Review Article

Life Sciences, Medicine and Biomedicine
Open Access

Bioinformatics in the identification of microorganisms

Yeoh Chiann Ying and Cheah Yoke Kqueen*

Department of Biomedical Science, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia.

ABSTRACT

Rapid and accurate identification of microorganisms can be of great value for clinical management. For many fastidious and slow-growing microorganisms, the conventional method used for detection is time-consuming, costly and labour-intensive. Hence, the development of new and improved microbial identification methods are necessary to overcome this bottleneck. Current trend has shifted towards the use of new molecular technologies in genomics and proteomics for bacterial identification and characterization. This mini review will focus on summarizing different types of genotypic and proteomics identification methods, as well as bioinformatics tools used for rapid identification and characterization of microorganisms from various specimens.

Keywords: *Microbe identification; bioinformatics; genotypic methods; sequencing; proteomics technologies.*

INTRODUCTION

Microbial diseases remain a major public health burden worldwide that is associated with high morbidity and mortality rates in many regions, particularly in developing countries. Therefore, the development of rapid and sensitive microbial identification methods is of tremendous advantage for correct diagnosis, efficient treatment, environmental health and food safety. In clinical laboratories, phenotypic tests such as microbiological culture and biochemical methods are widely used for laboratory identification and confirmation of microorganisms. These methods include gram-staining, culture and growth characteristics, serological based methods and biochemical profiling with the API identification schema and BBL-crystal system (Millar, Xu, & Moore, 2007). However, this phenotypic-based conventional methods used for laboratory diagnosis are time-consuming, costly and labour-intensive. The complete identification tests may take several days and sometimes the result is inconclusive. Although some automated microbial identification systems such as Biolog Microplate (Biolog, Inc.) and Sherlock (MIDI, Inc.) are commercially available, they often require prolonged growth for fastidious microorganisms and a proportion of unusual organisms are unidentified or misidentified (Tang et al., 1998). Consequently, genotypic-based microbial identification methods have been developed and widely used to identify microbe from various specimens.

Genotypic-based methods are cost-effective, easy to implement and provide highly discriminatory data as compared to phenotypic-based methods (Foxman, Zhang, Koopman, Manning, & Marrs, 2005). Molecular methods of identification such as DNA sequencing are often used in identifying an unknown species. Apart from genotypic methods, proteomics-based identification approaches such as mass spectrometry (MS), have also been widely explored for rapid microbe identification. A huge amount of data can be efficiently generated based on genomics and proteomics based experimental approaches for microbiological studies. However, the analysis of sequence data generated by different microorganisms often involves the application of various bioinformatics tools, methods and parameters.

* **Correspondence**
Cheah Yoke Kqueen
Department of Biomedical Science,
Faculty of Medicine and Health Sciences,
Universiti Putra Malaysia,
43400, Serdang, Selangor, Malaysia.
ykcheah@upm.edu.my
Tel: +603 9769 2343

Received: 17 July 2020
Revised: 7 September 2020
Accepted: 16 September 2020
Published: 6 November 2020

doi
https://doi.org/10.28916/lsmb.4.9.2020.64

The advance development of bioinformatics based on user-friendly web-based resources has been extensively used for data processing and computational resource management, such as wet-lab data analysis, genome sequencing, database creation, and data-mining. In recent decades, bioinformatics and wet-lab techniques are increasingly interdependent on each other for rapid identification and characterization of microorganisms. Bioinformatics enables researchers to study efficiently on microbial diversity, microbe identification and characterization, molecular taxonomy and community analysis patterns of both culturable and unculturable organisms (Tabish et al., 2013). Thus, proper understanding of computational methods is needed to retrieve genetic information and manipulate the massive amount of available data. The current chapter highlights the use of some molecular techniques in conjunction with bioinformatics tools for microbe identification.

GENOTYPIC IDENTIFICATION METHODS

Current genotypic identification methods can be divided into two categories, 1) fingerprint-based techniques, and 2) sequencing-based techniques. Pattern or fingerprint-based techniques are often used for characterizing species-level relationships, but less reliable in establishing the phylogenetic relationships above the species or genus level. Sequence-based techniques have the advantage over fingerprint-based techniques that sequences can be classified based on taxonomy and function, and effective in establishing phylogenetic relatedness above the genus level (Lozupone, Hamady, Kelley, & Knight, 2007). Polymerase Chain Reaction (PCR) is the most widely used methods for nucleic acid amplification. Following the PCR reaction, various post-amplification methods are applied to evaluate the PCR product such as use of specific probes, direct sequence analysis and restriction enzymatic analysis (Tabish et al., 2013).

Fingerprint-based methods

In general, genetic fingerprinting techniques generate a profile or pattern of the microbial community diversity based on amplification of a specific gene followed by separation of DNA fragments by electrophoresis. The resulting reactions yield fingerprints with fragments of different sizes that allow discrimination of a wide variety of microbes. Different samples can then be compared using computer assisted cluster analysis by software packages such as GelComparII and BioNumerics (Rastogi & Sani, 2011). These techniques are rapid and relatively easy to perform in which multiple samples can be simultaneously analysed (Hamady & Knight, 2009). The predominant fingerprinting technologies that have been used for microbe identification include amplified fragment length polymorphism (AFLP), restriction fragment length polymorphism (RFLP), pulsed-field gel electrophoresis (PFGE), random amplification of polymorphic DNA (RAPD), ribotyping, repetitive element polymerase chain reaction (rep-PCR) and multiplex PCR (Emerson, Agulto, Liu, & Liu, 2008). Descriptions of these fingerprinting technologies are provided in Table 1.

