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

Bioinformatics Tools Enabling U-Statistics for Microarrays

Knut M. Wittkowski, Asifa Haider, Ephraim Sehayek, Mayte Suárez-Fariñas, Maurizio Pellegrino, Alexandre Peshansky, Cameron Coffran, and Sanford Coker

Abstract—**It is rare that a single gene is sufficient to represent all aspects of genomic activity. Similarly, most common diseases cannot be explained by a mutations at a single locus. Since complex systems tend to be neither linear nor hierarchical in nature, but to have correlated components of unknown relative importance, the assumptions of traditional (parametric) multivariate statistical methods can rarely be justified on theoretical grounds. Empirical "validation" is not only problematic, but also time consuming. Here we demonstrates how bioinformatics tools, ranging from spreadsheets to grids, can enable u-statistics as a non-parametric alternative for scoring multivariate ordinal data. Applications are shown to improve assessment of genetic risk factors, quality control of microarrays and signal value estimation, scoring genomic profiles that best correlated with complex risk factors (cardiovascular diseases), and complex responses to an intervention (treatment of psoriasis).**

I. Introduction

When applying statistical methods to complex phenomena, a single measure often does not reflect all relevant aspects to be considered, so that several measures of influences and/or outcomes need to be considered. When the definite measure is not easily obtained, surrogate measures have to be evaluated, 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 SNP and gene expression microarrays.

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. Thus, data are interpreted as points in a Euclidian space. The number of dimensions is reduced by assuming them to be related by a function of known type (linear, exponential, etc.), allowing one to determine for each point the Euclidian distance from a model hyperspace.

While mathematically elegant and computationally efficient, this approach has shortcomings when applied to real world data. Since neither the variables' relative importance and correlation nor their functional relationship with the immeasurable latent factor 'overall usefulness', 'efficacy', 'risk', or 'safety' are typically known, construct validity [1] 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. The diversity of scoring systems used attests to the subjective nature of this process.

Even when the assumptions of the linear model regarding contribution to and relationship with the underlying immeasurable factor are questionable, as in genetics and genomics, one can often assume that the contribution of a locus and the expression of a gene have at least an 'orientation', i.e., that, if all other conditions are held constant, presence of an additional mutation or increase in a gene's expression is either 'good' or 'bad'. The sign of this orientation can be known (hypothesis testing) or unknown (selection procedures).

When faced with the risk of anal vs. vaginal contacts for sexual transmission of HIV [2], we presented a partial ordering for dealing with graded and ungraded variables, which allowed to incorporate knowledge that anal contacts carry more risk than vaginal contacts. Using the marginal likelihood (MrgL) for this partial ordering, we developed a nonparametric method to assess overall risk of HIV infection based on different types of behavior [2] and overall protective effect of barrier methods [3]. More recently, we applied this approach to assessing immunogenicity in cancer patients [4]. 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 SNPs or genes.

Manuscript received Mai 1, 2006. This work was supported in part by the National Center for Research Resources at the NIH under GCRC grant M01-RR00102.

K. M. Wittkowski is with The Rockefeller University Hospital, General Clinical Research Center, New York, NY 10021 USA (phone: 212-327-7175; fax: 212-327-8450; e-mail: kmw@rockefeller.edu).

E. Sehayek is with The Rockefeller University, Laboratory of Biochemical Genetics and Metabolism, (e-mail: sehayee@rockefeller.edu)

A. Haider is with The Rockefeller University, Laboratory of Investigative Dermatology, New York, NY 10021 (e-mail: haidera@rockefeller.edu)

M. Suárez-Fariñas and M. Pellegrino are with the Rockefeller University, Center for Studies in Physics and Biology (e-mail: farinam@rockefeller.edu and mpellegri@rockefeller.edu)

A. Peshansky was with The Rockefeller University Hospital, General Clinical Research Center and is now with the University of Medicine and Dentistry New Jersey, General Clinical Research Center, Newark, NJ 07103 (e-mail: a.peshansky@umdnj.edu)

C. Coffran and S. Coker are with The Rockefeller University, Department of Information Technology, New York, NY 10021 USA (e-mail: scoker@rockefeller.edu and cameron@rockefeller.edu).

1-4244-0033-3/06/$20.00 ©2006 IEEE.

