Adjustment of systematic microarray data biases

Monica Benito[1], Joel Parker[2], Quan Du[3], Lambert Skoog[3], Annika Lindblom[3], Charles M. Perou[2#] and J. S. Marron[4#]

[1]Department of Statistics and Econometrics
University of Carlos III
Madrid, Spain

[2]Department of Genetics and Pathology
Lineberger Comprehensive Cancer Center
University of North Carolina
Chapel Hill, NC 27599-7264

[3]Department of Molecular Medicine
Karolinska Institutet
S 17176 Stockholm
Sweden

[4]Department of Statistics
University of North Carolina
Chapel Hill, NC 27599-3260

corresponding authors

ABSTRACT

Motivation: Systemic differences due to experimental features of microarray experiments are present in most large microarray data sets. Many different experimental features can cause these biases including different sources, different production lots of microarrays, or different microarray platforms. These systematic effects present a substantial hurdle to the analysis of microarray data.

Results: We present here a new method for the identification and adjustment of systemic biases that are present within microarray data sets. Our approach is based on modern statistical discrimination methods and is shown to be very effective in removing systematic biases present in a previously published breast tumor cDNA microarray data set. The new method of "Distance Weighted Discrimination" is shown to be better than the Support Vector Machines (SVM) and Singular Value Decomposition (SVD) for the adjustment of systematic microarray effects and is shown to be of general use as a tool for the discrimination of systematic problems present in microarray data sets including the merging of two similar breast tumor data sets.

Availability: ??? Could we make something available?? Yes, a (hopefully) pretty good Matlab version is now ready for Joel, and/or George to take for a test drive. It might be good if this could happen before we actually submit.

Contact: marron@email.unc.edu Since we are both corresponding authors, shouldn't your email be here, too?

Supplementary Information: the complete figures that represent the cluster diagrams discussed in Figure 11 are available at http://www-unc-something Let's decide soon if this goes on your site or mine. My first guess is that yours is better…. Also we should put the software there, too.

One more point: Sometimes we say "systemic" and sometimes "systematic". Those don't seem like exactly synonyms to me, and probably we should be using just one, but it is not clear to me which is best. Maybe it is clear to you when each of these should be used? If so, you might want to do an editor search to just check that these are right….

INTRODUCTION

DNA microarrays are a powerful tool for the study of complex systems and are being applied to many questions in the biological sciences. In particular, the study of human tumors using patterns of gene expression have identified many expression differences that can predict important clinical properties like the propensity to relapse (van't Veer et al. 2002) or predict the survival outlook for a patient (Sørlie et al., 2001). These types of clinical sample studies are particularly challenging as the microarray experiments are often performed over many months because sample collection is prospective, with most samples being assayed soon after they are collected. These studies also have the additional challenge of assaying samples/tumors that are collected and processed at different institutions, and hence, systematic biases due to different handling procedure are typically present in these studies.

For many reasons, different types of systematic biases can be identified within microarray data sets. These biases are manifested as differences in gene expression patterns when one set of microarrays is directly compared to a second set of microarrays. When using "supervised" statistical analyses, these systematic biases show themselves as a subset of genes that tend to be more "highly expressed" in one set of microarrays versus another, and a concomitant subset of genes that are lower in expression in one set versus the other. These biases can typically be identified because they perfectly correlate with non-biological properties like where the samples were isolated and processed (source bias), or what print batch of microarrays the samples were tested on (batch effect). As can be expected, these systematic biases compromise the integrity of the data, and are especially troubling in experiments in which many samples are assayed over a long time period as these studies get assayed on many different print batches of microarrays.

Others have used Singular Value Decompositions (SVD) to correct for systematic biases in a data set of yeast cell cycle experiments (Alter et al., 2000), and to correct for microarray batch bias in a data set containing many soft tissue tumors (Nielsen et al., 2002). We present here a new method, called "Distance Weighted Discrimination (DWD)", (Marron and Todd 2002), which can be used to adjust microarray data sets to compensate for systematic biases. We examined our previously published breast tumor data set (Perou et al., 2000 and Sørlie et al., 2001) containing 107 cDNA microarray experiments and identified two distinct experimental biases. To evaluate the robustness of this new analysis technique, we applied DWD to this data set and showed a significant reduction in the source bias, and in the microarray batch bias. We also present data which suggests that this approach can be used to make adjustments for other systemic biases including across platform effects, which suggests that DWD presents a new and powerful method for adjusting microarray data sets for systematic artifacts.

