Pareto-Gamma Statistic Reveals Global Rescaling in Transcriptomes of Low and High Aggressive Breast Cancer Phenotypes

Alvin L.-S. Chua, Anna V. Ivshina, and Vladimir A. Kuznetsov

Genome Institute of Singapore, 60 Biopolis Str. 02-01,
138672, Singapore
kuznetsov@gis.a-star.edu.sg

Abstract. We propose a novel mixture probability model for the probability distribution function (PDF) of microarray signals, which comprises a noise and a signal component. The noise term, due to non-specific mRNA hybridization, is given by a lognormal distribution; and the true signal, from specific mRNA hybridization, is described by the generalized Pareto-gamma (GPG) function. The model, applied to expression data of 251 human breast cancer tumors on the Affymetrix microarray platform, yields accurate fits for all tumor samples. We observe that (i) high aggressive cancers have, in general, broader right tails in the GPG than low aggressive cancers; (ii) the exponent parameter value of the GPG distribution is not constant and correlates strongly with ~4000 expressed genes and several "gold standard" clinical risk factors. These results can not be obtained from so-called "scale-free network" models. We conclude that an accurate parameterization of scale-dependent GPG function could provide robust prognostic benefits for cancer patients.

1 Introduction

Determining all biologically significant expressed genes from the transcriptome of different cell types, in particular, clinical tumor types presents substantial biologically and technologically challenges. Oligonucleotide microarrays (e.g. Affymetrix U133 A GeneChips) can be used to simultaneously detect the gene expression levels in RNA samples of about 22000 known human mRNA transcripts, allowing us to study the cell at the level of the whole human transcriptome. A typical statistical analysis of microarray data focuses on gene by gene comparisons among different classes (cell types, states of cells etc) and on identification of the list of genes (prognostic signature) differentially expressed in the classes. There are many such signatures in datasets, and its remains unclear which are really important in prognosis of a disease [13,15]. Fewer works focus on integrative or systemic signatures of gene expression profiles through the whole range of signal intensity values. To make these aim more real, it is important to separate the signal from the noise components [4,8,10].

In this study, we develop a novel statistical model of the gene expression level probability function (GELPF) for microarray transcriptomes, which allows us to separate the noise and fit the underlying distributions of the hybridization signal

J.C. Rajapakse, L. Wong, and R. Acharya (Eds.): PRIB 2006, LNBI 4146, pp. 49 – 59, 2006.

intensity values over the whole dynamical range. We proceed to show, for the first time, that clinical and histopathological variants of human breast tumors reproducibly correlate with "global genetic signatures" and can therefore be systematically categorized by certain parameter values of the GELPF. Specifically, we present a novel global pattern selection algorithm based on concept of scale-dependent attribute of exponent parameter of GPG distribution (opposite to claims of the "scale-free" and Zipf statistics [2,7,17]). This concept, in conjunction with gene ontology analysis, reveals fundamental differences in cell proliferation, regulatory, cell adhesion and other processes that can be used to categorize low and high aggressive phenotypes and genotypes of breast cancer. Finally, we demonstrate that these differences manifest as a global effect, affecting around 4000 transcripts, most of which are derived from regulatory genes.

2 Breast Cancer Data Set

We analyze the expression data of 22215 gene probe sets on Affymetrix U133A in 254 primary breast tumors. These breast tissue samples were derived from Uppsala County, Sweden. The breast cancer dataset comprised 70 low aggressive (grade 1), 126 moderate aggressive (grade 2), and 55 high aggressive (grade 3) tumors by the Elston-Ellis classification; 3 samples with normal breast tissue were also present in this collection. Original microarray data set and clinical information was kindly provided by Lance Miller (Genome Institute of Singapore, Singapore; see also NCBI GEO data set GSE3494). The summarization procedure used was that suggested by Affymetrix, the core of which is the one step Tukey bi-weight algorithm. We work with the data on the linear scale and do not perform any further normalization in any of our data analysis.

