Proc. of the Second Intl. Conf. on Advances In Bio-Informatics, Bio-Technology and Environmental Engineering- ABBE 2014.
Copyright © Institute of Research Engineers and Doctors, USA .All rights reserved.
ISBN: 978-1-63248-053-8 doi: 10.15224/ 978-1-63248-053-8-02

Approximation of Microarray Gene Expression Profiles by the Stable Laws

Viacheslav Saenko, Yurij Saenko

***Abstract*— At the present time reliably established that probability density functions of gene expression of microarray experiments possess a number of universal properties. First of all these distributions have power asymptotic and secondly the shape of these distributions are inherent for all organisms and tissues. This fact led to appearance of a number works where authors are investigating various probability distributions for approximation of empirical distributions of gene expression. In the work the gene expression of various organisms are investigated which were obtained from microarrays of various manufactures. The probability density functions of the gene expression levels are approximated by the fractional stable distribution. The parameters of the fractional stable distributions were statistically estimated according experimental data. It is shown that for all investigated samples the experimental distributions have the power-law asymptotic. At the same time the fractional stable distribution is a good approximation for the microarray gene expression manufactured by Affymetrix and Illumina Corporation.**

***Keywords*— gene expression, microarray, stable law**

I. Introduction

The technology of hybridization DNA microarrays of high-density has opened possibility to study the expression of all known genes in a single experiment. Studying the dynamics of the gene expression is one of priority trends in modern biology and medicine as it allows understanding the pathological mechanisms at the cellular level. Gene expression is a complex of coordinated process which depends from large number external and internal factors [1]. This means that knowledge of theoretical distribution opens outlooks in development of models of gene expression dynamics. Therefore, the probability methods most suitable for description of such processes.

Currently doesn't exist a fixed opinion about the kind of probability distribution which describes the profiles of gene expression of microarray experiments.

Reliably established that empirical distributions are one-sided distributions, they have power-law asymptotic and character of these distributions don't changes for wide area of tissues and organisms from E. coli to H. sapiens [2]. Such universality suggests fundamental nature of processes which leads to observable distribution of gene expression. Analogues conclusions have been obtained in works other authors [3]–[7] where gene expression of various organisms is also investigated.

Viacheslav Saenko, Yurij Saenko
Laboratory of molecular and cell biology, Technological Research Institute S.P. Kapitsa, Ulyanovsk State University, Ulyanovsk, Russian Federation

Power-law asymptotic of an experimental distribution means that theoretical distribution must have the asymptotic of following form

$$p(x) \propto x^{-\alpha-1}, x \to \infty \qquad (1)$$

In the above work [2] the same distribution was applied for approximation of profiles of gene expression various organisms under consideration and was showed that the parameter α is varying within limits from 0.69 to 1.09. In another work [3] have investigated more than 40 tissues for 6 organisms, and for all samples the power-law distribution was obtained. In the article [5] was marked that the best approximation among Poisson distribution, exponential distribution, logarithmic distribution, power-law distribution, parettolike distributions and mixtures of logarithmic and exponential distributions gives discrete Paretto distribution $p(m) = (m+b)^{-\alpha-1}/z$, where α is varying within limits from 0.974 to 1.88.

However the distribution (1), which is named Zipf-Pareto distribution, isn't the only distribution with power-law asymptotic. In the paper [4] was obtained that if to make logarithmic transformation of raw expression data and then align and standardize them $\xi = (\log s - \mu)/\sigma^2$, then distribution of transformed data, is well described by log-normal distribution. Here μ is mathematical expectation and σ^2 is variance of random variable log s, s is raw value of gene expression. In another article [7] authors suggest to use double Pareto-log-normal distribution. Besides Pareto-log-normal distribution the authors in the work tested such distribution as Zipf-Pareto distribution, log-normal distribution, log-gamma distribution, log logistic distribution, right-side Pareto-distribution. As a result, in the paper, the authors conclude that the best results are obtained with the double Pareto-lognormal distribution.

