Chapter 14

Introduction to molecular fingerprinting tools and bioinformatic analysis

Murugadas Vaiyapuri and Radhakrishnan V N

Antimicrobial resistance (AMR) is becoming a major concern for human health. The World Health Organization (WHO) has prioritised the diseases for which research and development programmes are urgently needed. Pathogens are prioritised depending on the determination to create new antibiotics or preserve current medications for treatment or control techniques. Pathogens are classified as critical, high, or medium priority. Priority 1 bacteria include carbapenem-resistant *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and ESBL-producing *Enterobacteriaceae*; priority 2 bacteria include vancomycin-resistant Methicillin-resistant *Enterococcus faecium* Vancomycin-intermediate-resistant *Staphylococcus aureus* Cephalosporin-resistant *Staphylococcus aureus*, fluoroquinolone-resistant Fluoroquinolone-resistant *Neisseria gonorrhoeae* Fluoroquinolone-resistant *Salmonellae* Clarithromycin-resistant *Campylobacter* spp. Priority 3 includes fluoroquinolone-resistant *Helicobacter pylori*. Ampicillin-resistant *Shigella* sp. Penicillin-resistant *Streptococcus pneumoniae*, *Haemophilus influenzae* (WHO, 2017). Gram-negative bacteria, such as *A. baumannii*, *P. aeruginosa*, ESBL producing Enterobacteriaceae (*Klebsiella, E. coli, Serratia*, and *Proteus*), *Campylobacter sp, Helicobacter pylori, Salmonella sp, Shigella sp*, N. *gonorrhoeae*, and *H. influenzae*, are the key targets for management. *S. aureus, Enterococcus faecium,* and *S. pneumoniae* are three more major Gram-positive bacteria that contribute to the urgency of AMR.

The challenge of regulating AMR begins with giving proof of its prevalence as well as recording the genotype of the prevalent bacteria. Molecular methods remain crucial for understanding regional and global epidemiology, as well as the point of genesis and transmission of infections based on clone relatedness and genetic diversity down to the strain level (Barrett et al., 2006; Vaiyapuri et al., 2019). The valuable evidence gathered

will be an important factor in developing methods for their control (Ranjbar et al., 2014). AMR resistance molecular approaches determine the existence of AMR genes or particular mutations linked with antibiotic resistance (WHO, 2019). Molecular approaches may supplement phenotypic methods by giving additional information, such as the particular gene or mutation behind a resistance phenotype, boosting our knowledge of the amount of resistance in a given situation as well as the underlying processes responsible for resistance (Bissonnette et al., 2017; Burnham et al., 2017; WHO, 2019).

Molecular fingerprinting tools are categorized into 4 based on the approach

1. Sequencing based method
2. Amplification based method
3. Hybridization based method
4. Restriction digestion-based method

1. Sequence based molecular fingerprinting tools

Multi-locus sequence typing

The molecular typing procedure which involves the sequencing analysis of more than a single locus is called multi-locus sequence typing (MLST). The first MLST procedure was demonstrated during 1998 for *Neisseria meningitides* and later on the MLST analysis was extended to more than 100 species or genera of bacteria (Enright & Spratt, 1999; Maiden *et al.*, 1998). The MLST analysis can be used for epidemiological investigations in public health, animal health and food borne disease investigations. This is one of the best molecular tools for fingerprinting of bacterial pathogens. Mostly implemented in global and local epidemiology. The MLST analysis is targeted on housekeeping genes of clinically important bacterial pathogens. The MLST analysis is divided into three steps, amplification of housekeeping genes, sequencing of the amplicons and analysis of sequences for assigning the sequence types. In general, the amplicons range between 450 to 500bp for analysis purposes. In exceptional cases, the amplicon size may vary. This tool has high reproducibility, more discriminating power, portability of data

as it is sequence based, speed of the completion, easy interpretation and inter-laboratory comparison. Example of *S. aureus* / MRSA- MLST- analysis. The MLST scheme for the *S. aureus* (MRSA and MSSA) was developed by Enright et al. (2000). The MLST can be performed by amplification of fragments, elution and sequencing, bioinformatics analysis.

