Microarray data analysis

Payam Behzadi[1], Elham Behzadi[2], Reza Ranjbar[1]

[1]Molecular Biology Research Center, Baqiyatallah University of Medical Sciences, Tehran, Iran;
[2]Department of Microbiology, College of Basic Sciences, Shahr-e-Qods Branch, Islamic Azadi University, Tehran, Iran.

Corresponding author: Dr. Reza Ranjbar, Molecular Biology Research Center, Baqiyatallah University of Medical Sciences;
Address: Shahid Nosrati alley, Sheikh Bahaee Avenue, Molla Sadra Street, Vanak Square, Tehran, Iran;
Telephone: +982188039883; E-mail: ranjbarre@gmail.com

Abstract

Aim: Inspite of significant progressions in microarray techniques and accurate management, the analysis and interpretation of raw data is a big challenge for the majority of researchers on global scale. In this mini-review the authors have described general parameters to overcome current errors and drawbacks occurred in management, analysis and interpretation of microarray raw data.
Methods: Visualization of correct data is related to researchers' knowledge about methodologies, experimental designing, appropriate platforms, upgrade softwares, suitable statistical tests and valuable databases. Hence, being up to date and skillful is considered a key factor for ensuring accurate data management, analysis and interpretation.
Results: Application of correlated methodologies, experimental designs, platforms, softwares, statistical tests and virtual databases, guarantees high quality management, analysis and interpretation of microarray raw data.
Conclusion: In accordance with new efforts in the field of databases and softwares, microarray data management, analysis and interpretation have been improved. The rise of microarray technology applications may lead to a significant decrease of the costs in the future.

Keywords: biological information, data analysis, microarray.

Introduction

Microarray technologies are one of the most advanced experimental methods in the field of clinical diagnostics. These techniques permit researchers to view the molecular biology strategies of microorganisms via tracing tens of thousands of genes at once. Microarrays are excellent approaches for evaluating various cells from dissimilar aspects of genetics and molecular biology. The basis of microarray is simple and it consists of a glass slide carrying determined spots of DNA molecules which are known as probes. The probes are located on a glass slide via different techniques of photolithography, inkjet printing and robot spotting. The DNA probes are commonly complementary strands of an entire genome or a gene. Hybridization between a single stranded DNA of a sample and another single stranded DNA belonging to a probe, constitutes the cornerstone of the function of microarray (1-6).

The preference of microarray architecture is the possibility of analyzing a huge number of genes simultaneously in a single quick molecular test (1,2). Reliability, reproducibility and the quality of microarray data promotes the accessibility to accurate biological information; however, the interpretation and analysis of a huge amount of microarray data are accounted as a big challenge for scientists and molecular biology experts (1,2,5). In this mini-review article we provide general knowledge about the various aspects relating to the process of microarray data analyzing.

The process of microarray data analysis

The interdisciplinary technologies like microarrays are known as data tsunami producing techniques. In these cases, a variety of data are provided by machines and the computational analyses must be interpreted by a human being. This process is known as a hard-work procedure and is a big challenge (7,8).

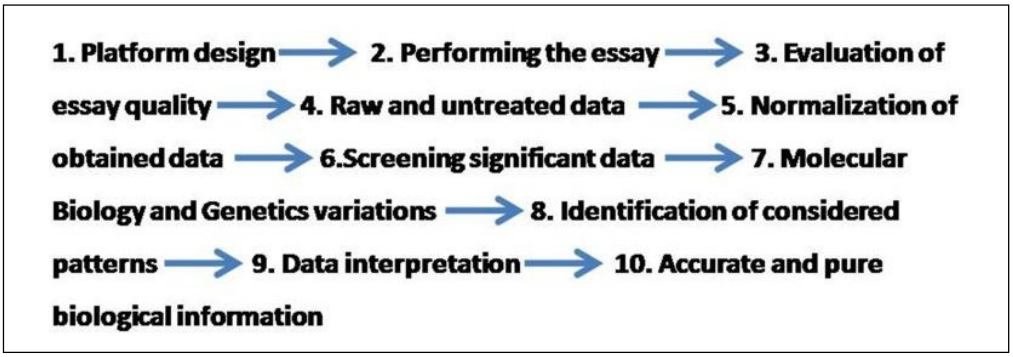
The data sets of microarrays include a huge raw information which must be understood via effective analysis methodologies. The quality of data analyses is considered as the main goal of researchers. Thus, the first step to have a qualified data analysis is designing a suitable experimental performance. Simultaneously, there are a large number of commercial and non-commercial software packages for guiding investigators to have a successful approach to standard microarray data analysis (3,6). Each data analysis must be screened via quality control, preprocessing of data, identification of genetic differentiations for converting into meaningful and acceptable biological information (9,10).

