for each detected peak, a retention time, a peak apex and a baseline; and

- extracts a mass spectrum based on data at each retention time and a base peak signal intensity derived from the peak apex and the baseline; and

- implement a fingerprinting system that:

- performs background correction by, for each detected peak, producing a resulting m/z value by subtracting mass over charge (m/z) values of the baseline from an original m/z value and, if the resulting m/z value is positive, restoring the original m/z value;

- 3 -

generates a compact data representation comprising, for each detected peak, the retention time, original m/z values and intensities; and
performs chemical fingerprinting for the raw LC-ESI-MS data based on the compact data representation.

Also disclosed is a method for analysing liquid chromatography electrospray ionisation mass spectrometry (LC-ESI-MS) data, comprising:

using a raw data processor to:

obtain raw LC-ESI-MS data from a source and extract a base peak chromatogram (BPC) from the raw data;
detect, from the BPC, a peak for each signal intensity above a first predetermined threshold and, for each detected peak, a retention time, a peak apex and a baseline; and
extract a mass spectrum based on data at each retention time and a base peak signal intensity derived from the peak apex and the baseline; and

using a fingerprinting system to:

perform background correction by, for each detected peak, producing a resulting m/z value by subtracting mass over charge (m/z) values of the baseline from an original m/z value and, if the resulting m/z value is positive, restoring the original m/z value;
generating a compact data representation comprising, for each detected peak, the retention time, original m/z values and intensities; and
perform chemical fingerprinting for the raw LC-ESI-MS data based on the compact data representation.

Also disclosed is a method of building a database, comprising:

for each of a plurality of base peak chromatograms (BPCs), at a data processor:

- 4 -

obtaining the BPC either directly or from LC-ESI-MS data;
detecting a peak for each signal intensity above a first predetermined threshold and, for each detected peak, a retention time, a peak apex and a baseline; and
extracting a mass spectrum based on data at each retention time and a base peak signal intensity derived from the peak apex and the baseline; and
at a fingerprinting system:
performing background correction by, for each detected peak, producing a resulting m/z value by subtracting mass over charge (m/z) values of the baseline from an original m/z value and, if the resulting m/z value is positive, restoring the original m/z value; and
generating a compact data representation comprising features, each feature being a said detected peak or, for each detected peak, the retention time, original m/z values and intensities;
storing the compact data representation in the database;
and
providing a query interface for:
receiving a query comprising, for a new BPC or new LC-ESI-MS data, a new compact data representation for the new BPC or new LC-ESI-MS data;
performing chemical fingerprinting by comparing of one or more features of the new compact data representation to the features of the compact data representations in the database; and
returning one or more of the compact data representations from the database, based on the chemical fingerprinting.

Also disclosed is a database built in accordance with the above database building method.

- 5 -

Brief description of the drawings

Embodiments of the present invention will now be described, by way of non-limiting example, with reference to the drawings in which:

Figure 1 illustrates a method for processing LC-ESI-MS data for fingerprinting chemicals and compounds.

Figure 2 – Representative total ion current (top) and base peak chromatogram (bottom) from LC-ESI-MS of a sample.

Figure 3 – Representative detected peak spectra (top), prior/post baseline spectra (mid), and corrected detected peak spectra (bottom).

Figure 4 – Sample mass spectra of valinomycin (top) and table of molecular ion peak intensity to intensity of the peak one m/z unit heavier ratios for three samples of valinomycin.

Figure 5 – (image A) Pivot table showing dynamic data summary. (image B) Adjustable fields for the pivot table. (image C) Current field settings for pivot table in (image A).

Figure 6 – computer system for implementing the method of Figure 1.

Detailed description

The present system and method provide a pipeline for automated high throughput processing and analysis of LC-ESI-MS data to provide actionable scientific insights. The pipeline, and thus the protocol implemented by that pipeline, presents a time- and resource-efficient method for chemical fingerprinting focusing on key feature extraction to parameterize complex data-rich LC-ESI-MS chromatograms. By parametrizing in this manner, data file size

- 6 -

can be minimised while processed information is retained for further utilization in new queries or dynamic data exploration.