Staphylococcal Protein A typing

Staphylococcal Protein A typing (SPA) is the single locus amplification and sequencing method. The SPA gene's variable region XR domain is the target area for SPA typing in *S. aureus* or MRSA. It is often used for Staphylococcal protein A (spa) typing (Frenay et al., 1999). This approach involves amplification, sequencing, and SPA type designation, and it is often employed in local epidemiology since it accumulates genetic alterations very slowly (Koreen et al., 2004). Open source or software from StaphType (Ridom GmbH, Wurzburg, Germany) and Based upon Repeat Pattern (BURP) are used to perform minimum spanning tree-based clustering of spa kinds (Harmsen et al., 2003; Sammeth et al., 2006; Aires de Sousa et al., 2006).

Whole genome sequencing

Whole-genome sequencing (WGS) is a thorough approach for studying bacterial isolates' whole genomes. There are many ways for sequencing whole genomes. In the late 1970s, Maxam and Gilbert's chemical cleavage approach and Sanger sequencing by chain-termination method were used to sequence viral and bacterial genomes (Maxam and Gilbert., 1977; Sanger et al., 1977). In 2008, a transition to a faster, automated sequencing approach was made, allowing for the sequencing of bigger bacterial and eukaryotic genomes. Following-Sanger sequencing technologies are referred to as 'next-generation sequencing.' It can generate massive volumes of sequencing data at a very cheap cost and time. 454-sequencing, pyrosequencing, Illumina sequencing, and Sequencing by Oligonucleotide Ligation and Detection are all examples of second-generation sequencing (SOLiD). While Sanger's sequencing operates on the basis of chain termination by the inclusion of di-deoxynucleotides (ddNTPs), A, T, G, and C, the output data is slightly less

than one kilobase (kb). NGS is the massively parallel sequencing of millions of DNA fragments at the same time, resulting in the generation of millions of nucleotide short reads. The most prevalent NGS technology is undoubtedly Illumina dye sequencing, which employs a "sequence by synthesis" strategy. The genomic DNA is split at random into small pieces and affixed to the inner surface of a flow cell, where sequencing will occur. A solid-phase PCR is then used to generate clusters of clonal populations from each of the individual DNA strands bound to the flow cell. At the end of each cycle, the incorporating nucleotide's identity is recorded using a photodetector by activating the fluorophores with suitable lasers, followed by enzymatic removal of the blocking fluorescent moieties and progression to the next location (Fedurco et al, 2006; Turcatti et al, 2008; Heather and Chain, 2016)

2. Amplification based methods

Repetitive extragenic palindromic-PCR (REP-PCR)

Repeating extragenic palindromic -PCR (Versalovic et al., 1991, 1994) is a fingerprinting technology established on the PCR foundation in a bacterial genome that is based on the amplification of these repetitive elements that are particularly distinct to strains within the species of bacteria. These repeating REP elements were found in a variety of *Enterobacteriaceae* and non-*Enterobacteriaceae* bacteria, and the REP sequences are palindromic, forming a stem-loop structure (Higgins et al., 1982; Stern et al., 1984). Two sets of primers targeting these REP elements, based on 38-bp sequences having degenerate sequences in six places with a 5-bp variable loop among both sides of a conserved palindromic stem, were used for typing in different bacteria. REP element-based typing of AMR bacteria such as *E. coli, Salmonella* sp., and others has been described (Qian and Adhya, 2017).

Enterobacterial repetitive intergenic consensus-PCR (ERIC-PCR)

The ERIC sequences, which include a 126-bp sequence that is highly conserved core inverted repeat and are situated in extragenic areas of the bacterial genome, are another type of repetitive DNA sequences utilised for bacterial typing. Initially detected in *E. coli* and *Salmonella Typhimurium*, the

typing approach based on the ERIC pattern is currently being extended to additional bacteria in the *Enterobacteriaceae* (Sharples et al., 1990).

