INTERACTIVE COLOR MOSAIC AND DENDROGRAM DISPLAYS FOR SIGNAL/NOISE OPTIMIZATION IN MICROARRAY DATA ANALYSIS

Jinwook Seo[1,2,4], *Marina Bakay*[1,4], *Po Zhao*[1], *Yi-Wen Chen*[1], *Priscilla Clarkson*[3], *Ben Shneiderman*[2], *Eric P Hoffman*[1]

[1]Research Center for Genetic Medicine, Children's National Medical Center
[2]Dept. of Computer Science & Human-Computer Interaction Lab, University of Maryland
[3]Exercise Science, University of Massachusetts- Amherst
[4]The first two authors contributed equally to this manuscript

ABSTRACT

Data analysis and visualization is strongly influenced by noise and noise filters. There are multiple sources of "noise" in microarray data analysis, but signal/noise ratios are rarely optimized, or even considered. Here, we report a noise analysis of a novel 13 million oligonucleotide dataset - 25 human U133A (~500,000 features) profiles of patient muscle biposies. We use our recently described interactive visualization tool, the Hierarchical Clustering Explorer (HCE) to systemically address the effect of different noise filters on resolution of arrays into "correct" biological groups (unsupervised clustering into three patient groups of known diagnosis). We varied probe set interpretation methods (MAS 5.0, RMA), "present call" filters, and clustering linkage methods, and investigated the results in HCE. HCE's interactive features enabled us to quickly see the impact of these three variables. Dendrogram displays showed the clustering results systematically, and color mosaic displays provided a visual support for the results. We show that each of these three variables has a strong effect on unsupervised clustering. For this dataset, the strength of the biological variable was maximized, and noise minimized, using MAS 5.0, 10% present call filter, and Average Group Linkage. We propose a general method of using interactive tools to identify the optimal signal/noise balance or the optimal combination of these three variables to maximize the effect of the desired biological variable on data interpretation.

INTRODUCTION

The instructional information for the formation and function of cells, tissues, and organisms is encoded in the shared genetic material of each cell (genes). Humans have about 40,000-50,000 genes. Each gene has the potential to be "expressed" into mRNA (transcription), and then these mRNAs are translated into the protein components of the cells and tissues. The entire human 3 billion letter genetic code is known and web-accessible, however functional roles have been assigned to only a small minority of the genes/proteins.

Research on the gene/mRNA/protein axis in the life sciences has traditionally been based on the study of single or small numbers of genes, mRNAs, or ptoteins, and these studies are not computationally intensive. Recent technological advances have enabled a highly parallel approach to biological data generation through the use of microarrays of nucleic acid molecules. Current implementations of microarrays take advantage of the ability of nucleic acids of complementary sequence to bind to each other to form the classic "double stranded" DNA molecule; one strand is placed on the microarray, and the complementary strand in solution then seeks out and binds the immobilized strand. RNA and DNA can form even stronger duplexes, so a DNA microarray is very efficient at querying a complex solution of mRNAs derived from tissues or cells. The most common implementation of microarrays is to produce one or more DNA probes for each gene, and then address these probes to specific places on a glass substrate (microarray). A complex solution of fluorescently labeled mRNAs from cells or tissues is then hybridized to the microarray, with the amount of binding to the specific feature on the array representing the relative concentration of that mRNA in the cell or tissue. Laser scanning of the microarray produces an image where the fluorescent intensity of each array feature is calculated as a concentration of that gene in the original solution. Simultaneous analysis of many thousands of genes on the microarray leads to an "expression profile" of the original cell or tissue. This profile represents the subset of the 40,000 genes that are being employed by that cell or tissue, at that particular point in time. As microarrays now contain up to 500,000 features, the large amount of data generated by microarray analysis of biological samples is providing a fertile ground for the application of theories in computer science, human/computer interaction, and information visualization to biological data sets.