SYSTEMS AND METHODS

2.1 Hypothetical discrimination based adjustments

One way of understanding the problems with SVD/PCA for removal of systematic effects is to recall that SVD/PCA seeks only to find "directions of greatest variation". When this goal is consistent with the systematic biases effect (meaning the systematic bias effect generates more variation than any other parts of the data, as measured by the sums of squares), then good results will be obtained using SVD/PCA. This appears to have driven the positive results reported by Alter, Brown and Botstein (2000) and Neilsen et al (2002). However, when the magnitude of the systematic effect variation is similar to other components of variation, as is seen in Figure 5 (or perhaps even smaller as seen in Figure 7), then this approach can easily fail. ??? Referencing Figures that appear later seems a little weird to me, but I get the idea that you prefer not to go into the detailed explanation of these at this point, which is OK for me… In these situations, where the first SVD/PCA direction is not appropriate for bias adjustment, a natural way to improve the analysis is to make full use of the systematic bias information (i.e. each case is known to belong to a particular batch, or known to be derived from a given source). Then instead of choosing directions to maximize variation in the full population (the goal of SVD/PCA), it is natural to choose directions to maximize separation of the bias. These

points are illustrated, using a toy example of source effect, in Figure 1. This toy example is only two dimensional (i.e. only two genes are considered), to make it easy to visualize the data "point cloud". Note that the two subpopulations (shown in red and blue) are quite separate from each other, and also have similar distributions (i.e. the same population shape), so that a simple translation would be able to remove any differences between the populations. The main goal of this paper is to find effective ways of finding the direction (and magnitude) of this translation.

The direction vector of the first Principal Component (i.e. the SVD direction) for these data is overlaid as the long thick black line in Figure 1. Note that this direction is clearly wrong for our goal of removing the difference between these populations. In particular, when the data are projected onto this direction vector, the subpopulations will overlap. The reason is that the PC1 direction is the "direction of greatest variation in the data", which in this case is quite different from effective source adjustment. Also overlaid is the Fisher Linear Discrimination (FLD) direction. Note that this direction is correct for removal of the source effect. In particular, when each source is shifted in this direction, by an amount determined by the source subpopulation means, then the distributions will be indistinguishable. The reason that FLD works much better is that it exploits the source labels, which are ignored by PCA/SVD.

In addition to finding better directions for source effect adjustment, we recommend another important improvement over the SVD adjustment. Instead of completely subtracting all variation in the chosen direction (as is done with the usual SVD approach), we only subtract the subpopulation means of the data projected on the given direction. This preserves any variation in this direction that is not caused by source effects, instead of squashing out all structure in this direction as is done by subtracting the first PC direction (particularly dangerous in SVD contexts, since the first PC direction is chosen to contain "maximal interesting structure"). In Figure 1, this corresponds to shifting the subpopulations so that they overlap, instead of projecting the data onto a single line.

While FLD is very effective for the toy data shown in Figure 1, it has less desirable properties for more realistic data contexts like microarrays, as shown in Figure 2. In particular, FLD has poor performance in High Dimension, Low Sample Size (HDLSS) contexts. This problem not only arises for microarray data, but also appears in other statistical contexts such as medical image analysis, and chemometrics. HDLSS data pose a very serious challenge to most classical statistical multivariate analysis settings (such as FLD), because the first step in those analyses ("sphering the data" by multiplying by the root inverse covariance matrix) fails, since the covariance matrix is not full rank. This point is illustrated in Figure 2A, which shows a different toy example, this time in 50 dimensions. The data are all simulated Gaussian, with independent components and unit variance. All of the mean vectors are zero, except in the first component where there are 20 data points (shown as plusses) with mean +2.2, and 20 data points (shown as circles) with mean -2.2. The projections of these 50 dimensional vectors onto the first component is shown in Figure 4A, as "jitterplots" (meaning random heights are used to provide visual separation of the points, Tukey and Tukey 1990), with smooth histograms (see Wand and Jones 1995) overlaid. While the subpopulations are clearly separated in this

plot, it can be quite challenging to find this direction because of the relatively high noise level and high dimensionality (a familiar situation in microarray analysis).

