Observation of intermittency in gene expression on cDNA microarrays

Leif E. Peterson[1]* and Kwong Lau[2]

[1]*Departments of Medicine, Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030*
[2]*Department of Physics, University of Houston, Texas 77204*

May 14, 2002

Abstract

We used scaled factorial moments to search for intermittency in the log expression ratios (LERs) for thousands of genes spotted on cDNA microarrays (gene chips). Results indicate varying levels of intermittency in gene expression. The observation of intermittency in the data analyzed provides a complimentary handle on moderately expressed genes, generally not tackled by conventional techniques.

PACS: 87.10.+e

Scaled factorial moments have found widespread use in high-energy physics for detecting intermittency in particle production [1-10]. The presence of jet-like structures and perhaps quark-gluon plasma phase in particle production can result in clustering of data in bins leading to holes and spikes in the rapidity distribution. This investigation is based on the type of intermittency defined as nonstatistical fluctuations invariant over the scale of resolution of particle rapidity [11-18]. We do not consider, for example, the type of intermittency found in turbulence which produces non-Gaussian tails in temperature distributions [19]. In high energy physics, Bialas and Peschanksi [10] reported that the true bin probabilities can only be observed with infinite statistics. In the case of finite particles, the observed distribution of particles $Q(p_1, ..., p_M)$ smears out the Bernoulli component, as shown in (5.2) of [10]. To overcome this, scaled factorial moments of the observed data are used to measure the scaled moments of the true distribution. If only statistical fluctuations are present in the rapidity distribution of particles, then there will be no intermittency. The added value of scaled factorial moments is that they also remove the Poissonian noise to reveal dynamical fluctuations which may be present. This paper describes the use of scaled factorial moments to search for intermittency in gene expression on complimentary DNA (cDNA) microarrays (gene chips) which contain simultaneous expression levels for thousands of genes. Intermittency in this context implies that we are in search of abundances of gene expression values within the Gaussian-like distribution of expression. If there is no intermittency, then we would expect smooth Gaussian-like distributions of expression with only statistical variation.

During nucleic acid *transcription*, each "coding" gene synthesizes a messenger ribonucleic acid (mRNA) using a deoxyribonucleic acid (DNA) template. mRNA transits from the cell nucleus to the ribosome which resides in the cellular cytoplasm. In the ribosome, mRNAs are *translated* into proteins consisting of amino acids. For each gene, there are usually multiple copies of mRNA found in the cytoplasm. In the laboratory, mRNA is extracted from treated (diseased) and control (normal) cells. Reverse transcription (RT) is then used to generate a complimentary copy of DNA (cDNA) from each RNA. During RT, cDNAs from experimental cells are labelled with Cy5 dye which fluoresces in the red wavelength, and cDNAs from normal cells are labelled with Cy3 dye which fluouresces in the green wavelength. The labelled cDNAs are then aliquoted onto a cDNA microarray which has been spotted with DNA targets that are complimentary to the cDNA. During hybridization, the red and green-labelled cDNAs competitively bind with the spotted DNAs. Following drying, excimer laser scanning and computer image processing of a microarray, the pixel-averaged intensity of red and green signals for each spot (DNA) reveals the level of expression of a particular gene in the treated and normal cells. The logarithm of the ratio of intensities is known as the log expression ratio (LER), which compares gene expression in treatment tissue with normal tissue. (Spot intensities are normalized for total treated and normal mRNA used for the hybridization). Positive values of LER indicate greater gene expression (upregulation) in treated (or diseased) cells, whereas negative values of LER indicate lower expression (downregulation) in treated cells when compared with normal cells. In summary, cDNA microarrays use competitive hybridization to compare concentration levels of thousands of genes simultaneously expressed in treated and normal cells.

