Microarray long oligo probe designing for Bacteria: An in silico pan-genomic research

Payam Behzadi[1], Reza Ranjbar[1]

[1]Molecular Biology Research Center, Baqiyatallah University of Medical Sciences,Tehran, Iran.

Corresponding author: Dr. Reza Ranjbar
Address: Shahid Nosrati alley, Sheikh Bahaee Avenue, Molla Sadra Street, Vanak Square, Tehran, Iran;
Telephone: +982188039883; E-mail: ranjbarre@gmail.com

Abstract

Aim: Increasing the accuracy, reproducibility and flexibility of the output is the main goal for clinical and microbiological diagnostic techniques to control the progression of infection outbreaks and epidemiology. DNA microarray technology provides these goals via professional probe designing. In this research article we have shown the basis of microarray long oligo probe designing for detection and identification of different types of bacteria.
Methods: In this *in silico* investigation, a diversity of free online and offline softwares, databases, and tools were applied. Among a huge number of bioinformatics facilities, the authors have used AlleleID 7.7, BLAST/NCBI, GView Server, OligoAnalyzer 3.1, PanSeq Server.
Results: In this research project, the comparative pan-genomic methodology visualized an obvious and clear illustration regarding to the quality of microarray probes. By the help of *in silico* methodology, we were able to predict the quality of the designed probes. Two types of microarray long oligo probes with ‘Good/Best’ quality were designed and produced.
Conclusion: The use of effective bioinformatics facilities directly affects on the accuracy, flexibility, reliability, rapidity and reproducibility of microarray diagnosis. Best probes guarantee the unbiased and high quality outcomes.

Keywords: bacteria, bioinformatics, in silico, long oligo probe, microarray.

Introduction

There are several molecular biology tools and techniques like Polymerase Chain Reaction (PCR) which are used as a proper and routine clinical and microbiological diagnostic tool. In parallel with PCR application, the advanced, rapid and accurate tool of microarray technology is applied for 2 decades. However, microarray technology has been performed in research centers rather than clinical centers. Among different types of microarray technologies, the DNA microarray technology is a whole genomic analytical method which is able to detect, identify and quantify 1 ng of DNA molecules. Interestingly, the DNA microarray technology has the capability of detection and identification of several thousands of microbial genera, species and strains at once (1-7).

In parallel with progression of next generation sequencing (NGS) methods like DNA microarray technologies in recent decade, the cost of advanced NGS techniques and diagnostics has significantly decreased. These types of diagnostic methods involve comparative genomic and pan-genomic characteristics of microorganisms like bacteria. The accuracy of whole genomic methodologies is guaranteed; because there are some specific genomic sequences (up to 40%) which are completely different to a specific bacterial species/ strain (6,8,9).

As the bacterial pathogens may cause deathful and dangerous diseases and infections in contaminated people, an accurate, rapid and cost-effective diagnostic tool is needed. DNA microarray technology is an appropriate candidate for detection and identification of bacterial pathogens. DNA microarray technique is consisted of several different steps comprising microarray chip spotting, microarray probe designing, bacterial DNA labelling, hybridization, and microarray scanning. However, the last step may be replaced by fluorescence microscope, when the only goal is to detect and identify a particular bacterial agent (2,5,10-14).

According to previous studies, the type of platform and the kind of probe set used in microarray technologies determine the level of probable biases in final outcomes (2,4,15).

For this reason the aim of this *in silico* research article is to show how a microarray long oligo probe can be designed for detection and identification of pathogenic bacteria.

Methods

To design one or more microarray long oligo probes there is a vital need for pan-genomic similarities among the main bacterial strain and its close related bacteria. For example, among *Enterobacteriaceae* members there is a close relationship between *Escherichia coli* and *Salmonella enterica*/ *Shigella* spp.. So, these bacteria must be checked by comparative genomic methodology (16-19).

There are some processes which must be performed one by one to produce the best outcome in association with microarray long oligo probes. By the help of National Center for Biotechnology Information (NCBI) database (http://www.ncbi.nlm.nih.gov/), the genomic information relating to the bacterial strains were extracted. This database encompasses a wide range of choices (1,2,10). The computational method for probe designing was achieved through the following steps:

Genomic data collection

The NCBI FTP Site was selected (ftp://ftp.ncbi.nlm.nih.gov/). This site offers several services. The Genomes folder involves a huge mass of data (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/). This service has been modified (12/21/15, 5:16:00 PM) and some links and addresses have been changed (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/). Prior to the latest modification in NCBI, the main bacterial strain .gbk file (the main bacterial strain is the strain that from its genome sequence the microarray long oligo probes were designed and produced) and for close related bacterial strains, the .fna files were downloaded. All these files were zipped and used for the next step. The whole genome sequences are available throughout other websites including Sanger Institute (http://www.sanger.ac.uk/resources/downloads/bacteria/), National Institute of Allergy and Infectious Diseases (NAID) Genomic Centers for Infectious Diseases (GCID) (http://www.niaid.nih.gov/labsandresources/resources/dmid/gsc/Pages/default.aspx) and the Ilumina (http://www.illumina.com/) (4,10,20,21).