Figure 2B shows the results of FLD for these data. The implementation is done with a generalized inverse of the full sample covariance matrix. The shape of the projected data sets look quite different from the projections in Figure 2B, with all of the data from each class lying essentially on top of each other. This is because FLD seeks to find the direction that maximizes the separation of the classes, relative to the spread within the classes. Because there are only 40 data points in 50 dimensions, it is not surprising that this type of "perfect separation" is possible. However, note that the subpopulation shapes are much different from those in Figure 2A, which represents the optimal direction for discrimination (i.e. the direction that will work the best for discriminating new data). The angle of the FLD direction (i.e. 58 degrees), to the optimal is also shown. This shows that FLD has found a spurious direction, and is driven by sampling artifacts that will change completely for a different set of data. Essentially FLD is "feeling random artifacts in this particular data too strongly", and so this direction will suffer from poor generalizability as a discrimination rule. This problem can be viewed as over fitting of the data.

Another approach to this problem is to use Support Vector Machines (SVM), discussed in detail in Section 2.3. The performance of the SVM, for the 50 dimensional toy data, is shown in Figure 2C. Note the projected data are no longer completely piled up, and that the angle to the optimal is substantially better, reduced to 36 degrees. However, there is still substantial data piling at the margin (the interior points where data from both classes tend to accumulate), which is quite reminiscent of the over-fitting problem of FLD illustrated in Figure 2B. Again there is a suggestion that FLD can also be "feeling too many sampling artifacts".

Marron and Todd (2002) have addressed this problem by the development of Distance Weighted Discrimination (DWD), discussed in Section 2.4, and illustrated in Figure 2D. Note that now the subpopulations appear more spread (as for the optimal projection in Figure 2A), and also the direction has a smaller angle to the optimal direction, now only 26 degrees. Because of this strong performance in HDLSS situations, DWD is recommended for both this type of systematic artifact adjustment, and for other supervised learning (i.e. statistical discrimination) tasks for microarray data. An additional advantage of using DWD for systematic artifact adjustment is that the projected subpopulation shapes look more Gaussian, so that the subpopulation means, used in the adjustment, are more appealing as notions of "population center".

2.2 Microarray production, hybridizations and initial data processing.
All microarrays and samples used in this study have been previously published; the experiments used in Figures 3-8 were taken from the Stanford Microarray Database (SMD) and are described in Perou et al. 2000 and Sørlie et al. 2001. The remaining examples illustrate the effectiveness of DWD cross platform adjustment, where the goal is to combine this Stanford data set with data from van't Veer et al. 2002, which are available at the Rosetta Inpharmatics website.

We first performed a number of gene filtering steps before any analyses were done. First, for all data obtained from the SMD, we filtered all genes for a signal intensity of 50 or greater in both the red and green channels and insisted that this signal intensity criteria be present on 70% or more of the 107 experiments for each gene. Next, we took the log2 transformed normalized R/G ratio for each gene on the microarray. The missing values in this data table were imputed using the KNN-impute feature contained within the Significance Analysis of Microarrays excel plug-in program (Tusher et al., 2001 and Troyanskaya et al. 2001). This imputed data set was then used for all analyses.

2.3 Algorithms – Support Vector Machines (SVM)

The SVM is a powerful discrimination method initially proposed by Vapnik (1982, 1995). Also see Burges (1998) for an easily accessible introduction, Cristianini and Shawe-Taylor (2000) for a detailed introduction, and http://www.kernel-machines.org/. The essential idea is to find a hyperplane that separates the two classes (i.e. each systematic bias) as well as possible. When the data are "separable" (meaning prefect separation is possible), then the hyperplane is chosen to maximize the minimum distance of all of the data to the hyperplane. The minimizing distance is called the "margin". An interesting view comes from studying the normal vector of the separating hyperplane, and the projection of the data upon that. This is the view shown in Figure 2C. The interior points where the data pileup shows the margin. The SVM can be viewed as optimizing the direction vector to maximize the size of this margin. When the data are not separable, penalty terms (for those data points on the wrong side of the boundary) are added to the optimization problem, but it is still accessible to standard quadratic programming methods. The non-separable case is usually not particularly important in HDLSS situations such as microarray analysis. This projection of the data onto the SVM normal vector, for the data of Figure 5, is shown in Figure 3. The effect is perhaps surprisingly similar to Figure 2C. Again note that the use of the means of the projections shown in Figure 3, for adjustment in this direction, is not very attractive, because both distributions look quite skewed (in opposite directions). When means are subtracted, to adjust for the systematic effect, the population shape will be rather strange in this direction.