Here, we present computational tools based on a closely related approach, u-statistics, which is computationally more efficient. With u-statistics [5], individual analyses can often be performed using spreadsheet software. Screening for optimal subsets of explanatory variables becomes feasible without the restrictions imposed by commonly used hierarchical strategies, although larger sample sizes and, even more importantly, larger numbers of variables require a variety of bioinformatics strategies.

U-statistics let to a family of simple tests. For uncensored data, this includes stratified rank tests with MrgL block weights [6] in general, for binary data the stratified McNemar [7] test [8], for designs with two or more treatments the WMW [9], Kruskal-Wallis [10], and Friedman tests [11].

II. Methods

A. U Statistics

To develop a computationally efficient procedure to score multivariate ordinal data, 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 always 'good' or always 'bad'.

Each subject is compared to every other subject in a pairwise manner. For stratified designs, these comparisons will be made within each stratum (e.g., sex) only. When the genes of interest can be assumed to be correlated with the outcome, although not necessarily in a linear fashion, a partial ordering [12] 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 is 'superior'.

For univariate data, all pairs of observations can be decided, i.e., the resulting ordering is 'çomplete'. For multivariate data, however, the ordering is only 'partial', in general, because for some pairs of expression profiles the order may be undetermined. This is the case, for instance, if the expression of the first gene is higher in subject A, but that of the second gene is higher in subject B.

Although a partial ordering does not guarantee that all pairs of subjects can be ordered, typically all subjects can be scored. With I as an indicator function, one assigns a score to each subject by counting the number of subjects being inferior and subtracting the number of subjects being superior

$$\mathrm{u}\left(x_{jk}\right)=\sum\nolimits_{j'k'}\mathrm{I}\left(x_{j'k'}<x_{jk}\right)-\sum\nolimits_{j'k'}\mathrm{I}\left(x_{j'k'}>x_{jk}\right)$$

The lattice on Fig. 1 provides a graphical representation of a partial ordering for multivariate data, showing the main features of u-statistics. (1) Pairs are linked if the order is independent of any (non-zero) weights that could be assigned to the different variables. (2) Adding a highly correlated variable is unlikely to have any effect on the lattice structure. Relative importance and correlation do not even need to be constant, but may depend on the other variables.

Some applications may ask for specific partial orderings. Intervals, for instance, can only be ordered if they are disjoint. This leads to tests for censored data, including the tests of Gehan [13, 14] for Kaplan-Meier curves. Drawing on a general theory, yields a family of statistical methods for a variety of situations, including signal value estimation.

Pt	Ttl	rChol	rCamp	rSito
1	3.0	2.5	1.5	3.5
2	4.0	2.5	1.0	0.5
4	1.0	4.0	3.5	1.5
6	1.0	0.5	0.5	0.5
5	2.0	14.0	8.0	5.0
9	1.5	1.0	1.0	2.0
8	8.0	12.0	6.0	6.0

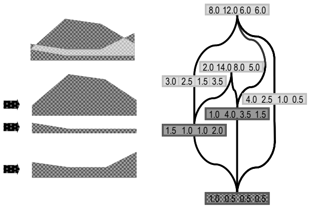

Fig. 1. Generation of a partial ordering from the blood cholesterol profiles (total cholesterol, cholestanol, camposterol, sitosterol) of seven patients. The profiles of patients 4, 6, and 9 are shown to the right of the data. The overlay above the individual profiles shows that both #4 and #9 have a higher profile than #6, but that the pair wise order between #4 and #9 cannot be decided without making assumptions about the relative importance of Ttl and rSito (higher in #9) vs rChol and rCamp (higher in #4). The complete partial ordering of the seven patients is depicted as a lattice, with lines indicating for which pairs the pairwise order can be decided. Patients #4, #6, and #9 are highlighted.

