Statistical Modelling 2017; **17(4–5)**: 319–326

Discussion on the paper 'Statistical contributions to bioinformatics: Design, modelling, structure learning and integration' by Jeffrey S. Morris and Veerabhadran Baladandayuthapani

Jeanine J Houwing-Duistermaat[1,2], Hae Won Uh[2] and Arief Gusnanto[1]
[1]Department of Statistics, University of Leeds, Leeds LS2 9JT, United Kingdom.
[2]Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands.

Abstract: Bioinformatics is an important research area for statisticians. This discussion provides some additional topics to the paper, namely on statistical contributions to detect differential expressed genes, for protein structure prediction, and for the analysis of highly correlated features in Glycomics datasets.

Key words: derived traits, directed networks, FDR, filtering, glycomics, protein structure

1 Introduction

Morris and Baladandayuthapani (M and B) discuss statistical contributions to bioinformatics in four different areas: design, modelling, structure learning and integration. The authors not only manage to highlight many relevant and important contributions of statistics to the field of bioinformatics but also illustrate cases where proper and rigorous statistical principles are not considered. We like to congratulate the authors with this very thorough and highly relevant piece of work. To complement the article, we wish to highlight several areas of bioinformatics in which statisticians have also made significant contributions. Moreover, we like to share our own experience in working with a novel type of omics, namely Glycomics. We finish with a discussion about the challenges for statisticians to work in the field of statistical bioinformatics.

2 Data cleaning, filtering and the effect on downstream analysis

Our experience with the statistical analyses of Glycomics datasets measured in epidemiological studies motivated our consortium MIMOmics (www.MIMOmics.eu),

Address for correspondence: Jeanine Houwing-Duistermaat, Department of Statistics, University of Leeds, Leeds LS2 9JT, United Kingdom.
E-mail: j.duistermaat@leeds.ac.uk

10.1177/1471082X17706135

which develops methods for integrated analysis of multiple omics datasets. First, we encountered that statistical analyses of Glycomics data were hampered by our lack of knowledge about high throughput technology and preprocessing steps which were taken before the data were given to us. Second, straightforward association analysis of one omic variable with an outcome was the state of the art methodology which fails to identify multi omics risk profiles associated with complex traits. Also, M and B discuss such elementwise modelling approaches (p. 43). While elementwise approaches are straightforward to implement, they ignore the correlation structure among the elements which results in suboptimal testing and inefficient estimators. In our opinion, it also reduces the insight in the complex biological mechanisms underlying the traits. Third, we wondered what the effect of data cleaning would be on downstream analysis and whether different statistical approaches needed different data cleaning procedures.

Glycomics is an emerging novel omics science of a much higher complexity than proteomics (Hart and Copeland, 2010). Although Glycomics changes are said to be a hallmark of virtually any human disease (Walt, 2012), the structural complexity of glycans together with technological limitations hampered understanding the nature of glycans and their effects on other biological processes. Only recently, high throughput methods for quantification of N-glycans for a large number of samples have been developed. In theory, data preprocessing minimizes contributions from unwanted biases and experimental variances. An essential step is normalization; any omics measurements or abundances need to be placed on comparable scales. Although improper normalization methods can significantly impair the data (Kohl et al., 2012; Dillies et al., 2013), their impact on a downstream analysis has not been well thought out. For instance, the mostly used normalization method in Glycomics data analysis is the Total Area (TA) normalization. TA scales each sample or row so that the sum of all glycan intensities in that sample equals 1 (or 100%). This results in a so-called compositional data (Aitchison, 2003), or sometimes called percentage data. The question is how appropriate this method is when applied to highly correlated Glycomics data—the median correlation coefficients are around 0.7. It is well known that TA results in changes of the correlation structure. Because of the sum to unity constraint, when the level of one glycan increases, the level of another must decrease. This, unfortunately, leads to the problem of losing the possibility to interpret the correlation coefficients between the original components, commonly referred to as the negative bias problem (Pearson, 1897).

As pointed out by M and B, networks provide a natural way of representing the dependency structure among variables providing more refined biological interpretations. Correlation networks are widely used to explore, analyze and visualize high-dimensional data. It is not at all clear how to perform a network analysis based on TA normalized data since this normalization changes the correlation structure. Moreover, there is a wish to derive a directional network in which not only the associations between nodes but also the directions are given (Krumsiek et al., 2011). Graphical Gaussian Model (GGM) is one of the methods to construct such directed networks. Causal relations are described by an underlying directed acyclic graph (DAG). However, for TA normalized glycans, the nodes are ratios of two random variables, namely the abundance of one glycan divided by the total abundances of all glycans. How to construct a DAG for such a variable? How to deal with confounders and mediators having a causal effect on only one of these variables? Note that also beyond data cleaning, within the field of glycomics and metabolomics derived traits are popular variables to study (see, e.g., Reiding et al., 2017). Derived traits are composite variables of glycans and metabolites which have similar structural and chemical properties. These traits might have a biological interpretation as a single unit, but within a regression framework interpretation is far from straightforward. An alternative for using derived traits or TA normalized glycans in a regression framework is to model the joint distribution of the single variables of such traits.

