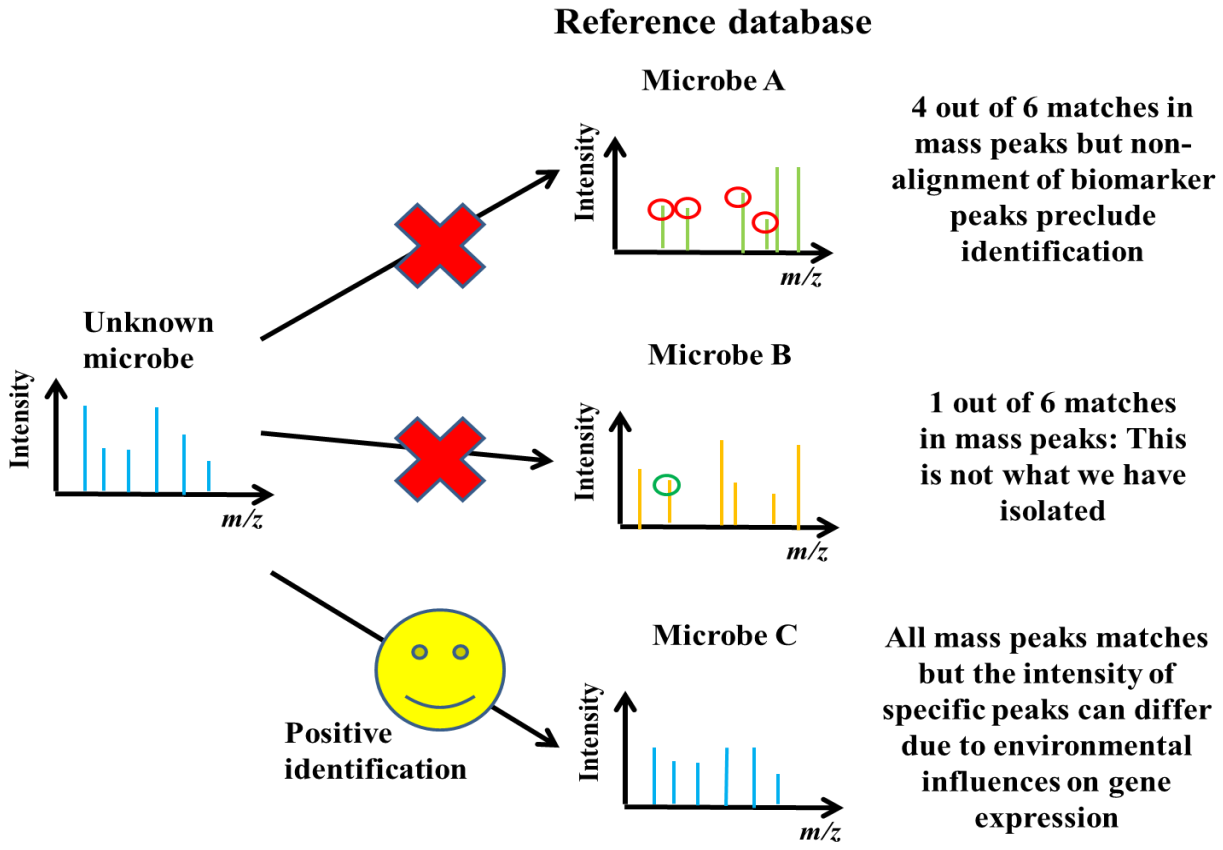
Graphical illustration for explaining mass spectrum fingerprinting in microbial identification

Wenfa Ng

Novena, Singapore, *Email Address*: ngwenfa@alumni.nus.edu.sg



Abstract

Pattern recognition is a common approach for identifying an unknown entity from a set of known objects curated in a database – and find use in various data processing applications such as microbial identification. Whether matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) or electrospray ionization tandem mass spectrometry (ESI MS/MS), mass spectrometry techniques are increasingly used for identifying microbes in the research and clinical settings via species- or strain-specific mass spectrum signatures. Although the existence of unique biomarkers - such as ribosomal proteins - underpins mass spectrometry-enabled microbial identification, lack of corresponding genome or proteome information in publicly accessible databases for a large fraction of extant microbes significantly hamper biomarker (and species) assignment. Nevertheless, the reproducible generation of species-specific mass spectrum across different growth and environmental conditions opens up the

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.580v1 | CC-BY 4.0 Open Access | rec: 2 Nov 2014, publ: 2 Nov 2014