Sequencing-based methods

Conventional cultivation methods have a major drawback in identifying a vast majority of microorganisms in environmental samples due to the presence of non-culturable microorganisms (Mocali & Benedetti, 2010). To date, sequencing-based methods are widely used for identification and characterization of unknown bacteria or novel pathogens isolates. These methods require the use of software tools such as Basic Local Alignment Sequence Tool (BLASTn) and FASTA to compare a query sequence to known sequences in the National Center for Biological Information (NCBI) GenBank for species identification. Other specialized databases such as MicroSeq (Applied Biosystem) and BIBI (Devulder, Perrière, Baty, & Flandrois, 2003), have also been developed for the characterization of microorganisms. Some of the most commonly used sequencing-based identification techniques are discussed below.

Ribosomal gene sequencing

Ribosomal gene sequencing is relatively rapid and well-established technique for characterisation of a number of microorganisms, especially for unusual, non-culturable, and slow-growing pathogens (Millar et al., 2007). Three rRNA genes are found in bacteria, i.e. 5S, 16S and 23S rRNA. However, the 16S rRNA gene has been extensively used in microbial detection due to several reasons, i) 16S rRNA gene is universally found in all bacteria with well known function, ii) 16S rRNA gene evolves at a slow and constant mutation rate over time. The gene sequence has both conserved and variable sequence motifs, which are useful for phylogenetic analysis, and iii) The gene has appropriate sequence size (around 1500 bp), which make it easy to sequence and large enough for identification and analysis of phylogeny (Clarridge & Alerts, 2004). The rRNA gene sequences has been widely used to identify and classify microbial species, and investigate microbial diversity in a range of environments. For instance, the phylogenetic relationship of the actinobacteria (Girard et al., 2013) and fungus (Borneman & Hartin, 2000) isolated from environmental samples could be successfully determined based on rRNA gene homology. Recently, the 16S-23S rRNA intergenic transcribed spacer (ITS) has also been employed for identification due to its high sequence variability, which is useful for species recognition. These ITS regions surrounded by conserved sequences (16S/23S and 5.8S/18S/28S) can be amplified using universal bacterial or fungal primers (Tabish et al., 2013).

Analysis of rRNA begins by amplifying the gene coding for rRNA followed by sequencing and database searches. The calculation of pairwise of sequence similarity for the target rRNA gene can be carried out using the Ez-Taxon-e server (Kim et al., 2012). Sequence reads can be taxonomically assigned using bioinformatics tools, such as the Naïve Bayesian Classifier tool (Wang, Garrity, Tiedje, & Cole, 2007) by comparing the sequence with known taxonomically classified sequences deposited in public universal databases. For instance, Ribosomal Database Project (Wang et al., 2007), Greengenes (DeSantis et al., 2006), SILVA (Quast et al., 2013) and ribosomal RNA operon copy number database (rrnDB) (Stoddard, Smith, Hein, Roller, & Schmidt, 2015). In addition, a database of molecular markers, namely The Targeted Loci Project, has been developed for phylogentic analysis and identification of bacteria, archae and fungi (Tatusova et al., 2015). Lastly, a phylogenetic tree can be constructed using the Molecular Evolutionary Genetics Analysis (MEGA) software 5.0 to examine evolutionary relationships (Tamura et al., 2011).

Multilocus sequencing

In recent decades, multilocus sequencing has become one of the powerful tools used for typing of microbial species. Two multiple sequencing techniques are currently used, 1) multilocus sequence typing (MLST) and multilocus sequence analysis (MLSA). MLST is a well-established technique, which identifies internal fragments (usually 400 to 500 bp) of multiple housekeeping genes using specific primers to allow amplification and sequencing of the products. Different sequences within a bacteria species are assigned as distinct alleles for each housekeeping gene and a unique combination of alleles at each locus define the allelic profile that specifies the sequence type (ST). The allelic profile can then be compared with a public accessible database (www.mlst.net) to determine the genetic relatedness of the bacterial strains. Lastly, a dendrogram is constructed to determine the relationships among different ST using the matrix of pairwise differences between allelic profiles. Various bioinformatics tools such eBURST (Feil, Li, Aanensen, Hanage, & Spratt, 2004), SeqMan (DNASTAR, USA), BioNumerics (Applied Maths, Belgium) and MEGA (Tamura et al., 2011), have facilitated researchers to analyse and process MLST data.

Several different MLST strategies have been applied in bacteria

Table 1: Common fingerprinting methods used in microbe identification

Method	Description	Database
Amplified fragment length polymorphism (AFLP)	The use of two restriction enzymes to digest chromosomal DNA followed by linking of adapters to the restriction sites	User generated
Restriction fragment length polymorphism (RFLP)	The use of restriction enzymes to digest chromosomal DNA followed by separation of the restriction fragments by electrophoresis. Fragments are then transferred to a membrane filter by southern blot procedure and hybridize with label DNA probe	User generated
Pulsed-field gel electrophoresis (PFGE)	Chromosomal DNA is separated in a gel matrix by applying an electrical current	Public universal database PulseNet (www.cdc.gov/pulsenet)
Random amplification of polymorphic DNA (RAPD)	Short random primers (typically 10-mer primers) are used to randomly amplify segments of target DNA under low annealing temperature.	User generated
Ribotyping	The use of restriction enzyme to digest genomic DNA into small fragments followed by probing for rRNA gene using southern blot hybridization	User generated; Commercial universal database available Eg. DuPont's RiboPrinter system (www.dupont.com)
Repetitive element polymerase chain reaction (rep-PCR)	Repetitive DNA elements in the chromosomes of bacteria are amplified by PCR, which produces fingerprinting pattern specific to each strain	User generated; commercial individual database available Eg. DiversiLab system (www.biomerieux-diagnostics.com)
Multiplex PCR	Several different target DNA sequences can be amplified simultaneously by using multiple primer pairs in a mixture	User generated