Amplified fragment length polymorphisms (AFLP)

The amplified fragment length polymorphism (AFLP) approach is excellent for fingerprinting DNAs of any origin and complexity. AFLP offers various benefits over other DNA fingerprinting techniques. The ability to check a full genome for polymorphism and its repeatability are the most crucial. AFLP has the potential to become a universal DNA fingerprinting technique since it can be used to any DNA sample, including microbial DNA, human, animal, and plant DNA. The polymerase chain reaction is used to selectively amplify restriction fragments from a digest of total genomic DNA in the AFLP approach. Zabeau and Vos were the first to create this approach (1993). The AFLP process consists of four basic steps: DNA digestion, ligation, amplification, and gel analysis. Two restriction enzymes initially degrade genomic DNA. The DNA fragments are ligated using double-stranded oligonucleotide adapters that are identical to one of the 5' or 3' ends formed during restriction digestion. The ligated DNA fragments are amplified by PCR using primers that are complementary to the adaptor and restriction site sequences, as well as extra selected nucleotides at the 3' end. The employment of selective primers minimises the mixture's complexity. Under precise annealing conditions, only fragments with complimentary nucleotides extending beyond the restriction site will be amplified by the selective primers. AFLP is a RAPD variant that may discover restriction site polymorphisms without previous sequence information by employing PCR amplification for restriction fragment detection. Janssen et al., (1996) found considerable support for the use of AFLP in bacterial taxonomy by comparing newly acquired data to findings produced by well-established genotypic and chemotaxonomic approaches like as DNA hybridization and cellular fatty acid analyses. Screening of DNA markers connected to genetic characteristics and microbiological typing, diagnostics of genetically inherited disorders, pedigree analysis, forensic typing, and parentage analysis are some of its possible uses.

Randomly amplified polymorphic DNA (RAPD)

The RAPD approach is a PCR-based discriminating method in which short arbitrary primers anneal to several random target sequences, resulting in the formation of the test organism's fingerprint profile. The target sequence to be amplified is unknown in RAPD analysis, and a 10base random sequence primer is utilised in the experiment to construct the RAPD profile. The low-stringency annealing conditions required for the RAPD-PCR reaction result in the amplification of randomly sized DNA fragments. The RAPD-PCR multiple band patterns is followed by dendrogram analysis to provide fingerprint profiles for the test organism. This technique may be used to identify clonal variants in bacterial strains. Because RAPD patterns are not always reproducible, hence this approach must be used under carefully controlled settings. The RAPD approach was utilised to identify enteropathogenic *E. coli, Salmonella, Shigella, Vibrio, Aeromonas*, and *Listeria* in food and water samples. RAPD has been used by multiple organisations to identify and characterise LAB strains from diverse sources, including human, food, and milk samples.

3. Hybridization based method

Microarray

A microarray is made up of regularly organised target DNA sequences that are connected to a solid substrate such as glass, silicon wafers, nylon membranes, or other functionalized substrates. The sample's DNA is fluorescently labelled and put to the array (hybridization). A fluorescent microarray detector and a computer application can then identify and evaluate several distinct AMR genes (Holzman, 2003). Fink et al. (2019) created a microarray-based AMR chip that identifies massive ARGs for -lactams and vancomycin. Using 14 probes, an array chip was built for six key classes of antibiotics, including -lactams, macrolides, aminoglycosides, tetracyclines, sulphonamides, and trimethoprim (Card et al., 2014) and found a total of 14 distinct resistance genes conferring resistance to six antibiotic classes. A microarray chip was created for 166 ARGs found in major Gram positive and negative bacteria (Garneau et al., 2010).

Fluorescent in situ hybridization (FISH)

FISH is a method that uses fluorescently labelled oligonucleotide probes to hybridise to the complementary DNA sequences of resistance genes. After the hybridization procedure is completed, any leftover probes are rinsed away. The signal from the bounded probes for ARGs is captured using epifluorescence or confocal laser scanning microscopy (Levsky and Singer, 2003). FISH probe designed to detect ampicillin, macrolide, and chloramphenicol resistance in *Escherichia coli, Helicobacter pylori*, and *Bacillus cereus* (Demiray and Yilmas, 2005; Juttner et al, 2004; Laflamme et al, 2009; Lee et al., 2019).