This technology has been demonstrated on an LC-ESI-MS dataset comprising of fermentation extracts from 2,554 strains in 3 to 5 media for a total of 8,734 LC-ESI-MS sample data files. It has also been employed to extract insights for (1) prioritization for antimicrobial screening to minimize experimental effort and accelerate discovery through targeted screening, (2) correlating genetic and chemical information to infer production pathways for natural products that can be exploited for sustainable manufacturing, and (3) molecular networking for discovery of analogues with potential for interesting functional activity. The methods provided herein enable new antimicrobial compounds to be identified, by characterizing and annotating compounds present in a sample for correlation with antimicrobial properties as quantified via biological assays.

The pipeline is reflected in Figure 1, which shows a method 100 for analysing LC-ESI-MS data. The method 100 is broadly separated into two groups of steps. A first group of steps is performed by a raw data processor, to take raw LC-ESI-MS data and extract a mass spectrum. The second group of steps is performed by a fingerprinting system to take data derived by the raw data processor, and process it to perform chemical fingerprinting.

Following the method 100 through, the raw data processor obtains raw LC-ESI-MS data from a source (102a) and extracts a base peak chromatogram (BPC) from the raw data (102b). The source may be a mass spectrometer such that raw data is obtained directly. In other embodiments, the LC-ESI-MS data is stored elsewhere (e.g., in a database) and is extracted through known means.

The BPC can be extracted by any known means. In some embodiments, the BPC is extracted from a directory of raw LC-ESI-MS data into text files. This can be done in an automated way, using third party software such as Mestrenova,

- 7 -

which can read LC-ESI-MS raw data from the mass spectrometers of multiple vendors as well as from files of open formats (e.g. mzData or mzXML).

The method 100 uses BPC since it reduces background noise by representing only the most intense m/z peak at every point, unlike total ion current (TIC) which represents the combined intensity of all m/z peaks. This allows for easier peak picking and recognition in the following steps. The difference between the two data sets is reflected in Figure 2, in which the top image shows TIC for a sample and the bottom image shows BPC for the same sample. The ellipses 200, 202 in the TIC image contain a significant number of peaks that are not reflected in the BPC image. These additional peaks can obscure the importance of peaks that have greater influence on the identification of compounds.

The Mestrenova script implemented for present purposes automates extraction of the BPC graph as x-axis (retention time) and y-axis (signal intensity) coordinates in a text file. In other embodiments, a neural network may be trained on raw paired LC-ESI-MS data and BPCs, to map from the former to the latter.

The raw data processor takes the BPC and detects a peak for each signal intensity above a predetermined threshold (104a). Even when fingerprinting a complex mixture of compounds, the individual compounds in a sample can be separated and detected by LC-ESI-MS as peaks in the BPC. Relatedly, the liquid chromatography portion of the analytical method differentiates compounds based on retention time (x-axis of the BPC). Hence, each peak in the BPC represents the elution of a compound or mixture of compounds from the sample being analysed.

Simple formulae can then be used to detect the peaks. As retention time increases, signal intensity varies. In the signal intensity gradient becomes zero, the signal intensity has either plateaued, in the case that neighbouring signal intensity peaks are the same or very similar, or peaked, where the gradient of

- 8 -

the signal intensity becomes negative. The data can thus be analysed using simple scripts such as a combination of Excel formulae and Visual Basic for Applications (VBA) scripts, to automatically recognize peaks through gradient changes.

To remove peaks that provide little discriminating information, a background threshold is used, below which the readings on the BPC are assumed to constitute background noise. That threshold may be a fixed value for all LC-ESI-MS data – e.g., a BPC signal intensity of 30,000 for peaks and 15,000 for background noise – or may be set based on the heights of the peaks anticipated to be acquired, or based on the peak heights of a family of compounds sought to be identified in the BPC. The parameters for peak picking such as set threshold sensitivity (i.e., signal intensity of the cut-off threshold – first predetermined threshold) and background threshold (i.e., signal intensity threshold below which readings are assumed to be background noise – second predetermined threshold) can be adjusted to tune sensitivity.

The process is thus:

for step 104a:

- (i) Y-axis (intensity) changes on the BPC within a set threshold (default: 15,000) are taken as background noise and averaged out.
- (ii) If y-axis (intensity) continually increases more than the set sensitivity (default: 30,000) then a peak is annotated.
- (iii) For each peak, a peak apex is identified based on the gradient of the signal intensity over retention time.

for step 104b:

- (iv) In each case of step (iii), the peak apex (i.e. when intensity stops increasing) occurs either at a plateau (i.e., gradient is zero or close to zero) or a sharp point (i.e., gradient sign changes). If intensity plateaus, the x- and y-values of the mid-point of the plateau are assumed to correspond to the peak and its retention time. If intensity begins to decrease (i.e. a sharp point) then the highest y-value and

- 9 -

its corresponding residence time (x-value) are taken to be the peak and its retention time. Moreover, the highest peak over the LC-ESI-MS data is taken to be the base peak of the BPC.