Two types of experimental platforms are frequently used for generating microarray data. One, "spotted microarrays", involves the physical spotting of solutions of relatively large (~1,000 nucleotide) cDNA clones or large oligonucleotides (~70 nucleotide) to glass slides. Spotted microarrays are typically presented with two solutions of labeled mRNAs; one from one tissue, and one from a second tissue, each labeled with a different fluorescent molecule. The data generated by spotted microarrays is a simple ratio of the two fluorescent colors at each feature. The ratio provides a relative concentration of that mRNA in the two mRNA solutions used. The second type of microarray involves the *in situ* synthesis of 25 nucleotide probes within defined 20 μm^2 areas on glass (Affymetrix arrays) (http://www.affymetrix.com). A major distinction between spotted cDNA (and also emerging spotted 70-mer oligonucleotide arrays) and the Affymetrix microarrays is that the former provides a single ratio measurement for the difference between two tissues or cells, while Affymetrix arrays conduct between 16 and 40 distinct measurements for each mRNA in one solution. This generates a redundant, stand-alone profile that can be databased and compared to other experiments in other laboratories (see http://microarray.cnmcresearch.org/).

New human Affymetrix microarrays use 1 million oligonucleotide probes to query most (~40,000) human mRNAs in two small (1.28 cm^2) glass arrays. Importantly, Affymetrix arrays have intrinsic redundancy of measurements for each gene, with 16 "perfect match" probes for different regions of each gene sequence, with each perfect match paired with a similar

"mismatch" probe with a single destabilizing nucleotide change in the center of the 25 nucleotide sequence. This mismatch is meant to serve as a noise filter; labeled mRNA binding to the "mismatch" is considered to represent non-specific binding, and thus a measure of "noise" for the corresponding perfect match. The complete set of 16 probe pairs is called the "probe set" for any single gene.

A key step in the analysis of Affymetrix arrays is the interpretation of the complete probe set for a gene, with the derivation of a single "signal" value representative of the different probes within the set. Given the relatively large amount of data (16 perfect match probes [PM], with 16 paired mismatch controls [MM]), there are many possible algorithms that can be employed to derive the "signal" value from the 16 PM/MM pairs. As might be expected, there is considerable debate concerning the appropriate use of the MM signal as a noise filter when interpreting the entire probe set. The Affymetrix algorithm for "signal" is calculated using the One-Step Tukey's Biweight Estimate resulting in a weighted mean across all probe pairs in the probe set [1]. This algorithm gives substantial weight to the mismatch, using it as representative of non-specific "noise" of the perfect match.

Others claim that the penalty for the mismatch is too severe; in many instances, the mismatch signal is a composite of true hybridization to the correct RNA, as well as non-specific noise. In one increasingly popular method, termed RMA [2], normalized and log-transformed perfect match values are used without a strong penalty for mismatch signals. Normalization is done across many microarrays within a "project", rather than the stand-alone normalization used by Affymetrix [2, 3]. Irizarry *et al* used a log scale linear additive model for probe level data across arrays after appropriate background removal and normalization. A robust procedure is used to fit the model and get the estimated log scale measure of expression. The RMA method performs very well with known "spike in" RNAs, providing greater sensitivity and more stable "signals" from probe sets. However, the greater sensitivity of the RMA method would be expected to come at a cost of specificity; the less weight given to the mismatch "noise" filter by RMA would be expected to lead to greater signal/noise problems in complex solutions.

In addition to the signal/noise issues regarding probe set interpretation and normalization, there is an issue of whether a confidence filter should be superimposed on the data. As the signal intensity of a probe set decreases, it approaches the threshold "noise" or "background" level of the microarray. Clearly, including all probe sets in an analysis, regardless of the signal/noise ratio, gives greater potential sensitivity. However, this sensitivity comes at a cost of increasingly poor signal/noise ratios. With the Affymetrix algorithms, the relatively high weight placed upon the mismatch penalty enables the assignment of a continuous p value variable to each signal intensity, based upon the confidence that the perfect match and mismatch ratios are significantly different from noise, or not. The RMA algorithms, by largely ignoring the MM probes, are less able to provide a confidence regarding "acceptable" signal/noise thresholds

There are two outputs from the Affymetrix noise calculations; one is the continuous p value assignment, and the other is a simple "present/absent" threshold. When the probe set detection p value reaches a certain level of significance, then the probe set is assigned a "present" call, while all those probe sets with less robust signal/noise ratios are assigned an "absent" call. This enables the use of a "present call" threshold noise filter. In the examples reported here, we used a "10% present call" noise filter. This means that any specific probe set was required to show at least 3 "present" assignments in the 25 microarrays in the project (>10% "present" calls). We have previously reported this "data scrubbing" method in a series of publications, but had not systemically analyzed the effect of this filter on data interpretation and conclusions [4,5,6,7,8].