Let N represent the total number of genes spotted on a cDNA microarray. Let the range of LERs (y) on a microarray be $\Delta y = y_{max} - y_{min}$. Consider M non-overlapping equally-spaced bins with width $\delta y = \Delta y/M$. The qth ($q = 2, ..., 5$) scaled factorial moment [10] is defined as

$$F_q = M^{q-1} \sum_{m=1}^{M} \frac{n_m(n_m-1)\cdots(n_m-q+1)}{N(N-1)\cdots(N-q+1)}, \qquad (1)$$

where M is the total number of bins, n_m is the number of genes whose LER value falls within bin m, and $N =$

*Corresponding author.
*E-mail address:*peterson@bcm.tmc.edu (L. Peterson)

$\sum_m n_m$ is the total number of genes on the array.

We considered two sets of data for our analysis. The first was based on expression of 2,466 genes in the yeast *S. Cerevisiae* at different times following various experimental treatments (79 arrays) available at web site http://genome-www4.stanford.edu/MicroArray/SMD/publications [20]. These data reflect gene expression of *S. Cerevisiae* during experimental treatment with alpha factor arrest ("alpha"), centrifugal elutriation ("elu"), temperature sensitive mutation ("cdc15"), sporulation ("spo"), high temperature ("heat"), reducing agent dithiothrietol ("dtt"), low temperature ("cold"), and diauaxic shift ("diau"). The second data set consisted of expression for 9,706 genes in 60 cancer cell lines available at web site http://discover.nci.nih.gov/nature2000 [21]. Cancers represented are melanoma ("ME"), lung ("LC"), central nervous system ("CNS"), colorectal ("CO"), leukemia ("LE"), ovarian ("OV") , renal ("RE"), prostate ("PR"), and breast ("BR").

Calculations began by first determining "base" bin counts for the maximum number of bins possible for each array, $M_{max} = \Delta y/0.01$, where 0.01 was the precision of the data. We observed that F_2 increases rapidly when the bin size is smaller than 0.01, mostly likely due to round-off error in the creation of the LER values from the raw data. Round-off error can create artificial holes and spikes in the data. We, therefore, only consider bin sizes larger than 0.01. F_q was calculated for observed and simulated LERs at total bin numbers $M = M_{max}/L$ ($L = 2, 3, ..., M_{max}/2$). A lower bound of 30 was used for M in all calculations. Bin counts n_m for observed LERs were tabulated using M equally spaced bins of width $\delta y = 0.01 \times L$. Bin counts for simulated data were based on kernel density estimation [22] in the form

$$f(m) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{LER_i - y_m}{h}\right), \tag{2}$$

where $f(m)$ is the simulated bin count for the mth bin, N is the total number of LERs, $h = 1.06\sigma N^{-0.2}$ is the bandwidth, and σ is the standard deviation of LERs on the array. K is the Epanechnikov kernel function [23] defined as

$$K(u) = \begin{cases} \frac{3}{4}(1-u^2) & |u| \le 1 \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $u = (LER_i - y_m)/h$, LER_i is the value of each LER, and y_m is the lower bound of the mth bin. The simulation is essentially a smoothed function of the data with attendant statistical fluctuation. A second round of F_q calculations were made after subtracting 0.1 from y_{min} and y_{max}, redetermining bin cutoffs, and recalculating n_m and $f(m)$ in order to shift the scale of the phase space. This allowed us to look more closely at statistical fluctuations and also provided twice as many values of F_q for observed and simulated LERs. Plots of $\ln F_q$ vs. $\ln M$ were constructed for each array and each value of q. The difference between slopes for observed and simulated data was based on fitting the linear model

$$\ln F_q = \beta_0 + \beta_1 (\ln M), \tag{4}$$

separately for observed and simulated data, where β_0 is the y-intercept and β_1 is the slope. Slope difference was calculated as

$$\Delta\beta_1 = \beta_{1,\text{observed}} - \beta_{1,\text{simulated}}. \tag{5}$$

Positive values of $\Delta\beta_1$ implies intermittency among the observed LERs.