- (v) The nearest baseline time points prior to, and after the detected peak are also sampled for background correction. A baseline time point corresponds to the retention time points at which the signal intensity crosses the background threshold. The m/z values of the baselines, being the m/z values at the time points of the nearest prior- and post-peak crossings of the background threshold, can be recorded as set out below.

Using base peak signal intensity avoids the need to determine the "area under curve", being the basis for peak detection and analyses used by existing technologies. Area under curve algorithms for measuring compound abundance require the shape of the peak to be recognised, as well as navigating interaction between neighbouring peaks that are closely positioned in the BPC (i.e. start and end of peak, unresolved peaks). Contrastingly, base peak signal intensity from peak apex mass spectra is representative of compound abundance and could be used for further analysis such as fold change comparisons (Table 1).

Strain	A33995	A100046	A100046	A100046	A100047	A100047	A100050	A100050	A100050
Media	CA09LB	CA02LB	CA09LB	CA10LB	CA02LB	CA09LB	CA02LB	CA09LB	CA10LB
BP Signal Intensity	154,307	1,586,009	1,045,232	1,308,268	1,643,396	619,278	1,477,011	833,034	1,059,790
Area Under Curve	6,865	69,240	43,721	56,050	72,716	23,945	65,489	33,232	45,385
BP fold change	1.00	10.28	6.77	8.48	10.65	4.01	9.57	5.40	6.87
Area fold change	1.00	10.09	6.37	8.16	10.59	3.49	9.54	4.84	6.61
error%	0%	2%	6%	4%	1%	13%	0%	10%	4%
error	0.00	0.19	0.41	0.31	0.06	0.53	0.03	0.56	0.26

Table 1 – Comparison of base peak (BP) signal intensity and area under curve as indicators of compound abundance in LC-ESI-MS analytical data

In Table 1, the wildtype (parent) strain is indicated in **bold**, with all other strains being mutant strains. Ratios were calculated by dividing mutant strain BP signal intensity or area under curve by parent strain values. Error = BP fold change – Area fold change. Error% = Error / BP fold change. This shows that the disagreement between the two calculation methods is low, which is not to say that the area under curve is correct when there is greater disagreement.

This shows that the present methods also facilitate metabolite mapping (upregulation), to map the quantity and diversity of compounds detected in different mixtures obtained from wild type and engineered (mutant) organisms. Such a mapping facilitates analysis of increased production of certain high value compounds by engineered organisms – i.e., “upregulation”.

This shows that detecting peaks as a single point at the apex of the peak for its retention time and mass spectra, can be used to minimize processing time and data storage requirements. The sampled time points of detected peaks and associated baselines are then taken forward to the next step for further processing.

Step 106 involves extracting a mass spectrum based on data in the BPC at each retention time, and the base peak signal intensity derived from the peak apex and the baseline. This can be achieved by extracting the mass spectrum at each time point from the raw data. Since the retention time of each peak is extracted in step 104b, the mass spectrum can be extracted for each time point corresponding to a retention time (i.e., if the retention time is 5 seconds, then the corresponding time point will be 5 seconds). A Mestrenova script, or other automated script, can be used to reference the time points annotated in each sample (e.g., the retention time of each peak, or of each peak and associated baseline time points) to automatically extract the mass spectra from the

- 11 -

relevant files – extracting may be done into text files. Mass spectra are recorded as a series of mass over charge (m/z) values and their associated intensities detected. In some embodiments, this data is extracted for all time points for which the intensity exceeds a predetermined threshold – e.g., 5,000 – to minimize recording of background noise. Thus, m/z values that meet the minimum intensity threshold (i.e., 5,000 with respect to each individual peak within a single mass spectrum) are retained to minimize recording of background noise.

The absolute values of the thresholds can vary depending on experimental parameters, and equipment configuration. Moreover, the values may vary between different types of equipment – for example, variations in mass spectrometer materials, carriers, delivery mechanism and the like can affect the residence time and m/z values. The thresholds specified herein are therefore applicable to a particular system, though can be expected to vary for reasons set out above.