Note that the SVM direction represents an improvement over anything based on SVD, with the two sources far more separated than can be seen in any PC direction in Figure 5 (especially in the PC1 direction where there is considerable overlap. Thus a major improvement of SVM over SVD for source adjustments, is demonstrated for this data set. This comes from the fact that SVM is essentially aggregating over all useful directions. In Section 2.4, a further improvement, based on Distance Weighted Discrimination, is proposed. This method finds a direction with a similar large spread between the batches, and gives subpopulations with a more attractive Gaussian-type shape, as suggested in Figure 2D.

2.4 Algorithms - Distance Weighted Discrimination (DWD)

Distance Weighted Discrimination was initially proposed by Marron and Todd (2002). The goal is to improve the performance of the SVM in HDLSS contexts, as illustrated in Figure 2C. The main idea is to improve upon the criterion used for "separation of classes" in the SVM. The SVM has data piling problems along the margin, because it is

maximizing the minimum distance to the separating plane, and there are many data points that achieve the minimum. A natural improvement is to replace the minimum distance by a criterion that allows all of the data to have an influence on the result. DWD does this by maximizing the sum of the inverse distances. This results in directions that are less adversely affected by spurious sampling artifacts, as shown in Figure 2D.

Figure 4A shows the projection of the data onto the DWD direction for the same data as used in Figures 5 and 3. As one would expect from Figures 2D and 3, the sources are still well separated. A careful look at the horizontal scales shows that the "average population separation" is even larger in Figure 4A than it is in Figure 3. Furthermore these subpopulations now look much more symmetric (even more Gaussian), so the subtraction of respective subpopulation means in this direction will remove the source effect in an appealing manner.

The specifics of the batch adjustment (thinking of the data as vectors with entries corresponding to genes) are:

i. The DWD direction vector is found
ii. The subpopulations (e.g. respective source subsets) are all projected in that DWD direction.
iii. The subpopulation projected means are computed
iv. Each subpopulation is shifted in the DWD direction, by an appropriate amount, through the subtraction of the DWD direction vector multiplied by each projected mean.

Figure 4B checks the performance of DWD as a systematic bias effect removal tool, by applying the same DWD based method to the source adjusted data. Note that this time DWD does not even find a direction where the data are separated. Another verification of the good performance of DWD is the elimination of the source effect shown in Figure 6, where the different sources appear to be randomly intermingled. The relative behavior of SVM and DWD shown here is very typical of a number of other examples that we have studied. Some of these are shown in Section 3 and include adjustments for microarray print batch effects, and even for microarray experiments based on different platforms.

3.1 Implementation of DWD to adjust for sample source bias

We identified in our previous microarray data set, a set of genes whose expression nearly perfectly correlated with where the samples came from (i.e. Stanford University or Norway); we do not believe that this set of genes is due to true biological differences, but that it is instead, due to the systematic differences in how the sample RNAs were prepared. Useful views of this data can be based upon Singular Value Decompositions (SVD), which is equivalent to Principal Component Analysis (PCA). Straightforward understanding of this analysis comes from thinking about the vectors of gene expressions, for each case, as points in a high dimensional point cloud. SVD and PCA can be viewed as finding "interesting directions" for understanding the structure of the point cloud. More precisely, they find "directions of greatest variability". A view that makes the source effect problem more clear is shown in Figure 5. This figure shows a matrix of plots of one and two dimensional PCA projections. The plots on the diagonal show the 1-

d projections (commonly called "principal component scores") of the data onto each of the first four eigenvectors (i.e. the directions of interest in the point cloud). The individual microarray experiments are shown as colored dots, where the colors indicate the two different sources of breast tumors used in our previous studies (i.e. Norway or Stanford). The horizontal axis shows the PC scores (an axis with the numerical values is not shown, because these numbers are not particularly interpretable), and the vertical axis shows a random height used for visual separation (the same "jitter plot" visualization used in Figures 3 and 4). The black curves in the 1-d diagonal projection plots are smoothed histograms (again as in Figures 3 and 4). The off diagonal graphics all show the 2-d projections onto different pairs of eigenvectors (directions in the point cloud space) as scatterplots, with the x-axis corresponding to the component whose 1-d projection is directly above or below, and with the y-axis corresponding to the component whose 1-d projection is directly to the right or left. Thus Figure 5E is a "flip about the 45 degree line" of Figure 5B, and both of these show how the first PC direction relates to the second.