To obtain accurate data, there is an urgent need for appropriate and suitable experimental designing. Design issues are important parameters to have a good procedure design. Furthermore, management of data and selection of an accessible platform are other issues that have a direct contribution in the process of microarray data analysis (Figure 1) (2,6). Despite the presence of determining and progressive commercial microarray platforms, the conversion of raw data into meaningful biological information is still a typical challenge for the majority of investigators. According to different studies, replacing the automatic handling of microarray data analysis instead of manual performance may lead to provide high quality, proper and meaningful data interpretation in a facilitated manner; for this reason, the Microarray Gene Expression Database (MGED) was established to facilitate the interpretation of microarray data. The great secret in data interpretation is: looking for the input process of the caught output data (2,7-14).

The quality of microarray data analysis

The significant progressions and advances in data analysis methodologies and tools have led to difficulties in selecting a proper approach for researchers and specially for amateur users. There are different commercial microarray platforms made by Affymetrix, Agilent, Amersham Biosciences, Illumina, NimbleGen, Applied Biosystems Inc. (ABI), Xeotron and Febit which differ in experiment design, probe sets and protocols. Therefore, it is important to exploit the right microarray

Figure 1. The 10-stage cycle pertinent to the process of microarray data analysis



platform for correlated aims; otherwise it may lead to incorrect raw data and false interpretations (10,13,15,16).

The development of microarray technologies has accommodated, optimized and validated cross-platforms for maximizing the accuracy of interpretation of raw data. Therefore, the combination of suitable platforms and softwares in the format of a unite workflow may remarkably increase the realization of a successful and qualified data analysis (2,17,18).

Besides, the quality and accuracy of microarray data are depending on probes' identities, sensitivity, specificity, type of chips, and target labelings. Also, a proper filtering and calculating normalization factors are needed to coordinate variables relating to target labeling whether with Cy3 or Cy5; however the one-channel or one-color microarray is more accurate and useful to obtain a qualified microarray data analysis (2,10,14,19-21).

The design of microarray experiments has a direct effect on qualification of microarray data analysis. Relating to conditions, there are several statistical tests including t-test ([the Welch's test] [the Wilcoxon Rank Sum] for two comparable groups) and analysis of variance (ANOVA) ([one-way ANOVA/the Kruskal-Wallis test] [two-way ANOVA/Factorial ANOVA]) for determining the level of statistical significance (p-value) (2,22,23). Availability, facility, reliability, and the concordance of techniques used in microarrays, guarantee the high level quality of data analysis (2).

Management of microarray data analysis

Incorrect and unsuitable microarray data may lead to failed interpretation of data analysis. Therefore, a software is needed for checking the quality metrics of raw data throughout the quality parameters (2,5,24).

Microarray assays are able to generate and deliver huge amounts of raw data via employing platforms and techniques. Hence, in an individual microarray practice, hundreds of thousands of data points are produced that must be arranged, managed and interpreted (2,13).

A good management of microarray data is resulted from an appropriate applied design, effective softwares, and proper practical process in arrays. Simultaneously, the available public databases are important technical supporters for retrieving and interpreting microarrays raw data to have a standard microarray data analysis. The accessible internet supplements help researchers to analyze data sets in a facilitated manner. SOURCE (http://source.stanford.edu) is one of the most common databases which is used by investigators. Additionally, the microarray experiment checklist database or the Minimum Information About a Microarray Experiment (MIAME) and its supporter, the public genomics database of Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) belonging to National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov) are important available repositories which minimize the drawbacks of microarray raw data manage-

ment; however, there are several electronic databases like the European Bioinformatics Institute (EBI), which provide important means for microarray data management and analysis (2,7,10,19,21,25-29).