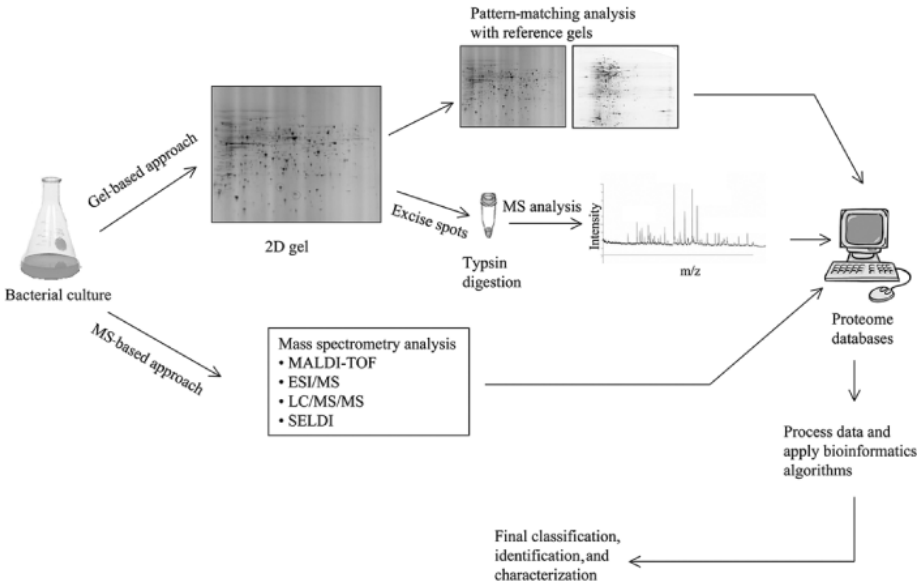

Figure 1: Overview of proteomics methods in microbial identification

isolated from the environment or human, as for instance, *Salmonella* spp. (Sun et al., 2014), *Streptomyces* spp. (Rong & Huang, 2010), and *Enterococcus* spp. (Ruiz-Garbajosa et al., 2006). Alternatively, MLSA can also be used to determine the phylogenetic relationships between closely related bacterial species. MLSA involves DNA sequencing of multiple fragments of conserved protein encoding genes followed by a comparison of concatenated sequences. This method uses a smaller subset of genes (≤ 6) as compared to MLST (Gevers et al., 2005). Recently, a ribosomal multilocus sequence typing (rMLST) approach has been developed, which provides a universal reference point to complement existing multilocus sequencing methods. This method uses the 53 ribossomal protein subunits (*rps* genes) that allows universal characterization of bacteria from domain to strain (Jolley et al., 2012).

Whole microbial genome sequencing

Since the first report on the complete microbial genome of *Haemophilus influenza* in 1995 (Fleischmann et al., 1995), currently over hundreds of microbial genome have been sequenced and archived in the NCBI genome database. More than 10000 microbial genome assemblies have been publicly accessible in the year of 2014, in which nearly 30000 prokaryotic genomes have been successfully sequenced (Tatusova et al., 2015). The rapid advent of the high-throughput next generation sequencing technologies such as pyrosequencing, Illumina, and bioinformatics tools have dramatically reduced the cost and time needed for whole microbial genome sequencing projects. The major impact of the bioinformatics in genome sequencing include, 1) development of automated sequencing techniques 2) combining the sequences of smaller fragments (contigs) together to create a complete whole genome sequence, and 3) the prediction of promoters and protein coding regions of the genome (Bansal, 2005).

Genome sequencing provides a more comprehensive view of microbial genetic diversity by analysing all the genetic information present in a target sample. Briefly, the procedures used in microbial sequencing involve, 1) DNA extraction of target sample, 2) random fragmentation of target DNA into small fragments, 3) ligation and cloning, 4) sequencing of DNA fragments, 5) sequences alignment using specialized programs such as MEGAN Metagenome analyser (Huson,

Auch, Qi, & Schuster, 2007), 6) the annotation of sequences in open reading frame (ORFs) to predict encoded protein and gene functions (Rastogi & Sani, 2011). Whole genome sequences of multiple organisms can be compared by using specialized databases such as Integrated Microbial Genomes, IMG (Markowitz et al., 2012), and Microbial Genome Database, MBGD (Uchiyama, Mihara, Nishide, & Chiba, 2015). Bioinformatics tools based on whole genome polymorphism comparisons have proved to be useful in identifying marker sequences for the identification and differentiation of microbial species (Nagarajan, Loh, & Swarup, 2013).

Metagenomics

Metagenomics is a new promising technique in which the genome contents of a microbial community are sequenced for the determination of the microbial diversity and microbial functional ability in the environment. Also, metagenomics approach can be used to identify microbial strains or genes of biotechnological interest for several biotechnology applications such as discovering new antibiotics and enzymes, and the remediation of natural and artificial pollutants (Abbasian, Lockington, Megharaj, & Naidu, 2015).

One major advantage of metagenomics is the direct sequencing of the samples without the need of cultivation or any prior knowledge of the gene sequence. However, the application of metagenomics generates a huge amount of complex data that can only be analysed or processed using powerful bioinformatics tools. Consequently, several computational tools such as MG-RAST (Meyer et al., 2008), IMG/M (Markowitz et al., 2012), CAMERA (Sun et al., 2011), have been developed to facilitate the analysis of complex metagenomic data sets. Metagenomics analysis based on high-throughput sequencing approach is applicable to any environment. For example, gene samples from nine microbial communities of distinct environments were obtained by pyrosequencing, followed by a comparative analysis to determine different metabolic requirements characteristic to each habitat (Dinsdale et al., 2008).

The employment of high-throughput sequencing (HTS) coupled with powerful bioinformatics tools for metagenomics data analysis has prompted the creation of large scale metagenomic projects. For example, the Human Microbiome Project (HMP) funded by the National Institutes of Health (NIH), is mapping all the microbial communities to study human associated microbes (Tumbaugh et al., 2007; Aagaard et al., 2013). A total of 178 microbial genomes had been completely annotated under The Human Microbiome Jumpstart Reference Strains Consortium (Proctor, 2011).