Note that in Figure 5A, the red and blue points are somewhat separated. The approach suggested by Alter et al. (2000) and Nielsen et al., (2002) is to remove this source effect by subtracting this PC direction from the data. However, for this data set, there is substantial overlap of source effects in the PC1 direction, suggesting that deeper investigation would be useful. A stronger suggestion that this is the case comes from Figure 5B, which compares the first and second eigen directions (i.e. PC1 and PC2). Note that better separation between the red and blue subpopulations is possible when using a diagonal separating line, than using a horizontal line that would be entailed from using only the PC1 direction. This casts doubt on the approach of simply removing the first principal component from the data; in particular, removal of some linear combination of the first and second directions (i.e. a slanted line in the plot) should provide a better source adjustment. This opens the question of finding other directions, which may be more appropriate for source adjustment.

A main goal of this paper is to present some improved approaches to finding directions which better separate the data than the single first PC. The result of our "source effect" removal using DWD, is shown in Figure 6. Now the colors, representing the two sources, are very well mixed, meaning that the systematic sample source effects in the data have been very effectively removed. The same is true for higher order PC components (we have looked at orders up to 8, but these are not shown to save space). Our result is better than that where just the first eigen vector is removed, as recommended by Alter, Brown and Botstein (2000) and Neilsen et al (2002), which is summarized in Figures 5 F, G, H, J, K, L, N, O and P, i.e. the plots below the top row and to the right of the first column in Figure 5. For example, Figure 5 H shows a strong systematic effect still present in the data. The good results in Figure 6 can be viewed as appropriately summarizing all of the directions in Figure 5, that show a need for adjustment, as well as many other directions not shown here. This summarization effect is why the visual separation apparent in Figure 4 is much more than any seen in Figure 5.

3.2 Implementation of DWD to adjust for other systemic biases

In this section, additional examples are considered, which show that the superiority of DWD for source adjustment over SVD approaches is not a fluke of the particular data set under consideration. The first of these is another systematic microarray bias, known as the "batch effect". Most spotted DNA microarrays, particularly those produced at academic facilities, are physically produced in groups of 100-200 due to the number of locations that are available on the microarray robot printing platter (see the "M guide" at http://cmgm.stanford.edu/pbrown/mguide/index.html for robot details). A given "print run" or "batch" of microarrays tends to show a "batch bias", which is manifested as a set of genes whose high or low expression perfectly correlates with what batch the sample was assayed on. This effect can be relatively small on some batches and very significant on others, however, it has been our experience that nearly every batch of microarrays shows some systematic batch bias.

Figure 7 shows essentially the same PCA scatter plots as in Figure 5, using the same set of 107 breast tissue experiments, except this time the data points are colored according to microarray "batch" (three batches or different print runs of microarrays were used). As in Figure 5, it is clear that there is a systemic effect of batch on the structure of the data. However, note that this time, the effect appears most markedly in the 4th eigen direction, Figure 7P. It is clear that in this case the classical SVD batch adjustment (based on only the first eigen direction) would be ineffective at removing this batch bias.

All of the methods discussed above apply to two class discrimination, but this data set came from three different batches, i.e. three different classes. To address this additional level of complexity, which is common in many microarray data sets (for example samples coming from three different sources), we took a step-wise approach. An inspection of Figure 7 shows that in the PC4 direction, the very small Batch 1 (red) appears more consistent with Batch 2 (green). Hence, we first made a batch adjustment between Batches 1 and 2 (combined) and Batch 3 (blue). Next we apply the same method to the adjusted data, to separate Batch 1 from Batch 2. Because these data also have a source effect, as illustrated in Figure 1, a third step, removing that source effect as well, is also sensible. The result of the three step process, shown in Figure 8, reveals subpopulations that are now well intermingled (i.e. the batch effect has been successfully removed). Analogs of Figures 3 and 4, for these adjustments, show quite similar lessons: the DWD gives excellent separation and good subpopulation shapes, whereas the SVM separated similarly well, but with the same resulting less appealing skewed projected subpopulation shapes. Because the lessons are so similar, these plots were not included.