After cleaning and normalization steps, a next necessary step might be filtering. Filtering might be applied for dimensional reduction purposes, but might also provide more efficient downstream analyses of the data. For example, among tens of thousands of probes in microarray, not all of them are expected to be expressed at a biologically meaningful level. Among probes that are expressed at a biologically meaningful level, only some of them are expected to be differentially expressed between experimental conditions. This is a difficult issue to deal with since these probes that are expressed at very low level will only add noise and can even introduce bias. In the end, they can severely harm the sensitivity to detect truly differential expressed probes. When we can identify them beforehand, we can improve the situation and expect to increase the sensitivity in detecting differentially expressed genes.

Currently, there are no general guidelines on how to perform this filtering process. One thing that the scientists generally agree is that the filtering methods to use is (microarray) technology-specific and sometimes experiment-specific. For oligonucleotide arrays, Calza et al. (2007) proposed a filtering method on prove-level data. They develop a robust linear model on the probes and identify problematic probes from their array-to-array variation. They demonstrate that the filtering method increases the sensitivity to detect differentially expressed probesets.

Finally, as also mentioned by M and B in Section 6, researchers often wish to integrate multiple omics datasets in one model since measurements from different platforms contain different related biological information. The statistical challenge is to deal with the different size, structure, sparsity, scale and measurement error of omics datasets. Within our consortium, we have considered various approaches. With regard to prediction, stacking of omics datasets appears not to perform well. Models based on stacked omics sources might perform even less well than a model based on a single omic predictor. Whether to use a parallel or a sequential approach when combining omics datasets depends on the question as well as on the characteristics of the omics datasets (Rodriguez-Girondo et al., 2017). There is a need to obtain more insight in the relationship between omics datasets. For this purpose, we have considered partial least square (PLS) methods within MIMOmics. This method identifies subspaces of the two omics datasets which are highly related. Unfortunately, when omics datasets have a substantial specific part which is often the case, PLS methods do not provide the necessary insight. An alternative might be O2PLS, which identifies common, dataset specific and residual subspaces (Trygg and Wold, 2003; El Bouhaddani et al., 2016).

3 Finding differential expressed genes

Most of the paper of M and B is on preprocessing and modelling of omics data, while especially data from microarray experiments have motivated statisticians to developed novel testing strategies. When a scientist conducts a microarray gene expression experiment, they expect a priori some probes to be significant in the experiment. This assumption has some implications on the statistical inference. One of the most famous example is the introduction of the positive false discovery rate (pFDR) by Storey (2002), which operates under the assumption that some probes are differentially expressed, as an alternative to the Benjamini and Hochberg's (1995) false discovery rate (FDR), which operates under the assumption that none of the probes are differentially expressed.

Further, implication of the above assumption is in the seminal work by Efron et al. (2001) in the empirical Bayes analysis of microarray data. The assumption allows a probabilistic statement be derived on a test statistic, in particular the posterior probability that a probe is differentially expressed given its test statistic. The method proposed by Efron et al. is so important that we can attach a probilistic statement to *any* test statistic since the distribution of the test statistics under the null is obtained through permutation. In this concept, the distribution of test statistics under the null distribution is not a point mass at zero, but a distribution obtained for the dataset at hand. The posterior probability for non differential expression can be considered as *local* false discovery rate (fdr), as oppose to the Benjamini and Hochberg's *global* FDR. In the latter, the control of (global) FDR applies across different experiments, while the former controls FDR locally in the experiment at hand.

The formulation of Efron et al. (2001) allows an extension to multi-dimensional local fdr (Ploner et al., 2006). In the proposed method, the additional second dimension in the calculation of local fdr is the standard error of the difference between groups. This safeguards us from probes that appear to be significant simply because the standard error is underestimated by chance.

When we have tens of thousands of probes in the analysis, the multiplicity problem does not only occur in the context of hypothesis testing but also in effect estimation. Gusnanto et al. (2005) consider the estimation of log fold-changes between experimental groups as a mean to identify differentially expressed probes. This is done by assuming that linear models' parameters are assumed to follow a mixture of three normal distributions, corresponding to probes that are differentially expressed in both direction and non-differentially expressed. This enables identification of differentially expressed probes at the same time as the effect estimation. M and B (p. 46) advocate to use functional regression. To detect differentially expressed regions, these functions can be regression on outcomes as case-control status. To deal with the high dimensionality, M and B developed Bayesian methods.

