U Statistics for Microarrays: Normalization, Signal Value Estimation, Gene Expression Profiles

Mayte Suárez Fariñas, Asifa Haider, Knut M. Wittkowski

Summary

When assessing genomic profiles, it is rare that a single gene is sufficient to represent all aspects of genetic activity. Since complex systems tend to be neither linear, nor hierarchical in nature, but correlated and of unknown relative importance, the assumptions of traditional multivariate statistical methods can often not be justified on theoretical grounds. Establishing validity through empirical validation is not only problematic, but also time consuming. This paper proposes the use of u-statistics for scoring multivariate ordinal data and a family of simple non-parametric tests for analysis. The scoring method is demonstrated to be applicable to scoring profiles genomic that best correlated with complex responses to an intervention (treatment of psoriasis). Finally, we will demonstrate that the same methodology also leads to less biased estimates for (low-level) signal value estimation in microarray data. We apply this approach to correlating activity of anti-inflammatory drugs along genomic pathways with disease severity of psoriasis based on both clinical and histological parameters.

Key words: multivariate, rank test, ordinal, genomic, profile, risk, microarray

1. INTRODUCTION

When analyzing complex phenomena by means of statistical methods, a single measure does often not appropriately reflect all relevant aspects to be considered, so that several measures of influences and/or outcomes need to be considered. Sometimes the definite measure is not easily obtained, so that a set surrogate measures has to be evaluated. At other times, e.g., when the aim is to ameliorate a complex phenomenon, a definitive measure may not even exist. Such problems may arise in many applications, although here we focus on low and high level gene expression analysis.

In our main example, we will focus on the effect of treatment on chronic diseases, in general, and psoriasis, in particular. Psoriasis is a skin disease caused by activation of multiple cell types including keratinocytes, vascular cells, and various types of leukocytes. Treatment efficacy can be measured by histological criteria, by intradermal expression of inflammatory cytokines, or by clinical characteristics, such as redness (vascular response) and scaling (keratinocyte response). Since the advent of micro arrays, researchers are now interested in genes whose expression is controlled in a concerted fashion and related to the response.1

Most multivariate methods are based on the linear model, either explicitly, as in regression, factor, discriminant, and cluster analysis, or implicitly, as in neural networks. One scores each variable individually on a comparable scale, either present/absent, low/intermediate/high, 1 to 10, or z-transformation, and then defines a global score as a weighted average of these scores. In other words, data are interpreted as points in a Euclidian space of (independent) dimensions. The number of dimensions is reduced by assuming the dimensions to be related by a specific function of known type (linear, exponential, etc.), allowing one to determine for each point the Euclidian distance from a hyperspace.

While mathematically elegant and computationally efficient, this approach has shortcomings when applied to real world data. Since the relative importance of the variables, the correlation among them, and the functional relationship of each variable with the immeasurable latent factor ‘efficacy’, ‘safety’, ‘risk’, or ‘overall usefulness’ are typically unknown, construct validity cannot be established on theoretical grounds. Instead, one needs to resort to empirical ‘validation’, choosing weights and functions to provide a reasonable fit with a ‘gold standard’ when applied to a sample. While this allows for a comparison between studies where the researchers agreed to use the same scoring system, the diversity of scoring systems used attests to the subjective nature of this process.

Even when the assumptions of the linear model regarding the contribution to and the relationship with the underlying immeasurable factor are questionable, as in genomics, it is often reasonable to assume that the expression of each gene has at least an ‘orientation’, i.e., that, if all other conditions are held constant, an increase in this gene’s expression is either ‘good’ or ‘bad’. The direction of this orientation can be known (hypothesis testing) or unknown (selection procedures). A higher expression of several related genes may indicate increased disease activity.