Restriction digestion based fingerprinting methods

Pulsed-field Gel electrophoresis (PFGE)

PFGE is a "Gold standard" approach for bacterial pathogen molecular subtyping. The approach involves in-situ macro-restriction of isolated genomic DNA in an agarose plug and digestion with restriction endonucleases (Barrett et al., 1994). Later, as part of the PulseNet approach, the gold standard this technique is widely used for disease outbreak analysis from food (Swaminathan et al., 2001). Initially, the Tenover (1995) guideline was employed, and subsequently, software for character-based analysis was created to evaluate genomic DNA based on the band number and positions of the band that emerged in the gel electrophoresis (Barrett et al., 2006). These approaches were also employed for source tracing, as well as local and worldwide epidemiology (Vernile et al., 2009).

Restriction Fragment Length Polymorphism (RFLP)

The restriction enzyme restricts the microbe's chromosomal DNA in RFLP. The approach was originally created and utilised to build linkages in the human genome (Botstein et al., 1980). The same approach may be used to produce bands from any DNA, including PCR products and tagged probes with restriction sites. The fingerprint or banding pattern created by agarose gel electrophoresis shows the availability and distribution of restriction sites throughout the chromosome. The RFLP technique may be used to compare strains within species, and using rare cutting restriction enzymes minimises the number of bands generated when compared to using frequent cutting restriction enzymes. The banded pattern was then utilised in probe

hybridization. As a result of restrictions such as time and labour demanding work for pure DNA extraction, restriction digestion and probe-based hybridization, and recording and analysis of bands, this approach has lost its relevance (Ben-Ari and Lavi, 2012; Ranjbar et al., 2014).

Multi Locus Enzyme Electrophoresis

MLEE is a system established using restriction enzymes for multiple loci of housekeeping genes prior to the introduction of MLST. The banding pattern, including the number and location of the bands, is also examined in this manner (Zahner et al., 1994). The use of MLEE approaches has decreased significantly since the development of the MLST scheme (Kotetishvili et al., 2003).

Bioinformatic analysis of fingerprint data

Bioinformatic analysis of fingerprint data is performed as character-based analysis or sequence-based analysis. In the character-based analysis, the number of bands and position of bands are taken into consideration for the normalization and analysis purpose. In this type of character-based analysis, the image quality is very crucial. There is much software used for character-based analysis, bio numeric is the software paid version and GelJ is the open-source software used in the character-based analysis. In the sequence-based analysis, the sequence data is either compared to the public domain or assigned value for the fingerprint data.

Example for character-based analysis ERIC PCR fingerprint analysis. Using GelJ software normalise the banding pattern visually. Construct the phylogenetic tree of the selected isolates with GelJ software. Keep the DNA ladder (100bp) for the normalization of banding position. Construct the phylogenetic tree based on the similarity calculated by Pearson correlation between the fingerprints with the tolerance of 1%, and grouping of the fingerprints with the help of the unweighted pair group method using arithmetic averages algorithm (UPGMA) (Rasschaert et al. 2005). Example for sequence-based data is multi-locus sequencing typing data. The assignment of alleles number and allelic profile has to be carried out after analysing the quality of the sequences obtained and after assembling. After assigning the allelic profiles, the sequence type's identification will be

carried out in the PubMLST public domain. Assigning the clonal complexes and relationship to the already existing bacterial strains in the public domain can be carried using the Minimum spanning tree development based on algorithm called BURST (based upon related sequence types) developed by Mellman *et al.* 2008. This will bring you the information on how many evolutionary events happened for your strain and what is the diversity of clones prevalent in your local region. As per the information on https://pubmlst.org/organisms/staphylococcus-aureus, there are now 902, 1091, 954, 607, 937, 855, 1024 alleles were identified and available in public domain for *arc, aroE, glpF, gmk, pta, tpi and yqiL* loci respectively. Detailed information on MLST and analysis for bacterial strains can be sought from https://pubmlst.org/organisms/staphylococcus-aureus., MRSA from seafood (Murugadas et al., 2017; Vaiyapuri et al., 2019).

****