As was noted above there are two facts which allow us to make an assumption about fractional stable nature of distribution of a gene expression. First of all the distribution of gene expression has power-law asymptotic $p(x) \propto x^{-\alpha-1}$. The fractional stable distribution (FSD) has exactly the same asymptotic. Secondly, the shape of the gene expression distribution is very similar to the shape of the FSD. Consequently, the next step is testing the hypothesis about fractional stable nature of gene expression distribution. There are more fundamental reasons which may lead to power-law distributions. The gene expression in a cell is coordinated process and large groups of genes may change their expression in dependence from expression of others genes. Most genes in a cell are grouped into special groups - signaling or metabolic pathways. At present are revealed more than 2100 signaling and metabolic pathways. If at some time moment particular

SEEK
DIGITAL LIBRARY

Proc. of the Second Intl. Conf. on Advances In Bio-Informatics, Bio-Technology and Environmental Engineering- ABBE 2014.
Copyright © Institute of Research Engineers and Doctors, USA .All rights reserved.
ISBN: 978-1-63248-053-8 doi: 10.15224/ 978-1-63248-053-8-02

gene activates and it begins to synthesize its mRNA then activates immediately a set of genes associated with this gene. As a result a concentration of a connected set of mRNA may sharply increase and as consequence the intensity of emission sharply grows. At the same time expression of another group of genes may be suppressed. Such variation of gene expression must lead to power-law distributions.

In present work the FSD [8] are used for approximation gene expression profiles. The FSD belong to the class of infinitely divisible distributions and, in addition, they are the limit distributions of sums of independent identically distributed random variables.

II. Methods

Data were obtained from microarrays of three manufacturers: Affymetrix, Agilent и Illumina. For microarrays of the Affymetric company were processed of gene expression the following organisms mammals (human and rat), bird (chiken), worms (C. elegans), plant (rice and Arabidopsis thaliana), insect (drosophila), bacterium (P. aeruginosa). For microarrays of the Agilent company were processed of gene expression the following organisms: mammals, fish, bird, plant, insect, bacterium and fungus. For microarrays on the Illumina company were processed three organisms: human and rat. All experimental data were obtained from free databases ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) and Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/).

We were interested of data which had not been exposed any primary processing (data from RAW files). The channels PM and AM were processed separately for microarrays of the Affymetrix company. The red and green channels were processed for the Agilent microarrays. In particular the following channels were processed: red median signal (rMedianSignal), green median signal (gMedianSignal), red mean signal (rMeanSignal), green mean signal (gMeanSignal). For microarrays of the Illumina company RAW data were processed. All expression which were processed weren't undergo any preliminary normalization or processing.

The process of processing looks as follows. Expression for the organism under consideration from processed channel is considered as sample of independent identical distributed random variable $Z_1, Z_2, \dots, Z_N$. The parameters $\alpha, \beta, \theta, \lambda$ of the FSD are estimated under this sample by algorithm which has been described in [9]. After that a sample of fractional stable random variables $Z(\alpha, \beta, \theta, \lambda)$ were simulated with estimated values of the parameters $\hat{\alpha}, \hat{\beta}, \hat{\theta}, \hat{\lambda}$. For simulation random variables $Z(\alpha, \beta, \theta, \lambda)$ the algorithm described in [9] was used. Next histogram was constructed. At the same time on sample $Z_1, Z_2, \dots, Z_N$ a histogram of gene expression levels was constructed. After this theoretical and empirical distributions were compared. In the case when these distributions differed insignificantly then for the χ^2 Pirson's criterion was applied for checking the hypothesis about coincidence of these two distributions.

III. Results

The results of approximation of gene expression profiles for microarrays of the Affymetrix company are shown on the Fig. 1. On the figure the probability density functions (PDFs) for various organisms are depicted. Diamonds and crosses correspond to PDFs of gene expression for PM and MM channels respectively. Solid line and dashed line correspond to the FSDs are calculated for estimated values of the parameters for PM and MM channels respectively. It is seen from the figure more satisfactory agreement is achieved for a gene expression of a human, a rat, a chicken and rice both for PM channel and for MM channel. For C. Elegans and P. aeruginosa a satisfactory agreement between theoretical and empirical distributions is achieved only for MM channel. However when testing the hypothesis of acceptance of two distributions the χ^2 criterion rejects the hypothesis about fractional stable nature distribution of gene expression for all processed organisms. For others results which depicted on Fig. 1 difference of empirical and theoretical distributions are clearly seen.