When we were faced with the analysis of anal vs. vaginal contacts as risk factors for sexual transmission of HIV,3 we presented a partial ordering for dealing with graded and ungraded variables, which allowed to incorporate preexisting knowledge that anal contacts carry more risk without having to ignore the number of vaginal contacts reported. Using the marginal likelihood principle with this partial ordering, we developed a non-
parametric method to assess the overall risk of HIV infection based on different types of behavior or the overall protective effect of barrier methods against HIV infection. More recently, we applied this approach to assessing immunogenicity in cancer patients. In short, one determines all rankings compatible with the partial ordering of the observed multivariate data and then computes a vector of scores as the average across these rankings. While this constituted the first objective approach to the analysis of multivariate ordinal data, because it did not rely on questionable assumptions, it lacked computational efficiency. The computational effort required could be prohibitive even for moderately sized samples, let alone micro arrays with thousands of genes being considered.

Here, we propose a closely related approach based on u-statistics, which is computationally more efficient. With this approach, individual analyses can often be performed even using spreadsheet software and screening for optimal subsets of explanatory variables becomes feasible without the restrictions imposed by commonly used hierarchical strategies. This approach entails a family of simple non-parametric statistical tests. For censored data, the resulting tests reduce to those of Gehan, Schemper, Finkelstein-Schoenfeld and Moye. The proposed family of tests applies to stratified designs with two or more treatments, including the Wilcoxon/Mann-Whitney (WMW) test, the Kruskal-Wallis test, and the Friedman test. It also allows for Scheffe-type multiple comparisons.

2. U SCORES

Our aim is first to develop a computationally efficient procedure to score multivariate ordinal data. We then present simple non-parametric tests for comparing these scores between groups, with an option for stratification. We will not make any assumptions regarding the functional relationships between variables and the latent factor, except that each variable has an orientation, i.e., that if all other variables are held constant, an increase in this variable is either ‘good’ or ‘bad’.

For the proposed scoring mechanism, each subject is compared to every other subject in a pairwise manner. For stratified designs, these comparisons will be made within each stratum only. When a the genes of interest can be assumed to be correlated with the outcome, although not necessarily in a linear fashion, a partial ordering among the subjects is easily defined. If the second of two subjects has values at least as high among all variables, but higher in at least one variable, it will be called ‘superior’.

Even though a partial ordering does not guarantee that all subjects can be ordered on a pairwise basis, they can all be scored. With I as an indicator function, one can assign a scores to each subject by simply counting the number of subjects being inferior and subtracting the number of subjects being superior

\[ u(x) = \sum_{j/k} I(x_{jk} < x_{jk}) - \sum_{j/k} I(x_{jk} > x_{jk}) \]

For multivariate data, this ordering is ‘partial’, because for some pairs of expression profiles the order may be indeterminate. This is the case, for instance, if the first gene is higher in subject A, but the second gene higher in subject B. Some applications may ask for specific partial orderings. Intervals, for instance, can only be ordered, if they are disjoint, thus their pairwise order may be undetermined. When estimating the signal value for a particular gene on a microarray from a probe set of pairs of perfect and mis-matches, several parametric and semi-parametric (‘robust’) methods have been proposed. A mis-match (MM) differs from a perfect match (PM) in that a single nucleotide is exchanged for its Watson-Crick complement to estimate the non-specific portion of the binding. With low expression levels it is to be expected that random errors mismatches have occasionally higher expression levels than perfect matches \( x_{pk} < x_{pk,MM} \). To allow for a linear model to be used based on the logarithms of the differences, it has been suggested by one manufacturer, Affymetrix, to artificially decrease \( x_{pk,MM} \) of such probe pairs to a heuristically motivated level that ensures each difference to be positive. Of course, this causes a severe bias for genes with low expression levels, because even a gene that is not expressed at all is guaranteed to yield a positive estimate. When using u statistics, this bias can easily be overcome by employing the following partial ordering:

\[ \{ x_k < x_j \} \Leftrightarrow \{ x_{pk} < x_{pj} \} \land \{ -x_{pk,MM} < x_{pj,MM} \} \]

From this, one selects the pair with a score of zero as the most ‘typical’, or, if necessary, the average or median among those closest to zero. As this guarantees ‘outliers’ to be excluded, the believed need for taking logarithms has been overcome. If one is now to request that this estimate be non-negative, the resulting bias would be much lower than if one decreases \( x_{pk,MM} \) for each pair where \( x_{pk} < x_{pk,MM} \).