3 Parametric Model

3.1 Basic Functions

A simple power law given by the standard Pareto distribution [14] can be used as the GELPF to describe the empirical distribution function (EDF) of oligonucleotide microarray data, thus reflecting the distribution of original RNA sample. However, this model is only accurate around the tails of the EDF, where signal intensities are high, and therefore cannot be used to study the distribution in the low and medium intensity regions.

The lognormal probability distribution function is another popular choice for the GELPF [3,7]. This model gives betters fits for microarray data than a simple power law particularly for low and moderate signal intensity values, however it fails to fit the tails of the EDF accurately [3].

Here, we assume that the GEPLF is significantly affected by different sources of noise process and explicitly take in to account in our probabilistic model the additive and multiplicative noises. We model the observed EDF of the hybridization signal by choosing a mixture probability distribution function [6], which is a weighted sum of

two terms; the first due to non-specific hybridization and the second due to specific hybridization of mRNA molecules affected by multiplicative noise:

$$p(x) = \alpha p_s(x) + (1-\alpha)\, p_n(x)\,, \tag{1}$$

where $p_s(x)$ and $p_n(x)$ represent the probability distributions for specific and non-specific hybridization signals respectively and $\alpha \in (0,1]$ is the probability of state s (background noise). The mixing parameter α is treated as an extra parameter to be fitted. In microarray studies, the specific cases of (1) have been considered [4,10].

Here, we observe that low intensity hybridization signals for different samples vary significantly, even if they originate from the same type of tissue or cell line. For this reason, we assume that the EDF attributed to the non-specific hybridization dominates in at low signal intensity values. We describe this part of the EDF using a lognormal distribution, given by

$$p(x;q,m,s) = \frac{1}{\sqrt{2\pi}s(x-q)} \exp\left[-\frac{(\ln(x-q)-m)^2}{2s^2}\right], \tag{2}$$

where the shape parameter is $s \in \mathrm{R}^+$, the shift parameter is $q \in \mathrm{R}$ and the scale parameter is $m \in \mathrm{R}$. The domain of the distribution is $x \in (q,\infty)$.

To model the specific hybridization process, we observed statistics of genes expression level from microarray data as well as other systems like SAGE. The SAGE platform provides a quantitative estimate for the number of transcripts of an expressed gene and thus, given a sufficiently large SAGE library we can, by appropriately filtering the noise, observe real transcription levels. For SAGE transcriptome data, the EDF is well fitted using the generalized discrete Pareto (GDP) distribution as the GELPF [8-10]. As such, we assume that the continuous analog of the GDP function [9] can be used to model and analyze microarray transcriptome data. It is described by the function

$$p(x;k,a,b) = k\,\frac{(a+b)^k}{(a+x)^{k+1}}\,, \tag{3}$$

where the $k \in \mathrm{R}^+$ is the exponent, the scale parameter is $a \in \mathrm{R}$, the minimum value is given by $b \in (-a,\infty)$ and the domain of the distribution is $x \in (b,\infty)$ [6,9]. The global property of the distribution is characterized by the exponent parameter k, which describes a power law tail. A smaller value for the exponent represents a distribution with broader tails. Also, the cumulative distribution function is given by

$$f(x;k,a,b) = 1 - \left(\frac{b+a}{x+a}\right)^k\,, \tag{4}$$

For high intensity signals the GEPLF follows a weak power law $p(x) \sim x^{-(k+1)}$. However, consider the region where signal intensities are low. Here, the GEPLF is significantly affected by the multiplicative noise process, where the overall effect is to smear out the generalized Pareto probability (GPP) function. This can be represented by a convolution of GPP function with a smearing function. The exact form of the

smearing function is not crucial, as long as we have sufficient control over the shape and scale of the distribution. For this, we have chosen the gamma (Γ) probability distribution, for which we are able to obtain an analytical form for the convolution with the GPP function. This affords us a significant speed up in the fitting algorithm. It is important to note that despite this smearing effect, the final form for the GELPF after convolution should retain the same exponent as the Pareto distribution in the tails (see below). The smearing distribution is given by

$$p(x;\beta,\gamma)=\frac{1}{\beta\Gamma(\gamma+1)}\left(\frac{x}{\beta}\right)^{\gamma}\exp(-x/\beta), \tag{5}$$

where $\Gamma(x)$ is the gamma function, $\gamma\in[-1,\infty)$ is the shape parameter, and $\beta\in(0,\infty)$ is the scale parameter. The domain of the distribution is $x\in(0,\infty)$. In particular, we choose ad hoc to convolve the Γ distribution with γ set to unity with the Pareto distribution. As such, the smearing effects are determined by the shape parameter β alone. We will validate this choice by evaluating the final results of the fit.