When estimating the signal value for a particular gene on a microarray from a probe set of pairs of perfect (PM) and mis-matches (MM), several parametric and semi-parametric ('robust') methods have been proposed. An MM differs from a PM in that a single nucleotide is exchanged for its Watson-Crick complement to estimate non-specific binding. For genes that are not expressed, it is to be expected that mismatches have higher expression levels than perfect matches ($x_{k,\mathrm{PM}}<x_{k,\mathrm{MM}}$) in 50% of all probe pairs. To allow for a linear model based on the logarithms of the differences, it has been suggested [15] to artificially decrease $x_{k,\mathrm{MM}}$ of such probe pairs to a heuristically motivated level that ensures each difference to be positive. Of course, this decreases sensitivity for genes with low expression levels. When using u statistics, this bias can easily be overcome by employing the following partial ordering:

$$\left\{x_k<x_{k'}\right\}\Leftrightarrow\left\{\left(x_{k,\mathrm{PM}}<x_{k',\mathrm{PM}}\right)\wedge\left(-x_{k,\mathrm{MM}}<-x_{k',\mathrm{MM}}\right)\right\}$$

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 perceived need for taking logarithms is overcome. Even if one is to request that this estimate be non-negative, the resulting bias would be much lower than if one decreases $x_{k,\mathrm{MM}}$ for each pair where $x_{k,\mathrm{PM}}<x_{k,\mathrm{MM}}$.

B. Bioinformatics tools

When Hoeffding formalized this concept of u-statistics in 1948 [5], he allowed for multivariate observations, yet the potential of u-statistics for the analysis of multivariate data was not fully recognized, most likely because the computational effort to handle multivariate data was prohibitive, in general, and no algorithm was presented that would have allowed application of the method at least for small sample sizes. When Gehan [13, 14], in 1965, applied u statistics to censored observations, he viewed them as univariate observations (x_{jk1}: time under study), accompanied by an indicator of precision (x_{jk2} = 1: event, x_{jk2} = 0: censoring).

Even with today's computers, dealing with large quantities of complex data requires a combination of bioinformatics strategies to develop computationally feasible tools.

Methods: First and foremost, a method had to be found that was not np-hard. Moving from the MrgL principle to u statistics, while foregoing some second order information, formed the basis for developing algorithms that were computationally efficient.

Algorithms: The u-test, except for a missing variance term, had already been published in 1914, 33 years before Mann and Whitney, by Thomas Deuchler [16]. Unfortunately, he presented his ideas more verbally, which made the results less accessible internationally. On the other hand, being a psychologist, he laid out a scheme for computations that, had it been more widely known, could potentially have given u-statistics an equal footing with methods based on the linear model. Based on his work, we developed algorithms that, besides growing with the square of the number of subjects only, are easily implemented.

			2	1	2	0	0	1	0	
			1	2	0	0	1	0	0	
			2	0	0	1	0	0	0	
2	1	2	0	0	1	1	1	1	1	5
1	2	0	0	0	0	0	1	1	1	3
2	0	0	-1	0	0	0	0	1	1	1
0	0	1	-1	0	0	0	0	0	1	0
0	1	0	-1	-1	0	0	0	0	1	-1
1	0	0	-1	-1	-1	0	0	0	1	-2
0	0	0	-1	-1	-1	-1	-1	-1	0	-6

Fig. 2. Computation of u-scores from a data set with the lattice structure of Fig. 1. The seven profiles are written both to the left and above a square array. Cells are filled according to the following rule: If the row profile is higher then the column profile, enter "1", if it is smaller "-1", otherwise "0". Once all cells are filled, the u-scores are obtained as the row sums of the array.

For small samples, spreadsheets for different partial orderings can be downloaded from muStat.rockefeller.edu. While clearly not the suggested implementation for routine applications, this demonstrates the ease and computational simplicity of the method.

Language: The power of S (www.insightful.com) or R (www.r-project.org) as a bioinformatics tool lies in the ease with which statistical concepts can be expressed in the code. The downside of having a conceptually simple language is lack of computational efficiency.

Implementation: To deal with a large number of subjects, several subroutines had to be written in C for computational efficiency.

Environment: 'Screening' thousands of expression profiles or epistatic sets to find the profile or set whose scores correlate best with the scores of a complex phenotype, can easily become impractical even on a fast computer or traditional 'beowolf' style cluster. Thus, we have formed a grid of PC work stations at The Rockefeller University Hospital. Data is uploaded to a Web front-end and then passed to a dispatcher that splits the job in dozens of work units, which are then sent to the work stations to be analyzed in parallel.