3. ASYMPTOTICS AND TEST STATISTICS

When Mann-Whitney, in 1947, proposed their version of what is now known as the Wilcoxon/Mann-Whitney test, it was one of the first uses of u-statistics. Hoeffding formalized this concept in 1948 for the one-sample case and in 1951, Lehmann considered the two sample case. Originally, observations were allowed to be multivariate. In 1990, Lee explicitly stated that “there is nothing in the above theory that requires [the random variables to take values in \( \mathbb{R} \)], and in fact they may take values in any suitable space.” When Gehan, in
1965, applied u statistics to censored observations, however, he viewed them as univariate observations \((x_{i1};\) time under study), accompanied by an indicator of precision \((x_{i2} = 1: \text{event}, x_{i2} = 0: \text{censoring}),\) rather than as multivariate data. Thus, the potential of u-statistics for the analysis of multivariate data has yet to be fully realized, most likely because, at that time the method was developed, the computational effort to handle multivariate data in general, was prohibitive. Today, of course, these limitations no longer exist.

Basing the analysis of genomic profiles on a general theory, obviously, has many advantages. In particular, by drawing on the work of others, we obtain a family of statistical methods for a variety of situations, such as stratified designs, and more than two groups. For censored data, this test reduces to a stratified rank test with marginal likelihood block weights, in general, and for binary data to the stratified McNEMAR test, as a replacement for the TDT. For censored data, the unstratified version of this test reduces to GEHAN’S generalizations of the WILCOXON/MANN-WHITNEY and KRUSKAL-WALLIS tests, with additional longitudinal measures, to the test proposed by FINKELSTEIN-SCHOENFELD.

4. LOW LEVEL ANALYSIS

High density oligonucleotide expression arrays are widely used in biomedical research, being Affymetrix chips one of the most popular. Affymetrix chips use short oligonucleotides to probe for genes in an RNA sample, using 11-20 pairs of probes. Pairs are composed by perfect match intended to hybridize with transcript for the intended gene and mismatched with try to quantitate non specific hybridization. Probes pairs are summarized using different algorithm to obtain the expression measurement of the genes. Here we used the u-statistics to summarize the probe-level intensity data for affy chips. As in MAS 5 (affymetrix current default algorithm) we used the mismatches probes to correct for not specific hybridization.

Figure 1 demonstrates how to perform signal value estimation in a spreadsheet. While clearly not the suggested implementation for routine applications, it demonstrates the ease and computational simplicity of the method.

![Figure 1: Signal value estimation with u-statistics](image1.png)

Figure 2: MAS 5.0 bias for genes with low expression levels.

To assess the performance of the U-statistics in the summarization of Affymetrix probe level data, we used the R-package AffyComp package available from the Bioconductor Project (www.bioconductor.org). As benchmark data set we use data from Affymetrix SpikeIn study (http://www.affymetrix.com/analysis/download_center2.affx) for the HUG133a chips. This data set consists of 3 technical replicates of 14 separate hybridizations of 42 spiked transcripts in a complex human background at concentrations ranging from 0.125pM to 512pM arrayed in a Latin Square format. The concentration of the 14 gene groups in the first experiment is 0, 0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, and 512pM. Each subsequent experiment rotates the spike-in concentrations by one group; i.e. experiment 2 begins with 0.125pM and ends at 0pM, on up to experiment 14, which begins with 512pM and ends with 256pM. We keep background correction and scale normalization step identical as in MAS 5 algorithm.
Figure 3. Nominal concentrations vs observed expression in log2-scale a) MAS 5. b) Ustat.

Figure 4 a shows the mean of the observed expression v. nominal expression. The slope of the regression of observed values vs nominal expression (Signal detect slope) is reported in Table 1. As we can see the U-statistic with a happen to be more precise than RMA and MAS5, having a slope of 0.86 against 0.68, 0.77 for MAS and RMA. This gap is still bigger in the range of low intensity signal, were MAS5 outperform RMA (0.57 vs 0.2) and Ustat outperform MAS. The small bias for the Ustat estimation is clear in Figure 4 b.