PeerJ PrePrints

possibility of identifying unknown microbes via comparing peak positions between mass spectra, without requiring knowledge of biomarker molecular identities. Thus, the mass spectrum fingerprinting (or pattern recognition) approach circumvents the need for biomarker information. Alignment of as many mass peaks as possible (particularly, those of phylogenetic significance) between spectra is the basis of mass spectrum fingerprinting. In contrast, variation in gene expression and metabolism (and hence, biomolecules' abundances) with environmental and nutritional factors, meant that alignment of peak intensities, though desired, is not a strict requirement for identification. With large diversity of biomolecules present in each microbial species, mass spectrometry-based microbial identification is inherently data-intensive; thereby, requiring statistical tools and computational implementation of the pattern recognition approach, which is incorporated in software packages of microbial typing instruments. Nevertheless, relegation of algorithmic details of pattern recognition to the backend of software obfuscates the approach's conceptual underpinnings and hinders students' understanding. More important, mathematics-centric approaches for explaining the conceptual basis of pattern recognition, though useful, are generally less pedagogically accessible to life science students relative to visual illustration techniques. This short primer describes a simple graphical illustration (featuring three examples common in mass spectrometry-based biotyping workflows) that attempts to explain the conceptual underpinnings of mass spectrum fingerprinting, and highlights caveats for avoiding misidentification.

Keywords: profiling microbes; ribosomal proteins; biomarker; phylogeny; microbial ecology; taxonomy; bioinformatics; pattern recognition; MALDI-TOF MS; mass spectrum fingerprinting;

Subject areas: microbiology; education; ecology; bioinformatics; biodiversity

Conflicts of interest

The author declares no conflict of interest.

Author's contributions

Wenfa Ng developed the pedagogical idea and wrote the manuscript.

Funding

No funding was used in this work.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.580v1 | CC-BY 4.0 Open Access | rec: 2 Nov 2014, publ: 2 Nov 2014

Identifying an object by comparing it with other items is an intuitive method for determining the extent of similarity – the basis for classifying the myriad objects around us along the continuum from totally different to similar and identical. Similarly, microbial identification initially relied on morphological and phenotypic characteristics such as cell shape, types and colours of pigment secreted, growth rate, and cell spreading behavior, for classifying and identifying different microorganisms present in the environment. Realisation that the relatively small microbial morphological and phenotype space severely constrains the identification of the vast diversity of extant microorganisms potentiates the development of biochemical, and more recent, nucleic-acids methods (in particular, 16S rRNA gene sequencing) for identifying microbes. Limitations such as long time between sample and result, complicated sample preparation, and possible bias in analysis results, in biochemical and nucleic acids identification methods, motivated the development of numerous alternative techniques for identifying microorganisms from diverse sample matrixes. One approach increasingly gaining importance, both in clinical diagnostics and research, is mass spectrometric profiling of cellular content of microorganisms where, subsequent annotation of mass spectra data with species-specific biomarkers affords identification. Specifically, the high sensitivity, low limit-of-detection, and generic nature of mass spectrometry profiling facilitate its use in identifying many microorganisms. To this end, the two most commonly used mass spectrometry-based microbial identification approaches are polymerase chain reaction coupled to electrospray ionization mass spectrometry (PCR-ESI MS), and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS).

PCR-ESI MS is an extension of the nucleic acids based approach where PCR is used in amplifying a target gene - thought to yield meaningful phylogenetic information - for subsequent sequence determination via ESI MS/MS. On the other hand, MALDI-TOF MS relies on the gentle (i.e., without inducing fragmentation) ionization of biomarker molecules such as ribosomal proteins in identifying an unknown microbe. Specifically, by profiling the set of ribosomal proteins unique to a species, and subsequently comparing the mass spectra data (particularly, mass peaks position) with those curated in a reference database, positive identification of the unknown microbe is possible if a match (of high similarity) can be obtained. In particular, identification of microbes from MALDI-TOF MS data is usually effected by the biomarker-based approach, where phylogenetically significant biomolecules such as ribosomal proteins serve as identifiers for particular species and strains. Nevertheless, the lack of sufficient publicly available annotated genomic, proteomic and metabolomics information for a sizeable fraction of extant microbial species severely hampers the utility of the biomarker-based approach; a problem which may be alleviated in the future with greater accessibility of high throughput sequencing and proteomics techniques at declining cost as well as accompanying high accuracy automated functional annotation workflows.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.580v1 | CC-BY 4.0 Open Access | rec: 2 Nov 2014, publ: 2 Nov 2014