DNA microarray

An intermediate method between fingerprinting and sequencing techniques is the use of DNA microarray. This approach allows simultaneous identification of specific microbes and providing ecological context for the phylogenetic resolution and functional structure of a given microbial community (Emerson et al., 2008). Briefly, microarray works on the general principle of spotting DNA fragment (probes) for thousands of genes onto a surface of glass or plastic and subsequently bind to complementary DNA or RNA strand. Hybridization is quantified by detection of a fluorescent-labelled targets using laser scanners. The scanner generates a digital image that is further analysed by specialized software to transform it into a numerical reading to determine the relative concentrations of DNA in a sample (Tabish et al., 2013). A whole genome DNA microarray can be created for a comprehensive genetic analysis of an organism. Genomic variation (e.g. amplications, deletions, insertions, rearrangements, and base-pair changes) can be detected using microarray-based approaches (Gresham, Dunham, & Botstein, 2008).

Two types of DNA microarray are currently recognized. They are cDNA arrays and oligonucleotide arrays. In general, cDNA arrays are made by printing a double stranded cDNA on a solid surface using robotic pins whereas oligonucleotide arrays are produced by synthesizing specific oligonucleotides in a specific alignment on a solid support using photolithography (Singh & Kumar, 2013). DNA microarray used in microbial ecology could be classified into two categories, 1) PhyloChip, which has been developed based on the small-subunit ribosomal gene (Wilson et al., 2002); 2) Functional gene arrays such as GeoChip, which has been developed to identify microbes involved in essential biogeochemical processes (He et al., 2007).

PROTEOMICS TECHNOLOGIES

The advent of new proteomics technologies creates an excellent complement to genomics-based methods for microbial identification. Microbial samples can be analysed either using gel-based one-or two-dimensional sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) or different MS-based techniques. In recent decades, mass spectrometry (MS) has become a popular tool for rapid microbial characterization based on the identification of protein biomarkers using experimental mass spectra data. The predominant mass spectrometry techniques that have been utilized for microbial identification and characterization include matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS), electrospray ionization mass spectrometry (ESI-MS), surface-enhanced laser desorption/ionization mass spectrometry (SELDI), or the combination of mass spectrometry, gel electrophoresis and bioinformatics (Emerson et al., 2008). The unknown microbial sample will be identified using bioinformatics analysis (database search and computer algorithm analysis) either via comparison of mass spectra data with a proteomics database containing mass spectra of known organism, or through protein sequence matching in the publicly available protein database.

Overview of proteomics methods in microbial identification is summarized in Figure 1 (Emerson et al., 2008).

Gel-based proteomics methods

SDS-PAGE is one of the most established gel-based technique in which microbe can be differentiated based on their cellular protein contents. First, microbial lysate is prepared and run on SDS-PAGE to separate their whole cell protein content. The SDS-PAGE can then be analysed by comparing it with publicly available reference gel patterns in a protein database. Protein spots of interest can be excised, and subjected to mass analysis to determine their molecular weight (Pandey & Mann, 2000). Another type of approach known as two-dimensional gel electrophoresis (2DE), which uses the combination of isoelectric focusing (IEF) and SDS-PAGE (Magdeldin et al., 2014). Complex mixtures consisted of several thousands of different cellular proteins can be resolved in a single gel analysis. Proteins are first separated via isoelectric focusing (IEF), which separates proteins with a pH gradient according to their isoelectric point, followed by the second dimension SDS-PAGE, which separates each protein according to their molecular weight. A 2DE map generated by an unknown sample can be further analysed by comparing it with the reference database for identification (Magdeldin et al., 2014). Although gel-based techniques are long established, it is still labour-intensive and not suitable for point of care applications.

Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS)

Advance mass spectrometry techniques such as MALDI-TOF MS, has emerged as a powerful, rapid and accurate tool for the identification and characterization of microorganisms. It has been used for the routine clinical diagnosis of pathogens, and identification of environmental organisms, in particular fastidious and slow growing organisms (Biswas & Rolain, 2013). In practice, the sample is spotted on a MALDI-TOF sample target with a matrix solution and allowed to air dry. The plate is then inserted in the MS and bombarded with a laser to create gas phase ions. The released ions travel through a vacuum tube and the time-of-

flight of the ions is precisely measured (Glish & Vachet, 2003). A mass of spectral fingerprint is generated, which contains mass-to-charge ratios of the molecules detected for each species. An unknown species can be identified by comparing the obtained unique spectra with an empirically compiled mass spectral reference database of known organisms. Several studies have been carried out in which MALDI-TOF MS has been successfully used for the identification of a variety of bacteria, including *Salmonella* spp. (Dieckmann & Malorny, 2011), *Staphylococcus* spp. (Wolters et al., 2011), anaerobic bacteria (Barba et al., 2014), *Streptomyces* spp. (Arango et al., 2018), *Enterococcus* spp. and Escherichia coli (Santos et al., 2015).

Electrospray ionization mass spectrometry (ESI-MS)

Over the last decade, ESI-MS has emerged as a sensitive, rapid and reliable tool for identifying complex, non-volatile and thermally labile biological sample. ESI uses high voltage electrical energy to transfer ion from the liquid into the gaseous phases, followed by mass spectrometric analysis (Glish & Vachet, 2003). A major breakthrough in ESI-MS is the development of tandem mass spectrometry (MS/MS). Tandem mass spectrometry involves two stages of mass analysis to examine the fragmentation of the protein of interest (Glish & Vachet, 2003). Protein fragment sequence information can then be generated and subjected to a database search to identify that specific protein. A major application of MS/MS based technologies is the use of liquid chromatography mass spectrometry (LC-MS/MS) for quantitative analysis of compounds in pharmaceutical studies (Haneef et al., 2013). The structure of a compound can be determined based on the molecular ion peaks and fragmentation patterns. A vast amount of protein and peptide data created by tandem mass spectrometry needs to be properly managed for better outcomes. Thus, various bioinformatics software analysis tools have been developed to address this task. Peptide identification algorithms can be broken into two broad categories: (1) database search, which identify peptide by matching the unknown amino acid sequences against a protein database such as UniProt (Apweiler et al., 2004), and (2) de novo search, which infers peptide sequences without the need of genomic data. Several peptide searching algorithms have been successfully established. For example, Mascot (Perkins, Pappin, Creasy, & Cottrell, 1999), SEQUEST (Diament & Noble, 2011), MassWiz (Yadav et al., 2011), and PEAKS DB (Zhang et al., 2011).