The grid is controlled by a Linux server acting that acts as dispatcher, verifies integrity of returned work units, and notifies the requester when the job is done. This server runs an IBM DB2 database, secure Apache web server for a management console, and grid software from United Devices (www.ud.com). The client nodes consist of mixed x86 Microsoft Windows workstations running the United Devices agent, which processes the work unit at low priority whenever the workstation is idle. As the agent is centrally customized to include an installation of the S-Plus application, it suffices for the work units to include the subset of the data to be analyzed and the S-Plus script to do the analysis.

These tools and services, which for the first time make u-statistics for multivariate data more widely available, can be accessed through muStat.rockefeller.edu

III. Applications

A. Genetic Data from Trios (Case and Parents)

In a study on the genetics of cardiovascular diseases [17], we are analyzing genetic data of hypercholesteremia patients and their parents. Each parent transmits one of two alleles at each locus and we would like to find loci where one form is transmitted more often to affected children than the other.

In 1993, the sign test formula resurfaced in the Transmission Disequilibrium Test (TDT) [18] with p_T and q_T counting the number of rare and common alleles, respectively, transmitted to a diseased child. Since then, the TDT has become one of the most frequently used methods in genetics.

Although non-parametric tests require fewer assumptions to be made than parametric tests, they still require that observations are independent. While each parent transmits its allele independently, the effects of the two alleles transmitted to the same child are not independently observed. For a dominant locus, where one copy of the disease allele is sufficient to cause the disease, heterozygous children of two heterozygous parents contribute evidence associating both alleles with the disease. If one simply counts alleles, the contribution of these children cancel each other out in the effect estimate, but inflate the variance term, reducing the power of the TDT to detect dominant diseases.

Fig. 3 demonstrates how even exact tests, where all possible permutations of data need to be considered, can be implemented in as few as four lines of native S language code.

```
#----------------------------------------------------------------------
# pP,   qP = number of PP,   PQ children of PP~PQ parents
# pX,xX,qX = number of PP,PQ,QQ children of PQ~PQ parents
# pQ,   qQ = number of PQ,   QQ children of PQ~QQ parents
#----------------------------------------------------------------------

O3  <- function(X1,X2,X3,Op) matrix(outer(outer(X1,X2,Op),X3,Op))
Est <- function(pP,qP,pX,qX,pQ,qQ) O3(pP-qP, (2^1)*(pX-qX),pQ-qQ,"+") # (1)
Var <- function(pP,qP,pX,qX,pQ,qQ) O3(pP+qP, (2^2)*(pX+qX),pQ+qQ,"+") # (2)

asymp.SMN.pvalue <- function(...)  1-pchisq( Est(...)^2/Var(...) ,1)  # (3)

exact.SMN.pvalue <- function(pP,qP,pX,qX,pQ,qQ) {
   b0  <- function(n) if (n==0) 1 else dbinom(0:n, n, .5)
   Dst <- function(nP,nX,nQ) O3(b0(nP),(2^0)*b0(nX),b0(nQ),"*")
   tb <- cbind(
      Dst(nP<-pP+qP, nX<-pX+qX, nQ<-pQ+qQ),
      Est(0:nP,nP:0, 0:nX,nX:0, 0:nQ,nQ:0)^2)
   return(1-sum(tb[tb[,2]<c(Est(pP,qP,pX,qX,pQ,qQ)^2),1])) }

pT <- pP+(2*pX+xX)+pQ
qT <- qP+(2*qX+xX)+qQ)

asymp.TDT.pvalue <- function(pT,qT) asymp.SMN.pvalue(pT,qT, 0,0,0,0)
exact.TDT.pvalue <- function(pT,qT) exact.SMN.pvalue(pT,qT, 0,0,0,0)
```

Fig. 3. S (also R) code for both the asymptotic and the exact versions of the stratified McNemar test and the TDT. The numbers in parentheses refer to the equation numbers in [8]. Eclipses ("…") are part of the code. Note that `asympt.TDT.pvalue(pT,qT) = 1-pchisq((pT-qT)^2/(pT+qT),df=1)`

B. Microarray Quality Control

For univariate data, the median and the range between the 25% and the 75% quantile are simple application of u-statistics. Interestingly, they are conceptually much simpler than mean and standard deviation. Neither need exclusion of 'outliers' to be considered nor needs a justifications to be sought for taking the squares (rather than, for instance, the absolute value) of an observation's distance from the 'center' (mean or median, respectively) to determine the deviation from a model hyper plane.