Pan-genomic comparative analysis

Pan-genomics shows a 95% nucleotide sequence similarity between close strains pertaining to the same bacterial species. Despite this close similarity, there is a huge variability within genomic pools of close bacterial strains (22-25).

Because of the vast genomic diversity, there is an appropriate methodology to illustrate and visualize the similarities and dissimilarities within the gene content of several close bacterial strains. This online tool is available via GView Server (https://server.gview.ca/). GView is a free Java application which visualize the prokaryotic (Archaeal and Bacterial) genome maps in both forms of circular and linear. This bioinformatics application provides a complete, accurate and clear image to show the genome map, gene content, similarities and dissimilarities found between compared bacterial genomes (21,26-28).

For applying GView Server services, the user has to register for this free tool. After the registration is completed, the user can use GView Server.

To compare several genomes by GView Server, the unique genome option was selected as the analysis type. The main bacterial strain sequence data file (.gbk file) as the reference genome was uploaded. The GenBank ans EMBL file formats are permitted to upload for the main bacterial sequence (21,27).

At the end, the email address is needed to push the continue button and upload the .gbk file. In the next page the .fna files (more than 1 with FASTA format) belonging to close related bacterial strains were uploaded. There are some default parameters comprising value cutoff, genetic code, alignment length cutoff and percentage identity cutoff which must be regulated in accordance with the related aim (21,27). In addition to GReview Server as a powerful pan-genome visualizer tool which is used for different purposes such as comparative pan-genomic analysis, there is the online tool of PanSeq Server which is a free access and rapid pan-genome sequence analysis program (https://lfz.corefacility.ca/panseq/) (29). However, there are other pan-genome sequence analyzer including m GenomeSubtractor (http://202.120.12.134/mGS2/) (30) and nWayComp (31). The PanSeq is capable to extract the unique regions in a genome or compared genomes. The outcomes resulted from PanSeq are easy to use. The Novel Region Finder (NRF) modulates the input sequences. The main bacterial strain (main pan-genome sequences) must be selected from Query Strain and added to selected Query. The other bacterial pan-genome sequences were selected from Reference Strains and added to Selected Reference. All the input sequences were chosen from complete genome choices. There are various default parameters which can be modulated in accordance with the final purpose. In the process of multiple sequence comparison by NRF module, the output of sequence analysis is provided as a file in FASTA format (21,29).

Pan-genomic alignment algorithm

The unique sequences resulted from PanSeq analysis were downloaded into FASTA format and blasted by BLAST search service/NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi). The more repetitive sequences within the output unique sequences, the longer takes time to compute the matched sequences (21,29,32-34).

Probe designing software

The conserved sequences in a bacterial genome are proper genomic regions for designing microarray probes. Furthermore, an ideal conserved sequence using for microarray probe lacks mutation, and possesses a suitable biological melting temperature.

These strain/species specific conserved sequences were detected precisely. Thus, an appropriate software like AlleleID is needed for designing and producing effective and flexible microarray long oligo probes. The selected conserved and unique sequences was added into Microarray tab/New sequence page in ALLELEID software.Then the sequence was analyzed by analyze/Probe Search options. The probe search page includes several default options such as design probes of length (in bp). For designing a long oligo probe, the 55-64 bp was selected. other defaults were selected relating to the research goal. The search button on probe search page produces and designs microarray long oligo probes (2,32,35,36).

Probe alignment algorithm

The designed and produced probes were tested again. This process was done by BLAST/NCBI. The BLAST aligns the sequence of designed probes (2,34,35).