Figure 1 shows the frequency histogram for $N = 2,402$ LERs binned in $M = 106$ bins of width $\delta y = 0.02$ for the `spo0` microarray in the sporulation experiments on *S. Cerevisiae* [20]. Several large spikes and holes are visible in the LER distribution, suggesting clustering of LERs at the scale of the bin size $\delta y = 0.02$. The greatest contribution to F_q is from the spikes near the central peak of the distribution. Holes near the tails contribute little to F_q. Figure 2 illustrates plots of $\ln F_q$ vs. $\ln M$ for the

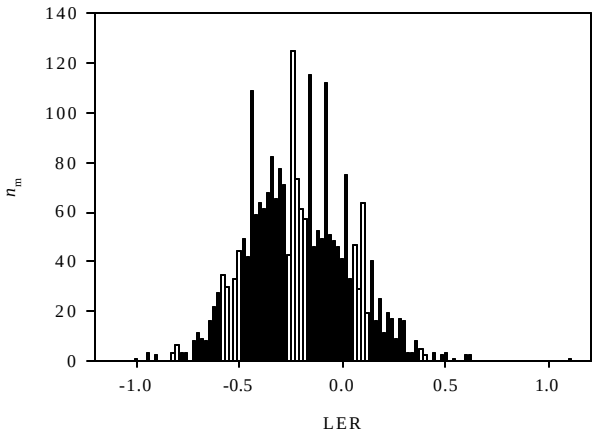

Figure 1: Binned LERs from the `spo0` microarray in the sporulation experiments on *S. Cerevisiae* [20].

same `spo0` array whose binned LERs are shown in Figure 1. Replicate values of $\ln F_q$ are also plotted at various values of $\ln M$ as a result of calculating F_q after a 0.1 shift in the scale of y. The spread in the data shows the level of bias in the choice of the binning. The slope of simulated data is usually very small, indicative of lack of intermittency in the simulated data. Slightly negative slopes for simulated data were also observed for smooth gaussian-like distributions, where $F_q \sim Constant$ which changes with M.

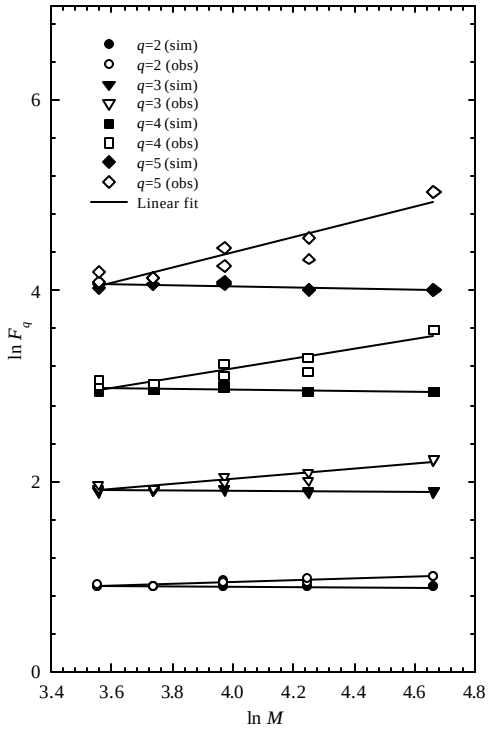

Figure 2: Log scaled factorial moments, $\ln F_q$, vs. natural log of total bin number, $\ln M$, for observed (obs) and simulated (sim) LERs for the `spo0` array [20]. Replicate values of $\ln F_q$ can be observed at various values of $\ln M$ as a result of calculating F_q after a 0.1 shift in the scale of y.