An alternative approach, pattern recognition (also known as mass spectrum fingerprinting), is capable of analyzing the obtained mass spectra for identifying unknown microbial species without requiring knowledge of the mass peaks' molecular identities. Though simple in concept and amenable to be performed by visual observation for low-dimensional (i.e., small) datasets, the increasing prevalence of high dimensional datasets in microbial ecology studies necessitates the use of statistical techniques for determining the similarity between two entities. In essence, pattern recognition compares the mass spectra of the unknown microbe with those present in a curated reference database/library in search of a match, and thus, positive identification. More specifically, by comparing the presence/absence of mass peaks of microbes cultivated and prepared under identical (or, at least similar conditions), pattern recognition identifies microbes through answering the similarity question at the aggregate level relative to the fine-grained biomolecule specific approach underlying biomarker-based techniques. Taken together, pattern recognition abstracts the identification question from the individual biomarker to an aggregate consortium of mass peaks; thereby, avoiding the need for biomarker identities.

Though pattern recognition is a generic technique, qualitative differences exist in applying the approach to PCR-ESI MS and MALDI-TOF MS. Specifically, the former is primarily focused on reconstructing the nucleotide sequence of a target biomarker gene, while the latter identifies a microbe by comparing a broad set of mass peaks between the mass spectra of unknown and known microorganisms. The rest of this primer will focus on explaining the concept of pattern recognition in relation to MALDI-TOF MS microbial typing.

Environmental stressors trumps repair mechanisms in generating genetic mutations and potentiating divergence of species and strains as microorganisms adapt to fluctuating environmental and nutritional conditions throughout Earth's evolution. The functional role of ribosomal proteins as phylogenetic biomarkers accrues to their high conservation across species - given their important roles as part of the ribosomal complex which mediates protein translation. Thus, with evolutionary pressure negatively selecting for high mutation rates in housekeeping proteins, ribosomal protein genes afford their use as molecular chronometers cataloguing the gradual accumulation of mutations, which, as a set, measures the evolutionary distance between species and strains. The observed small differences in ribosomal protein sequences between strains and species are, however, sufficient to translate into mass differences resolvable by most modern MALDI-TOF mass spectrometers. Thus, detecting small differences in mass peak positions (i.e., *m/z* ratio) often belonging to biomarker proteins forms the conceptual basis upon which mass spectrometry instruments and accompanying software identifies microbial species.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.580v1 | CC-BY 4.0 Open Access | rec: 2 Nov 2014, publ: 2 Nov 2014

Two important caveats hinder the use of a single biomarker for unique identification as well as utilizing the available mutation space at the sequence level for identifying microbes. Specifically, identification of microorganisms via profiling a single ribosomal protein is practically impossible since specific ribosomal protein share high degree of similarity between and within species. On the other hand, silent mutations do not manifest as changes in amino acid sequences - and detectable mass differences - thus, obviating their use for cataloguing the sequence divergence between microorganisms. In contrast, MALDI-TOF MS circumvents the aforementioned conundrum by performing a wide spectrum "scan" over the entire range of *m/z* ratios within the instrument's operating window for yielding a biomolecule map of a particular microbe. Such a map typically captures a large fraction (though not all) of biomolecules present - especially those with abundances significantly above the mass spectrometer's detection limits. Given the importance of ribosomal proteins in maintaining organismal health, their high relative abundances meant that they are usually more prevalent and distinguishable, by either machine or visual observation techniques, and helps anchor their use in mass spectrometry biotyping. Thus, by aggregating over the multitude of different ribosomal proteins present in each microbe, sufficient number of ribosomal proteins would likely exhibit mass differences with their counterparts in another microorganism that, as a holistic set, helps differentiate one species from another. Or, viewed from another perspective, after MALDI-TOF MS profiling of cellular content, each microbial species generates a unique mass spectrum fingerprint comprising species-specific ribosomal proteins and other species-independent biomolecules peaks. Collectively, through mass spectrum comparisons (i.e., position and number of mass peaks), the pattern recognition approach is theoretically capable of identifying an unknown microbe (if a match cannot be obtained with curated mass spectra of known microbes), or yields a positive identification (in the case of a match with sufficient similarity).

Modern MALDI-TOF mass spectrometers generate large datasets per microorganism and given the diversity of microbial species on Earth, computational analysis is the only feasible approach for implementing the pattern recognition method in interrogating the large number of mass spectra typical of microbial identification. Nevertheless, analysis delegated to the computer does not mean that students should view them as black boxes in the workflow; especially without adequate comprehension of the conceptual basis of the approach used. Thus, using a simple conceptual diagram, the present primer attempts to explain the concept underlying the application of pattern recognition in mass spectrometry-based microbial identification.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.580v1 | CC-BY 4.0 Open Access | rec: 2 Nov 2014, publ: 2 Nov 2014

Example illustrating the conceptual basis of pattern recognition in microbial biotyping

