Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Assessing the Information Content of Short Time Series Expression Data

Eric H. Yang and Ioannis P. Androulakis*
Department of Biomedical Engineering
Rutgers University, Piscataway, NJ 08854

***Abstract*— Due to experimental constraints, the sampling of biological system with microarray data is severely constrained. In similar fashion to sampling theory of signals, the under-sampling of a system oftentimes leads to sub-optimal results from which it is difficult to draw proper conclusions. In our work we create a mathematical framework which will show that the sampling methodology for short time series microarray data may lead to data whose ability to distinguish non-random behavior within the biological system is severely constrained.**

I. Introduction

The advent of the micro-array has been hailed as a revolution in molecular biology[1]. It ushered in the era of high throughput gene expression analysis, allowing researchers to interrogate the expression levels of many genes at once. The next step in this revolution is temporal expression profiling where gene expression levels are measured at multiple time points. The motivation for taking time dependent samples was that the temporal evolution of genes offers a better perspective upon the underlying mechanisms that govern an organism's response to an external stimulus. However, given serious experiment constraints, especially in mammalian systems, the high throughput nature of microarrays applies only to the space of measured genes and not the time domain limiting the sampling rate for the experiments. For instance, the majority of data sets within SGD(Saccharomyes Genomic Database) are between 4 and 6 time points in length[2].

Recently, work has been done utilizing hashing for the selection and classification of temporal expression data. Two of the most recent algorithms are Short Time-series Expression Miner(STEM)[2] and SeLection of INformative Genes via Symbolic Hashing of Time Series(SLINGSHOTS)[3]. These algorithms both perform fine grained clustering though the use of a hashing function which assigns similarly shaped expression profiles to a specific motif, and work under the assumption that highly populated motifs are more relevant to an organism's underlying response to an external stimulus than motifs which were sparsely populated. One of the intermediate results from this class of algorithms is a histogram which gives the populations of the different motifs **Figure 1**. Some of the shorter time series illustrated population dynamics which are very similar to those obtained when using randomly generated data i.e. showed an exponential distribution which is characteristic of the performance of an idealized hashing algorithm upon randomly generated data, whereas experiments with a greater number of time points did not. This suggests that there is a level of ambiguity that can arise due to insufficient sampling.

Sampling theory states that one must sample at twice the highest frequency which exists in the data set. This is known as the Nyquist Sampling Rate[4]. The implication of this is that the number of required samples should be twice the intrinsic time constant of the processes being studied. This in practical terms is usually infeasible given experimental constraints, particularly in the case where animal studies are involved.

The obvious consequence to this is that one is unable to capture the true response of the system without first knowing some information about the response before the experiment is carried out. Many researchers therefore set up experiments with rough ideas as to the underlying dynamics. However, oftentimes these rough estimates of the underlying response are insufficient and lead to a sub-optimal sampling strategy[5]. This sub-optimal sampling strategy often leads to a data set which is indistinguishable from a randomly generated data set. This consequence is important because the driving goal behind systems biology is the creation of mathematical models which can explain the *non-random* responses of an organism to an external challenge[6]. To

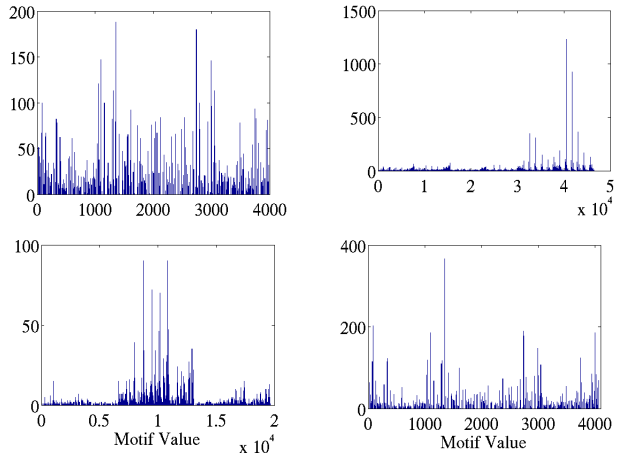

Fig. 1. 4 motif histograms generated via SLINGSHOTS. The first motif is the histogram of a randomly generated data set with five time points. The second motif distribution was generated from a 6 time point data set obtained from Calvano *et al.*[7]. The third motif distribution was generated from a 17 time point data set (GEO: GDS253)[8], and the last motif distribution was generated from a 5 point data set (GEO: GSE802)[9]