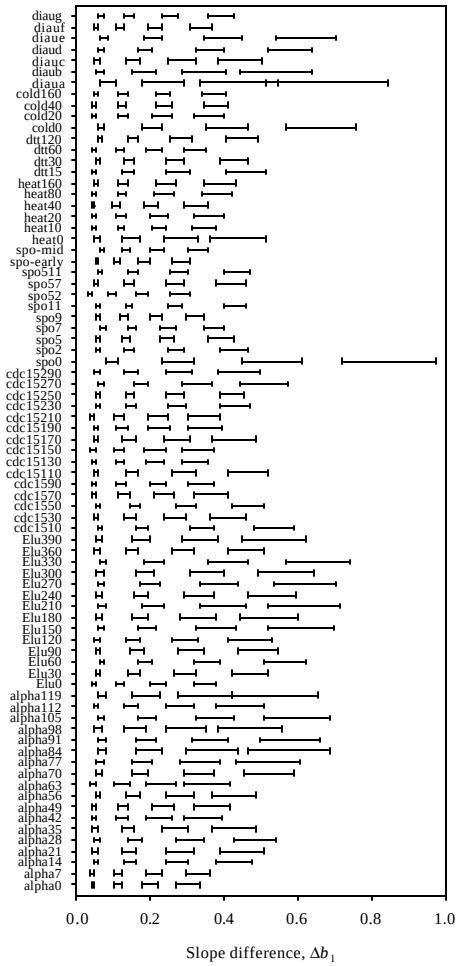

Figure 3: $\Delta\beta_1$ and uncertainty for observed and simulated data for the 79 arrays in the *S. Cerevisiae* data [20]. Ranges shown are for $q = 2, ..., 5$, with the smallest range at $q = 2$.

Figure 3 shows values of $\Delta\beta_1$ for the 79 arrays in the *S. Cerevisiae* data set [20]. Overall, the $\Delta\beta_1$ values are positive, with an average value of 0.03-0.04, indicative of intermittency in the data. The greatest signals were observed in the experiments for sporulation (`spo0` array) and diauxic shift (`diaua` array). For the alpha factor arrest (alpha) experiment, a slightly elevated signal was seen early on, dropped, and then picked up again toward the latter time points. Interestingly, the signal fluctuated over time periods in the centrifugal elutriation experiment to the extent that periodicity can be noticed. For the experiment involving the temperature-sensitive mutant `cdc15`, the signal was similar to that observed in the centrifugal elutriation experiment. During sporulation, the signal was largest at the beginning, dropped thereafter, and continued to be jumpy until the cold and diauxic shift experiment. There is no discernible trend within a given treatment, and there is no abrupt transition from one array to the next. The different treatments are not ordered in any particular way in Figure 3. The higher factorial moments follow the same trend as that of F_2, only with higher values. This is not surprising because $\Delta\beta_1$ for the qth moment is about $q/2$ times that of F_2 if the only source of intermittency is bin-by-bin fluctuations. The plotted error bars for each data point in Figure 3 are based on the error in the fitting of the slope, as well as the error of the simulated data. As one can see, the significance of the signal is usually more than 3 standard deviations.

Figure 4 shows the $\Delta\beta_1$ for the 60 cancer cell lines. Among them, the greatest intermittency signal was detected for the colon cancer cell line `CO-HT29` and central nervous system cancer cell line `CNS-U251`, since they both had the greatest values of $\Delta\beta_1$ for all values of q. Wide variation in $\Delta\beta_1$ can be observed across all of the data, with no consistent signal occurring for each type of cancer. The most jumpy transition in $\Delta\beta_1$ was seen for cen-