For the convolution, we extend and take into account the domains of each of the two distributions carefully. The resulting generalized Pareto-gamma (GPG) distribution is given by

$$\begin{aligned}p(x;k,a,b,\beta)=&\frac{k}{\beta}\left(-\frac{a+b}{\beta}\right)\exp(-\frac{x+a}{\beta})[q\Gamma\left(-k,-\frac{a+b}{\beta},\frac{x+a}{\beta}\right)+\\&\Gamma\left(1-k,-\frac{a+b}{\beta},\frac{x+a}{\beta}\right)]\theta(x-b)\end{aligned} \tag{6}$$

and the cumulative distribution function by

$$\begin{aligned}f(x;k,a,b,\beta)=&\left[1-\left(\frac{a+b}{a+x}\right)^{k}-k\left(-\frac{a+b}{\beta}\right)^{k}\left(\exp\left(-\frac{a+b}{\beta}\right)\Gamma\left(1-k,-\frac{a+b}{\beta},-\frac{\mathrm{x}+\mathrm{a}}{\beta}\right)\right.\right.\\&\left.\left.+\Gamma\left(q,\frac{\mathrm{x}+\mathrm{a}}{\beta}\right)\Gamma\left(-k,-\frac{a+b}{\beta},-\frac{\mathrm{x}+\mathrm{a}}{\beta}\right)\right)\right]\theta(x-b),\end{aligned} \tag{7}$$

where $\Gamma(a,x)$ is the incomplete gamma function [6] and $\Gamma(a,x,y)=\Gamma(a,x)-\Gamma(a,y)$. The ranges for each of the parameters for the GPG distribution is the same as that of the original distributions, and the domain is $x\in(b,\infty)$.

Notice that described above probabilistic model has not been considered in the statistical literature.

3.2 Fitting Procedure

The fitting problem can be cast as a minimization problem where the best fit criterion is regarded as the set of parameters that minimizes a residue function, which compares the GELPF to the EDF. The lognormal probability distribution that describes the noise has three parameters: the scale parameter m, shape parameter *s* and shift *q*. Also, the GPG distribution is described by four other parameters: the exponent *k*, the scale parameter *a*, the lower bound cutoff *b* and the smearing

parameter β. The final parameter that we use for our fits is the mixing parameter α, as described by equation (1). Studying the data, we find that the shift parameter q for the lognormal distribution is small and as such, for our data set we set this parameter to zero.

It is also difficult to treat the mid signal region, where the lognormal distribution decays rapidly and the GPG distribution starts to set in, as there are several different conformations which give a reasonable fit. However, we find that the lower cutoff to the GPG distribution depends crucially on the location of the start of the decay. In our simulations we choose ad hoc to fix this parameter at the mode of the lognormal distribution given by $exp(m\text{-}s^2)$. Again, our final results validate this choice.

For the fitting procedure, we apply the Anderson-Darling statistic [1] where fits are made comparing the GELPF with the EDF. To construct the EDF first consider the sequence of the signal intensities in ascending order, represented by s_i. That is, the sequence has the property $s_i \geq s_j$ whenever $i \geq j$. Also, let us extend the sequence by defining s_0 to be zero and s_{N+1} the infinitely large limit, where N represents the number of probes on the microarray. The EDF [16] for the data set is then defined as

$$F(x) = \frac{i}{|E|} I_{[s_i, s_{i+1})}(x), \tag{8}$$

where $i \in \{0,\ldots,N\}$ and $I_{[a,b)}$ is an indicator function which is unity is the domain $x \in [a,b)$ and zero otherwise.