Nevertheless it should be noted what this difference may be consequences both of hardware restriction and imperfection of algorithms selection of point glow and their digitization during process of translating them from image to a data file. One evidence of the presence of hardware constraints may serve Fig. 2 (left panel). On this figure gene expression of human genome is depicted but at the same the empirical distribution has been plotted in all range of values. Here it should be noted what on the Fig. 1 PDFs are plotted not for all range of gene expression. It is seen from the Fig. 2 (left panel) at large values of expression $\gtrsim 10^4$ a power law dependence is broken and PDF rapidly goes to zero. Such effect is called an effect of truncation and may be consequence of the hardware restriction at large values of gene expression intensity.

Let consider now the results of processing microarrays of the Agilent company. In the RAW files four channels correspond to gene expressions results. These channels differ by color and by the method of calculation of gene expression. In technological process of these microarrays red and green dye are used and two method of calculation of gene expression value are also used. The first method consists in calculation of mean value of intensity obtained from all pixels a probe under investigation. The second method consists in choosing median value of intensity of gene expression at processing of all pixels of the probe. According to this here and after we will denote: gMeanSignal (rMeanSignal) is mean signa of green (red) channel; gMedianSignal (rMedianSignal) is median signal in green (red) channel. During the process of processing it was obtained what PDFs of median and mean signal from same color almost coincide with each other. It is clearly seen from the Fig. 2 (right panel) on which PDFs of gene expression are depicted for a genome of a rice (Oryza sativa). From the figure we can see that PDFs of gene expression for mean and median signals for both channels practically coincide with each other. Same conclusions were obtained for all processed experimental data. Therefore in this work we will be used only median signal from red and green channels.

SEEK DIGITAL LIBRARY

Proc. of the Second Intl. Conf. on Advances In Bio-Informatics, Bio-Technology and Environmental Engineering- ABBE 2014.
Copyright © Institute of Research Engineers and Doctors, USA .All rights reserved.
ISBN: 978-1-63248-053-8 doi: 10.15224/ 978-1-63248-053-8-02

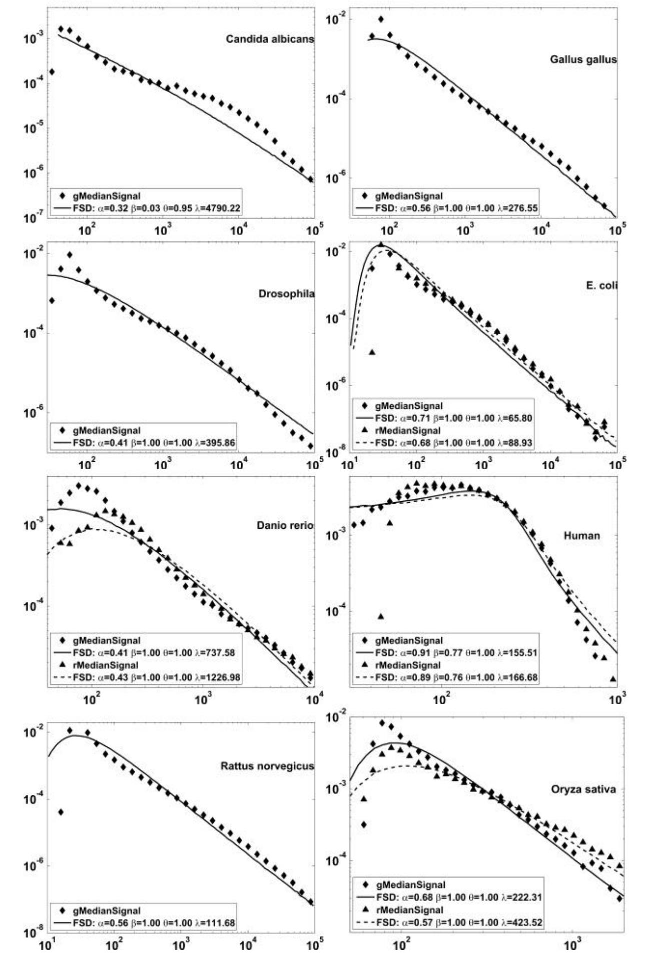

Figure 4. Distribution of gene expression of microarray experiments for various organisms obtained from the Agilent microarray chip. Names of organisms are showed on the pictures. Black points are empirical

