Gene expression data analytics

Arsen Arakelyan
Institute of Molecular Biology
Yerevan, Armenia
aarakelyan@sci.am

Levon Aslanyan
Institute for Informatics and Automation Problems
Yerevan, Armenia
lasl@sci.am

Anna Boyajyan
Institute of Molecular Biology
Yerevan, Armenia
aboyajyan@sci.am

ABSTRACT
Nowadays, technology of the gene expression measurement raises new issues related to proper data analytics and accurate interpretation of results. Gene expression measurements are performed on random populations of cells; the physical measurement process is related to hybridization and so to the false positive and false negative expression components that are stochastic. The measured expression value of a condition dependent gene in normal state $G_n = B + N$, and under the studied condition $G_d = B + D + N$ may at times coincide due to the fact that a very large set of genes are very low expressed. In addition, two different conditions can bring two different genes to the same expression easily. Statistically, in a population, not only gene expressions but also the gene expression profiles (vectors) become randomly evaluated. We prove that there is a real information limitation to provide the knowledge extraction about the state and properties of cells and we refer again to [1] that initiates the additional use of the functional pathway framework.

Keywords
Gene expression, probability distribution, classification, feature selection

1. INTRODUCTION
Identification of key genes responsible for tissue/cell differentiation and disease development has been under intensive research during the recent decades. Nowadays, technology of the gene expression measurement has changed dramatically from single gene observations to massively parallel measurements for whole transcriptome profiles, such as microarray or RNA-sequencing experiments [1]. However, these approaches have raised new issues related to proper data analytics and accurate interpretation of results. To date, the main pipeline for gene expression data analysis includes identification of differentially expressed genes among different spatial or temporal states, followed by search of functional gene sets from the obtained gene list that can be fittingly interpreted. It is clear that the logic behind the mentioned strategy, although implemented in huge number of studies, has its shortcomings inherited from the nature of gene expression and measurement techniques.

Expression of the given gene in a cell can be defined as the number of mRNA transcripts of that gene present at a given time point (at the time of measurement). In fact, the expression of each gene in a cell is not constant but is represented by a temporal profile of mRNA counts, which depends on the rate of mRNA synthesis and degradation. These rates depend on complex factors regulating expression of the gene based on its essence for cell function. In general, gene expression measurements are performed on populations of cells. The true expression value of a given gene will thus represent a random variable sampled from the pooled distribution of cell-specific temporal gene profiles of the studied cell population. In reality mRNA transcript count cannot be directly quantified (even with RNA sequencing) and requires additional steps for sample preparation and detection which introduce noise in measurements. Thus, the measured gene expression value is a stochastic variable, derived from the summed distribution of pooled true gene expression profiles and noise distribution.

2. PROBABILISTIC MODEL OF GENE EXPRESSION MEASUREMENTS
Let's assume that the measured gene expression in normal state (G_n) is a random variable which can be represented as a sum of background gene expression in normal state (B) and a random variable sampled from noise distribution (N): $G_n = B + N$. If there is a departure from normal state (diseased or treatment-associated conditions) the measured gene expression is represented as $G_d = B + D + N$, where D is the distribution of departure values associated with the state, and is equal to 0 if the gene expression is not dependent on the state and is different from 0 otherwise. While attempts have been made to approximate the distribution of gene expression under normal conditions as well as noise distribution (Table 1), the statistical properties of condition specific gene expression remains unclear.

It can be assumed that the genes associated with any particular condition (condition dependent genes) comprise a very small proportion of the whole genome. In this case, in high-throughput experiments, the expression profiles of condition dependent genes across states are not easily distinguishable from the large numbers of condition independent genes, whose expression profiles, being independently distributed, may coincide with condition specific ones. On the other hand, measured expression values of a single condition dependent gene in normal state ($G_n = B + N$) and under the studied condition ($G_d = B + D + N$) may at times coincide due to the fact that the distributions of G_n and G_d may overlap. This may result in both false positive (genes that have expression ratio over the threshold but are not associated with the studied condition) and false negative (disease associated genes that have expression ratio lower the threshold) results. The most of available gene expression analysis algorithms are not able to deal with this problem and simply choose differentially expressed genes on the basis of thresholds and cutoffs. This may be a reason for unsatisfactory correlation between data obtained using different platforms [25].