*Corresponding Author: yannis@rci.rutgers.edu

rigorously quantify the "informative" nature of a data set, we propose the creation of a clustering/selection algorithm independent metric that quantifies the ability of data set to provide information regarding a non-random process. This is due to the assumption that microarrays are measuring the effects of a coordinated process responding to external stimuli[10]. We will leverage a fundamental property of random time series, namely the unique shape of its auto-correlation function. What we will illustrate is the effect of a long data set upon the autocorrelation function, the effect of a typical short time series data set, and the effect of having a large number of genes with very similar motifs.

II. Data

We will be utilizing three data sets for our initial determination, first a 17 time point corticosteroid data set (GDS253)[8] the second is a burn data set (GSE802)[9], and finally a bacterial endotoxin induced sepsis data set containing 6 time points[7].

Before the data sets are evaluated, they will be normalized via the z-score:

$$X' = \frac{X - \mu}{\sigma} \tag{1}$$

This is done in order to remove the effects of signal scaling upon our metric.

III. Methods

We will define the information content of a data set as the ability for it to distinguish itself from a randomly generated data set. The property which we will be leveraging is the characteristic profile of the autocorrelation of a random expression profile.

The autocorrelation function is defined as:

$$\int_{-\infty}^{\infty} f(x) * f(x+t)dx \tag{2}$$

The reason we wish to use the autocorrelation function is because given a z-score normalized random Gaussian signal, it has the characteristic shape given in **Figure 2**. The characteristic shape has a large spike at time lag = 0 denoting the perfect correlation of a signal with itself. It then immediately falls to a low level symbolizing that the value at time point N has no relationship to the value at time point N-1. We will use this characteristic shape as the basis for comparison to distinguish between our informative and non-informative sets. The autocorrelation function works best at illustrating the difference between random and non-random data if the time series are long. Therefore, in order to obtain a long time series, we take the measurement for the N genes at M time points, and concatenate the expression profiles for the different genes in order to obtain a single time series that is NxM points long. Since we want to verify whether or not the randomness we observed was caused by the sampling rate, a random permutation of the genes must first take place. This is to assure that the non-randomness is due to the expression level measurements and not due to some outside influence. For instance had we sorted a randomly generated data set of 8800 genes by the first time

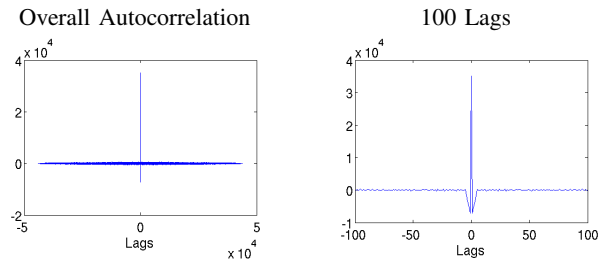

Fig. 2. The characteristic profile of the autocorrelation function after z-score normalization evaluated with a randomly generated time series. What we are interested is the low correlation values at $t \neq 0$.

point, then the second in a similar fashion to the radix sort[11] (Sorting via the Least Significant Bit, and moving to the Most Significant Bit), then we would obtain the profile shown in **Figure 3**. Although this data is made up of random sequences similar to the data used to generate **Figure 2**, it does not have the same autocorrelation function. This is because order is being imposed upon the data set via the sorting process. A random permutation removes order imposed by outside factors and focuses specifically upon whether the measurements at time N have any relationship to the measurements at N-1. After discarding the correlation

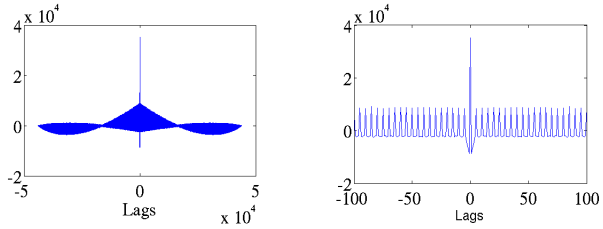

Fig. 3. The profile of a 8800 randomly generated expression profiles of 5 time points which have been sorted by the first time point, then the second, etc.