Regarding the specific thresholds specified herein, random or background noise is assumed to be below a background noise m/z abundance threshold. Presently, that threshold is set to 5,000, below which recordings are disregarded. The threshold is determined whereby random noise signals would have equivalent intensities and hence any values at or below this threshold are not informative.

The Base Peak Chromatogram (BPC) noise intensity threshold (presently 15,000) is used for averaging, being a “smoothing” process that avoids the issue of random noise signals creating artificial peaks. The Base Peak Chromatogram (BPC) peak intensity threshold (presently 30,000) helps to identify time points at which compounds are present and at which mass spectra should be extracted. All mass spectral data above the m/z noise abundance threshold (presently 5,000) are recorded and processed as part of the pipeline for chemical fingerprinting and generation of compact data representations.

- 12 -

The mass spectral data at identified time points above the background noise m/z abundance threshold (presently 5,000) is recorded, as well as the accompanying baselines to the sides for further noise correction. This is because aside from random noise, there can also be common contaminants present that exceed the random noise threshold (i.e. persistent noise). The baselines are recorded so that only sample data is captured in the mass spectra. This avoids the problem where the base peak recorded originates from a noise peak instead of an actual sample peak.

To facilitate efficient chemical fingerprinting, background noise is corrected (step 108) to enable production of a compact data representation of the mass spectrum (step 110). Background correction step 108 is performed by identifying baseline values and identifying peaks that exceed the baseline values. In some embodiments, background correction is performed for each detected peak, by subtracting mass over charge (m/z) values of the baseline from an original m/z value of each peak. If the resulting m/z value is positive, then the original m/z value is restored.

The An Excel VBA script (see Annex for details) automatically pulls the mass spectra values for each detected peak and does background correction by subtracting the intensities of similar m/z values of both prior and post baseline mass spectra from the detected peak mass spectra. While a single baseline can be subtracted, or an amount less than both baselines, subtracting both baselines ensures that all of the baseline is removed from the signal, even where there is fluctuation between baselines. When subtracting baselines from values around respective peaks, if a peak is missing a baseline (e.g., it is at the start or end of the recording or the sample BPC is too crowded with peaks) then the remaining baseline is used twice or doubled. If a peak is missing both baselines, then the nearest baseline from a previous peak is used twice. All m/z values in the detected peak mass spectra that are still positive are then retained and restored to their original intensities from the detected peak mass spectra.

- 13 -

These corrected values as well as the raw spectra are then stored as compact data files – e.g., in excel files (.xlsx) – comprising extracted key features of (a) retention time, (b) original m/z values, and (c) associated intensities of each detected peak. Where an "original" value or similar is stored, this nomenclature includes proxies for the original value, such as a baseline value and an amount by which a peak exceeds the baseline – the original m/z value can be reconstructed by adding the peak and baseline.

Using steps 102 to 110, processing times are dramatically reduced when compared with commercial software that calculates area under curve – e.g., a 320-sample run processes in a time of 7min 41s, being approximately 50x faster than the six hour period required to process the same run on commercial software Agilent MassHunter Profinder.

Chemical fingerprinting (step 112) can then be performed for new, raw LC-ESI-MS data, by comparing a compact data representation for the new LC-ESI-MS data to the compact data representations already stored in memory. If the raw data peaks match the spectrum of a stored, compact data representation, then the compounds are similar or the same. The data can then be annotated – annotations are done automatically via algorithms/scripts. The detected compounds are automatically grouped by determining if their retention time and base peak m/z values are within a specific threshold (adjustable – but default values are 0.2 min and 0.02 m/z) of known or predicted values – e.g., in a database. If so, then they are annotated with the compound corresponding to the known or predicted values.