Surface-enhanced laser desorption/ionization mass spectrometry (SELDI)

The SELDI mass spectrometry is a rapid and high-throughput technique for profiling protein mixtures. A major advantage of SELDI is its ability to analyse protein samples with high throughput capacity and minimal sample requirement. This technique utilizes chromatographic chip (ProteinChip) with modified surfaces that allows the separation of proteins based on their chemical and physical characteristics (i.e. hydrophobic, hydrophilic, acidic, metal affinity). The sample is applied directly to the surface and the bound proteins are profiled using the mass analyser. The integrated mass analyser will produce mass spectra data based on the mass-to-charge ratio of the proteins for further analysis (Nilsen et al., 2011). The most widely used bioinformatics approaches for SELDI spectra data analysis are decision tree-based ones and support vector machines (Smith et al., 2007). SELDI technology becomes an alternative approach for microbial identification and differentiation based on the comparison of protein profiling. This technology has also been widely used for biomarker discovery and protein profiling studies in the medical oncology field (Rodrigo et al., 2014; Smith et al., 2007).

CHALLENGES IN BIOINFORMATICS

The rapid advent of new microbial identification methods that are based primarily on molecular-based techniques and bioinformatics tools offer an excellent complement to conventional microbiological methods (Emerson et al., 2008). Though many molecular techniques and bioinformatics tools have been introduced for microbial identification, some challenges and limitations still exist. The major drawbacks of bioinformatics approaches are they often involve stimulation-based science. Despite the use of sophisticated bioinformatics tools to model an experiment and predict the outcomes, there is a need to perform wet-lab experiments to testify the predictions made. Thus, the progress in these techniques has to remain interdependent to facilitate the detection of microorganisms (Bansal, 2005).

Sequence-based identification methods become more cost-effective nowadays due to the development of next generation sequencing technologies. Based on the data collected from National Human Genome Research Institute, there is a substantial reduction in the cost of sequencing (Chun & Rainey, 2014). Advances in the field of genome sequencing have dramatically increased the amount of genome sequences stored in public databases. Consequently, data-mining of this vast amount of dataset for microbial identification will be challenging considering the cost of accessing the necessary bioinformatics hardware and software (Chun & Rainey, 2014). A greater challenge will be the establishment of integrated databases for rapid assembly of immense amount of sequences data. It is also difficult to maintain a large and rapidly growing database (Emerson et al., 2008). Many databases have been closed or terminated due to inability to deal with massive amount of new data and insufficient funding to maintain readily available databases. Though GenBank is a primary public database, its data entries may be mistakenly labelled and contaminated, leading to wrong interpretation (Tabish et al., 2013).

The development of bioinformatics research and applications has facilitated microbial characterization via automated analysis of huge number of microbial genomes. However, it has several limitations: (1) lack of available gene-functionality from the wet-lab data, (2) lack of computational methods to explore huge data with unknown functionality, (3) limited availability of protein-protein and protein-DNA interactions, and (4) limited knowledge of transient temporal behavior of genes and molecular pathways (Bansal, 2005). Remaining challenges include the lack of standardized methods for routine application of these techniques, and lack of sufficient experience scientists to carry bioinformatics research (Emerson et al., 2008).

CONCLUSION

The advance development of genomics and proteomics technologies, as well as the high-throughput bioinformatics tools have promoted the study of microbial biotechnology. It is becoming apparent that public accessible genomic and proteomics databases play a critical role in identifying microbe based on the wet-lab derived sequence information. Since molecular identification technologies are highly dependent on the bioinformatics, continuous improvement of the software and databases are critical for more accurate analysis. Although most of the molecular identification approaches available have some limitations for complex microbial communities structure and function analysis, a combination of several molecular techniques can be applied to increase the accuracy and reliability of test results. The development of high-throughput microbial identification techniques with bioinformatics capabilities are needed for the maintenance of public health.

DISCLOSURES

The authors declare no conflicts of interest in this work.

ACKNOWLEDGEMENTS

The authors are grateful to Yayasan Penyelidikan Antartika Sultan Mizan (YPASM) and Geran Putra Malaysia for the research funding and the Department of Biomedical Science, Faculty of Medicine and Health Sciences, Universiti Putra Malaysia for the facilities.