Since in microarray analysis we are usually interested in the gene expression differences when comparing with a different experimental condition, estimation of the fold change or log ratios (difference of log of the quantities) is also an important to assess the sensitivity of the expression ratios to total quantity of RNA plot. Figure 5 shows the performance of MAS5 and U-Stat in terms of fold-change estimation. Again the U-stat have better slope than the MAS5, RMA algorithms, especially in the low intensity range (Table 1, Intensities – fc slope rows).

Table 1

<table>
<thead>
<tr>
<th></th>
<th>RMA</th>
<th>MAS5</th>
<th>U-Stat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal detect slope</td>
<td>0.678</td>
<td>0.767</td>
<td>0.862</td>
</tr>
<tr>
<td>Signal detect R2</td>
<td>0.898</td>
<td>0.9062</td>
<td>0.8511</td>
</tr>
<tr>
<td>Low.slope</td>
<td>0.204</td>
<td>0.577</td>
<td>0.668</td>
</tr>
</tbody>
</table>

Figure 4. MAS 5.0 and U-Statistics. a) Slope b) bias

Figure 5. Observed Log ratio vs nominal log ratio for MAS 5 and U-Stat.

Figure 6: MAS 4.0, MAS 5.0, and U-statistics –

Figure 6 shows the Variance across replicates plot, where the variance of the expression values across replicates is assessed. As the median of the SD indicates in Table 1, the variance of the Ustat is slightly higher than MAS and both have higher variability than RMA especially for small intensities. So what the algorithm gain in accuracy (small bias), it lost in precision.
5. High Level Analysis

5.1. Overview

When trying to identify the factors that, by working together, cause a complex phenomenon such as disease susceptibility or treatment effect, we are faced with several problems. First, most complex phenomena cannot be ‘measured’ in the traditional sense, because of the lack of a physical scale. Instead, we are faced with several indicators. While it is often reasonable to assume that ‘more’ is ‘worse’ for each of them, it may not be easy to determine, how much ‘more’ is how much ‘worse’. Once the effect has been scored, we can identify the set of independent variables that indicate the most likely pathway or constellation causing the complex phenomenon. Again, a ‘measure’ has to be found to describe the contribution of several factors.

Psoriasis is a complex inflammatory disease characterized by hyperproliferation of keratinocytes and accumulation of activated T-cells in the epidermis and dermis of lesions. Treatments with various immunomodulatory or suppressive agents (e.g., cyclosporine and methotrexate) have a therapeutic index, which precludes long-term treatment. Therefore, there is an ongoing interest in reducing toxicity through targeting cells mediating this disease more specifically. Here, we demonstrate how this goal can be achieved by utilizing u statistics twice, first to score patients with respect to profiles of clinical outcome variables, and then to score various subsets of cytokines to identify the pathway by which the particular agent exerts its anti-inflammatory activity.

5.2. Scoring Clinical Outcomes

While the multiplicity of genes is obvious, it is often overlooked that the ‘other side’ to which the expression along genomic pathways is to be correlated is often also consists of several variables.

In Psoriasis, disease improvement in response to treatment is even less easily quantified. The PASI (Psoriasis Area Severity Index) and its variants, while frequently used, are crude measures at best. Like many linear scoring systems, it is computed by scoring thickness, red-

<table>
<thead>
<tr>
<th>Median slope</th>
<th>0.637</th>
<th>0.671</th>
<th>0.65</th>
</tr>
</thead>
<tbody>
<tr>
<td>High slope</td>
<td>0.79</td>
<td>0.76</td>
<td>0.74</td>
</tr>
<tr>
<td>Intensities -fc slope</td>
<td>0.677</td>
<td>0.765</td>
<td>0.862</td>
</tr>
<tr>
<td>(low) int.-fc slope</td>
<td>0.306</td>
<td>0.642</td>
<td>0.758</td>
</tr>
<tr>
<td>MedianSD</td>
<td>0.07</td>
<td>0.29</td>
<td>0.54</td>
</tr>
<tr>
<td>Null-log-fc-IQR</td>
<td>0.13</td>
<td>0.467</td>
<td>0.609</td>
</tr>
<tr>
<td>FC=2, AUC (FP&lt;100)</td>
<td>0.64</td>
<td>0.074</td>
<td>0.068</td>
</tr>
<tr>
<td>FC=2, AUC(FP&lt;25)</td>
<td>0.50</td>
<td>0.07</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Talk about the insertion of probe affinity information to reduce the variance.