Moreover, the above methodology can be used to build a database of compact data representations. After each new compact data representation is generated, if there is no corresponding compact data representation already stored in the database then the new compact data representation is stored in the database and can be labelled (e.g., by a user or by another machine learning model trained to label the compact data representations based on a BPC, LC-ESI-MS

- 14 -

data, compact data representation or other data) – i.e., when there is no compact data representation stored in the database that is the same as the new compact data representation for the new LC-ESI-MS data, or sufficiently similar (e.g., to within a threshold, or a retention time offset due to use of a different mass spectrometer). This involves implementing a query engine to compare mass spectra of samples by comparing their compact data representations. Comparing compact data representations comprises comparing features of those representations – e.g., peaks, m/z of respective base peaks, and retention times of the base peaks. The query engine may also filter the first mass spectrum, or its compact data representation, based on a specific peak intensity or specific m/z value, or on a peak intensity range or m/z value range. Thus, chemical fingerprinting can be undertaken by comparing new compact data representations with compact data representations stored in the database. Notably, a query may include a BPC or LC-ESI-MS data that is then converted to a BPC.

In summary, an embodiment of the methodology involves:

A) Raw data processing

- A1. Mestrenova script for automated raw data extraction
- A2. Peak apex sampling algorithm
- A3. Mestrenova script for automated extraction of mass spectra

B) Extracted data processing

- B1. Background correction and storage in excel files
- B2. Initial processing and chemical fingerprinting

C) Applications

- C1. Query algorithms
- C2. Dynamic data summary

Using steps 102 to 112, the present methodology provides:

- An ability to read raw LC-ESI-MS data from multiple instrument vendors.

- 15 -

- Tuneable peak picking algorithm. The intensity threshold can be adjusted to include or exclude more peaks.
- Peak apex mass spectrum sampling to minimize data size (10,000x smaller than raw file). Using the present threshold approach, where background and peak thresholds are selected and applied, data that has little or no effect on fingerprinting is excluded.
- Automated accelerated analysis (50x faster than commercial vendor software²). Comparison operations on much smaller data representations takes considerably less time than traditional comparisons.
- Dynamic summary with adjustable level of detail for easy exploration of data. For example, the output of the system can be a pivot table such as that shown in Figure 5. This enables dynamic summaries to be generated, with adjustable fields as desired. The lists are expandable and collapsible for different levels of detail and easy exploration.
- Processed data can be saved to facilitate future queries unlike vendor software. The new queries can operate on the stored data, without requiring re-processing of the raw data in response to a particular query.
- Scalable to 1,048,576 samples versus vendor software that is limited to 320-samples per run

Overall, the pipeline shown in Figure 1 accelerates the analysis of large LC-ESI-MS datasets by only extracting relevant key features in an automated, scalable fashion. This reduces the time required to extract actionable scientific insights. The workflow has been demonstrated on a dataset comprising of fermentation extracts from 2,554 strains in 3 to 5 media resulting in for a total of 8,734 LC-ESI-MS sample data files.

For fingerprinting, there is a number of parameters that can be used to define a peak. One or more, and preferably all, of these parameters is retained in the compact data representation. In particular, each detected peak may be characterized by the following parameters:

- i. Filename

- 16 -

- ii. Retention time (in minutes)
- iii. Number of m/z peaks above intensity of 50,000 (or other relevant threshold). Again, this threshold is determined empirically and can differ between recordings and equipment. The higher m/z peak threshold (e.g., 50,000 here) is used for significant peak detection for peaks that are characteristic of the mass spectra. The number of significant peaks can increase with molecular volatility – i.e., multiple large peaks suggests the molecule has fragmented due to lack of stability.
- iv. Base peak m/z (most intense m/z peak)
- v. Base peak intensity
- vi. Molecular ion peak m/z (either the largest most intense m/z or detected by isotopic abundance)
- vii. Calculated number of carbon atoms (by isotopic abundance)
- viii. Strain ID
- ix. Fermentation media

Carbon-13 has a natural abundance of 1.16% which can be leveraged to estimate the number of carbons (parameter vii) in a sample by looking at the ratio of the molecular ion peak intensity to the intensity of the peak one m/z unit heavier. Due to the inherent nature of carbon, the number of carbons present in the molecule can be estimated by comparing the intensities of certain m/z values in the mass spectra. Specifically, the molecular ion peak, M, and the peak that is 1 Da heavier, M+1, are used through the following formula:

$$85.2068966 * (M+1_intensity / M_intensity)$$

An example of this with three different samples containing valinomycin is shown in Figure 4, where the molecular ion peak intensity at 1128 m/z is compared to the intensity of the peak at 1129 m/z. Thus, the method 100 can involve identifying the base peak based on the highest signal intensity, and then determining if one or more molecular ion peaks exist in the mass spectrum, with

- 17 -

detectable isotopes in the mass spectrum. If molecular ion peaks are identified, the molecular ion peak with highest signal intensity is selected. The compact data representation will then include the molecular ion peak with highest signal intensity.