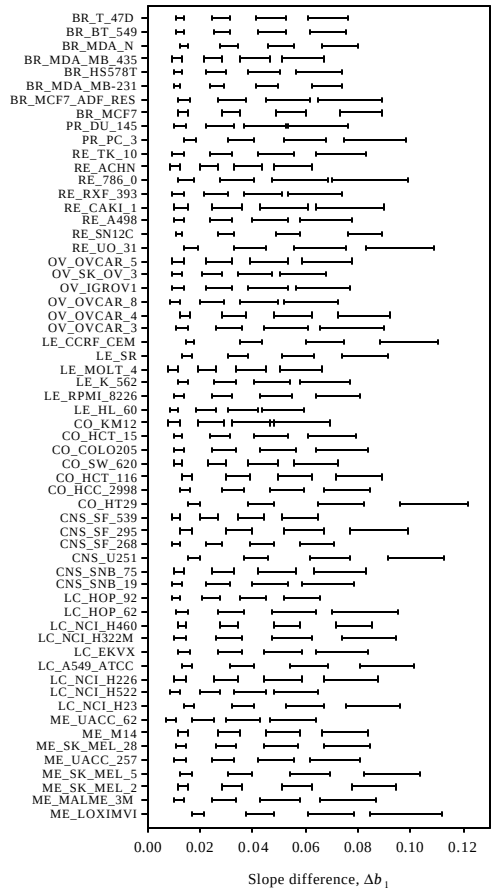

Figure 4: $\Delta\beta_1$ and uncertainty for observed and simulated data for the 60 cancer cell line arrays [21]. Ranges shown are for $q = 2, ..., 5$, with the smallest range at $q = 2$.

tral nervous system cancer cell lines. It is noteworthy that the average level of intermittency, as measured by $\Delta\beta_1$ for F_2, is about 0.01, about 3-4 times lower than that of the yeast data. The main difference is that there are about 10,000 genes in each cancer array, about 4 times more genes than the yeast array. The significance of the signal (signal/error), however, is comparable to that of the yeast data.

The observed intermittency in the data considered may suggest correlations in the abundances of expression levels within the Gaussian-like distributions of LERs. In the *S. Cerevisiae* sporulation experiments, a majority of genes whose LERs fell within the spikes in Figure 1 had dramatically altered expression values later on in the sporulation experiments. Chu et al. [24] reported temporal changes in expression among a large number of genes throughout the sporulation process. In the cancer cell lines whose LER distributions were investigated, changes in intermittency over the arrays are likely due to cancer-specific alterations in cell-cycle control, DNA repair, oncogenesis, tumor supression, apoptosis, and angiogenenesis, all of which affect tumor growth, severity and evasion from attack by the immune system [25]. Cancers vary in their cause and severity and there may be a wide range of unknown gene-gene and gene-environment interactions which impact gene expression.

Errors in reproducibility among the LERs considered were not provided by the groups that generated the data. However, several recent reports [26-30] give errors from various sources (probe preparation, spot size variability, scanning errors, software sophistication, etc.). Wildsmith et al. [26] reported a 28% standard error of the common logarithm of expression based on 64 replicate arrays containing 1248 duplicate spots. Lee et al. [28] reported a maximum misclassification of 9% based on three replicate arrays containing 288 genes. In this study, the spikes, for example, in Figure 1 are not likely due to misclassification. Error, on the other hand, affects the bin size and we have seen the effect throughout the bulk of bin-size range.

This study was a first step to search for intermittency without establishing biological relevance. There is a growing literature on the identification of microarray-based regulatory gene networks [31-34]. We have already begun looking at individual genes and their contribution to F_2 on a single array. Our current effort to develop microarray-based promoter models for co-expressed genes based on the Werner approach [33] will facilitate our understanding of regulatory control of genes with high contribution to F_2. Progress in this effort is limited by the rate at which we can manually select genes with high contribution to F_2, exon map the genes, fetch their upstream DNA promoter sequences, and then search for common transcription binding sites among the multiple promoter sequences in order to infer coregulation.

The observation of intermittency in the data analyzed provides a complimentary handle on moderately expressed genes, generally not tackled by conventional techniques. Biologists often focus on strongly downregulated or upregulated genes which are characterized by large negative and positive LERs. Our method of looking at intermittency in gene expression focused on the clustering of LERs independent of their absolute expression value. Thus, we were able to detect large density fluctuations among small LERs. As an example, spikes near the center of the binned distribution in Figure 1 whose LER-values were low greatly increased the factorial moments. Therefore, fold-change analysis, which focuses on large negative and positive LERs, or other multivariate statistical methods such as hierarchical cluster and principal component analyses, can miss unique density fluctuations at low LER values which are detected by factorial moments.