As one of our applications of u-statistics to microarrays, we have recently developed a tool, termed 'Harshlight' [19, 20] to identify localized defects on the surface of microarrays by plotting the distance of each location's log expression from the median across a set of chips. Since probes are randomly allocated on the chip, the shadowy circle on the left side of this u-filter image (Fig. 4b) is clearly an artifact, as are the isolated bright and dark spots close to the center and in the upper right corner. Fig. 4b demonstrates how Harshlight masks areas with localized defects, preventing them from interfering with subsequent analyses.

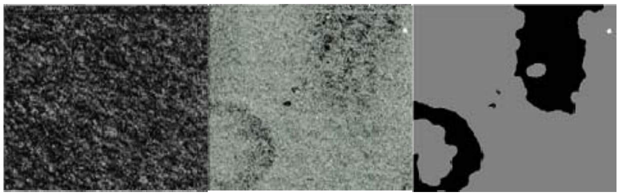

Fig. 4: (a): partial upper 50% HuU95av2 chip pseudo-image. (b): median filtered image (3 chips). (c) HarshLight mask

The justification for the choice of the arithmetic mean (average) as the measure of central tendency in linear models relies either on the law of large numbers and the central limit theorem or the assumption that the distribution of errors is symmetrical, in general, and Gaussian, in particular. Here, neither assumption is easily justified. Fig. 5 demonstrates that u-filtering causes less 'ghosting' than average filtering.

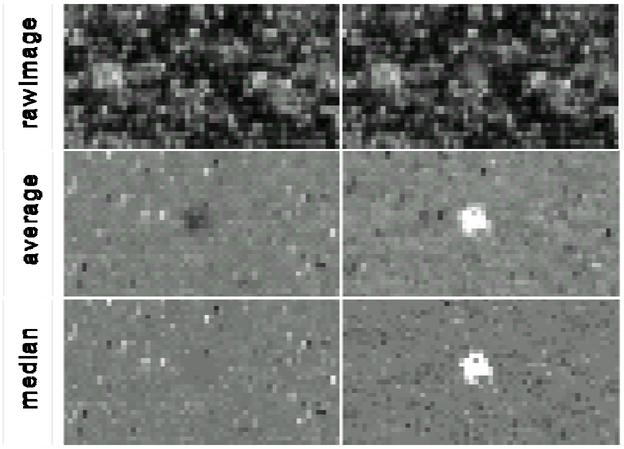

Fig. 5. 'bright spot' in the top right corner of Fig. 4. Top: raw image from the same area of two chips showing gene expression from the same sample under two experimental conditions. Center: average filtering, bottom: median filtering

C. Signal Value Estimation

On Affymetrix microarrays, standardized activity is often computed as log(PM-MM), assuming multiplicative effects and additive noise, although fluid dynamics close to surfaces are known to be highly non-linear.

When some genes are not expressed in the particular tissue, MM and PM reflect random noise only, so that MM is expected to be larger than PM in 50% of all cases. The logarithm of a negative number, however, is undefined, so for a formula based on the multiplicative model to be applicable, it has been argued that a probe with higher mis- than perfect match needs to be "background corrected". Fig. 6 uses data from one of our psoriasis patients to demonstrate the bias created by this "correction". By elevating all non-expressed genes to have signal value estimates between 1 and 100, it becomes virtually impossible to differentiate genes with 'true' expression levels within this range from unexpressed genes that were merely 'corrected'.

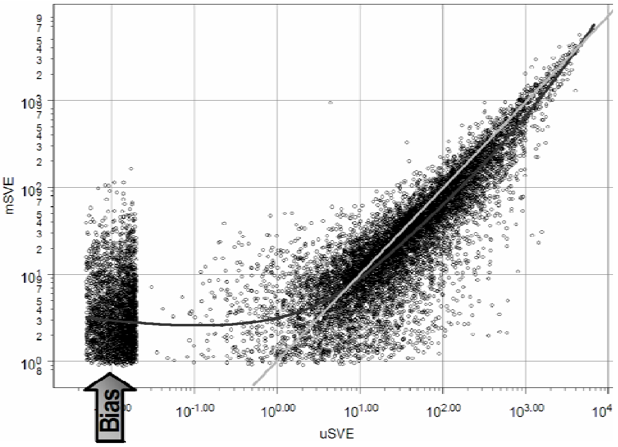

Fig. 6. MAS 5.0 bias for genes with low expression levels.