We expect residues to be a non-linear function in parameter space. There are several local minima which corresponds the possible conformations of fitting the two curves. Under these circumstances a Monte Carlo technique is appropriate for fitting the probably distribution with a residue function given by the Anderson-Darling statistic

$$A^2 = -N - \sum_i^N \frac{2i-1}{N} \left[\ln(f(S_i)) + \ln(1 - f(S_{N+1-i}))\right], \tag{9}$$

The crucial parameter in the generalized Pareto distribution is k, the exponent. This parameter characterizes the distribution over a few orders of magnitude. As such, a small change in the exponent alters the distribution throughout and therefore careful consideration must be given to the tail of the function. It is for this reason that we choose the Anderson-Darling statistic. The Anderson-Darling statistic weights tails favorably and allows us to fit the distribution in a robust way, giving us reliable estimates for the modeled GELPF.

The Metropolis algorithm [12] searches parameter space, making decisions based on differences in the Anderson-Darling statistic ΔA^2, due to a change in the parameters. The change is accepted when $\Delta A^2 \leq 0$ and with probability $exp(-\beta_B \triangle A^2)$ for any $\Delta A^2 > 0$. Here, the parameter β_B represents the inverse of the annealing temperature. We perform the search by doing a walk in parametric space starting at low positive value for β_B and systematically annealing the system by increasing β_B. The changes in the parameters are drawn from a Gaussian distribution with zero mean, and the variance is chosen with hindsight, base on how sensitive each parameter is with respect to the statistic A^2.

4 Results

4.1 Fitting Results

The overall results of the fitting procedure are good, thus validating the model that we propose. Figure 1 shows a typical fit of the GELPF to the EDF, and similarly Figure 2 shows the empirical histogram and decomposed best-fit density function for the same array. Note that our mixture model fits significantly better (by F-test) the empirical histograms of gene expression level than the log-normal probability function (data not presented).

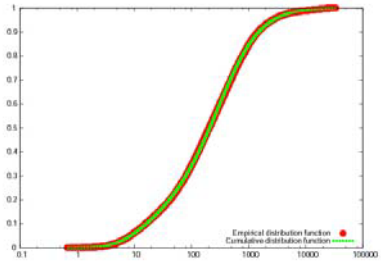

Fig. 1. The cumulative distribution functions of hybridization signal intensity values fits well to the empirical cumulative distribution function of representative microarray data

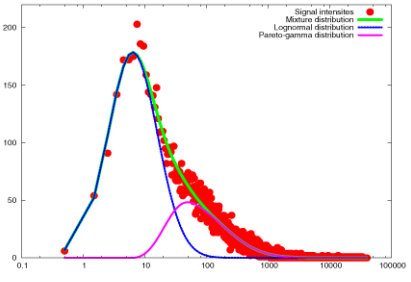

Fig. 2. The empirical density function and decomposition of the mixture best-fit density function (1). Binning size of the signal values is one unit.

We calculate the Pearson correlation coefficient between each of the parameters over all patients and observe that there is high correlation between scale parameter a and the exponent k of the Pareto distribution, with a correlation coefficient of 0.94. This is because both k and a play an important part in determining the shape of the distribution in the middle signal intensity region, where a large proportion of the signal is concentrated. As such, small changes in either parameter affect the Anderson-Darling statistic dramatically. We also see that there is a high correlation between the scale parameter of the lognormal distribution m, and that from the GPG distribution σ, with a correlation coefficient of 0.89. This implies that the source of noise that gives rise to both the lognormal distribution and the smearing out of the GPG distribution is the same.