Databases relating to microarray research

The most famous and common databases applied for microarray research were mentioned above; however, there are several public databases which can be helpful for researchers. These over-internet-scattered public databases are mostly accessible for free. Searching them one by one, may be a time consuming and overwhelming workflow for individual investigators around the world. Therefore, we have selected a number of active and available databases for microarray investigation as follow:

1.ENS transcriptome Genomic Service (http://www.transcriptome.ens.fr/sgdb/)

The Biology Department Genomic Service (SGDB) is a French database which offers different tools, softwares, or platforms for free (19,30).

2. GenBank (http://www.ncbi.nlm.nih.gov/genbank)

GenBank is a large and extensive nucleotide sequence database which belongs to the National Institutes of Health (http://www.nih.gov/) and supports biological annotation. GenBank is covered by NCBI. The International Nucleotide Sequence Database Collaboration (INSDC) (http://www.insdc.org/). INSDC is a giant multistructural database which operates between NCBI, DNA Data Bank of Japan (DDB) (http://www.ddbj.nig.ac.jp/) and the European Molecular Biology Laboratory (EMBL) (https://www.ebi.ac.uk/) (31).

3.The Enterez Databases (http://www.ncbi.nlm.nih.gov/books/NBK3837/#EntrezHelp.The_Entrez_Databases).

The Entrez Databases are a huge multisystem repository which includes 40 databases of molecular and literature disciplines. The most important databases are as below:

BioSystems, dbGaP, EST, Gene, Genome, Genome Projects, GEO datasets, GEO Profiles, GSS, HomoloGene, NCBI Web Site Search, NLM Catalog, Nucleotide, OMIM, PopSet, Probe, Protein, Protein Clusters, and UniGene (32).

a) BioSystems (http://www.ncbi.nlm.nih.gov/biosystems/): The BioSystems database is an applicative pipeline, which include biological pathways of Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.jp/kegg) resource and the EcoCyc (Escherichia coli K-12 MG1655) (http://ecocyc.org/), a subdivision of BioCyc repository (2,32).

b) Database of Genotypes and Phenotypes (dbGaP) (http://www.ncbi.nlm.nih.gov/gap/): dbGaP is an important repository which accommodates a wide range of biological information pertaining to molecular diagnostic tests (31,32).

c) Expressed Sequence Tag (EST) (http://www.ncbi.nlm.nih.gov/nucest/): The EST database is consisted of sequences (from GenBank) which are invaluable for assessing gene expression and annotating (31,32).

d) Gene (http://www.ncbi.nlm.nih.gov/gene): Gene is a wide public database which offers an appropriate knowledge about genes (32).

e) Genome (http://www.ncbi.nlm.nih.gov/genome): Within different parts of NCBI, Genome pipeline covers genomic annotations relating to eukaryotes (Eukaryotic Genome Annotation (EGA)) (http://www.ncbi.nlm.nih.gov/genome/annotation_euk/), prokaryotes (Prokaryotic Genome Annotation (PGA)) including bacterial and archaea (http://www.ncbi.nlm.nih.gov/genome/annotation_prok/), and viruses (PASC (PAirwise Sequence Comparison) (http://www.ncbi.nlm.nih.gov/sutils/pasc/viridty.cgi?textpage=overview). Genome database is a subdivision of the Entrez databases (32).

f) Genome Project (http://www.ncbi.nlm.nih.gov/bioproject): Genome Project which is linked to BioProject is a proper repository for researchers who work in the field of cellular organisms (32).

g) Gene Expression Omnibus (GEO) datasets (http://www.ncbi.nlm.nih.gov/gds): The GEO datasets are comprised of different biological tools and information. These datasets encompass those data which are introduced in the GEO database (32).

h) GEO Profiles (http://www.ncbi.nlm.nih.gov/geo): The Gene Expression Omnibus database aids MIAME as aforementioned (2,32).

i) Genome Survey Sequence (GSS) (http://www.ncbi.nlm.nih.gov/nucgss/): GSS is a subdivision database derived from NCBI which is analogous to EST (32).

j) HomoloGene (http://www.ncbi.nlm.nih.gov/homologene): HomoloGene is a proper database which shares homolog groups of eukaryotic completed genes (2,32).