From Fig. 7, using data from the commonly used 'spike in' dataset, we recognize an 'old friend' from statistical textbooks. As in many physical and biological systems, low concentrations are less reliably measured. Thus, variance increases as the concentration decreases. MAS 5.0 'flattens' the typical sigmoidal curve, but, as is to be expected, decreasing the variance for low concentrations comes at a price: a substantial bias.

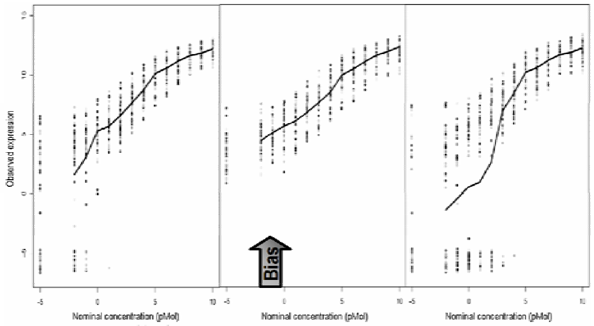

Fig. 7. MAS 4.0, MAS 5.0, and U-statistics – bias vs. variance stabilization.

D. Complex Dependent Variables

When trying to identify the factors that contribute to a complex phenomenon such as disease susceptibility or treatment effect, we are faced with several problems. First, most complex phenomena lack a physical scale to be 'measured' in the traditional sense. Instead, we are faced with several surrogate variables. 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'.

Psoriasis, for instance, is a complex inflammatory disease characterized by hyperproliferation of keratinocytes and accumulation of activated T-cells in lesional skin. 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.

The PASI (Psoriasis Area Severity Index) and its variants, while frequently used, are crude measures at best. As a linear scoring systems, it is computed by scoring thickness, redness, and scaling on a scale from 0 (none) to 4 (striking) for four body areas independently. The sum of these scores is then multiplied by the size of the area (legs: 40%, trunk: 30%, arms: 20%, head: 10%) and a score for the estimated percentage of skin involved from 0 (none) to 6 (90-100%). These weighted sums of individual scores are then added to an overall score. One characteristic of the linear model is that the difference between slight and no redness, for instance, is assumed to have the same meaning as the difference between moderate and striking scaling.

One of the advantages of u scores is that they are invariant to scale transformations (logarithms, weights, etc.). Moreover, independency is not required. 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. Thus, u scores are perfectly suited for dealing with complex phenotypes, using the analytical tools described above.

E. Genetic profiles and genomic pathways

Once the effect has been scored, we can identify the set of independent variables that indicate the most likely genomic pathway or genetic constellation causing the complex phenotype. Activity profiles along a genomic pathway can be scored in essentially the same fashion as response profiles, although we are faced with two additional levels of complexity.

First, if the size of the subset of relevant variables among a total of n is unknown, 2^n subsets need to be considered in an exhaustive search. Second, when scoring responses, it was reasonable to assume that we know, whether 'more' is 'better' or 'worse'. With genomic activity, this is typically not 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) with k components, all 2^{k-1} possible combinations of polarities are to be considered.

When fitting linear models, variables are frequently added or dropped sequentially, e.g., by selecting the most 'significant' variable in univariate analyses first, and then add more variables. As we have demonstrated [21], such strategies may not even come close to the optimum. Tree based approaches (CART [22]), are also sequential in nature. 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. Moving to random forests reduces the effect of outliers, but does not account for interactions.

With u-statistics, non-parametric analysis of large sample sizes m (number of subjects) are feasible, because the complexity is of order m^2 only, compared to $m!$ with the marginal likelihood principle. As there are "only" 30 000–35 000 human genes, looking into at least all pairs of genes before start pruning is within reach, but conducting 10^9 bivariate analyses takes several hours, even on a grid with 1000 PCs running at 2 GHz. Looking into sets of three genes among the subset of 1000 most "interesting" genes selected from the uni- and bi-variate analyses requires a similar effort.