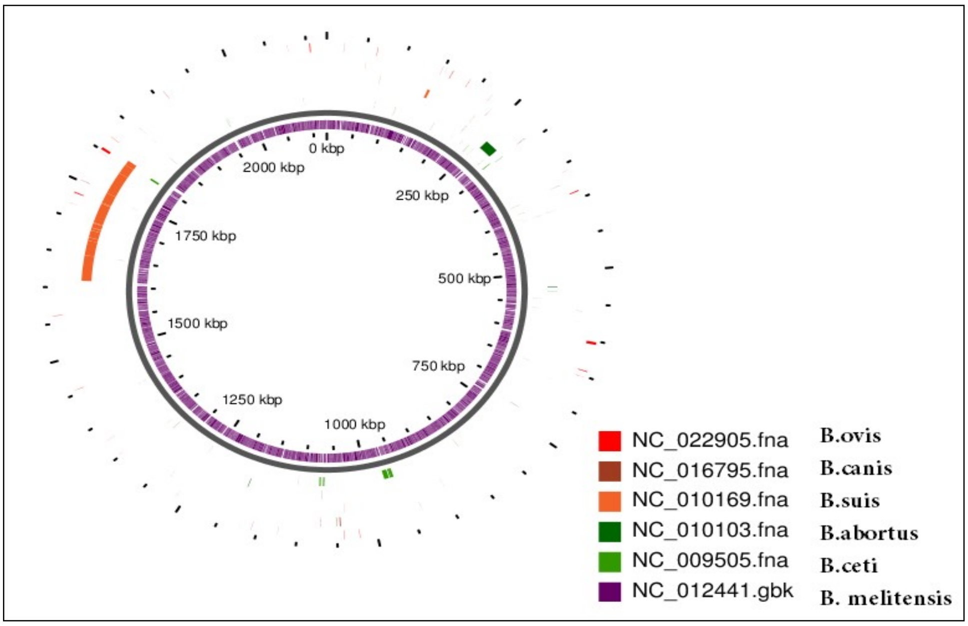
Probe physico-chemical characteristics

The rechecked designed microarray probes were studied by the important online tool of OligoAnalyzer (https://eu.idtdna.com/calc/analyzer). OligoAnalizer is a proper physico-chemical calculator which predicts a wide range of probe characteristics such as target type (DNA/RNA), concentrations (oligonucleotide, dNTPs, and ions of Na, K, and Mg). Other options including Analyze, Hairpin, Self-Dimer, Hetero-Dimer, NCBI BLAST, and Tm mismatch determine the biophysical properties regarding to designed probes (2,37).

Results

The output of Greview Server was an illustration which showed the relationship of the main bacterial strain with other compared strains. In Figure 1 the central circle belongs to the main bacterial strain and the outer circles pertain to other compared strains. The colored outer circles show the unique sequences within the main bacterial strain. This figure is related to *Brucella melitensis* (the main bacterial strain) and other *Brucella* spp. including *B.ovis, B.canis, B.suis, B.abortus, B.ceti*.

Figure 1. A pan-genomic comparison between six bacterial strains of *Brucella*[*]



[*]The main bacterial strain is *B.melitensis* (central circle) which is compared with five other strains (outer circles). The colored outer circles indicate the unique regions within *B.melitensis* pan-genome.

According to Figure 1, there is a very close relationship among aforementioned bacterial strains. In this case there will be a big problem for designing microarray probes; because the unique regions are limited for producing and designing best probes.

The output of PanSeq Server which shown the pan-genomic unique regions regarding to the main bacterial strain was confirmed by BLAST tool of NCBI. Then, the unique sequence is processed by AlleleID 7.7 which resulted in microarray long oligo probes. The produced probes were checked again by BLAST and OligoAnalyzer 3.1. Table 1 shows the designed long oligo probes in *Salmonella enterica* Typhi (Table 1). There is a close relationship between *Slamonella* spp, *Shigella* and *E.coli* strains. Thus, a comparative pan-genomic operation was done among *S.enterica* Typhi, other strains of *S.enterica*, *Shigella* spp. and *E.coli* bacteria (17).

Table 1. The designed microarray long oligo probes relating to *S.enterica* Typhi

Bacteria	Probes	Length	Quality
S.enterica Typhi	CGACCATTGAACCGACAATCTTGCTTATTCCATTACGACAATCACATTCATAGGATTCT	59	Best
S.enterica Typhi	GCCTGGCTTTCCTGGAGTCTCCTATTAAGTTACTATCAATATCCTTTGCTATGTCTTCTTCTA	63	Good

Discussion

The presence of a wide range of deathful and pathogenic bacterial agents provokes microbiologists to find out accurate, reliable, reproducible, rapid and cost effective diagnostic methods. As the previous studies show, the bacterial infectious diseases and the related epidemiology and outbreaks are in association with inappropriate diagnosis and incorrect treatment. Among different microbial diagnostics, molecular biology techniques and NGS technologies are known as effective and accurate methods (8,38,39).