Table 1. Sources and underlying distributions of randomization for microarrays experiments.

	Source	Distribution
Target measure	mRNA transcript count number in cell (synthesis and degradation) + variation of mRNA number in cell population	Negative binomial or Poisson [25]
Noise	Isolation + reverse transcription	Poisson distribution [26]
	PCR amplification	Poission distribution [29]
	Unspecific hybridization/Hybridization	Poisson [27]
	Image acquisition	Lorentzian or Gaussian [28]

3. THE INSTANT POWER OF GENE EXPRESSION DATA

3.1. Expression profiles and distribution

Discretization.

The nature of gene expression data depends on measurement techniques employed. Fluorescence measurement techniques such as PCR, quantitative PCR and microarrays produce continuous data, where as Serial measurement of gene expression (SAGE) and RNA-sequencing produce discrete count data [31]. However, even with continuous is "discretized", because the value is rounded to some precision. This can be done by a binning procedure like in [2].

The domain of Expression values.

Analysis of total mRNA transcript counts per cell showed individual cell contains 519,688 to 851,087 mRNAs 8,357 to 12,739 transcripts, expressed from 8,101 to 11,360 genes. The individual transcript levels vary in very broad range from 0.1 to 20,000 copies per human cell [20].The statistical distribution of gene expression values seems to depend on measurement techniques. Thus, SAGE measurement of gene expression produces a Pareto-like distribution model [4], two-color microarray follows Laplasian distribution [30], while RNA-seq can be approximated by Negative binomial distribution [25]. However, the common characteristic for all these distributions is severe skewness towards low-abundance transcripts. Empirical relative frequency distributions of the gene expression levels show the based domain of values are 1-100, and in each level there are still many genes expressed at that level. The maximum number of genes is expressed at very low levels ($\leq$ copy per cell).

Normal state expression.

Disease-specific genomic analysis (DSGA) employs analytics and comparison of condition specific to normal expression to extract data most closely associated with the disease. Specifically, DSGA defines a supervised step that mathematically transforms and simplifies expression data to highlight the pathologic component of expression. While retaining expression information about every gene, DSGA isolates and separates a disease-like and a normal-like portion of this expression. [3] exploits a supposition that the set of normal gene expressions is to be closed in a linear space. With the population increase this space covers the whole domain and then some models artificially try to optimize the real domain of normal expressions. [23], [24] show evidently, that the Logic Separation (LS) approach is more suitable for this. Considering two classes LS constructs edges (maximal subspaces) that contain element of one of these classes and do not contain elements of the other class. In binary case this brings us to the reduced disjunctive normal form of Boolean functions. Genome information belongs to this case. But gene expression data belongs to a multivalued grid so that edges constructed above the learning data are sub-grids. Having one class of normal expressions it is to construct convex closures by one or the set of directions/genes. Ideally it is to use the whole gene set but restrictions can help to lower computations and it is still to investigate if the information loss is acceptable. We use closure by the whole gene set, and the notation $G_n = B + N$, where B is the component in closure and N is the noise we mentioned above. Expressions of condition dependent genes we divide by $G_d = B + D + N$, and then B is the least square fit to the convex normal closure, N is the same noise and D represents the effective difference that corresponds to its condition dependent nature.

Gene expression profiles.

Gene expression profile is a column of expression data matrix discussed in point **Digitization**. Having the expression values digitized, and working under suppositions that:

- expressions are measured by the random sets of cells
- expression data contain a stochastic component that is comparable to the expression value of the majority of genes ($\leq$ copy per cell)
- in each expression level there are sensitively many genes as the whole genome length – the constant ratio of it
- conditions and condition dependent groups of genes appear independently and concurrently and the effective difference of expressions D become random

and applying the Chebyshev's inequality to this model we receive that with all sensitive gene profile there is a large number of the similar profile that corresponds to another condition or is a normal case of expression. We do not provide the detailed proof due to space limitation.

High Dimensional low sample size data and functional pathways.