One of the most pressing challenges in the microarray field is how to combine two data sets that came from two different groups, and which utilize different microarray platforms. In this scenario, many different systematic biases will be present including microarray batch effects (which in this case will be even greater due to different microarray platforms), source effects as each group will utilize a different source of experimental samples, different RNA extraction protocols, and other potentially unknown sources of systematic effects. As briefly discussed above, there are a number of studies that have used DNA microarrays and a two-color experimental design, to study the gene expression patterns coming from grossly dissected human breast tumors (Perou et al

2000/Sørlie et al. 2001 and van't Veer et al 2002); the combined data set of Perou and Sørlie was utilized in the earlier figures and consisted of 107 samples representing 78 grossly dissected breast tumors that were assayed using mRNA with direct labeling on cDNA microarrays produced at Stanford University (and which were assayed versus a cell line pool common reference sample). The van't Veer et al 2002 data set contained 117 grossly dissected breast tumor samples that were labeled using the linear amplification of total RNA, and which were assayed on Agilent long oligo DNA microarrays (and which were assayed versus a common reference that was a pool of 50 tumors).

Figure 9 shows the PCA representation of the raw data. Again these two data sets are so different that simple SVD adjustment appears to offer a reasonable adjustment. However, note that both the second and third eigen directions appear to suggest some improvement (again slanted lines give better separation than horizontal ones in Figures 9B and 9C), so improvement is expected from the DWD method. We next adjusted the data using DWD and one view of the adjusted data is shown in Figure 10. Note that the red and blue populations now have very good overlap, indicating a successful adjustment. Figure 10 also shows why earlier attempts at this adjustment, based on simple mean based methods, were not successful: there is a substantial outlier (visible in both the PC2 and PC3 projections). A strength of DWD, over mean based methods for bias adjustment, is its reduced sensitivity to such outliers.

One goal of our breast tumor studies was to identify the natural diversity of tumor subtypes present, and to accomplish this goal we identified a set of genes that we termed the "intrinsic" gene set (Perou et al. 2000), which when used to group breast tumors using hierarchical clustering analysis as implemented by Eisen et al. (1998), identified subsets of tumors/patients that predicted overall patient survival (Sørlie et al. 2001). The data displays presented in Figures 9 and 10 are suggestive of good integration, however, we wished to perform a combined hierarchical clustering analysis of the Stanford and van't Veer et al data sets because these two data sets represent similar microarray analyses, namely two-color microarray experiments done on grossly dissected human breast tumors.

In the combined data set cluster analysis, the common set of intrinsic genes across both data sets was determined (311/478 genes) ??? is it 311 or 478? ??? ; next each data set was separately imputed using the KNN-impute program of Troyanskaya et al. 2001, and then each gene was median centered within each data set. We next combined the data sets and performed a two-way average linkage hierarchical cluster analysis using the program "Cluster" and have displayed the data using "TreeView" (M. Eisen and colleagues (http://rana.lbl.gov/EisenSoftware.htm)). The "adjusted" and combined data set differed in that after each data set was imputed, we used DWD to adjust the Stanford to the van't Veer data set as shown in Figure 10, then we took the adjusted data and median centered each gene across all of the data and clustered.

As can be seen in Figure 11A, before adjustment, there was very little intermixing of the Stanford (Blue line) and van't Veer (Red line) samples as judged by examination of the

hierarchical cluster sample associated dendrogram (the full cluster diagrams, with complete gene names are available as supplementary materials Figure 12 and 13); even when there was mixing, these samples showed low correlations with the other samples in their dendrogram branches as evidenced by the length of the branches. After DWD adjustment, however, there was a great deal more intermixing of the Stanford and van't Veer samples (Figure 11B); in particular, the left most dendrogram branch in the unadjusted data (Figure 11A) contains many of the ER-positive tumors and was broken into two sub-branches, each of which was almost entirely composed of samples from one source. The corresponding ER-positive branch in the adjusted data (Figure 11B) was also on the left and shows a much greater degree of source intermixing, and the gene expression data itself showed more continuity across the luminal-ER positive expression cluster.

CONCLUSION

We have proposed a new method, based on Distance Weighted Discrimination, for adjustment of various differences between microarray experiment subpopulations. It is seen from several viewpoints that in many cases the new method can provide large improvement over previously proposed methods based on subtracting the first eigen direction from the data using SVD analysis. The new method worked well making adjustments for a number of distinct types of systematic effects including source and batch effects. An even more powerful application, however, was the use of DWD to remove or lessen, the many systematic biases that are present across similar data sets that were generated in different laboratories using different microarray platforms. The message observed from the PCA projection visualization, that DWD successfully removed this platform effect, is confirmed using clustering dendograms. We recommend DWD as a general approach for removing systematic bias effects from microarray data.