at Time Lag = 0 we need to quantify the differences in the distribution in the auto-correlation value. To do this we utilize the T-test and the F-test, two standard statistical tests which measure the difference in the mean and the difference in the standard deviation[12]. What we are looking for is either a large difference in the mean, or a large difference in the standard deviation between the two sets of data points. A large deviation in either of these two values will signal significant non-randomness present in the data set. Therefore when these two values are calculated, the informative score will be the max of these two values. In order to mitigate the effects of scaling upon these tests, the autocorrelation values will be linearly scaled so that the interval between the max value and the min value are 1.

$$\text{T-test:} \quad t = (\bar{X} - \bar{Y})\sqrt{\frac{n(n-1)}{\sum_{i=1}^{n}(\hat{X}_i - \hat{Y}_i)^2}} \tag{3}$$

$$\text{F-test:} \quad f = \frac{s_X^2}{s_Y^2} \tag{4}$$

X and Y are individual distributions

In addition to the value of the metric, we will also evaluate the effect of the informativeness upon the quality of the results derived from hierarchical clustering, a widely used clustering technique used to process biological data. We elected to use hierarchical clustering in order to provide independent verification of our metric, and to determine whether or not the initial observations made during our examination of the SLINGSHOTS and STEM algorithm were more widely applicable. The specific implementation which we will be using is CLUTO, a clustering package which implements hierarchical clustering along with various other optimization based clustering algorithms[13]. The application of any clustering algorithm upon temporal data works under the assumption of Guilt by Association [14]. Therefore assuming proper classification via expression profiles, we ought to see ontologies, or gene functions, differentially expressed within the clusters. Therefore in order to evaluate the quality of our clustering results, we will evaluate the distribution of ontologies among our different clusters. This is done by calculating enrichment [2]. The enrichment method that we will use calculates the p-value of an ontology being functionally enriched within a cluster via the formula given in **Equation 5**. This enrichment metric assumes that the presence of an ontology within a given cluster follows a hyper-geometric distribution, and calculates its enrichment accordingly.

$$P(M,m,n,N) = \frac{\begin{pmatrix} m \\ n \end{pmatrix}\begin{pmatrix} M-m \\ N-n \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} \quad (5)$$

M = Total number of genes

N = Total number of times ontology appears

m = genes in the cluster

n = genes in the cluster with the given ontology

An estimate as to the relative quality of the clustering results is the determination of how many ontologies are segregated by cluster. Provided that the assumption of "Guilt by Association" holds true, we ought to see large numbers of ontologies localized to a specific cluster. therefore, after we have calculated the enrichment for each ontology/cluster combination, we will make the determination of the quality of our clustering result by determining the number of ontologies which have localized to a single cluster with a p-value of less than .05. This is done by first taking the minimum p-value for each ontology over the different clusters, and then selecting those minimum cluster/ontology combinations for a p-value of less than .05. The ontologies for the dataset will be obtained from annotation files accompanying the microarrays used (RG-U34A, HG133A,B).

IV. Results/Discussion

In the case for our three data sets, the distribution values for the autocorrelation functions all had a mean close to zero. Since the metric was the max of the F-test and the T-test, we will not be reporting the t-test value for the three tests. It is important to note that this may not always be the case. For instance, it is possible for all of the expression profiles to be well correlated and therefore lead to cyclical spikes in the autocorrelation function leading to a non-zero mean. Therefore for the sake of completeness, we had included the t-test as part of our test metric. The autocorrelation functions for the three data sets are shown in **Figure 4**, with their corresponding motif distributions shown in **Figure 1**. What is clearly evident is that the 17 time point corticosteroid data set (GDS253) illustrates a significantly different distribution than the impulse response, whereas the 5 time point data set has a very similar distribution to the randomly generated data set. This was borne out when we calculated the F-statistic which came out to 4.56 for the GDS253 data set, and 1.11 for the GSE802 data set.