4.2 Correlation with Clinical Risk Factors

To study the bearing that GELPFs have on clinical risk factors, we used a popular distribution-free Kendall τ correlation coefficient to estimate the relationships of the parameters *k* and *a*, taken from the GPG distribution, with histological and "gold standard" pathobiological and prognostic markers of breast cancer available for our patients. We found that the exponent *k* anti-correlates highly with cell cycle data (Mib-1 Ab representing Ki-67 protein (τ=-0.280, $p<10^{-5}$)), as well as other conventional clinical risk factors including p53 mutation status (τ =-0.231, $p<10^{-5}$) and angiogenesis (vascular growth index (τ =-0.295, $p<10^{-5}$)). Moreover, parameters of the GPG decrease significantly with an increases in histological tumor grade signature (τ =-0.296, $p<10^{-5}$). It is interesting to note that we get only low correlations for the estrogen receptor index (τ=-0.071), progesterone receptor index (τ =-0.104) and tumor size (τ =-0.192).

Notice that the correlations profile of the exponent *k* and shift *a* parameters of the GPG distribution exhibit similar correlation profiles to histological grade and the clinical risk factors mentioned above. As such we can perhaps consider the parameters of the GPG distribution a novel prognostic index of *aggressiveness* and *proliferate activity* of the breast cancer .

4.3 Differences Between Aggressiveness of Tumors

In breast cancer, histological grade of tumor is an important integrative parameter of proliferate activity of the tumor which is used for classifying tumors into morphological and clinical subtypes and assessing patient risk. Grading seeks to integrate measurements of cellular differentiation and proliferate potential into a composite score that quantifies the aggressive behavior of the tumor.

We plotted the empirical cumulative distribution functions of estimated exponent for histologic grades 1, 2, and 3 of breast cancer patients. We found that there is a systematic shift from an average of *k*-value from 1.59 for grade 1 tumors to 1.48 for grade 3 tumors ($p<0.000001$ for medians by U-test). The decrease in the exponent indicates a broadening of the curve at the tail. This is consistent with the evidence that more aggressive cancers are due to a severe malfunctioning of the cell, where transcripts are produced in a more indiscriminate manner.

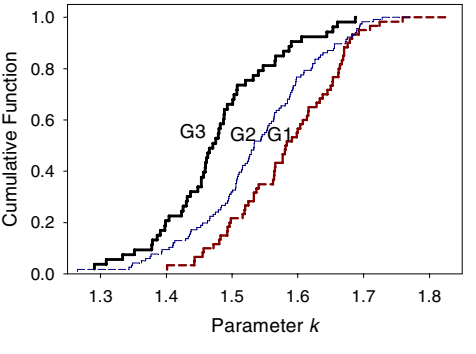

Fig. 3. The empirical distribution functions of the power law exponent for patients with tumor histologic grades 1(G1), 2(G2) and 3(G3)

4.4 Permutation Analysis

To further validate our claim that there is a significant change in the distribution affecting a large number of genes, we perform a non-parametric Kendall τ correlation test of each gene expression level with each of the fitted parameters. A straight forward test yields that a large number of probes correlate with all parameters. This is clearly not satisfactory, as we obtain a large number of false positives. Instead, we can search for systematic correlations only due to differences between grade 1 and grade 3 cancers.

Before applying the following test, we first use a BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat) procedure to search for probesets that do not match any human genes, as well as probsets that match to multiple target in genes, mRNAs and ESTs. These probes are removed from our list and thereby reducing the number from 22215 to 20873.

Consider the joint group with grade 1 and grade 3 patients. We permute all patients within the grade 1 and grade 3 subgroups, and calculate Kendall τ correlation values over the joint set. This procedure is applied using 500 iterations and the final Kendall τ coefficient considered is the average taken over all iterations. The permutations are designed destroy all correlations within each subgroup and only preserve correlations due to systematic differences between grade 1 and grade 3 tumors. The permutation test, over 500 iterations, is also used to distinguish a two tail $p<0.05$ for each parameter. Although in this process we may miss a large number of true positives, we believe it is a stringent test and will reflect important changes between the two cancer types. Table 1 shows results of the permutation test.

Table 1. Number of differentially expressed genes (G1 vs G3) as selected by the Kendall τ permutation test with respect to each parameter. Significant correlations with $|\tau|<0.12528$, $p<0.05$.