This paper is influenced by [1] which considers the case of mathematical analysis of High dimensional low sample size type data, which is a common case in genomics. The whole framework called GSS-PSF (Growing Support Set – Pathway Signal Flow) consists of several parts. The core two are the growing support set algorithm that is similar to data mining technique, particularly to association rule mining technique, and secondly, the pathway signal flow model that complements the expression data to achieve the required knowledge. In GSS-PSF there are several spots of research required to complete the study. This paper provides the solution to one of them.

3.2 Solution

The possible solution for the problem stated above is to perform knowledge driven analysis of gene expression data. First of all it should be clearly acknowledged that mRNA levels at the point of measurement presume only probable increase of protein levels (probable, because gene expression does not account for efficiency of protein translation, posttranslational modifications as well as protein functional state and degradation) after some time delay. However, if we assume that gene expression level is comparable to functional protein levels, we can expect increased levels of protein for overexpressed genes. And if levels of several genes are increased at

the time t_0, their functional protein levels will be simultaneously increased after t_{delay}. In this case the crucial protein-protein interactions become a crucial issue. If the levels of one interaction partner are increased but the other is not, the signal flow will not be effectively amplified. Thus, a functional effect may be expected only if the levels of both partners are deregulated. The biological pathway is a generalization of protein-protein interaction and is a directed and spatially defined sequence of bio-molecular physical and regulatory interactions that represent information (or signal flow) propagations leading to functional realizations of biological processes. As such if the pathway with one overexpressed protein results in much lower intensity of realization of biological process, that pathway with whole branch of mildly expressed proteins.

REFERENCES

[1] A. Arakelyan, L. Aslanyan, A. Boyadjyan. Current advances and limitations and future prospects in high-throughput gene expression analysis, Sequence and Genome Analysis II – Bacteria, Viruses and Metabolic Pathways. ISBN: 978-1-480254-14-5, iConcept press, 22p.,2012.

[2] Paul Pavlidis, Christopher Tang, and William Stafford Noble, Classification of genes using probabilistic models of microarray expression profiles, BIOKDD, page 15-21, 2001.

[3] Monica Nicolau, Robert Tibshirani, Anne-LiseBørresen-Dale, Stefanie S. Jeffrey, Disease-specific genomic analysis: identifying the signature of pathologic biology, Bioinformatics, vol. 23, pp. 957-965, 2007.

[4] V. A. Kusnetsov, G. A. Knott, R. F. Bonner,General statistics of stochastic process of gene expression in Eukaryotic cells, Genetics Society of America, vol 162, pp. 1321-1332, 2002.

[5] Johan Paulsson, Models of stochastic gene expression, Physics of Life Reviews, 2 (2005), pp. 157–175.

[6] H. Vikalo, A. Hassibi‡, B. Hassibi, Probabilistic modeling and estimation of gene expression levels in microarrays,NSF grant no. CCR-0133818, Office of Naval Research grant no. N00014-02-1-0578, and Caltech's Lee Center for Advanced Networking, 4p.

[7] Arakelyan, A., Boyajyan, A., Sahakyan, H., Aslanyan, L., Ivanova, K., &Mitov, I. (2010). Growing support set systems inanalysis of high-throughput gene expression data. In New trends in classification and data mining (pp. 47-53).

[8] Draghici, S., Khatri, P., Tarca, A.L., Amin, K., Done, A., Voichita, C., Georgescu, C., & Romero R. (2007). A systems biology approach for pathway level analysis. Genome Research, 17, 1537-1545.

[9] Gadbury, G.L., Page, G.P., Edwards, J., Kayo, T., Prolla, T.A., Weindruch, R., Permana, P.A., Mountz, J.D., & Allison, D.A.(2004). Power and sample size estimation in high dimensional biology. Statistical Methods in Medical Research, 13,325-338.

[10] Gershon, D. (2002). Microarray technology: An array of opportunities. Nature, 416, 885-891.

[11] Hall, P., Marron, J. S., &Neeman, A. (2005). Geometric representation of high dimension, low sample size data. Journal ofthe Royal Statistical Society: Series B (Statistical Methodology), 67, 427-444.

[12] Klebanov, L., Gordon, A., Xiao, Y., Land, H., &Yakovlev, A. (2006). A permutation test motivated by microarray dataanalysis. Computational Statistics Data Analysis, 50, 3619-3628.