Despite there being a variance of +/- 1 carbon, such carbon number calculations can give good estimates of the carbon content in detected compounds. However, due to possible formation of adducts, dimerization or even fragmentation, it may not always be possible to do such calculations. In addition, some of the above parameters are unique only in the context of strain fermentation extracts (e.g. strain ID and fermentation media) while other parameters provide no information about compound identity (e.g. filename and base peak intensity). Hence, in some embodiments a generalizable 4-parameter chemical fingerprint was employed to characterize detected peaks. These are:

1. **Retention time** – liquid chromatography separation method can differentiate compounds based on their elution time
2. **Number of m/z peaks above intensity of 50,000** – compound stability under ionisation conditions can result in differing degrees of fragmentation resulting in different number of significantly detected fragments
3. **Base peak m/z** – the most stable ion (or ion fragment) from ionisation of the compound will be detected as the base peak (most intense m/z)
4. **Molecular ion peak m/z** – the largest significant ion peak with detectable isotopes within the compound mass spectra, where significance is specified via a predetermined threshold. The molecular ion peak m/z is used for (1) characteristic identification of potential molecular structures, (2) part of the parameters to identify M intensity to calculate number of carbons.

Compounds are considered similar if they fulfil the 2 criteria of (1) base peak m/z within 0.02 m/z, and (2) retention time within 0.2 min. In this way,

- 18 -

unknown compounds can be fingerprinted, grouped, and annotated with arbitrary unique compound numbers for identification.

From the list of characterized detected peaks, a series of algorithms can be readily developed to automate analysis and querying. These algorithms can be written, for example, Excel VBA, and can include:

- i. m/z search – searching for a specific m/z value at a particular retention time within adjustable thresholds. This enables detection of a specific compound by m/z value.
- ii. base peak intensity search – filtering for all detected peaks with base peak intensities above a certain threshold. This will be generally relevant in cases where the same equipment is used, since the base peak intensity and retention times can vary between different types of equipment.
- iii. reference comparison – comparison between samples on similar detected peaks, their fold changes based on base peak intensity and listing new detected peaks

Query algorithms are written to quickly pull actionable scientific insights from the curated, compact data representation. These algorithms compare mass spectra of two samples (a first sample corresponding to first raw LC-ESI-MS data, and a second mass spectrum of a second sample corresponding to second raw LC-ESI-MS data), by comparing m/z values of respective base peaks, and retention times of the base peaks. If those two quantities are similar, the compounds for which the mass spectra were taken are assumed to be similar. This comparison can involve filtering the mass spectra based on a specific peak intensity or specific m/z value, a peak intensity range or m/z value range. For example, finding the best producing mutant strain in each fermentation media of a compound can be achieved by finding the highest base peak intensity for that compound. This can enable the mutant strain to be scaled up for isolation.

- 19 -

Data can be automatically arranged using dynamic data summary scripts. For example, pivot tables can be used to represent a data set as a dynamic summary – see Figure 5.

Curated data can thus be explored at varying levels of detail with expandable rows (e.g. different fermentation media for each strain – image A of Figure 5) and can be filtered (e.g. by minimum base peak intensity – image A of Figure 5). The set-up of the pivot table can also be manipulated with various possible fields (image B of Figure 5) and display options (image C of Figure 5). The current example in image A is set-up to show the highest intensity obtained for each compound. This facilitates easy understanding of how different strains and fermentation media affect production of each compound. Dynamic data summaries allow for quick exploration of a large dataset to gather desired scientific insights. Its dynamic manipulation enables diverse queries with the option for deeper insights while still maintaining a concise overview of the data.

Feedback from data analysis this way can be used to determine how effective particular engineering efforts are in increasing production of antimicrobial agents, or otherwise achieving predetermined goals.

This automated and high throughput LC-ESI-MS processing protocol set out with reference to Figure 1 presents a time- and resource-efficient method focusing on key feature extraction (i.e. peak apex sampling) to parameterize complex data-rich LC-ESI-MS chromatograms. Peak recognition issues are largely avoided by disregarding shape and instead using nearest baselines and peak intensity to define each peak regardless of shape. The compact data representation dramatically reduces file size. These two advantages in turn reduce processing and querying time.