There are a wide range of advanced molecular biology techniques like PCR which are cost effective, rapid, accurate and reliable. But the disadvantage of PCR technique is its limitation for processing several samples at once. PCR is not suitable for a huge number of samples (11,40,41). In contrast to PCR, DNA microarray is a proper tool for detecting and identifying tens of thousand genes at once. The principle of microarray technology is based on bioinformatics. So, the process of probe designing determines the level of accuracy and reliability of the results. Besides, suitable probe designing makes microarray flexible and comfortable (32,42,43).

For designing an effective, specific and reliable probe, there is a vital need for bioinformatics knowledge. A significant low content of G-C, low number and diversity of nucleotide sequences and abundance of repeated sequences decrease the accuracy, specificity and reliability of the probe. So, the use of powerful and easy to use tools, softwares, and databases guarantees the accuracy of the designed probe. Furthermore, the use of very close strains to the main bacterial strain declines the quality of designed probes. This makes probes incorrect, ineffective and inaccurate (2-4,16,21,44).

The products of AlleleID software are divided into Good probes and Best probes. As Lukajancenko and Ussery (16) have shown in their study, a high number of probes may be designed and produced for a specific bacterial strain. But the number of useful and effective probes is limited. The reason of this feature refers to the quality of the designed probes. To have an unbiased and accurate outcome, there is a must for choosing the best quality probes and avoiding Good ones.

In conclusion, DNA microarray technology is an accurate, reliable, rapid, cost effective and reproducible diagnostic technique. However, the level of biases in its outcomes is in association with probe quality and designing. The microarray technology has shortened the period of diagnosis time from two weeks into less than five days. Therefore, probe designing is an important process of microarray technique which guarantees the specificity, accuracy, reliability, flexibility and reproducibility of this rapid diagnostic tool.

Acknowledgement

This research was performed via grant facility offered by the Deputy of Research and Technology, Ministry of Health and Medical Education, Tehran, Iran.

Conflicts of interest: None declared.

References

1. Behzadi P, Behzadi E, Ranjbar R. Microarray data analysis. Alban Med J 2014;4:84-90.
2. Behzadi P, Behzadi E, Ranjbar R. Microarray probe set: Biology, bioinformatics and biophysics. Alban Med J 2015;2:78-83.
3. Bodrossy L, Sessitsch A. Oligonucleotide microarrays in microbial diagnostics. Curr Opin Microbiol 2004;7:245-54.
4. Sandberg R, Larsson O. Improved precision and accuracy for microarrays using updated probe set definitions. BMC Bioinform 2007;8:48.
5. Najafi A, Ram M, Ranjbar R. Microarray Principles & Applications. 1st ed. Tehran: Persian Science & Research; 2012.
6. Fournier PE, Dubourg G, Raoult D. Clinical detection and characterization of bacterial pathogens in the genomics era. Genome Med 2014;6:1-15.
7. Behzadi P, Behzadi E. Environmental Microbiology. 1st ed. Tehran: Niktab; 2007.
8. Didelot X, Bowden R, Wilson DJ, Peto TE, Crook DW. Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet 2012;13:601-12.
9. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. Curr Opin Genet Dev 2005; 15:589-94.
10. Behzadi P, Behzadi E, Ranjbar R. Basic Modern Molecular Biology. 1st ed. Tehran: Persian Science & Research; 2014.
11. Behzadi P, Ranjbar R, Alavian SM. Nucleic Acid-Based Approaches for Detection of Viral Hepatitis. Jundishapur J Microbiol 2015;8:e17449.
12. Behzadi P, Najafi A, Behzadi E, Ranjbar R. Detection and Identification of Clinical Pathogenic Fungi by DNA Microarray. Infectioro 2013;35:6-10.
13. Behzadi P, Behzadi E, Ranjbar R. The application of Microarray in Medicine. ORLro 2014;7:24-6.
14. Behzadi P, Behzadi E, Ranjbar R. Microarray and respiratory tract infections. ORLro 2015;29:34-6.
15. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 2003;31:e15.
16. Lukjancenko O, Ussery DW. Design of an Enterobacteriaceae Pan-Genome Microarray Chip. In: Computational Systems-Biology and Bioinformatics. Springer Berlin Heidelberg; 2010:165-79.
17. Gordienko EN, Kazanov MD, Gelfand MS. Evolution of pan-genomes of Escherichia coli, Shigella spp., and Salmonella enterica. J Bacteriol 2013;195:2786-92.
18. Skippington E, Ragan MA. Phylogeny rather than ecology or lifestyle biases the construction of Escherichia coli–Shigella genetic exchange communities. Open Biol 2012; 2:120112.
19. Sims GE, Kim SH. Whole-genome phylogeny of Escherichia coli/ Shigella group by feature frequency profiles (FFPs). Proc Natl Acad Sci USA 2011;108:8329-34.
20. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol 2008;11:472-7.
21. Sankarasubramanian J, Vishnu US, Sridhar J, Gunasekaran P, Rajendhran J. Pan-Genome of Brucella Species. Indian J Microbiol 2015;55:88-101.
22. Muzzi A, Donati C. Population genetics and evolution of the pan-genome of Streptococcus pneumoniae. Int J Med Microbiol 2011;301:619-22.
23. Konstantinidis KT, Tiedje JM. Towards a genome-based taxonomy for prokaryotes. J Bacteriol 2005;187:6258-64.