Kendal tau	m	s	k	a	beta	alpha
# original sign. correlations.	7652	7899	9958	8183	8775	4326
# sign. corr. after permutations	0	0	3934	3462	0	0

Of the 3934 probes that were selected with respect to the exponent parameter *k*, 1313 probes have been down regulated and 2621 probes up regulated in the grade 3 group compared to the grade 1 group. This finding is consistent with having broader tails in the GELPFs of grade 3 tumors. Both the parameters *k* and *a* yield significant correlations at $p<0.05$ as shown in table 1. None of the probes are significantly correlated with all other parameters at this p-value.

4.5 Gene Ontology (GO) Analysis

We run the set of probes that are selected with respect to the exponent *k* through the Panther classification system (http://www.pantherdb.org/). The top four categories listed by biological processes are cell cycle ($p<4\times10^{-17}$), cell cycle control ($p<10^{-11}$), nucleoside, nucleotide and nucleic acid metabolism ($p<2\times10^{-9}$), and cell proliferation

and differentiation ($p<2\times10^{-7}$). Similarly the first categories that appear by pathways are p53 pathway ($p<5\times10^{-4}$), ubiquitin proteasome pathway ($p<6\times10^{-4}$), p53 pathway feed back loops 2 ($p<10^{-3}$) and integrin signally pathway ($p<2\times10^{-3}$).

5 Discussion

It has been demonstrated that data analysis of gene expression measures based on simple statistical models of gene expression level distributions in the transcriptome can provide great improvements over the ad hoc procedures offered by large-scale gene expression methods (SAGE, microarrays, etc.) [4,8,11]. Our paper shows that global models can also be used to select a large statistically and biologically significant gene sets which is much bigger than the number of genes in prognostic "genetic signatures" (from few genes to few hundreds) [15]. In this study, we demonstrated that the proportion of genes which are differentially expressed among detected genes discriminated grade I and grade III breast cancer subtypes might be very large (~18% (3934/20873)).

Notice that the results based on popular "gene signature" approach published by different groups for the same disease are frequently non-concordant [13,15]. Perhaps a large proportion of the problems comes from different noisy processes leading to lack the concordant of the "gene signature" sets.

Several authors have claimed that the statistical distributions of gene expression levels are "scale-free" distributed [2,17]. A scale-free statistic is a specific kind of statistic of events which are distributed by the standard Pareto (SP) probability function [14] with a constant exponent parameter. The particular case when the exponent is unity is known as Zipf's law [6,11]. Based on the analysis of available raw SAGE [2] and microarray data sets [17] the authors claimed that Zipf's law is a universal statistical law of gene expression processes, and holds across all cell types, stages of differentiation and across the species. Due to different cell organisms, developmental processes and levels of biological complexity, it is difficult to trust to such a claim. From a statistical standpoint, an estimate of the parameters of the observed GELPF is sensitive to noise, sample size, binning procedure, and fitting procedure. Notice that for microarray data, a simple visual inspection of the EDF on the log-log scale shows essential deviations from a strict scale-free statistic in the low and medium signal intensity regions. For SAGE data, the GELPF is modeled well by the generalized discrete Pareto function [8-10].

Other authors favor the lognormal distribution function as a model for GELPF [3,7]. This model give reasonable fits for microarray data than a standard power law particularly for low and moderate signal intensity values; however it fails to fit the tails accurately [3,9].

In this study, used a novel mixture probabilistic model. We show that in the low intensity region of microarray GELPF is explicitly given by the lognormal distribution as well as the spread parameter β in the GPG distribution, in the mid intensity region, the term that dominates is the characteristic scale a of the GP distribution. We demonstrated that our model can be successfully applied not only to reduce noise component in microarray data, but it also allows us to select a large number highly reliable differentially expressed genes.

Expression profiles have been extensively used to classify the cancers and predict clinical outcome. A large number of statistically significant differentially expressed genes for breast human cancer have been reported [18]. In our work, for the first time a parametric probabilistic model of microarray data allows us to construct a global pattern selection algorithm that reveals a large number of highly significant probe sets that is biologically relevant. We confirm this statistically using gene ontology analyses. For both low and high aggressive tumor grade samples we show that there is a profound and statistically significant shift in the exponent parameter of the generalized Pareto-gamma (GPG) distribution.