Figure 6 shows a system for implementing the workflow of Figure 1. The system 100 includes memory 102. The memory stores instructions 104 in the form of computer program code. The memory may be a non-transitory computer

- 20 -

readable storage medium that stores program code that is executable to perform steps 102 to 110, and others. The instructions are interpreted by one or more processor 106, that then controls receipt, by a mass spectrum data input interface 108, of mass spectrum data (LC-ESI-MS data) either from internal memory, external memory 110 or directly from a mass spectrometer 112. The processor then implement the raw data processor 114 and fingerprinting system 116, to either generate compact data representations, or to query stored data in view of new LC-ESI-MS data.

It will be appreciated that many further modifications and permutations of various aspects of the described embodiments are possible. Accordingly, the described aspects are intended to embrace all such alterations, modifications, and variations that fall within the spirit and scope of the appended claims.

Throughout this specification and the claims which follow, unless the context requires otherwise, the word "comprise", and variations such as "comprises" and "comprising", will be understood to imply the inclusion of a stated integer or step or group of integers or steps but not the exclusion of any other integer or step or group of integers or steps.

The reference in this specification to any prior publication (or information derived from it), or to any matter which is known, is not, and should not be taken as an acknowledgment or admission or any form of suggestion that that prior publication (or information derived from it) or known matter forms part of the common general knowledge in the field of endeavour to which this specification relates.

Claims

1. A computer system for analysing liquid chromatography electrospray ionisation mass spectrometry (LC-ESI-MS) data, comprising:
memory; and
at least one processor, the memory storing instructions that, when executed by the at least one processor, cause the computer system to:

implement a raw data processor that:
 obtains raw LC-ESI-MS data from a source and extracts a base peak chromatogram (BPC) from the raw data;
 detects, from the BPC, a peak for each signal intensity above a first predetermined threshold and, for each detected peak, a retention time, a peak apex and a baseline; and
 extracts a mass spectrum based on data at each retention time and a base peak signal intensity derived from the peak apex and the baseline; and

implement a fingerprinting system that:
 performs background correction by, for each detected peak, producing a resulting m/z value by subtracting mass over charge (m/z) values of the baseline from an original m/z value and, if the resulting m/z value is positive, restoring the original m/z value; and
 generates a compact data representation comprising, for each detected peak, the retention time, original m/z values and intensities; and
 performs chemical fingerprinting for the raw LC-ESI-MS data based on the compact data representation.

- 22 -

2. The computer system of claim 1, wherein the raw data processor identifies, for each detected peak, the peak apex by determining if the peak comprises:
 - an intensity plateau, the peak apex comprising a midpoint of the detected peak; or
 - a gradient sign change to negative, the peak apex corresponding to a point at which the gradient sign changes to negative.
3. The computer system of claim 1 or 2, wherein the raw data processor detects, for each detected peak, the baseline by sampling a nearest baseline timepoint prior to the detected peak and a nearest baseline timepoint after the detected peak.
4. The computer system of claim 3, wherein the m/z values of the baseline comprise the m/z values of the nearest baseline timepoints.
5. The computer system of any one of claims 1 to 4, wherein the raw data processor:
 - detects background noise from the BPC based on an intensity of less than a second predetermined threshold; and
 - averages out the background noise.
6. The computer system of any one of claims 1 to 5, comprising:
 - identifying, as a base peak, a said peak of highest signal intensity;
 - determining if one or more molecular ion peaks exist in the mass spectrum, with detectable isotopes in the mass spectrum; and
 - selecting from the one or more molecular ion peaks, if any, a said molecular ion peak with highest signal intensity,wherein the compact data representation comprises the mass spectrum represented as a series of retention times, m/z values above the first predetermined threshold, base peak m/z , the base peak and molecular ion peak, if any, with highest signal intensity.