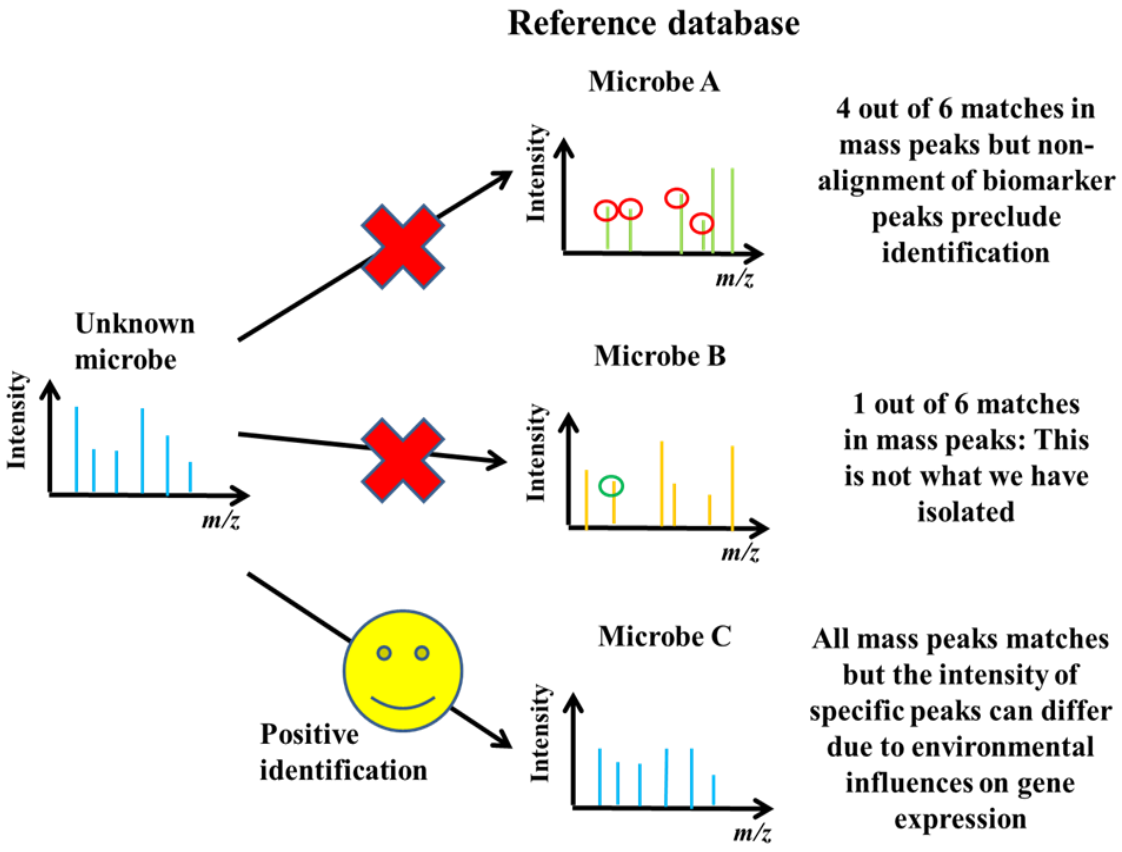


Figure 1: Graphical illustration explaining the conceptual basis of pattern recognition in mass spectrometry-based microbial identification with three examples that cover the range of cases typical of MALDI-TOF MS microbial typing workflows.

The ideal case of positively identifying a microorganism is complete 100% match in all peak positions between the mass spectra of the unknown and curated microbe. As depicted in the graphic (Figure 1), such cases are rare since fluctuations in environmental and nutritional conditions would likely alter gene expression patterns and cell metabolism that, in turn, manifests as the presence or absence of particular mass peaks of metabolites or proteins. Thus, the key criterion for positive identification lies in the close matching of as many biomarker peaks – particularly, those of ribosomal proteins – as possible, since biomarker peaks are given a greater weightage during calculation of the similarity score. For instance, in the example depicted, although 4 out of 6 peaks matches between the mass spectra of the unknown and microbe A, the poor alignment of the biomarker peaks (not explicitly shown) preclude identification. More important, why isn't a 66% match in number of mass peaks sufficient for positive identification? The answer lies in the high probability that various common or

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.580v1 | CC-BY 4.0 Open Access | rec: 2 Nov 2014, publ: 2 Nov 2014

PeerJ PrePrints

housekeeping proteins and metabolites may be accounting for a significant fraction of mass peaks matches, which given their prevalence in a large fraction of species and strains, are not unique microbial species or strain identifiers. It was mentioned earlier that the pattern recognition and mass spectrum fingerprinting approach is capable of identifying microbes without biomarker information; thus, in the above example, how do we know that there is no biomarker match? Given the unique role of ribosomal proteins as microbial species markers, microbial biotyping via MALDI-TOF MS essentially relies on this class of proteins for positive identification. Hence, given knowledge of the typical mass range of ribosomal proteins, a cursory scan of the obtained mass spectrum with the pattern recognition software would inform the presence/absence of ribosomal proteins. Specifically, there is no profiled ribosomal protein mass peak in the mass spectrum of microbe A in the example above; a possible case if microbe A is a virus.