What was surprising was that the bacterial endotoxin inflammation data set from Calvano *et al.* illustrated a highly non-random dynamic similar to the that of the much longer GDS253 data set despite being a short data set similar to the GSE802 burn data set. When we calculated the F-test statistic, we found that the value was 3.54. What this means is that the data has an expression dynamic which can be adequately captured with 6 time points, whereas the burn data set (GSE802) has a more complex behavior that requires more than 5 data points.

Examining the autocorrelation functions closer, we can discern the reasons for this difference. In both the bacterial endotoxin data set and the GDS253 corticosteroid data set, we can observe both a non-impulse like response at the very low lags ± 2, as well as the cyclical nature of the autocorrelation profile. The non-impulse like response suggests that enough sampling points have been taken to guarantee non-randomness of the individual expression profiles. Additionally, we can see periodic spikes within our autocorrelation functions suggesting that there is a significant degree of correlation within our data set, *i.e.* there are a relatively small number of true clusters within our data set.

None of these properties are evident within the GSE802 data set. The autocorrelation profile at ± 100 lags looks very similar to that of the randomly generated data. This means that insufficient time points have been measured to capture the inherent response of the system. In addition, there is no cyclical repines evident in the tail region, meaning that the true number of clusters is probably large in comparison to the other two data sets. If the assumption that an organism's response to an external stimuli is to bring online a large set of coordinated pathways is correct[2], then it would mean that the GSE802 data set does a poor job of capturing these dynamics.

When we used our metric to estimate the quality of the hierarchical clustering result, the GDS253 dataset with 17 time points showed 43% of its ontologies showed statistically significant localization to a specific cluster, while the bacterial endotoxin case also showed 43%. The GSE802 dataset on the other hand only showed 13.3% of its ontologies being

preferentially expressed in a single cluster. We believe that this illustrates a significant result because the less informative dataset showed a significantly lower amount of enrichment in agreement with its lower informativeness score. We performed on final test where we sub-subsampled the 17 time point dataset, taking measurements at [1,4,8,12,16] hours, and rerunning the comparisons, what we found was that the F-test reduced to 1.42, and it yielded 13.6% of total ontologies being statistically enriched. This results shows the degradation of the information content of a dataset and henceforth the results of the clcustering algorithm given an improper sampling strategy.

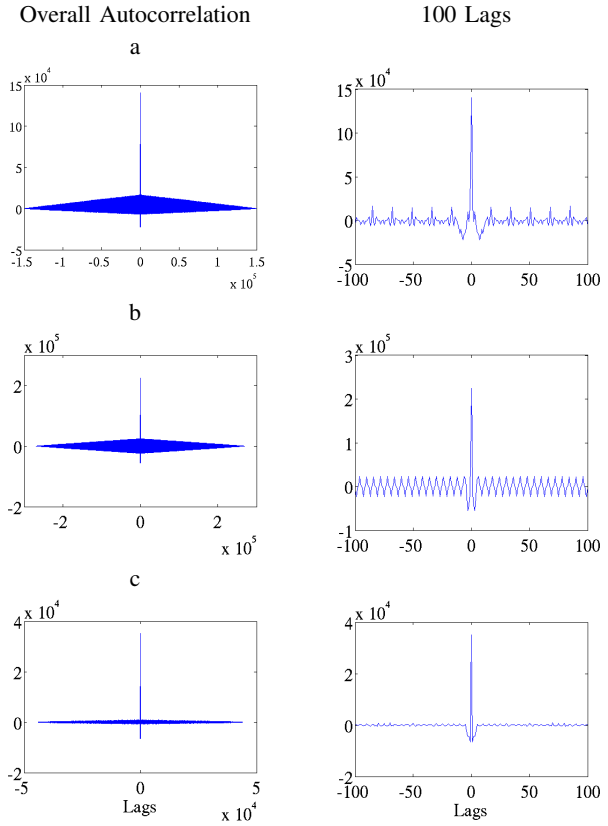

Fig. 4. The results of the auto-correlation function upon the three data sets we are using. From the top: a) 17 time point corticosteroid data set(GDS253), b) the 6 time point bacterial endotoxin data set, and c) the 5 time point burn data set (GSE802)

V. Conclusion

The primary message which we wish to convey is that while the initial observation upon the ability of a dataset to encapsulate a non-random process was uncovered during our examination of hashing algorithms, it is not an artifact of the algorithm itself. Our method independent metric for evaluating the informative nature of a dataset yielded an ordering which matched an ontology evaluation obtained with a commonly used clustering method namely hierarchical clustering. This means that the initial observation we made while studying the hashing based selection/classification algorithms is data dependent and not method dependent.