- 23 -

7. The computer system of any one of claims 1 to 6, further comprising a query engine, the query engine comparing a first mass spectrum of a first sample corresponding to first raw LC-ESI-MS data, and a second mass spectrum of a second sample corresponding to second raw LC-ESI-MS data, the query engine determining if the first raw LC-ESI-MS data and the second raw LC-ESI-MS data are for similar compounds by comparing m/z of respective base peaks, and retention times of the base peaks.
8. The computer system of claim 7, wherein the query engine is further configured to filter the mass spectrum based on a specific peak intensity or specific m/z value.
9. The computer system of claim 7 or 8, wherein the query engine is further configured to filter the mass spectrum based on a peak intensity range or m/z value range.
10. A method for analysing liquid chromatography electrospray ionisation mass spectrometry (LC-ESI-MS) data, comprising:
 - using a raw data processor to:
 - obtain raw LC-ESI-MS data from a source and extract a base peak chromatogram (BPC) from the raw data;
 - detect, from the BPC, a peak for each signal intensity above a first predetermined threshold and, for each detected peak, a retention time, a peak apex and a baseline; and
 - extract a mass spectrum based on data at each retention time and a base peak signal intensity derived from the peak apex and the baseline; and
 - using a fingerprinting system to:
 - perform background correction by, for each detected peak, producing a resulting m/z value by subtracting mass over

- 24 -

charge (m/z) values of the baseline from an original m/z value and, if the resulting m/z value is positive, restoring the original m/z value;
generate a compact data representation comprising, for each detected peak, the retention time, original m/z values and intensities; and
perform chemical fingerprinting for the raw LC-ESI-MS data based on the compact data representation.

11. The method of claim 10, wherein identifying, for each detected peak, the peak apex comprises determining if the peak comprises:
 - an intensity plateau, the peak apex comprising a midpoint of the detected peak; or
 - a gradient sign change to negative, the peak apex corresponding to a point at which the gradient sign changes to negative.
12. The method of claim 10 or 11, wherein detecting, for each detected peak, the baseline comprises sampling a nearest baseline timepoint prior to the detected peak and a nearest baseline timepoint after the detected peak.
13. The method of claim 12, wherein the m/z values of the baseline comprise the m/z values of the nearest baseline timepoints.
14. The method of any one of claims 10 to 13, further comprising using the raw data processor to:
 - detect background noise from the BPC based on an intensity of less than a second predetermined threshold; and
 - average out the background noise.
15. The method of any one of claims 11 to 16, further comprising:
 - identifying, as a base peak, a said peak of highest signal intensity;

- 25 -

determining if one or more molecular ion peaks exist in the mass spectrum, with detectable isotopes in the mass spectrum; and

selecting from the one or more molecular ion peaks, if any, a said molecular ion peak with highest signal intensity,

wherein generating the compact data representation comprises representing the mass spectrum as a series of retention times, m/z values above the first predetermined threshold, base peak m/z , the base peak and molecular ion peak, if any, with highest signal intensity

16. The method of any one of claims 10 to 15, further comprising implementing a query engine to compare a first mass spectrum of a first sample corresponding to first raw LC-ESI-MS data, and a second mass spectrum of a second sample corresponding to second raw LC-ESI-MS data, the query engine determining if the first raw LC-ESI-MS data and the second raw LC-ESI-MS data are for similar compounds by comparing respective compact data representations corresponding to the first mass spectrum and second mass spectrum.

17. The method of claim 16, further comprising causing the query engine to filter the first mass spectrum based on a specific peak intensity or specific m/z value.

18. The method of claim 16 or 17, further comprising causing the query engine to filter the first mass spectrum based on a peak intensity range or m/z value range.

19. A method of building a database, comprising:

for each of a plurality of base peak chromatograms (BPCs), at a data processor:

obtaining the BPC either directly or from LC-ESI-MS data;

detecting a peak for each signal intensity above a first predetermined threshold and, for each detected peak, a retention time, a peak apex and a baseline; and

- 26 -

extracting a mass spectrum based on data at each retention time and a base peak signal intensity derived from the peak apex and the baseline; and

at a fingerprinting system:

performing background correction by, for each detected peak, producing a resulting m/z value by subtracting mass over charge (m/z) values of the baseline from an original m/z value and, if the resulting m/z value is positive, restoring the original m/z value; and

generating a compact data representation comprising features, each feature being a said detected peak or, for each detected peak, the retention time, original m/z values and intensities;

storing the compact data representation in the database;

and

providing a query interface for:

receiving a query comprising, for a new BPC or new LC-ESI-MS data, a new compact data representation for the new BPC or new LC-ESI-MS data;

performing chemical fingerprinting by comparing of one or more features of the new compact data representation to the features of the compact data representations in the database; and

returning one or more of the compact data representations from the database, based on the chemical fingerprinting.

20.A database built in accordance with the method of claim 19.