For microarrays of the Agilent company were selected experimental data for mammals (Homo sapiens, Rattus norvegicus), bird (Gallus gallus), fish (Danio rerio), plant (Oryza sativa), insect (Drosophila melanogaster), fungus (Candida albicans) and bacterium (E. coli). The empirical PDFs for the median signal from red and green channels and PDFs of FSDs are shown on the Fig. 4. It is clearly seen that empirical PDFs aren't FSDs. Disagreement of empirical and theoretical distributions is very substantially. Nevertheless, let distinguish some properties which inherent to all the processed data. It is clearly seen that the asymptotic of the experimental PDFs haven't power law dependence $p(x) \propto x^{-\alpha-1}$. Most likely we can talk about dependence which closes to power-law behavior. Such behavior differs from the results obtained by using microarrays the Affymetrix manufacture (see Fig. 1). An existence of hardware distortions and distortions of algorithms of translating of intensity from an image file to numerical value can serve causes of deviation from the power-law dependence.

The PDFs of gene expression of human (Homo sapience) and rat (rattus norvegicus) for Illumina microarrays are shown on the Fig. 5. On the figures diamonds are experimental PDFs and solid line are FSD. It is seen from the figures the satisfactory agreement between experimental and theoretical PDFs is observed only for human genome. However usage χ^2 Pirson's criterion for checking correctness the hypothesis about fractional stable nature of the experimental distributions leads to necessity to reject this assumption. Nevertheless, it is seen from the figure that the FSD is good approximation for PDF of gene expression for human genome. For another genome is presented here the experimental distribution aren't belong to the class of FSDs. As well as in the previous case

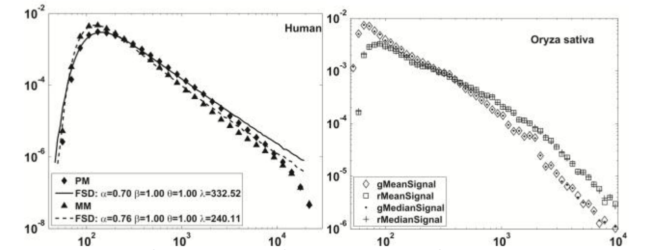

Figure 2. Left panel: PDF of gene expression for a human genome obtained by microarray of the Affymetrix company. Notation same as on the Fig. 1. Right panel: PDF of gene expression obtained by microarray of Agilent company. Diamonds (points) are mean (median) signal of green channel, squires (crosses) are mean (median) signal from red

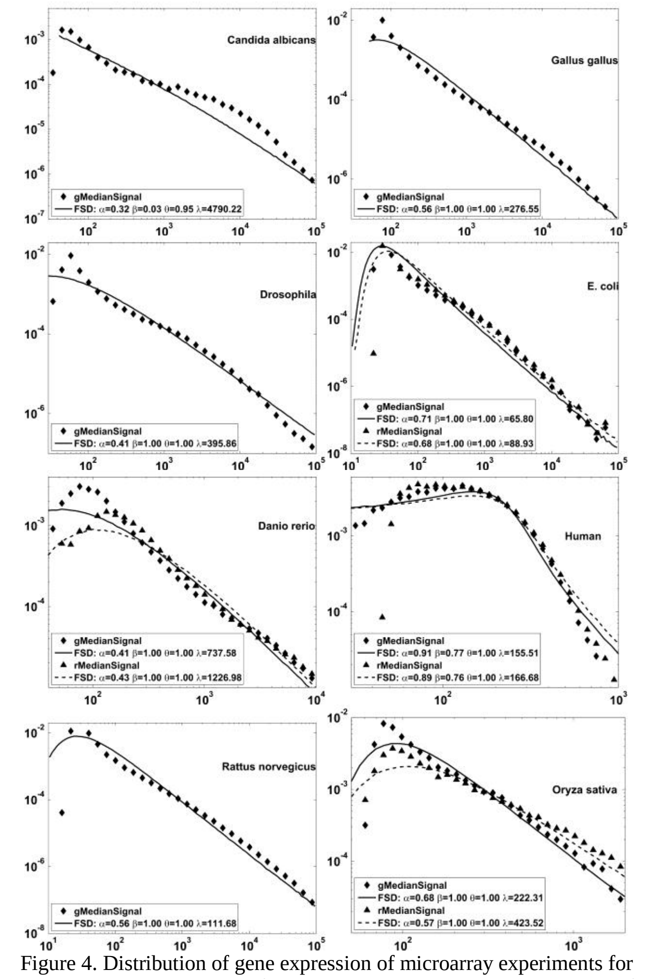

Figure 4. Distribution of gene expression of microarray experiments for various organisms obtained from the Agilent microarray chip. Names of organisms are showed on the pictures. Black points are empirical distribution and solid curve is fractional stable distribution. Parameters of FSD are showed on the figures.