We hypothesized that it would be possible to identify the most appropriate probe set analysis and noise filter methods by conducting permutational analysis of probe set "signal" algorithms, and "present call" noise filters. The goal was to use unsupervised hierarchical clustering to find the combination of methods that maximized the separation of the "known" biological variable, while minimizing confounding "noise". Interactive exploration techniques available in a visualization tool for microarray data analysis – the Hierarchical Clustering Explorer (HCE) [9] helped us quickly investigate the impact of different linkage methods to the analysis results. HCE load the data set produced by MAS or RMA, apply a filter, let users select a linkage method, and perform the hierarchical clustering. HCE shows an informative overview of the clustering results using color mosaic and dendrogram. At first, users can filter out less significant genes for grouping samples by pulling down a dynamic query control called 'minimum similarity bar' to increase the gene cluster tightness [9]. Then, users can investigate the clustering results of samples by looking at the dendrogram. A compact color mosaic overview of the clustering result and interactive dendrogram display enable the rapid determination of the best methods for signal/noise optimization in the microarray data analysis.

RESULTS AND DISCUSSION

We decided to test a relatively "noisy" experimental system in this study. This consisted of human muscle biopsies from patients with two defined types of muscular dystrophy, and normal controls. All muscular dystrophy patients had a molecularly-defined diagnosis, namely either Duchenne muscular dystrophy (mutations in the dystrophin gene), or Limb-girdle muscular dystrophy (homozygous missense mutation in the FKRP gene). Normal controls were from normal volunteers. We have previously shown that tissue heterogeneity is the dominant variable in human muscle profiles; tissue heterogeneity is a greater variable than the combined effect of inter-individual variation, age, sex, and technical variables (RNA isolation and labeling, microarray hybridization and scanning) [6]. For this reason, all profiles were generated from distinct pieces of muscle. The arrays included 10 Duchenne dystrophy profiles from 10 patients, 7 FKRP profiles from 4 patients, and 8 normal muscle profiles from 8 volunteers. We used stringent quality control criteria, including thresholds for appropriate scaling factors (SF), percent present calls (% PC), and controls for RNA integrity (GAPDH 3'/5' and HSACO7 3'/5' ratios). Both a pre-amplification (stain 1), and post-SAPE amplification (stain 2) scans were studied, with a replacement of any saturated probe sets using a PMT saturation detection and replacement protocol.

Among the 22,283 probe sets on the Affymetrix U133A microarray (~500,000 features), we found a consistent percentage of "present" calls for each of the 25 cRNA samples tested (DMD, 10 arrays, 39.4%±5.7%; FKRP, 7 arrays, 35.5%±2.2%; controls 8

arrays, 35.8%±4.6%). These arrays detected approximately half of the genes in the human genome.

All arrays were analyzed using both Affymetrix MAS 5.0 default settings, and also RMA methods. Data from each profile was converted into a spreadsheet with five columns: probe set name, Affymetrix signal, Affymetrix present/absent call, Affymetrix probe set detection p value, and RMA signal. The present/absent call assignment is based upon the probe set detection p value; the present/absent is a binary threshold value, while the probe set p value is a continuous variable. For the studies below, we only used the present/absent threshold value. Future implementations of our work will use the continuous probe set p value as a "weighting" function.