REFERENCES

Aagaard, K., Petrosino, J., Keitel, W., Watson, M., Katancik, J., Garcia, N., Patel, S., Cutting, M., Madden, T., Hamilton, H., Harris, E., Gevers, D., Simone, G., McInnes, P., & Versalovic, J. (2013). The human microbiome project strategy for comprehensive sampling of the human microbiome and why it matters. FASEB Journal, 27(3), 1012-1022. https://doi.org/10.1096/fj.12-220806

Abbasian, F., Lockington, R., Megharaj, M., & Naidu, R. (2015). The integration of sequencing and bioinformatics in metagenomics. Reviews in Environmental Science and Biotechnology, 14(3), 357-383. https://doi.org/10.1007/s11157-015-9365-7

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., & Yeh, L.-S. L. (2004). UniProt: The Universal Protein knowledgebase. Nucleic Acids Research, 32, 115-119. https://doi.org/10.1093/nar/gkh131

Arango, C., Acosta-Gonzalez, A., Parra-Giraldo, C. M., Sánchez-Quitian, Z. A., Kerr, R., & Diaz, L. E. (2018). Characterization of actinobacterial communities from Arauca river sediments (Colombia) reveals antimicrobial potential presented in low abundant isolates. The Open Microbiology Journal, 12, 181-194. https://doi.org/10.2174/1874285801812010181

Bansal, A. K. (2005). Bioinformatics in microbial biotechnology--a mini review. Microbial Cell Factories, 4, 19. https://doi.org/10.1186/1475-2859-4-19

Barba, M. J., Fernández, A., Oviaño, M., Fernández, B., Velasco, D., & Bou, G. (2014). Evaluation of MALDI-TOF mass spectrometry for identification of anaerobic bacteria. Anaerobe, 30, 126-128. https://doi.org/10.1016/j.anaerobe.2014.09.008

Biswas, S., & Rolain, J. M. (2013). Use of MALDI-TOF mass spectrometry for identification of bacteria that are difficult to culture. Journal of Microbiological Methods, 92(1), 14-24. https://doi.org/10.1016/j.mimet.2012.10.014

Borneman, J., & Hartin, R. J. (2000). PCR primers that amplify fungal rRNA genes from environmental samples. Applied and Environmental Microbiology, 66(10), 4356-4360. https://doi.org/10.1128/AEM.66.10.4356-4360.2000

Clarridge, J. E., & Alerts, C. (2004). Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clinical Microbiology Reviews, 17(4), 840-862. https://doi.org/10.1128/CMR.17.4.840-862.2004

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., & Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and Environmental Microbiology, 72(7), 5069-5072. https://doi.org/10.1128/AEM.03006-05

Devulder, G., Perrière, G., Baty, F., & Flandrois, J. P. (2003). BIBI, a bioinformatics bacterial identification tool. Journal of Clinical Microbiology, 41(4), 1785-1787. https://doi.org/10.1128/JCM.41.4.1785-1787.2003

Diament, B. J., & Noble, W. S. (2011). Faster SEQUEST searching for peptide identification from tandem mass spectra. Journal of Proteome Research, 10(9), 3871-3879. https://doi.org/10.1021/pr101196n

Dieckmann, R., & Malorny, B. (2011). Rapid screening of epidemiologically important Salmonella enterica subsp. enterica serovars by whole-cell matrix-assisted laser desorption ionization-time of flight mass spectrometry. Applied and Environmental Microbiology, 77(2), 4136-4146. https://doi.org/10.1128/AEM.02418-10

Dinsdale, E. A., Edwards, R. A., Hall, D., Angly, F., Breitbart, M., Brulc, J. M., Furlan, M., Desnues, C., Haynes, M., Li, L., McDaniel, L., Moran, M. A., Nelson, K. E., Nilsson, C., Olson, R., Paul, J., Brito, B. R., Ruan, Y., Swan, B. K., Stevens, R., Valentine, D. L., Thurber, R. V., Wegley, L., White, B. A., & Rohwer, F. (2008). Functional metagenomic profiling of nine biomes. Nature, 452(7187), 629-632. https://doi.org/10.1038/nature06810

Emerson, D., Agulto, L., Liu, H., & Liu, L. (2008). Identifying and characterizing bacteria in an era of genomics and proteomics. Bioscience, 58(10), 925. https://doi.org/10.1641/B581006

Feil, E. J., Li, B. C., Aanensen, D. M., Hanage, W. P., & Spratt, B. G. (2004). eBURST: Inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. Journal of Bacteriology, 186(5), 1518-1530. https://doi.org/10.1128/JB.186.5.1518-1530.2004

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., Mckenney, K., Sutton, G., Fitzhugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L. I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., Mcdonald, L. A., Small, K. V, Fraser, C. M., Smith, H. O., & Venter, J. C. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science, 269(5223), 496-512. https://doi.org/10.1126/science.7542800

Foxman, B., Zhang, L., Koopman, J. S., Manning, S. D., & Marrs, C. F. (2005). Choosing an appropriate bacterial typing technique for epidemiologic studies. Epidemiologic Perspectives and Innovations, 2, 10. https://doi.org/10.1186/1742-5573-2-10

Gevers, D., Cohan, F. M., Lawrence, J. G., Spratt, B. G., Coenye, T., Feil, E. J., Stackebrandt, E., Van de Peer, Y., Vandamme, P., Thompson, F. L., & Swings, J. (2005). Opinion: Re-evaluating prokaryotic species. Nature Reviews Microbiology, 3(9), 733-739. https://doi.org/10.1038/nrmicro1236

Girard, G., Traag, B. A., Sangal, V., Mascini, N., Hoskisson, P. A., Goodfellow, M., & van Wezel, G. P. (2013). A novel taxonomic marker that discriminates between morphologically complex actinomycetes. Open Biology, 3(10), 130073. https://doi.org/10.1098/rsob.130073

Glish, G. L., & Vachet, R. W. (2003). The basics of mass spectrometry in the twenty-first century. Nature Reviews Drug Discovery, 2(2), 140-150. https://doi.org/10.1038/nrd1011

Gresham, D., Dunham, M. J., & Botstein, D. (2008). Comparing whole genomes using DNA microarrays. Nature Reviews Genetics, 9(4), 291-302. https://doi.org/10.1038/nrg2335

Hamady, M., & Knight, R. (2009). Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. Genome Research, 19(7), 1141-1152. https://doi.org/10.1101/gr.085464.108

Haneef, J., Shaharyar, M., Husain, A., Rashid, M., Mishra, R., Parveen, S., Ahmed, N., Pal, M., & Kumar, D. (2013). Application of LC-MS/MS for quantitative analysis of glucocorticoids and stimulants in biological fluids. Journal of Pharmaceutical Analysis, 3(5), 341-348. https://doi.org/10.1016/j.jpha.2013.03.005