4 Protein bioinformatics

One of the unsolved main problems in bioinformatics which is ignored by M and B is the problem of protein structure (folding) prediction. Given a sequence of amino acids, are we able to identify the three-dimensional structure of the protein? The problem arises from the fact that it can take a long time and is very expensive to determine a three-dimensional structure of a single protein using, for example, X-ray crystallography.

Many of the methods proposed in the protein structure prediction are using deterministic physical principles from scratch. In particular, it involves global minimization of an energy function (Dill et al., 2007). Other proposed methods work with previously solved structures or 'template' (Floudas et al., 2006). The latter approach is based on the fact that the number of protein structure motif is limited and far less than the number of known proteins.

The contribution of statistics in this topic is by approaching the problem from probabilistic point of view. The key underlying aspect is protein geometry. The main part of the protein is its backbone, which can be considered as a sequence of amino acids that occupy a three-dimensional space. A protein structure can be represented by either a configuration of unlabelled points (atoms) in three dimensions or by the two dihedral angles (conformational angles). These dihedral angles are the two angles created between successive amino acids in the backbone. From known structures, we can have a collection of dihedral angles and their corresponding amino acids from, for example, the RSCB protein database (Berman et al., 2000).

The method proposed by Boomsma et al. (2008), for example, assumes that the dihedral angles from selected library of protein structures follow a bivariate von Mises distribution. They then model the sequential dependencies along the backbone by using a dynamic Bayesian network (a generalization of a hidden Markov model), which emits angle pairs, amino acid labels and secondary structure labels (whether the sequence, is part of alpha helix or beta sheet). This model samples directly structures that are compatible with a given sequence and resamples parts of a structure while maintaining consistency along the sequence of amino acids. They claim that the model makes it possible to evaluate the likelihood of any given structure. This method is shown to work well for local structure prediction. A book titled *Bayesian Methods in Structural Bioinformatics*, edited by Hamelryck et al. (2012), has been published that documents further Bayesian methods for protein structure prediction.

5 Final remarks

M and B discuss two examples where spurious associations were identified due to inappropriate experimental design and statistical analysis. We like to emphasize that many statisticians are not equipped to analyze this type of omics data either, which is in line with our own experience when we started to analyze Glycomics datasets. It is also illustrated by a paper which had to be withdrawn from Science due to inappropriate analysis of an unmatched case control study with regard to

Affymetrix and Illumina arrays. This work was later published in *Plos One* (Sebastiani et al., 2012). Generally, biostatisticians working with clinical data have no or little experience with noisy omics datasets. Many programmes in biostatistics do not offer this topic in their curriculum. Many chemists also have no idea how much data cleaning steps might influence downstream analysis and therefore do not inform statisticians. Within our recently funded training programme 'Innovative training in methods for future datasets', we will train early stage researchers in high throughput and statistical methodology jointly. We believe that this initiative will fill the current gap. For the current generation of statisticians, it is however essential that all the work which is needed to obtain clean data is recognized. If not, involvement of statisticians in the analyses of omics datasets will be hampered because statisticians would prefer to work in a clinical setting.

Another problem with omics datasets is data sharing. We had to postpone submission of several methodological papers because the clinical paper had to be published first. This is one of the reasons to reuse freely available datasets from repositories. Unfortunately, interesting clinical outcomes are often not available and a good description of the study design is often lacking. Selection of participants might bias the results. For example, the family data available for the participants of the genetic analysis workshop 2017 (www.gaworkshop.org) are partly multi case families, but details about the selection process were not provided. Especially when high dimensional datasets are analyzed without corrections for outcome dependent sampling, an increase in FDR might be expected. Some work has been done for analysis of secondary phenotypes (omics data) in case control studies (Monsees et al., 2009), but this work is not well known within the omics field. Recently, we extended this work to a flexible method for correction of selection of families (Tissier et al., 2017) by joint analysis of the primary and the secondary (omics) outcomes.

To conclude, we agree with the conclusion of M and B that recent developments in biology and medicine provide great opportunities for the statistical community to play a fundamental role in pushing the science forward, as we equip other scientist with the tools they need to extract the valuable information they contain. The opportunity to work at all levels of omics data analysis, from upstream to downstream, are still wide open. However, analysis closer to the upstream needs more attention from biostatisticians. This will inevitably become critical when future development in biological and medical technologies becomes a reality. To this end, the challenges in data cleaning need to be fully addressed and acknowledged, and the characteristics of the measured datasets as well as details of the study design need to be taken into account when analyzing the data. For many studies, it is our recommendation to involve multiple statisticians to cover the required clinical, epidemiological and bioinformatic expertise for appropriate statistical analyses of the data.

Acknowledgements

Part of the work has received funding from the European Union's Seventh Framework Programme (FP7-Health-F5-2012) under grant agreement number 305280 (MIMOmics).