To compare the effect of the nature of input data (probe set analysis method; noise filtering), and the effect of specific clustering linkage methods (see below), we used our recently described Hierarchical Clustering Explorer (HCE) program, an interactive visualization tool for the hierarchical agglomerative clustering results. Spreadsheets corresponding to each profile were then loaded into a customized version of HCE (*http://www.cs.umd.edu/hcil/multi-cluster/user_manual.html*). Unsupervised hierarchical clustering of the profiles was done using permutations of analysis method (Affymetrix MAS 5.0 vs. RMA), with and without a noise filter (10% present calls).

Hierarchical clustering algorithms have been widely used to analyze expression profile data sets. Among many kinds of hierarchical clustering algorithms, the agglomerative algorithm is a de facto standard for microarray experiment data analysis. Hierarchical agglomerative clustering algorithm is summarized as follows. Let's assume that we want to cluster m data items. Initially, each data item occupies a cluster by itself. Among the current clusters, we find a pair of clusters whose similarity value is the highest, and merge them to make a new cluster. Then, we update the similarity values between the new cluster and the remaining clusters. We continue the merge and update until there remains only one cluster of size m. When the agglomerative hierarchical clustering algorithm updates the similarity values between the newly merged cluster and remaining clusters, many different methods can be used, which are called 'linkage methods.' Our HCE program supports 5 different linkage methods: UPGMA, Average Group, Complete, Single, and One-by-one linkage [9]. For our data set, we studied the effect of 3 methods on our data set. We explain these three methods as follows. Let C_n be a new cluster, a merge of C_i and C_j. Let C_k be a remaining cluster.

1. Average Linkage (UPGMA : Unweighted Pair Group Method with Arithmetic Mean):
$Dist(C_n,C_k)=Dist(C_i,C_k)*|C_i|/(|C_i|+|C_j|)+Dist(C_j,C_k)*|C_j|/(|C_i|+|C_j|)$
2. Average Group Linkage: $Dist(C_n,C_k)=Dist(Mean(C_n),Mean(C_k))$
3. Complete Linkage: $Dist(C_n,C_k)=Max(Dist(C_i,C_k),Dist(C_j,C_k))$

For each gene expression measurement method (Affymetrix MAS 5.0, and RMA), we ran HCE with 3 different linkage methods mentioned above without applying any noise filter. In a second set of runs, we applied a noise filter (10% "present calls") to the data set and re-ran HCE with the same 3 linkage methods. We visualized the unsupervised clustering of the data set to determine the method that provided the best clustering according to our "known" biological variable (specific biochemical defect; patient diagnosis), and thus was most effective in reducing undesired noise.

Overall, we found that the Affymetrix MAS 5.0 probe set analyses were more successful in grouping profiles into the biologically appropriate clusters than RMA. RMA with no noise filters showed poor clustering into the relevant biological groups, suggesting that the higher sensitivity of RMA was leading to a very high degree of noise in the data analyses, and this was independent of linkage method (Figure 1). Imposing a 10%