A strongly significant shift in the exponent value suggests two fundamentally different tumor subtypes, affecting many structures, regulatory processes and pathways. These findings support the view of A.I. Ivshina et al. [5], that low and high grade breast cancer can be defined genetically, reflecting stable genetic independent pathobiological subtypes rather than a continuum of cancer progression.

The weak power law only sets in for the high intensity region, which corresponds to genes that produce a large number of copies of transcripts, typically house-keeping and some tissue-specific genes. In biological networks, we believe that it is important to pay due attention to the low and medium expression regions where the most of the regulatory mechanism is found [8,9].

Our gene selection procedure based on permutation analysis and Kendal τ statistics allows us to select the genes expressed at very low expression levels which can be clearly identified due to the fact that their expression distribution is stable within compared groups and distinguishable from the random pattern of additive noise.

References

1. Anderson, T.W., Darling, D.A.: Asymptotic theory of certain 'goodness-of-fit' criteria based on stochastic processes. Ann. Math. Stat., 23 (1952) 193-212
2. Furusawa, C., Kaneko, K.: Zipf's law in gene expression. Phys. Rev. Lett. 90(8) (2003) 088102
3. Hoyle, D.C., Rattray, M., Jupp, R., Brass, A.: Making sense of microarray data distributions. bioinformatics. Bioinformatics 18(4) (2002) 576-584
4. Dozmorov, I. et al.: Neurokinin 1 receptors and neprilysin modulation of mouse bladder gene regulation. Physiol. Genomics 12 (2003) 239-250
5. Ivshina, A.V. et al.: Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. In: Liu, E.T., Colman, A., C., Harris, C., Nishikawa, S.-I., Reddel, R. (eds.): Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. In: Stem cells, Senescence and Cancer. Keystone Symposia on Mol. Biol. (Singapore) (October 2005) p.76
6. Johnson, N. L., Kotz, S., Balakrishnan, N.: Continuous Univariate Distributions, Vol. 1 and Vol. 2. 2nd edn. Wiley-Interscience (1993)
7. Konishi, T.: Three-parameter lognormal distribution ubiquitously found in cDNA microarray data and its application to parametric data treatment. BMC Bioinformatics, Jan 13, 5 (2004) 5
8. Kuznetsov, V.A.: Distribution associated with stochastic processes of gene expression in a single eukaryotic. EURASIP J. App. Signal Processing 4 (2001) 258-296

9. Kuznetsov, V.A.: Mathematical Analysis and Modeling of SAGE Transcriptome. Horizon Science Press (2005) 139-179
10. Kuznetsov, V.A., Knott, G.D., Bonner, R.F.: General statistics of stochastic process of gene expression in eukaryotic cells. Genetics 161(3) (2002) 1321-1332
11. Li, W., Yang, Y.: Zipf's law in importance of genes for cancer classification using microarray data. J. Theor. Biol. 219(4) (2002) 539-551
12. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.H., Teller, A.H. Teller, E.: Equation of state calculations by fast computing machines. J. Chem. Phys. 21(6) (1953) 1087-1092
13. Michiels, S., Koscielny, S., Hill, C.: Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet 365(9458) (2005) 488-492
14. Pareto, V.: Cours d'economie Politique, Vol. II. Lausanne: F. Rouge (1897)
15. Reis-Filho, J.S., Westbury, C, Pierga, J.Y.: The impact of expression profiling on prognostic and predictive testing in breast cancer. J. Clin. Pathol. 59(3) (2006) 225-231
16. Stephens, M.A.: Statistics for goodness of fit and some comparisons. J. Amer. Stat. Ass. 23 (1974) 193-197
17. Ueda, H.R. et al.: Universality and flexibility in gene expression from bacteria to human. PNAS 101(11) (2004) 3765-3769
18. Zucchi, I., Mento, E., Kuznetsov, V.A. et al.: Gene expression profiles of epithelial cells microscopically isolated from a breast-invasive ductal carcinoma and a nodal metastasis. Proc Natl Acad Sci U S A 101(52) (2004) 18147-18152