On the other hand, the matching of one peak between the mass spectra of the unknown and microbe B rules out the possibility of a species identification since even if the mass peak belongs to a ribosomal protein, common occurrence of ribosomal protein of identical amino acid sequence across a variety of bacterial species precludes the positive identification of a microbe based on a single ribosomal protein match. In the final case, even though some of the peak intensities between the mass spectra of the unknown and microbe C are dissimilar, the close matching of all mass peaks positions in the two mass spectra is sufficiently robust for positive identification. Though peak intensity offers a second dimension of information potentially useful in microbial identification, practical use of peak height (intensity) for microbial identification is limited given that biomolecule abundance fluctuates depending on the extent in which various nutritional and environmental factors influence cell physiology, metabolism and gene expression. Thus, only in extremely rare cases would there be perfect match in both peak intensity and peak position between two mass spectra – an idealized case that may not be realized even if there is close matching of various sample preparation, instrument analysis and cell cultivation conditions. In general, though potentially useful, peak intensity remains of secondary importance in microbial identification relative to peak position. Taken together, the guiding criterion for a positive identification remains the close alignment of as many mass peaks of phylogenetic significance as possible.

Caveats for avoiding misidentifications

Though simple in concept and demonstrably useful in identifying microbes from MALDI-TOF mass spectra data in the absence of supporting genomic and proteomic information, several caveats exist in using the pattern recognition approach for microbial identification. Firstly, mass spectrum fingerprinting is a statistical approach, since given possible

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.580v1 | CC-BY 4.0 Open Access | rec: 2 Nov 2014, publ: 2 Nov 2014

PeerJ PrePrints

errors and problems with sample preparation etc., the ideal case of perfect match in all peak positions does not exist in practical implementation of the technique. Thus, a score is typically calculated to assess the degree of similarity between two mass spectra, and an arbitrary threshold is used in discriminating between positive and negative identification. Use of an arbitrary threshold necessarily introduce an element of uncertainty in the analytical workflow – but the approach remains valid if the threshold used has been validated in studies covering a wide range of microbes profiled using different mass spectrometers.

Secondly, given that myriad factors such as sample preparation, cell culture conditions, instrument calibration and maintenance, and even the type and make of the MALDI-TOF mass spectrometer used have an impact on the quality of the mass spectra and the mass peaks profiled, close similarity in instrument analysis conditions, sample preparation, and growth parameters of microbes are necessary for removing unwanted bias and reducing incidences of false-positives or false-negatives in identification.

Conclusions

Altogether, using a simple graphic depicting three examples typical of MALDI-TOF MS microbial typing, the present primer attempts to explain the conceptual basis of applying pattern recognition to mass spectrometry-based microbial identification. Specifically, close alignment of all mass peaks of phylogenetic significance between mass spectra, though desired, is not a practical possibility. Nevertheless, the goal remains to align as many biomarker peaks as possible given their higher weightage in the similarity score used for determining identification. Additionally, while desirable, alignment of peak intensity of biomarkers between mass spectra is not a strict criterion for positive identification since various environmental and nutritional factors influence gene expression and cell metabolism, which translates into differing abundance of particular biomolecules even for the same species and strains. Finally, caveats important to avoiding misidentification such as the need for ensuring close similarity in the sample preparation, instrumental analysis and culture conditions of the unknown and curated microbes, as well as the recognition that pattern recognition in general and mass spectrum fingerprinting in particular is a statistical (and probabilistic) technique are highlighted. In particular, the probabilistic nature of identifying microbes by pattern recognition of mass spectra, meant that analysis of the results obtained should consider the uncertainty involved as well as presence of any possible bias, both at the instrumental analysis and sample preparation level. Collectively, the primer should be useful as a supplement to either a microbiology or bioinformatics course for helping students better understand the conceptual underpinnings of pattern recognition's utility in mass spectrometry-based microbial identification.

PeerJ PrePrints

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.580v1 | CC-BY 4.0 Open Access | rec: 2 Nov 2014, publ: 2 Nov 2014