SEEK DIGITAL LIBRARY

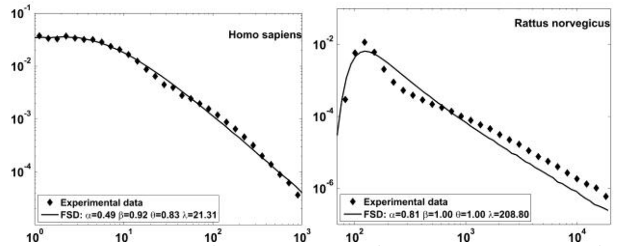

Figure 5. Distribution of gene expression of microarray experiments for various organisms obtained from the Illumina microarray chip. Names of organisms and the estimated parameters on the FSDs are showed on the pictures. Diamonds are empirical distribution and solid curve is fractional stable distribution. Parameters of FSD are showed on the figures.

the asymptotic of the experimental PDFs isn't described by power-law dependence. The power-law dependence is observed in mean but on this dependence some distortions are imposed.

IV. Discussion

An object of investigation was selected several organisms belonged to various classes: mammals, birds, fishes, plants, fungus, bacterium. Since in the present time there are many companies which produce microarrays, then the interesting question appears: how relate PDFs of gene expression between each other which have been obtained by microarrays of various companies? From fundamental understanding it is clear; since genes expression is proportional their concentration then law of distribution must be invariant towards manufacturer of microarray platform. In this work microarrays of three manufacturers (Affymetrix, Agilent и Illumina) were selected.

The results of comparison of theoretical end empirical densities are presented on the Figs. 1, 4, 5. As seen from the figures the law of distribution of gene expression for microarrays of different manufacturers is different. The PDFs of gene expression for microarrays the Affymetrix manufacture have clearly marked power-law asymptotic. However the effect of truncation is observed at large value of intensity of gene expression (see. Fig. 2) which breaks the power-law asymptotic. Clearly marked power-law asymptotic doesn't observes for PDFs of gene expression for microarrays the Agilent and Illumina manufacture (see Fig. 4 and 5). Is observed some decreasing which resembles the power-law dependence. Therefore for these data we can't talk about power-law asymptotic. By our opinion the differences in used algorithms of processing of initial data at their reading from microarray and subsequent translating there from image file to a numerical value are causes of divergence between the results of different platforms.

Approximation of PDFs of gene expression by FSDs has showed that the best agreement is achieved for gene expression of mammals and plants for microarrays the Affymetrix manufacture. However χ^2 criterion rejects hypotheses about coincidence of these two distributions. For PDFs of gene expression of microarrays the Agilent and Illumina manufacture the situation is absolutely different. There is clear difference between experimental distributions here and FSDs and in this case we can't talk about of coincidence of these distributions.

Nevertheless, the FSD good enough approximates empirical distribution both in the central part and in the tail part for gene expression of mammals and plants genomes. As we can see the values of the parameter α lie within interval $0.62 \leq \alpha \leq 0.83$. This values are in good agreement with results of works [2], [3], [5]. A value of second characteristic parameter of FSD β little differs from unit. This means that distribution of gene expression belongs to the class of stable laws which is a subclass of FSDs. As we can see the FSD good approximate empirical data of gene expression.

The fact that empirical distribution of gene expression is described by FSD allows making some assumption about character of background processes. The FSD is the limit distributions of sums of independent identically distributed random variables. Physical interpretation of this sum is a trajectory of particle undergoing a random walk which named Continuous Time Random Walk (CTRW) [10]. In the work [11] was shown that limit distribution of particle coordinate in framework of CTRW process is expressed through FSD. As consequence we can assume that the basis of the processes leading to the observed distribution of gene expression levels, are the processes described scheme CTRW.