References

Aitchison J (2003) *The Statistical Analysis of Compositional Data*. Caldwell, NJ: Blackburn Press.

Benjamini Y, and Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**, 289–300.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN and Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Research*, **28**, 235–42.

Boomsma W, Mardia KV, Taylor CC, Ferkingho-Borg J, Krogh A and Hamelryck T (2008) A generative, probabilistic model of local protein structure. *PNAS*, **105**, 8932–37.

El Bouhaddani S, Houwing-Duistermaat J, Salo P, Perola M, Jongbloed G and Uh HW (2016) Evaluation of O2PLS in Omics data integration. *BMC Bioinformatics*, **17**, 11.

Calza S, Raelsberger W, Ploner A, Sahel J, Leveillard T and Pawitan Y (2007) Filtering genes to improve sensitivity in oligonucleotide microarray data analysis. *Nucleic Acids Research*, **35**, e102.

Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel, D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloë D, Le Gall C, Schaëffer B, Le Crom S, Guedj M, and Jaffrézic F (2013) A comprehensive evaluation of normalization methods for illumina high-throughput RNA sequencing data analysis. *Brief Bioinform*, **14**, 671–83.

Dill KA, Ozkan SB, Weikl TR, Chodera JD and Voelz VA (2007) The protein folding problem: When will it be solved? *Current Opinion in Structural Biology*, **17**, 342–46.

Efron B, Tibshirani R, Storey JD and Tusher V (2001) Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, **96**, 1151–60.

Floudas CA, Fung HK, McAllister SR, MŁonnigmann M and Rajgaria R (2006) Advances in protein structure prediction and de novo protein design: A review. *Chemical Engineering Science*, **61**, 966–88.

Gusnanto A, Ploner A and Pawitan Y (2005) Fold-change estimation of differentially expressed genes using mixture mixed-model. *Statistical Applications in Genetics and Molecular Biology*, **4**, article 26.

Hamelryck T, Mardia K and Ferkingho-Borg J, eds (2012) *Bayesian Methods in Structural Bioinformatics*. Berlin, Germany: Springer.

Hart GW and Copeland RJ (2010) Glycomics hits the big time. *Cell*, **143**, 672–76.

Monsees GM, Taqmimi RM and Kraft P (2009) Genome-wide association scans for secondary traits using case-control samples. *Genetic Epidemiology*, **33**, 717–28.

Kohl SM, Klein MS, Hpchrein J, Oefner PJ, Spang R, and Gronwald W (2012) State-of-the art data normalization methods improve nmr-based metabolomic analysis. *Metabolomics*, **8**, 146–60.

Krumsiek J, Suhre K, Illig T, Jerzy Adamski, J, and Theis FJ (2011) Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Systems Biology*, **5**, 21.

Pearson K (1897) Mathematical contributions to the theory of evolution: On a form of spurious correlation which may arise when indices are used in the measurements of organs. *Proceedings of the Royal Society of London*, **60**, 489–98.

Ploner A, Calza S, Gusnanto A and Pawitan Y (2006) Multidimensional local false discovery rate for microarray studies. *Bioinformatics*, **22**, 556–65.

Reiding KR, Ruhaak LR, Uh HW, El Bouhaddani S, Van Den Akker EB, Plomp R, et al. (2017) Human plasma N-glycosylation as analyzed by matrix-assisted laser desorption/ionization-fourier transform ion cyclotron resonance-MS associates with markers of inflammation and metabolic health. *Molecular and Cellular Proteomics*, **16**, 228–42.

Rodriguez-Girondo M, Kakourou A, Salo P, Perola M, Mesker WE, Tollenaar RAEM, et al. (2017) On the combination of omics data for prediction of binary outcomes.

In Susmita Datta, and Bart JA Mertens eds, *Statistical Analysis of Proteomics, Metabolomics, and Lipidomics Data Using Mass Spectrometry*, pages 259–75. Berlin, Germany: Springer-Verlag.

Sebastiani P, Solovie N, DeWan AT, Walsh KM, Puca A, Hartley SW, et al. (2012) Genetic signatures of exceptional longevity in humans. *Plos One*, 7, e29848. URL http://dx.doi.org/10.1371/journal.pone.0029848 (last accessed 17 April 2017).

Storey JD (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, **64**, 479–98.

Tissier R, Tsonaka R, Mooijaart SP, Slagboom E and Houwing-Duistermaat JJ (2017) Secondary phenotype analysis in ascertained family designs: Application to the Leiden longevity study. To appear in *Statistics in Medicine*.

Trygg J and Wold S (2003) O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter. *Journal of Chemometrics*, **17**, 53–64.

Walt D (2012) *Transforming Glycoscience: A roadmap for the future*. Committee on assessing the importance and impact of glycomics and glycosciences. Washington, DC: The National Academis Press.