Figure 7: u scores for clinical outcome in the Toronto dataset.

Here, responses were measured after treatment as expression of K16 mRNA (Km), epidermal thickness (ET), and K16 histology (KH, 0: negative, 1: positive). Clearly, these outcomes are not independent. One of the advantages of u scores is that independency is not required. Moreover, as can also be seen from the definition, u-scores are invariant to scale transformations (logarithms, weights, etc.). Finally, adding highly correlated variables has little effect on the results. If the correlation is ‘perfect’, an additional variable would not have affected the results at all.

To demonstrate, how the proposed method extends the well-known procedure of Gehan and to show how it leads to easily computed test statistics, we will use the “Toronto” data set as an example. In this study, clinical response was measured as both time to recurrence and time to death, both being censored. Figure 7 demonstrates, how subjects can be scored by u-statistics, and how a test statistic is obtained. In this case, the difference between the two groups (ADC and SQCC) is not ‘significant’, as one would expect after looking at the distribution of scores in the two leftmost columns.
5.3. Scoring clinical outcomes

Clearly, activity profiles along a pathway can be scored in essentially the same fashion as response profiles, except for one additional level of complexity. When scoring responses, it was reasonable to assume that we know, whether ‘more’ is ‘better’ or ‘worse’. With activity, this is not necessarily true. If treatment were to shift activity from one pathways to the other ‘alternative’, less effective pathway, ‘better’ effects may be associated with less activity along the former pathway and more activity along the other. On the other hand, if pathways are synergistic, more activity on either pathway may be ‘worse’. Thus, one may wish to allow for various combinations of signs (polarities) to be associated with each set of activity variables. To allow for this, for each pathway (subset of genes), all possible combinations of polarities are to be considered.

Correlating Activity Pathways with Response Profiles

The variations among patients in their response to treatments can now be used to better characterize the genes directly regulated by the various experimental antibodies. Thus, to gain a better understanding of the potential of cytokines and receptors to contribute to inflammation in psoriatic lesions, we have studied mRNA levels for relevant family members In addition; we measured the concentration of epidermal CD3⁺ cells (ECD3). In the first step on the analysis, we identified five candidate genes for inflammatory activity. To demonstrate the impact of the statistical method being employed on the interpretation of the study, we will now demonstrate a sub-study, where we looked into these genes in detail.

Testing for a monotone relationship implies Spearman-type rank correlation for u-scores. For each set of independent variables, Table 2 contains one row. Within each of these sets, the all polarities are considered independently for each response variable and the result for the polarity giving the best correlation is given.

U-scores of K16 mRNA expression (RKm) and epidermal thickness (RET) have the highest correlation with a pathway score consisting of IL12, iNOS, and epidermal CD3⁺ cells (RKm: 0.789 – 0.790 when IL12 is also included, RET: 0.789). When K16 mRNA and epidermal thickness are evaluated together, the highest correlation (0.815) is seen for the same set of inflammatory factors. That the correlation is higher for the combination than for each variable alone further supports the standing hypothesis that changes in these three inflammatory factors affect both response characteristics in a ‘concerted action’. K16 histology in the response profile reduces the correlation for the set (IL12, iNOS, ECD3), although only marginally from 0.815 to 0.814. For RKH alone, however, a higher correlation (0.849) is seen for a different set of inflammatory factors (IFNg, iNOS, Stat4).

Table 2: Selected pathways of inflammatory genes and number of CD3⁺ cells in the epidermis, and correlation of their multivariate (1–6 variables) inflammation u scores with multivariate (1–3 variables) u score for response sorted by response score U2. Right part: The highest correlation per column is indicated in bold, all correlations at least as high as the smallest among them (0.789) are shaded. Left part: Pathways with the highest correlation with U2 by multivariate u-scores, forward selection, and univariate analysis are boxed. The pathways with the next highest correlations by bi- and univariate correlation are shaded.

This suggests that K16 histology is related to a different pathway than K16 mRNA expression and epidermal thickness, a pathway which may be independent of IL12. For instance, K16 histology may reflect effects that preceded the effects measured by acute K16 mRNA expression.