[13] Liu, Y., Hayes, D.N.N., Nobel, A., &Marron, J.S. (2008). Statistical Significance of Clustering for High-Dimension, Low-SampleSize Data. Journal of the American Statistical Association, 103, 1281-1293.

[14] Mayor, C., Brudno, M., Schwartz, J.R., Poliakov, A., Rubin, E.M., Frazer, K.A., Pachter, L.S., Dubchak, I. (2000). VISTA:visualizing global DNA sequence alignments o farbitrary length. Bioinformatics, 16, 1046-1047.

[15] Ruffalo, M., Koyut¨urk, M., Ray, S., &LaFramboise, T. (2012). Accurate estimation of short read mapping quality for next generation genome sequencing. Bioinformatics, 2012, 28(18), i349-i355.

[16] Schena, M., Shalon, D., Davis, R.W., & Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 270, 467-470.

[17] Schwartz, S., Oren, R., &Ast, G. (2011). Detection and removal of biases in the analysis of next-generation sequencingreads. PLoS ONE, 6(1), e16685.

[18] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., &Golub, T.R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proceedings of the National Academy of Sciences of the United States of America, 96, 2907–2912.

[19] Carter, M.G., Sharov, A.A, VanBuren, V., Dudekula, D.B., Carmack C.E,, Nelson C., Ko, M.S.ranscript copy number estimation using a mouse whole-genome oligonucleotide microarray. Genome Biol. 6(7):R61, 2005.

[20] Velculescu, V. E., Madden S. L., Zhang L., Lash A. E., Yu J. et al., Analysis of human transcriptomes, Nat. Genet. 23: 387–388, 1999.

[21] L. Aslanyan, H. Sahakyan, Chain Split and Computations in Practical Rule Mining, International Book Series, Information Science and Computing, Book 8, Classification, forecasting, Data Mining, pp. 132-135, 2009.

[22] A. Arakelyan, A. Boyajyan, L.Aslanyan, D. Muradyan, H. Sahakyan, Algorithmic analysis of functional pathways affected by typical and atypical antipsychotics, Computer Science and Information Technologies Conference, Yerevan, Sept. 28 – Oct. 2, pp. 361-363, 2009.

[23] Aslanyan L., Castellanos J., Logic based Pattern Recognition - Ontology content (1), International Journal "Information Technologies and Knowledge" (IJ ITK), Vol. 1/2007, ISSN 1313-0455.

[24] L. Aslanyan, V. Ryazanov Logic Based Pattern Recognition - Ontology Content (2), Information Theories and Applications, ISSN 1310-0513 2008, Volume 15, Number 4, pp. 314-318.

[25] Anders, S. and Huber W. Differential Expression Analysis for Sequence Count Data. Genome Biology 11:R106, 2010.

[26] Tan, P.K., Downey, T.J., Spitznagel. E.L.Jr, Xu. P, Fu. D., Dimitrov, D.S., Lempicki R.A., Raaka, B.M., Cam,M.C. Evaluation of gene expression measurements from commercial microarray platforms. Nucleic Acids Res. 31(19):5676-84, 2003.

[27] Tu, Y., Stolovitzky, G., Klein, U. Quantitative noise analysis for gene expression microarray experiments. Proc NatlAcadSci U S A. 99(22):14031-6, 2002.

[28] Brody, J.P., Williams, B.A., Wold, B,J,, Quake,S.R.Significance and statistical errors in the analysis of DNA microarray data.Proc NatlAcadSci U S A. 1;99(20):12975-8, 2002.

[29] Hassibi, A., Kakavand, H., Lee, T. H.A Stochastic Model and Simulation Algorithm for Polymerase Chain Reaction (PCR) Systems. In Workshop on Genomics Signal Processing and Statistics, 2004.

[30] Purdom, E., Holmes, S.P. Error distribution for gene expression data. Stat Appl Genet Mol Biol. 4:Article16, 2005.

[31] Costa, V.,Angelini, C., De Feis, I., Ciccodicola, A. Uncovering the Complexity of Transcriptomes with RNA-Seq. Journal of Biomedicine and Biotechnology, vol. 2010, Article ID 853916: 19 pages, 2010.