k) NCBI Web Site Search (http://www.ncbi.nlm.nih.gov/books/NBK3837/#EntrezHelp.The_Entrez_Databases): NCBI Web Site Search is a powerful search database which shows the online sources present on NCBI website.

l) National Library of Medicine (NLM) Catalog (http://www.ncbi.nlm.nih.gov/nlmcatalog/): NLM Catalog is a great collection of different resources such as softwares and other electronic repositories (32).

m) Nucleotide (http://www.ncbi.nlm.nih.gov/nuccore): the Nucleotide database is an effective and widespread repository which offers sequence data released from GenBank, DDB, EMBL (the INSDC partners). Therefore, biomedical investigators have special attention to the Nucleotide public database (32).

n) Online Mendelian Inheritance in Man (OMIM) (http://www.ncbi.nlm.nih.gov/omim): OMIM pipeline shares genomic data relating to human being. This useful database updates every day (32).

o) PopSet (http://www.ncbi.nlm.nih.gov/popset/): A unique dataset of DNA sequences which supports the evolutionary relationships between populations in microarray evolutionary genomics studies (32).

p) Probe (http://www.ncbi.nlm.nih.gov/probe/): Probe is a public database which is helpful for probe designing and determining the efficacy of designed probes.

q) Protein (http://www.ncbi.nlm.nih.gov/protein/): Protein database is an invaluable repository for protein sequences collected from different databases including GenBank, RefSeq, Swiss-Prot, Protein Research Foundation (PRF), Protein Data Bank (PDB), and the Protein Information Resource (PIR) (29,32).

r) Protein Clusters (http://www.ncbi.nlm.nih.gov/proteinclusters/): Protein Clusters pipeline shares Archaeal, Bacterial, Fungal, Herbal, Protozoan and Viral protein sequences produced by either prokaryotic genomes and plasmids or eukaryotic genomes and organelles. The Protein Clusters offers biological information and analysis tools (32).

s) UniGene (http://www.ncbi.nlm.nih.gov/unigene): UniGene database gives valuable information relating to gene or Pseudogene clusters. This topic is considerable in microarray assays (2,32).

4. FlyBase (http://flybase.org/)
The FlyBase database covers genes and genomes belonging to Drosophila, which can be useful in different aspects of studies; from a genetic experimental test on the insect to biological modeling information about human diseases (19,33).

5.Gene Ontology (GO) (http://geneontology.org/)
Gene Ontology is a multi-database consortium involving biological pathways, cellular compositions,

and molecular operations relating to genes, gene products, and their sequences (2,34).

Conclusion

Microarray technologies are going to overcome obstacles hindering data analysis and data interpretations. The improvement of softwares, protocols, and platforms promise significant refinement in microarray data analysis and data interpretations. On the other hand, microarray databases grow up increasingly as powerful and effective repositories for supporting data management. The public microarray databases provide us a vast field of data sets in a proper order which may lead to maximize the quality of data analysis and interpretation. Moreover, the progression of microarray application on global scale guarantees the reduction of software malfunctions and incorrect data analysis and interpretations. This condition provides the possibility of classification in association with databases, softwares, platforms, protocols, and tools. Having an appropriate classification of standards may lead to determine a logical approach. Furthemore, the rise of microarray technology applications may lead to a significant decrease of the costs in the future.

Conflicts of interest: None declared.