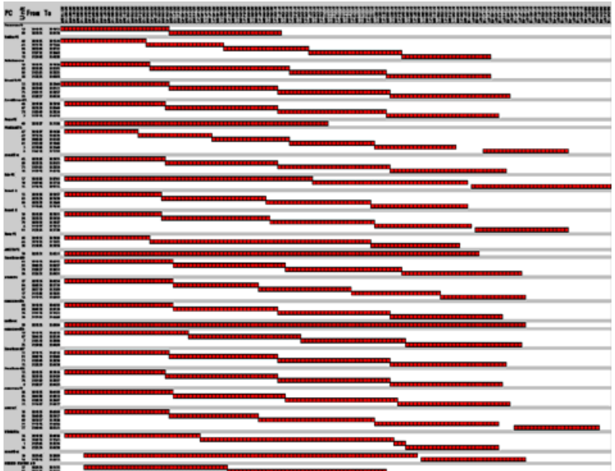

Fig. 8. Distribution of work units across a grid. Rows are work units by node (28 nodes) columns are time points (minutes). Depending on processor speed, concurrent applications, etc., the first node processed two work units, the second node five, and so on.

With SNP arrays, the number of variables is larger (currently at 100K), but the knowledge about the sequence on the chromosome helps with reducing complexity.

However, the large amount of data created on the grid regarding evidence for epistatic interaction between diplotypes is not easily interpreted. Thus, after the results from the nodes are collected, we are using S-Plus to generate maps highlighting evidence for epistatic interaction (Fig. 9).

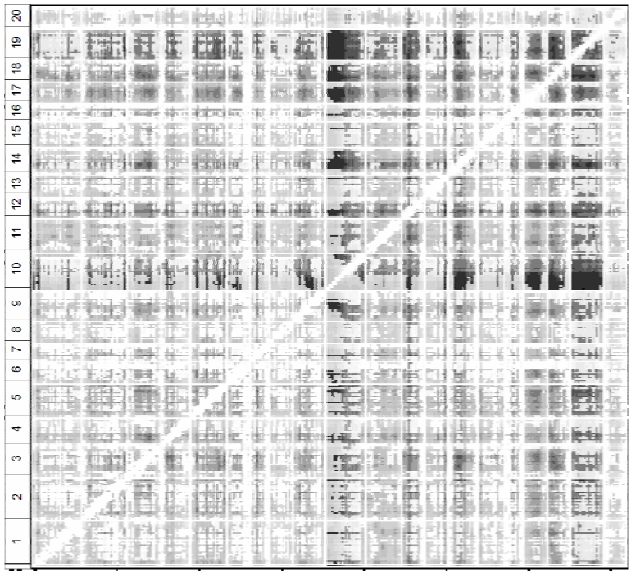

Fig. 9. Epistasis “heatmap” for genetic information from mice, confirming known risk loci on chromosomes 10 and 19, but also providing evidence for loci on chromosomes 12 and 14, which may act primarily through interaction with other risk loci.

IV. Discussion and Conclusions

Multivariate ordinal data are often used to assess semi-quantitative characteristics. Traditional approaches for combining measures into a utility function require that relative weights be assigned to the measures. Typically, neither the transformation (linear, exponential, polynomial, …), nor the weights are easily justified, but results based on inappropriate models easily may be misleading.

A frequently used attempt to resolve this dilemma is to use a ‘training set’ to determine transformations and weights within this set, and then to check, if this scoring system is ‘reasonably good’ when 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 and the number of possible combinations of functions and weights is infinite.

U statistics overcome the limitations of many approaches. No assumptions other then monotonicity need to be made. This, in turn, allows for u-statistics to drive a new generation of systems for decision support, in general, and predictive medicine, in particular. Because no empirical validation is needed, scoring systems can be created *ad hoc*, opening a range of applications, such as diagnostic support fine tuned to the particular characteristics of a patient and risk assessments in homeland security, where intelligence suggests indicators for imminent threats, but the relative importance of these indicators would only be known *post mortem* [23].

References

[1] L. J. Cronbach and P. E. Meehl, "Construct validity in psychological tests," *Psychological Bulletin*, vol. 52, pp. 281-302, 1955.

[2] E. Susser, M. Desvarieux, and K. M. Wittkowski, "Reporting sexual risk behavior for HIV: a practical risk index and a method for improving risk indices," *American Journal of Public Health*, vol. 88, pp. 671-674, 1998.