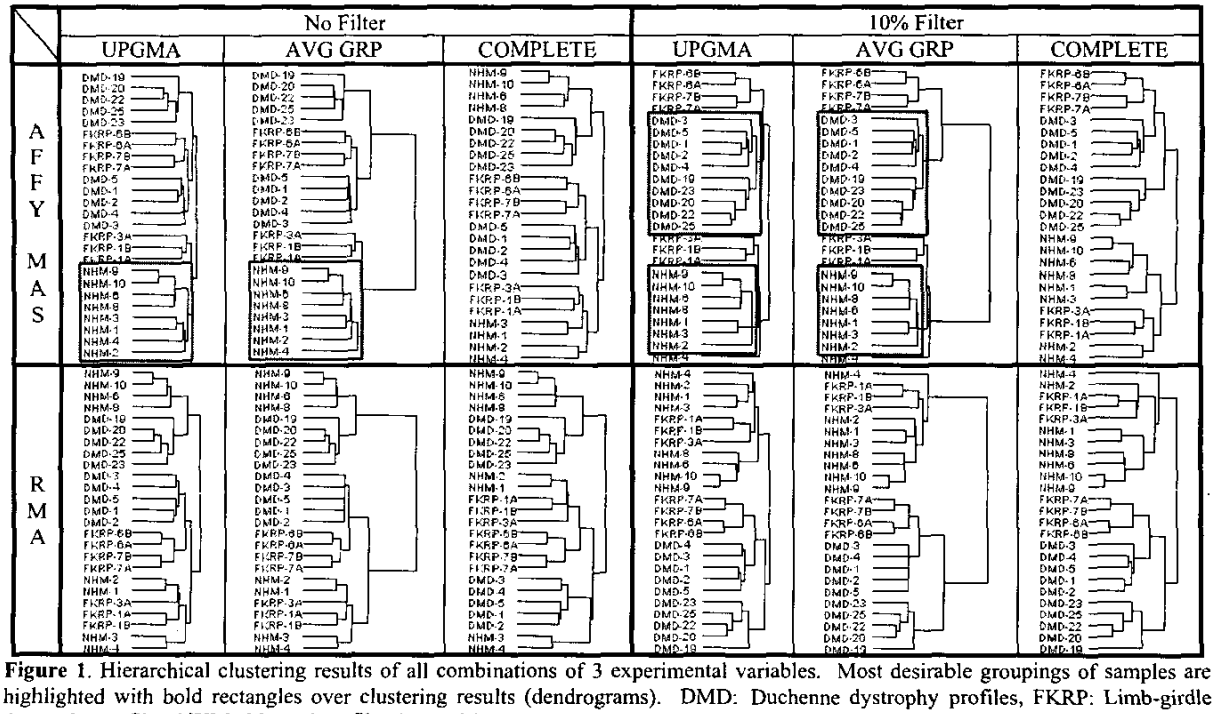
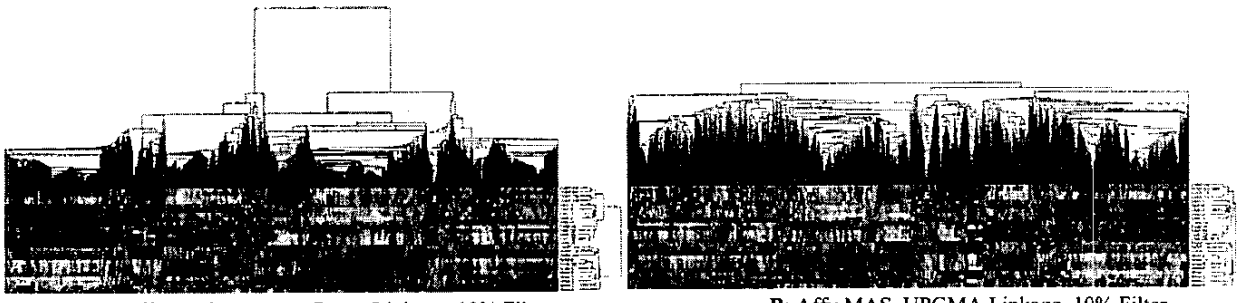


Figure 1. Hierarchical clustering results of all combinations of 3 experimental variables. Most desirable groupings of samples are highlighted with bold rectangles over clustering results (dendrograms). DMD: Duchenne dystrophy profiles, FKRP: Limb-girdle dystrophy profiles, NHM : Normal profiles (controls).



A: Affy MAS, Average Group Linkage, 10% Filter

B: Affy MAS, UPGMA Linkage, 10% Filter

Figure 2. Two most successful combinations of experimental variables and their clustering results (dendrograms). Average group linkage (A) shows better separation between distinctive clusters than UPGMA linkage (B).

present call noise filter did indeed improve the performance of RMA, although the Duchenne dystrophy patients never resolved into a single branch (Figure 1).

Of the three linkage methods, the Average Group Linkage proved the most appropriate for the unsupervised recognition of the biological variable. UPGMA and Average Group Linkage produced basically the same clusters, but the latter showed more clear separation between clusters (Figure 1, Figure 2). The noise filter (10% present calls) was also successful in improving the effect of the biological variable relative to technical noise. This method resulted in correct grouping of both Duchenne dystrophy and normal individuals, however no method was able to group all FKRP patients in one sub-cluster. Of importance is the fact that the FKRP profiles (FKRP-1A,1B,3A) clustered with normal controls showed the least histopathology by microscopic analysis of the tissue, and these patients also showed the least physical disability. Thus, the clustering agrees with the additional variable of "severity" of disease, in addition to the molecular diagnosis.