References

1. Selvaraj S, Natarajan J. Microarray data analysis and mining tools. Bioinformation 2011;6:95.
2. Olson NE. The microarray data analysis process: from raw data to biological significance. NeuroRx 2006;3:373-83.
3. Piatetsky-Shapiro G, Tamayo P. Microarray data mining: facing the challenges. ACM SIGKDD Explorations Newsletter 2003;5:1-5.
4. Behzadi P, Najafi A, Behzadi E, Ranjbar R. Detection and Identification of Clinical Pathogenic Fungi by DNA Microarray. Infectio.ro 2013;35:6-10.
5. Najafi A, Ram M, Ranjbar R. Microarray Principles & Applications. 1st ed. Tehran: Persian Science & Research Publisher; 2012.
6. Reimers M. Making informed choices about microarray data analysis. PLoS Comput Biol 2010;6:e1000786.
7. Karacapilidis N, Christodoulou S, Tzagarakis M, Tsiliki G, Pappis C, editors. Strengthening collaborative data analysis and decision making in web communities. Proceedings of the companion publication of the 23rd international conference on World wide web companion; 2014: International World Wide Web Conferences Steering Committee.
8. Agrawal D, Bernstein P, Bertino E, Davidson S, Dayal U, Franklin M, et al. Challenges and Opportunities with Big Data 2011-1. 2011.
9. Kapetis D, Clarelli F, Vitulli F, de Rosbo NK, Beretta O, Foti M, et al. AMDA 2.13: A major update for automated cross-platform microarray data analysis. BioTechniques 2012;53:33.
10. Hardiman G. Microarray platforms-comparisons and contrasts. Pharmacogenomics 2004;5:487-502.
11. Stafford P, Tak Y. Biological Interpretation for Microarray Normalization Selection. Methods in Microarray Normalization. 2012:151.
12. Guzzi PH, Di Martino MT, Tradigo G, Veltri P, Tassone P, Tagliaferri P, et al. Automatic summarisation and annotation of microarray data. Soft Computing 2011;15:1505-12.
13. Korenberg MJ. Microarray data analysis: Methods and applications: Springer; 2007.
14. Rudy J, Valafar F. Empirical comparison of cross-platform normalization methods for gene expression data. BMC Bioinformatics 2011;12:467.
15. Tan PK, Downey TJ, Spitznagel Jr EL, Xu P, Fu D, Dimitrov DS, et al. Evaluation of gene expression measurements from commercial microarray platforms. Nucleic Acids Res 2003;31:5676-84.
16. Servant N, Gravier E, Gestraud P, Laurent C, Paccard C, Biton A, et al. EMA-AR package for Easy Microarray data analysis. BMC Res Notes 2010;3:277.
17. Barnes M, Freudenberg J, Thompson S, Aronow B, Pavlidis P. Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. Nucleic Acids Res 2005;33:5914-23.
18. Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J. Independence and reproducibility across microarray platforms. Nat Methods 2005;2:337-44.
19. Moreau Y, Aerts S, Moor BD, Strooper BD, Dabrowski M. Comparison and meta-analysis of microarray data: from the bench to the computer desk. Trends Genet 2003;19:570-7.
20. Knight J. When the chips are down. Nature 2001;410:860-1.

21. Diehn M, Sherlock G, Binkley G, Jin H, Matese JC, Hernandez-Boussard T, et al. SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. Nucleic Acids Res 2003;31:219-23.

22. Knudsen S. A biologist's guide to analysis of DNA microarray data: John Wiley & Sons; 2011.

23. McDonald JH. Handbook of biological statistics: Sparky House Publishing Baltimore, MD; 2009.

24. Reimers M, Heilig M, Sommer WH. Gene discovery in neuropharmacological and behavioral studies using Affymetrix microarray data. Methods 2005;37:219-28.

25. Larsson O, Wennmalm K, Sandberg R. Comparative microarray analysis. OMICS 2006;10:381-97.

26. Quackenbush J. Computational analysis of microarray data. Nat Rev Genet 2001;2:418-27.

27. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 2005;21:3439-40.

28. Engreitz JM, Morgan AA, Dudley JT, Chen R, Thathoo R, Altman RB, et al. Content-based microarray search using differential expression profiles. BMC Bioinformatics 2010;11:603.

29. Behzadi P, Behzadi E, Ranjbar R. Basic Modern Molecular Biology. 1st ed. Tehran: Persian Science & Research Publisher; 2014.

30. ENS transcriptome platform web site [Internet]. Biology Department Genomic Service. 2012 [cited 2014]. Available from:http://www.transcriptome.ens.fr/ sgdb/ .

31. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Res 2012:gks1195.

32. Entrez Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2005-. 2006 Jan 20 [Updated 2014 Apr 9]. Available from: http:/ / www.ncbi.nlm.nih.gov/ books/ NBK3837/.

33. Pierre SES, Ponting L, Stefancsik R, McQuilton P. FlyBase 102—advanced approaches to interrogating FlyBase. Nucleic Acids Res 2014;42:D780-D8.

34. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet 2000;25:25-9.