Provided that temporal microarray experiments ought to provide a glimpse into a coordinated process, we believe that the information content of a raw microarray dataset should reflect some aspect of non-randomness. Therefore, we believe that our metric provides a method for evaluating how well a given dataset can capture the temporal evolution of some unknown underlying process. We believe that this proposed method can help researchers evaluate the quality of datasets provided online, and then make the decision whether more or less data points are needed in order to capture the experimental dynamics which they are interested in.

VI. Acknowledgments

The authors would like to thank Dr. John Semmlow for his help and insightful comments concerning the paper, financial support from the NSF under the NSF-BES 0519563 Metabolic Engineering Grant.

References

[1] V. Cheung, M. Morley, F. Aguilar, A. Massimi, R. Kucherlapti, and C. G., "Making and reading microarrays." *Nature Genetics*, 1999.

[2] J. Ernst, G. J. Nau, and Z. Bar-Joseph, "Clustering short time series gene expression data," *Bioinformatics*, vol. 21 Suppl 1, pp. i159–i168, Jun 2005.

[3] E. Yang, F. Berthiamume, M. Yarmush, and I. Androulakis, "An integrative systems biology approach for analyzing liver hypermetabolism," *Proceedings of the Joint 9th International Symposium. Processing Systems Engineering and 16th European Symposium*, 2006.

[4] J. Semmlow, *Biosignal and Biomedical Image Processing: Matlab-based applications*. Marcel Dekker, 2004.

[5] I. Simon, Z. Siegfried, J. Ernst, and Z. Bar-Joseph, "Combined static and dynamic analysis for determining the quality of time-series expression profiles," *Nat Biotechnol*, vol. 23, no. 12, pp. 1503–1508, Dec 2005, evaluation Studies.

[6] J. Hu, M. Kapoor, W. Zhang, S. R. Hamilton, and K. R. Coombes, "Analysis of dose-response effects on gene expression data with comparison of two microarray platforms," *Bioinformatics*, vol. 21, no. 17, pp. 3524–3529, Sep 2005, evaluation Studies.

[7] S. E. Calvano, W. Xiao, D. R. Richards, R. M. Felciano, H. V. Baker, R. J. Cho, R. O. Chen, B. H. Brownstein, J. P. Cobb, S. K. Tschoeke, C. Miller-Graziano, L. L. Moldawer, M. N. Mindrinos, R. W. Davis, R. G. Tompkins, and S. F. Lowry, "A network-based analysis of systemic inflammation in humans," *Nature*, vol. 437, no. 7061, pp. 1032–1037, Oct 2005, clinical Trial.

[8] R. R. Almon, D. C. DuBois, K. E. Pearson, D. A. Stephan, and W. J. Jusko, "Gene arrays and temporal patterns of drug response: corticosteroid effects on rat liver," *Funct Integr Genomics*, vol. 3, no. 4, pp. 171–179, Dec 2003.

[9] A. Jayaraman, M. L. Yarmush, and C. M. Roth, "Evaluation of an in vitro model of hepatic inflammatory response by gene expression profiling," *Tissue Eng*, vol. 11, no. 1-2, pp. 50–63, Jan 2005, evaluation Studies.

[10] L. Soinov and M. Kapushesky, "Unraveling Nature's Networks," *Genome Biol*, vol. 4, no. 10, p. 341, 2003, congresses.

[11] D. Knuth, *The Art of Computer Programming, Volume 3: Sorting and Searching*. Reading, MA: Addison-Wesley, 1973.

[12] S. Ross, *Introduction to Probability and Statistics for Engineers and Scientists*. New York, NY: Elsevier Academic Press, 2004.

[13] Y. Zhao and G. Karypis, "Clustering in life sciences," *Methods Mol Biol*, vol. 224, pp. 183–218, 2003.

[14] C. J. Wolfe, I. S. Kohane, and A. J. Butte, "Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks," *BMC Bioinformatics*, vol. 6, p. 227, 2005.