[3] K. M. Wittkowski, E. Susser, and K. Dietz, "The protective effect of condoms and nonoxynol-9 against HIV infection," *American Journal of Public Health*, vol. 88, pp. 590-596, 972, 1998.

[4] J. Banchereau, A. K. Palucka, M. Dhodapkar, S. Kurkeholder, N. Taquet, A. Rolland, S. Taquet, S. Coquery, K. M. Wittkowski, N. Bhardwj, L. Pineiro, R. Steinman, and J. Fay, "Immune and clinical responses after vaccination of patients with metastatic melanoma with CD34+ hematopoietic progenitor-derived dendritic cells," *Cancer Research*, vol. 61, pp. 6451-8, 2001.

[5] W. Hoeffding, "A class of statistics with asymptotically normal distribution," *Annals of Mathematical Statistics*, vol. 19, pp. 293-325, 1948.

[6] K. M. Wittkowski, "Friedman-type statistics and consistent multiple comparisons for unbalanced designs," *Journal of the American Statistical Association*, vol. 83, pp. 1163-1170, 1988.

[7] Q. McNemar, "Note on the sampling error of the differences between correlated proportions or percentages," *Psychometrica*, vol. 12, pp. 153-157, 1947.

[8] K. M. Wittkowski and X. Liu, "A statistically valid alternative to the TDT," *Human Heredity*, vol. 54, pp. 157-64., 2002.

[9] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics*, vol. 1, pp. 80-83, 1954.

[10] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, pp. 583-631, 1952.

[11] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*, vol. 32, pp. 675-701, 1937.

[12] K. M. Wittkowski, "An extension to Wittkowski," *Journal of the American Statistical Association*, vol. 87, pp. 258, 1992.

[13] E. A. Gehan, "A generalised two-sample Wilcoxon test for doubly censored samples," *Biometrika*, vol. 52, pp. 650-653, 1965.

[14] E. A. Gehan, "A generalised Wilcoxon test for comparing arbitrarily singly censored samples," *Biometrika*, vol. 52, pp. 203-223, 1965.

[15] E. Hubbell, W.-M. Liu, and R. Mei, "Robust estimators for expression analysis," *Bioinformatics*, vol. 18, pp. 1585-1592, 2002.

[16] G. Deuchler, "Über die Methoden der Korrelationsrechnung in der Pädagogik und Psychologie," *Z. pädagog. Psychol.*, vol. 15, pp. 114-131, 145-159, 229-242, 1914.

[17] E. Sehayek, H. J. Yu, K. von Bergmann, D. Lutjohann, K. M. Wittkowski, M. A. Levenstien, D. Gordon, M. Stoffel, L. Garcia-Naveda, J. Salit, M. Blundell, J. M. Friedman, and J. L. Breslow, "Genetics of cholesterol absorption and plasma plant sterol levels on the Pacific island of Kosrae," *Circulation*, vol. 110, pp. 720, 2005.

[18] R. S. Spielman, R. E. McGinnis, and W. J. Ewens, "Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)," *American Journal of Human Genetics*, vol. 52, pp. 506-16., 1993.

[19] M. Suárez-Fariñas, A. Haider, and K. M. Wittkowski, "“Harshlighting” small blemishes on microarrays," *BMC Bioinformatics*, vol. 6, pp. 65, 2005.

[20] M. Suarez-Farinas, M. Pellegrino, K. M. Wittkowski, and M. O. Magnasco, "Harshlight: a “corrective make-up” program for microarray chips," *BMC Bioinformatics*, vol. 6, pp. 294, 2005.

[21] K. M. Wittkowski, E. Lee, R. Nussbaum, F. N. Chamian, and J. G. Krueger, "Combining several ordinal measures in clinical studies," *Statistics in Medicine*, vol. 23, pp. 1579-1592, 2004.

[22] L. Breiman, *Classification and regression trees*. Belmont, CA: Wadsworth, 1984.

[23] K. M. Wittkowski, "Novel Methods for Multivariate Ordinal Data applied to Genetic Diplotypes, Genomic Pathways, Risk Profiles, and Pattern Similarity," *Computing Science and Statistics*, vol. 35, pp. 626-646, 2003.