Our permutation study of probe set analysis (two methods), noise filtering (two methods), and clustering linkage method (3 methods) found that this particular data set was classified most accurately with:

- **Affymetrix probe set analysis method**
- **10% present call noise filter**
- **Average Group Linkage**

It is important to stress that we chose our data set as one that was particularly fraught with uncontrolled variables. Thus, the intrinsic noise in this data set is sufficiently high such that analysis is best done with more stringent criteria, at the concomitant loss of sensitivity. Other data sets that have fewer confounding variables (less sources of noise) are likely to benefit from the greater sensitivity of the RMA method, and may not require the noise filter (10% present calls).

CONCLUSION

In conclusion, we feel that each project should undergo a "signal/noise" analysis with interactive visualization tools that help researchers understand the result with their rapid possibilities for exploration, as we have presented here. By using permutations of probe set signal analysis methods, and noise reduction filters (either % present call thresholds, or future implementations of continuous variable probe set p values), with unsupervised clustering, the analysis method that most faithfully assembles profiles into the appropriate biological groups should maximize the signal from the biological variable, while minimizing the confounding noise intrinsic to the project. This results in a balanced signal/noise assay that should provide the best balance between sensitivity and specificity. Our future plans are to implement an interactive project analysis environment that combines a more extensive and automated project analysis methods with interactive visualization tools, where researchers can systemically try combinations of variables to achieve the best clustering into the desired biological groupings.

Acknowledgements: This work was supported by N01 NS-1-2339 from the NIH.

REFERENCES

[1] E. Hobbell, "Estimating signal with next generation Affymetrix software," In: Gene Logic Workshop on Low Level of Affymetrix GeneChip data, 2001. http:/www.stat.berkeley.edu/users/terry/zarray/Affy/GL_Workshop/genelogic2001.html

[2] B.M. Bolstad, R.A. Irizarry, M. Astrand, and T.P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, Vol. 19 No. 2, pp. 185-193, 2003.

[3] R.A. Irizarry, B.M. Bolstad, F. Collin, L.M. Cope, B. Hobbs, and T.P. Speed, "Summaries of Affymetrix GeneChip probe level data," *Nucleic Acids Res.*, Vol. 31 No. 4 e15, 2003.

[4] P. Zhao, S. Iezzi, V. Sartorelli, D. Dressman, and E.P. Hoffman, "Slug is downstream of myoD: Identification of novel pathway members via temporal expression profiling," *J Biol Chem*; Vol. 277, Issue 33, pp. 30091-30101, Aug 16, 2002.

[5] M. Bakay, P. Zhao, J. Chen, and E.P. Hoffman, "A web-accessible complete transcriptome of normal human and DMD muscle," *Neuromuscular Disorders*, Vol. 12, pp. S125-S141,2002.

[6] M. Bakay, Y.W. Chen, R. Borup, P. Zhou, K. Nagaraju, and E.P. Hoffman, "Sources of variability and effect of experimental approach on expression profiling data interpretation," *BMC Bioinformatics*; Vol. 3, pp. 4-15, 2002.

[7] S. DiGiovanni, S.M. Knoblach, C. Brandoli, S.A. Aden, E.P. Hoffman, and A.I. Faden, "Temporal gene profiling after experimental spinal cord injury identifies cell cycle genes associated with neuronal damage and cell death," *Annals. Neurol.*, in press.

[8] D.S. Hittel, W.E. Kraus, and E.P. Hoffman, "Skeletal muscle dictates the fibrinolytic state after exercise training in overweight men with characteristics of metabolic syndrome," *J Physiol.*, in press.

[9] J. Seo and B. Shneiderman, "Interactively Exploring Hierarchical Clustering Results," *IEEE Computer*, Vol. 35, No. 7, pp. 80-86, July 2002.