References

Alter, O., Brown, P. O. and Botstein, D. (2000) Singular value decomposition for genome-wide expression data processing and modeling, Proceedings of the National Academy for the Sciences, U S A, 97, 10101.10106.

Burges, C. J. C. (1998) A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery, 2, 955-974, see also web site: http://citeseer.nj.nec.com/burges98tutorial.html.

Cristianini, N., Shawe-Taylor, J. (2000) Introduction to Support Vector Machines, Cambridge University Press, Cambridge, United Kingdom.

Eisen, M.B. and Brown, P.O. (1999) DNA arrays for analysis of gene expression. Methods Enzymol, 303, 179-205.

Gollub, J. et al. (2003) The Stanford Microarray Database: data access and quality assessment tools. Nucleic Acids Res, 31, 94-6.

Marron, J. S. and Todd, M. J. (2002) Distance Weighted Discrimination, unpublished manuscript, internet available at: http://www.optimization-online.org/DB_HTML/2002/07/513.html.

Nielsen, T. O., West, R. B., Linn, S. C., Alter, O., Knowling, M. A., O.Connell, J. X., Zhu, S., Fero, M., Sherlock, G., Pollack, J. R., Brown, P.O., Botstein, D., van de Rijn, M. (2002) Molecular characterization of soft tissue tumours: a gene expression study, Lancet, 359, 1301-1307.

Perou, C.M. et al. (2000) Molecular portraits of human breast tumours. Nature, 406, 747-52.

Sørlie, T. et al. (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy for the Sciences, USA, 98, 10869-74.

Troyanskaya, O. et al. (2001) Missing value estimation methods for DNA microarrays, Bioinformatics, 17, 520-5.

Tukey, J., and Tukey, P. (1990). Strips Displaying Empirical Distributions:Textured Dot Strips. Bellcore Technical Memorandum.

Tusher, V.G., Tibshirani, R. & Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy for the Sciences, U S A, 98, 5116-21.

van .t Veer, L.J. et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer, Nature, 415, 530-6.

Vapnik, V. N. (1982) Estimation of Dependences Based on Empirical Data, Springer Verlag, Berlin (Russian version, 1979). 18

Vapnik, V. N. (1995) The Nature of Statistical Learning Theory, Springer Verlag, Berlin.

Wand, M. P. and Jones, M. C. (1995) Kernel Smoothing, Chapman and Hall, New York. 19

Figure 1: Toy example showing how PCA directions can be wrong for batch adjustment, motivating methods based on discrimination ideas.

Figure 2: 50 dimensional Gaussian toy example, to illustrate HDLSS failing of FLD, and superior performance of DWD over SVM.

Figure 3: Projection of data from Figure 5, onto the normal vector of the SVM separating plane. Shows good separation of subpopulations, but data are piled up at margin.

Figure 4: Application of DWD to same data as in Figures 3, 5 and 6. Shows both good separation, and also reasonable subpopulation shape for mean shift adjustment.

Figure 5: PCA projection scatterplot matrix, showing 1-d (diagonal) and 2-d projections of data onto Principal Component directions, of raw Stanford data. Groupings of colors indicate serious source effect problems.

Figure 6: Scatterplot matrix of PCA projections, after DWD adjustment, of Stanford data. Random dispersion of colors (instead of clustering as in Figure 5) shows that source adjustment was effective.

Figure 7: PCA projection scatterplot matrix of raw Stanford data, using batch colorings. Groupings of colors this time indicate serious batch effect problems, in a way that leaves conventional PC1 adjustment completely ineffective.

Figure 8: Scatterplot matrix of PCA projections, after DWD adjustment, of Stanford data. Random dispersion of colors again indicates adjustment was effective.

Figure 9: PCA projection scatterplot matrix of raw combined Stanford - van't Veer et al data. Strong grouping by colors highlight the major differences between these platforms.

Figure 10: Scatterplot matrix of PCA projections, for the adjusted Stanford - van't Veer data. Overlap of the color groups shows an effective adjustment.

Figure 11: Clustering dendogram analyses, with the the van't Veer et al cases shown in red, and the Stanford cases blue. Figure 11A shows that simple median recentering, provides inadequate mixing across platforms, resulting in red-green patterns driven by batch effect. However, the DWD platform adjustment, resulting in Figure 11B, does gives excellent mixing of the platforms, resulting in red green patterns of biological significance.