He, Z., Gentry, T. J., Schadt, C. W., Wu, L., Liebich, J., Chong, S. C., Huang, Z., Wu, W., Gu, B., Jardine, P., Criddle, C., & Zhou, J. (2007). GeoChip: A comprehensive microarray for investigating biogeochemical, ecological and environmental processes. The ISME Journal, 1(1), 67-77. https://doi.org/10.1038/ismej.2007.2

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. Genome Research, 17(3), 377-386. https://doi.org/10.1101/gr.5969107

Jolley, K. A., Bliss, C. M., Bennett, J. S., Bratcher, H. B., Brehony, C., Colles, F. M., Wimalarathna, H., Harrison, O. B., Sheppard, S. K., Cody, A. J., Maiden, M. C. J. (2012). Ribosomal multilocus sequence typing: Universal characterization of bacteria from domain to strain. Microbiology, 158, 1005-1015. https://doi.org/10.1099/mic.0.055459-0

Kim, O. S., Cho, Y. J., Lee, K., Yoon, S. H., Kim, M., Na, H., Park, S.-C., Jeon, Y. S., Lee, J.-H., Yi, H., Won, S., & Chun, J. (2012). Introducing EzTaxon-e: A prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. International Journal of Systematic and Evolutionary Microbiology, 62, 716-721. https://doi.org/10.1099/ijs.0.038075-0

Lozupone, C. A., Hamady, M., Kelley, S. T., & Knight, R. (2007). Quantitative and qualitative diversity measures lead to different insights into factors that structure microbial communities. Applied and Environmental Microbiology, 73(5), 1576-1585. https://doi.org/10.1128/AEM.01996-06

Magdeldin, S., Enany, S., Yoshida, Y., Xu, B., Zhang, Y., Zureena, Z., Lokamani, I., Yaoita, E., & Yamamoto, T. (2014). Basics and recent advances of two dimensional- polyacrylamide gel electrophoresis. Clinical Proteomics, 11(1), 16. https://doi.org/10.1186/1559-0275-11-16

Markowitz, V. M., Chen, I. M. A., Chu, K., Szeto, E., Palaniappan, K., Grechkin, Y., Ratner, A., Jacob, B., Pati, A., Huntemann, M., Liolios, K., Pagani, I., Anderson, I., Mavromatis, K., Ivanova, N. N., & Kyrpides, N. C. (2012). IMG/M: The integrated metagenome data management and comparative analysis system. Nucleic Acids Research, 40(1), 123-129. https://doi.org/10.1093/nar/gkr975

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J., & Edwards, R. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics, 9(1), 386. https://doi.org/10.1186/1471-2105-9-386

Millar, B. C., Xu, J., & Moore, E. (2007). Molecular diagnostics of medically important bacterial infections. Current Issues in Molecular Biology, 9, 21-40.

Mocali, S., & Benedetti, A. (2010). Exploring research frontiers in microbiology: The challenge of metagenomics in soil microbiology. Research in Microbiology, 161(6), 497-505. https://doi.org/10.1016/j.resmic.2010.04.010

Nagarajan, K., Loh, K.-C., & Swarup, S. (2013). Bioinformatics and molecular biology for the quantification of closely related bacteria. Applied Microbiology and Biotechnology, 97(14), 6489-6502. https://doi.org/10.1007/s00253-013-4943-5

Nilsen, M. M., Uleberg, K.-E., Janssen, E. A. M., Baak, J. P. A., Andersen, O. K., & Hjelle, A. (2011). From SELDI-TOF MS to protein identification by on-chip elution. Journal of Proteomics, 74(12), 2995-2998. https://doi.org/10.1016/j.jprot.2011.06.027

Pandey, A., & Mann, M. (2000). Proteomics to study genes and genomes. Nature, 405(6788), 837-846. https://doi.org/10.1038/35015709

Perkins, D. N., Pappin, D. J., Creasy, D. M., & Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis, 20(18), 3551-3567. https://doi.org/10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2

Proctor, L. M. (2011). The human microbiome project in 2011 and beyond. Cell Host and Microbe, 10(4), 287-291. https://doi.org/10.1016/j.chom.2011.10.001

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glockner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Research, 41(1), 590-596. https://doi.org/10.1093/nar/gks1219

Rastogi, G., & Sani, R. K. (2011). Molecular Techniques to Assess Microbial Community Structure, Function, and Dynamics in the Environment. In I. Ahmad., F. Ahmad., & J. Pichtel (Eds), Microbes and Microbial Technology, pp. 29-58. New York: Springer. https://doi.org/10.1007/978-1-4419-7931-5_2

Rodrigo, M. A. M., Zitka, O., Krizkova, S., Moulick, A., Adam, V., & Kizek, R. (2014). MALDI-TOF MS as evolving cancer diagnostic tool: A review. Journal of Pharmaceutical and Biomedical Analysis, 95, 245-255. https://doi.org/10.1016/j.jpba.2014.03.007

Rong, X., & Huang, Y. (2010). Taxonomic evaluation of the Streptomyces griseus clade using multilocus sequence analysis and DNA-DNA hybridization, with proposal to combine 29 species and three subspecies as 11 genomic species. International Journal of Systematic and Evolutionary Microbiology, 60, 696-703. https://doi.org/10.1099/ijs.0.012419-0

Ruiz-Garbajosa, P., Bonten, M. J. M., Robinson, D. A., Top, J., Nallapareddy, S. R., Torres, C., Coque, T. M., Canton, R., Baquero, F., Murray, B. E., del Campo, R., & Willems, R. J. L. (2006). Multilocus sequence typing scheme for Enterococcus faecalis reveals hospital-adapted genetic complexes in a background of high rates of recombination. Journal of Clinical Microbiology, 44(6), 2220-2228. https://doi.org/10.1128/JCM.02596-05

Santos, T., Capelo, J. L., Santos, H. M., Oliveira, I., Marinho, C., Gonçalves, A., Araújo, J. E., Poeta, P., & Igrejas, G. (2015). Use of MALDI-TOF mass spectrometry fingerprinting to characterize Enterococcus spp. and Escherichia coli isolates. Journal of Proteomics, 127, 321-331. https://doi.org/10.1016/j.jprot.2015.02.017

Singh, A., & Kumar, N. (2013). A review on DNA microarray technology. International Journal of Current Research and Review, 5(22), 1-5.