On the other hand it is known that asymptotic behavior of CTRW process is described by generalized diffusion equation [10] expressed through fractional derivatives

$$\frac{\partial^\beta p(x,t)}{\partial t^\beta} = -D(-\Delta)^{\alpha/2} p(x,t) + \frac{t^{-\beta}}{\Gamma(1-\beta)}\delta(x) \qquad (2)$$

Here $\partial^\beta / \partial t^\beta$ is the Riemann-Liuville fractional derivative and $(-\Delta)^{\alpha/2}$ is the Laplace operator of fractional order [12], D is diffusion constant. Solution of this equation is expressed through FSD [11]

$$p(x,t) = \left(Dt^\beta\right)^{-1/\alpha} q\left(|x|\left(Dt^\beta\right)^{-\frac{1}{\alpha}}; \alpha, \beta, 0, 1\right)$$

where $q(x; \alpha, \beta, \theta, \lambda)$ is FSD. At the same time the parameters α and β simultaneously are exponents of fractional power of derivatives in the equation (2). Thus, from this facts, we can conclude, that processes leading to observed gene expression can be described by using equation in fractional derivatives. But the question about nature and main characteristics of these processes remains open.

Acknowledgment

The work was supported by Ministry of Education and Science of Russian Federation (grant No 6.1617.2014/K) and Russian Foundation or Basic Research (grant No. 12-01-00660)

References

[1] L. T. Macneil and A. J. M. Walhout, “Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression.,” *Genome Res.*, vol. 21, no. 5, pp. 645–57, May 2011.

Copyright © Institute of Research Engineers and Doctors, USA .All rights reserved.
ISBN: 978-1-63248-053-8 doi: 10.15224/ 978-1-63248-053-8-02

[2] H. R. Ueda, S. Hayashi, S. Matsuyama, T. Yomo, S. Hashimoto, S. A. Kay, J. B. Hogenesch, and M. Iino, “Universality and flexibility in gene expression from bacteria to human.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 11, pp. 3765–9, Mar. 2004.

[3] C. Furusawa and K. Kaneko, “Zipf’s Law in Gene Expression,” *Phys. Rev. Lett.*, vol. 90, no. 8, pp. 8–11, Feb. 2003.

[4] D. C. Hoyle, M. Rattray, R. Jupp, and A. Brass, “Making sense of microarray data distributions.,” *Bioinformatics*, vol. 18, no. 4, pp. 576–84, Apr. 2002.

[5] V. A. Kuznetsov, G. D. Knott, and R. F. Bonner, “General statistics of stochastic process of gene expression in eukaryotic cells.,” *Genetics*, vol. 161, no. 3, pp. 1321–1332, Jul. 2002.

[6] L. S. Liebovitch, V. K. Jirsa, and L. A. Shehadeh, “Structure of genetic regulatory networks: evidence for scale free networks,” in *Complexus Mundi - Emergent Patterns in Nature*, 2006, pp. 1–8.

[7] C. Lu and R. D. King, “An investigation into the population abundance distribution of mRNAs, proteins, and metabolites in biological systems.,” *Bioinformatics*, vol. 25, no. 16, pp. 2020–7, Aug. 2009.

[8] V. E. Bening, V. Y. Korolev, T. A. Sukhorukova, G. G. Gusarov, V. V. Saenko, V. V. Uchaikin, and V. N. Kolokoltsov, “Fractionally stable distributions,” in *Stochastic Models of Structural Plasma Turbulence*, V. Y. Korolev and N. N. Skvortsova, Eds. Utrecht: Brill Academic Publishers, 2006, pp. 175–244.

[9] V. Saenko and Y. Saenko, “Application of the fractional stable distributions for approximation of gene expression profiles,” no. arXiv:1406.7114 [math.ST], pp. 1–7, Jun. 2014.

[10] R. Metzler and J. Klafter, “The random walk’s guide to anomalous diffusion: a fractional dynamics approach,” *Phys. Rep.*, vol. 339, no. 1, pp. 1–77, Dec. 2000.

[11] V. V. Uchaikin, “Montroll–Weiss problem, fractional equations, and stable distributions,” *Int. J. Theor. Phys.*, vol. 39, no. 8, pp. 2087–2105, 2000.

[12] S. G. Samko, A. A. Kilbas, and O. I. Marichev, *Fractional Integrals and Derivatives -Theory and Application*. New York: Gordon and Breach, 1973.

SEEK
DIGITAL LIBRARY