Interestingly, IL12, by itself, has a very low correlation with U2, the correlation of 0.480 being the second lowest among all pathways. The correlation of iNOS alone (0.582) is only the third highest with respect to R2. If one had selected the inflammatory parameters based on univariate correlations, the average of the RKm and RET scores, one would have chosen (IL8, Stat1) with a correlation of only 0.661, rather than (IL12, iNOS) with a correlation of 0.741. Notably, the set selected by screening all possible sets of inflammatory factors (IL12, iNOS) was disjoint from the set that would have been selected by univariate methods (IL8, Stat4). Which pathway to choose has tremendous implications for biological processes? One would predict that both Stat-1 and IL-8 mRNA could be transmitted by the Stat-1 transcription factor activated by IFNg. However, a set including IL-12 and iNOS pair implicates NFXb and Stat-4 as transcription factors. As an alternative to the ex-
haustive search through all sets of inflammatory factors, one might have first selected the most important factor in univariate analysis, which, in this case, is Stat. Among the bivariate sets including Stat, one would have selected Stat/IL8 and stopped there. Thus, a hierarchical analysis would have had no advantage over the simple univariate analyses and, like with those, the interpretation would have been misleading.

DISCUSSION

Multivariate ordinal data are frequently used to assess semi-quantitative characteristics, such as genomic profiles. Traditional approaches for combining different measures into a utility function require that a relative weight be assigned to each measure. Typically, neither the choice of a transformation (linear, exponential, polynomial, ...), nor the choice of weights \( w_j \), is easily justified. With biological systems, the complexity of the system makes such a justification problematic. If an inappropriate model is chosen an analysis based on such a utility function may be misleading. "This is not very reassuring considered that most models are chosen for their mathematical convenience rather than their biological plausibility",[23 p.1352]

In signal value estimation, for instance, logarithmic transformation, chosen mainly for mathematical convenience, created a problem with negative pairs, which, consequently, were believed in need for a "background correction". This assumption that negative observations are 'impossible', however, is based on a misunderstanding of basic statistical methodology. In fact, for genes that are not expressed in the particular tissue, one would expect half of the pairs to have higher mis- than perfect matches. Trying to "correct" this, obviously, creates a bias, because estimates for genes with expression level zero have signal value estimates as high as genes with low, but positive expression levels.

When fitting linear models, variables are frequently added or dropped sequentially. For instance, one may look for the most "significant" variable in univariate analyses first, and then add more variables in a 'step-up' fashion. Such a strategies, however, may not even come close to the optimum, as we have demonstrated. Tree based approaches (CART[24]), are an alternative, where subjects are separated by the most significant variable first, and each subset is then separated by another subset-specific variable. While this may result in easily communicated decision strategies, step-functions are not more easily justified on theoretical grounds than linear, exponential, or polynomial functions.

A frequently used attempt to resolve this dilemma is to use a "training set" to determine transformations and weights that yield optimal results within this set, and then to check, if the results are 'reasonably good' when this specific scoring system is applied to an 'evaluation set'. If not, one selects another family and/or optimality criterion and tries again. Of course, a set of functions and weights that seems to be 'reasonably good' in the evaluation set is not guaranteed to be optimal. Thus, it has also been suggested that "if [a] method is to be used, its statistical properties should be examined under different, biologically plausible, alternative distributions by simulation."[23 p.1352]

The approach proposed here overcomes the limitations of the above approaches. The advantage of the proposed approach is that no additional assumptions need to be made and validated. Thus, no empirical evaluation is needed. Since no assumption regarding the functional form of the relationship is made, u scores are scale independent. Moreover, no assumptions need to be made regarding relative importance of variables or correlation among the variables. Relative importance and correlation do not even need to be constant, but may vary with the level of the underlying latent factor. If the variables describe different risk indicators, for instance, other variables may be relevant for low risk subjects, than for high-risk subjects. Adding a highly correlated variable is unlikely to affect any of the existing pair-wise orderings and, thus, has little or no effect on the scores.

REFERENCES


18. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 1947; 18:50-60.