L.E.P. acknowledges the support of grant CA-78199-04 of the National Cancer Institute, and C. Aime, F. Kun, A. Loctionov, M. Patra, R. Peschanki, and E. Sarkisyan-Grinbaum for helpful discussions on intermittency. K.L. is supported in part by the U.S. Department of Energy, Grant no. DE-FG03-96ER41004, and in part by the Texas Advanced Research Program, Grant no. 3652-0023-1999.

References

[1] OPAL Collaboration, Eur. Phys. J. C, 11, 239 (1999).

[2] EHS/NA22 Collaboration, Phys. Lett. B, 382, 305 (1996).

[3] EMU-01 Collaboration, Z. Phys. C, 76, 659 (1997).

[4] SLD Collaboration, SLAC-PUB-95-7027 (1996).

[5] WA80 Collaboration, Nuc. Phys. A, 545, 311c (1992).

[6] F. Kun, H. Sorge, K. Sailer, G. Bardos, W. Greinber, Phys. Lett. B, 355, 349 (1995).

[7] Z. Jie and W. Shaoshun, Phys. Lett. B, 370, 159 (1996).

[8] J.W. Gary, Nuc. Phys. B, 71S, 158 (1999).

[9] P. Bozek and M. Ploszajczak, Nuc. Phys. A, 545, 297c (1992).

[10] A. Bialas and R. Peschanksi, Nuc. Phys. B, 273, 703 (1986).

[11] A. Bialas and R. Peschanksi, Nuc. Phys. B, 308, 857 (1988).

[12] A. Bialas, Nuc. Phys. A, 525, 345c (1991).

[13] M. Blazek, Int. J. Mod. Phys. A, 12, 839 (1997).

[14] P. Bozek, M. Ploszajczak, R. Botet, Phys. Rep., 252, 101 (1995).

[15] J. Fu, Y. Wu, L. Liu, Phys. Lett. B, 472, 161 (2000).

[16] W. Shaoshun, L. Ran, W. Zhaomin, Phys. Lett. B, 438, 353 (1998).

[17] I. Sarcevic, Nuc. Phys. A, 525, 361c (1991).

[18] E.A. DeWolf, I.M. Dremin, W. Kittle, Phys. Rep., 270, 1 (1996).

[19] L.P. Kadanoff, Physics Today, August, 34, (2001).

[20] P.T. Spellman, G. Sherlock, M.Q. Zhang, et al., Mol. Biol. Cell., 12, 3273 (1998).

[21] D.T. Ross, U. Scherf, M.B. Eisen, et al., Nat. Genet., 24, 208 (2000).

[22] D. Fadda, E. Slezak, A. Bijaoui, Astron. Astrophys. Suppl. Ser., 127, 335 (1998).

[23] V.A. Epanechnikov, Theor. Prob. Appl., 14, 163 (1969).

[24] S. Chu, J.L. DeRisi, M.B. Eisen, et al., Science, 282, 699 (1998).

[25] S.B. Baylin, Science, 277, 1948 (1997).

[26] S.E. Wildsmith, G.E. Archer, A.J. Winkley, et al., Biotechniques, 30, 202 (2001).

[27] A. Brazma, P. Hingcamp, J. Quackenbush, et al., Nature Genet., 29, 365 (2001).

[28] M-L.T. Lee, F.C. Kuo, G.A. Whitmore, et al., P.N.A.S., 97, 9834 (2000).

[29] C.S. Brown, P.C. Goodwin, P.K. Sorger, P.N.A.S., 98, 8944 (2001).

[30] X. Wang, S. Ghosh, S-W. Guo, Nucl. Acids. Res., 29, e75 (2001).

[31] A. Wuensche, Pac. Symp. Biocomp., 3, 89 (1998).

[32] D.J. Lockhart, E.A. Winzeler, Nature, 405, 827 (2000).

[33] T. Werner, Biomol. Eng., 17, 87 (2001).

[34] Y. Pilpel, P. Sudarsanam, G.M. Church, Nat. Genet., 29, 153 (2001).