Smith, F. M., Gallagher, W. M., Fox, E., Stephens, R. B., Rexhepaj, E., Petricoin, E. F., Liotta, L., Kennedy, M. J., & Reynolds, J. V. (2007). Combination of SELDI-TOF-MS and data mining provides early-stage response prediction for rectal tumors undergoing multimodal neoadjuvant therapy. Annals of Surgery, 245(2), 259-266. https://doi.org/10.1097/01.sla.0000245577.68151.bd

Stoddard, S. F., Smith, B. J., Hein, R., Roller, B. R. K., & Schmidt, T. M. (2015). rrnDB: Improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. Nucleic Acids Research, 43(1), 593-598. https://doi.org/10.1093/nar/gku1201

Sun, J. F., Ke, B. X., Huang, Y. H., He, D. M., Li, X., Liang, Z. M., & Ke, C. W. (2014). The molecular epidemiological characteristics and genetic diversity of Salmonella Typhimurium in Guangdong, China. PLoS One, 9(11), e113145. https://doi.org/10.1371/journal.pone.0113145

Sun, S., Chen, J., Li, W., Altintas, I., Lin, A., Peltier, S., Stocks, K., Allen, E. E., Ellisman, M., Grethe, J., & Wooley, J. (2011). Community cyberinfrastructure for advanced microbial ecology research and analysis: The CAMERA resource. Nucleic Acids Research, 39, 546-551. https://doi.org/10.1093/nar/gkq1102

Tabish, M., Azim, S., Hussain, M. A., Rehman, S. U., Sarwar, T., & Ishqi, H. M. (2013). Bioinformatics Approaches in Studying Microbial Diversity. In A, Malik., E. Grohmann., & M. Alves (Eds), Management of Microbial Resources in the Environment, pp. 119-140. Dordrecht: Springer. https://doi.org/10.1007/978-94-007-5931-2_6

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutioanry distance, and maximum parsimony methods. Molecular Biology and Evolution, 28(10), 2731-2739. https://doi.org/10.1093/molbev/msr121

Tang, Y. -W., Ellis, N. M., Hopkins, M. K., Smith, D. H., Dodge, D. E., & Persing, D. H. (1998). Comparison of phenotypic and genotypic techniques for identification of unusual aerobic pathogenic gram-negative bacilli. Journal of Clinical Microbiology, 36(12), 3674-3679. https://doi.org/10.1128/JCM.36.12.3674-3679.1998

Tatusova, T., Ciufo, S., Federhen, S., Fedorov, B., McVeigh, R., O'Neill, K., Tolstoy, I., & Zaslavsky, L. (2015). Update on RefSeq microbial genomes resources. Nucleic Acids Research, 43(1), 599-605. https://doi.org/10.1093/nar/gku1062

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-liggett, C., Knight, R., & Gordon, J. I. (2007). The human microbiome project: Exploring the microbial part of ourselves in a changing world. Nature, 449(7164), 804-810. https://doi.org/10.1038/nature06244

Uchiyama, I., Mihara, M., Nishide, H., & Chiba, H. (2015). MBGD update 2015: Microbial genome database for flexible ortholog analysis utilizing a diverse set of genomic data. Nucleic Acids Research, 43(1), 270-276. https://doi.org/10.1093/nar/gku1152

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology, 73(16), 5261-5267. https://doi.org/10.1128/AEM.00062-07

Wilson, K. H., Wilson, W. J., Jennifer, L., Desantis, T. Z., Viswanathan, V. S., Kuczmarski, T. A., Andersen, G. L., & Radosevich, J. L. (2002). High-density microarray of small-subunit ribosomal DNA probes. Applied and Environmental Microbiology, 68(5), 2535-2541. https://doi.org/10.1128/AEM.68.5.2535-2541.2002

Wolters, M., Rohde, H., Maier, T., Belmar-Campos, C., Franke, G., Scherpe, S., Aepfelbacher, M., & Christner, M. (2011). MALDI-TOF MS fingerprinting allows for discrimination of major methicillin-resistant Staphylococcus aureus lineages. International Journal of Medical Microbiology, 301(1), 64-68. https://doi.org/10.1016/j.ijmm.2010.06.002

Yadav, A. K., Kumar, D., & Dash, D. (2011). MassWiz: A novel scoring algorithm with target-decoy based analysis pipeline for tandem mass spectrometry. Journal of Proteome Research, 10(5), 2154-2160. https://doi.org/10.1021/pr200031z

Zhang, J., Xin, L., Shan, B., Chen, W., Xie, M., Yuen, D., Zhang, W., Zhang, Z., Lajoie, G.A., & Ma, B. (2011). PEAKS DB: De Novo sequencing assisted database search for sensitive and accurate peptide identification. Molecular & Cellular Proteomics, 11(4), M111.010587. https://doi.org/10.1074/mcp.M111.010587

Citation:

Yeoh, C. Y. ., & Cheah, Y. K. (2020). Bioinformatics in the identification of microorganisms. *Life Sciences, Medicine and Biomedicine*, *4*(9). https://doi.org/10.28916/lsmb.4.9.2020.64

Life Sciences, Medicine and Biomedicine
ISSN: 2600-7207