24. Rodriguez-Valera F, Ussery DW. Is the pan-genome also a pan-selectome? F1000Res 2012;1.

25. Lawrence JG, Hendrickson H. Genome evolution in bacteria: order beneath chaos. Curr Opin Microbiol 2005;8:572-8.

26. Petkau A, Stuart-Edwards M, Stothard P, Van Domselaar G. Interactive microbial genome visualization with GView. Bioinformatics 2010;26:3125-6.

27. GView Server Guide. Available from: https://server.gview.ca/guide#unique (Accessed: 10 February, 2016).

28. Su H-C, Khatun J, Kanavy DM, Giddings MC. Comparative genome analysis of ciprofloxacin-resistant Pseudomonas aeruginosa reveals genes within newly identified high variability regions associated with drug resistance development. Microb Drug Resist 2013;19:428-36.

29. Laing C, Buchanan C, Taboada EN, Zhang Y, Kropinski A, Villegas A, et al. Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. BMC Bioinform 2010;11:461.

30. Shao Y, He X, Harrison EM, Tai C, Ou H-Y, Rajakumar K, et al. mGenomeSubtractor: a web-based tool for parallel in silico subtractive hybridization analysis of multiple bacterial genomes. Nucleic Acids Res 2010;38:W194-200.

31. Yao J, Lin H, Doddapaneni H, Civerolo EL. nWayComp: a genome-wide sequence comparison tool for multiple strains/species of phylogenetically related microorganisms. In Silico Biol 2007;7:195-200.

32. Jahandeh N, Ranjbar R, Behzadi P, Behzadi E. Uropathogenic Escherichia coli virulence genes: invaluable approaches for designing DNA microarray probes. Cent European J Urol 2015;68:452-8.

33. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389-402.

34. Zhang J, Madden TL. PowerBLAST: a new network BLAST application for interactive or automated sequence analysis and annotation. Genome Res 1997;7:649-56.

35. Apte A, Singh S. AlleleID. PCR Primer Design. Springer; 2007:329-45.

36. López-Campos G, Martínez-Suárez JV, Aguado-Urda M, López-Alonso V. Bioinformatics in Support of Microarray Experiments. Microarray Detection and Characterization of Bacterial Foodborne Pathogens. Springer; 2012:49-92.

37. Owczarzy R, Tataurov AV, Wu Y, Manthey JA, McQuisten KA, Almabrazi HG, et al. IDT SciTools: a suite for analysis and design of nucleic acid oligomers. Nucleic Acids Res 2008;36:W163-9.

38. Feero WG, Guttmacher AE, Relman DA. Microbial genomics and infectious diseases. N Engl J Med 2011;365:347-57.

39. Chan VL. Bacterial genomes and infectious diseases. Pediatr Res 2003;54:1-7.

40. Speers DJ. Clinical applications of molecular biology for infectious diseases. Clin Biochem Rev 2006;27:39.

41. Edwards MC, Gibbs RA. Multiplex PCR: advantages, development, and applications. Genome Res 1994;3:S65-75.

42. Patil MR, Bhong CD. Veterinary Diagnostics and DNA Microarray Technology. Int J Livest Res 2015;5:1-9.

43. Lévêque N, Renois F, Andréoletti L. The microarray technology: facts and controversies. Clin Microbiol Infect 2013;19:10-4.

44. Severgnini M, Cremonesi P, Consolandi C, De Bellis G, Castiglioni B. Advances in DNA microarray technology for the detection of foodborne pathogens. Food Bioprocess Tech 2011;4:936-53.