Journal of
Data Mining in Genomics & Proteomics

Tan et al., J Data Mining Genomics Proteomics 2015, 7:1
DOI: 10.4172/2153-0602.1000183

Short Communication Open Access

Generalized Measure of Dependency for Analysis of Omics Data

Qihua Tan[1,2]*, Martin Tepel[3], Hans C Beck[4], Lars M Rasmussen[4] and Jacob von Bornemann Hjelmborg[2]

[1]Unit of Human Genetics, Department of Clinical Research, University of Southern Denmark, Odense, Denmark
[2]Epidemiology, Biostatistics, and Biodemography, Department of Public Health, University of Southern Denmark, Odense, Denmark
[3]Department of Nephrology, Odense University Hospital, and University of Southern Denmark
[4]Department of Clinical Biochemistry and Pharmacology, Odense University Hospital, and University of Southern Denmark

As a popular measure of association, the Pearson's correlation coefficient has been frequently used in omics data analysis e.g. in feature selection process during prediction model building using high dimensional gene expression data [1] and proteomics data [2]. However, Pearson's correlation coefficient captures only linear relationships which greatly limit its use in situations of nonlinear association. Statistical modeling for dealing with nonlinear patterns can be complicated [3] and requires intensive computation in case of high dimensional data such as microarray data or genome sequence data. In the analysis of omics data, high dimension means that there can be diverse patterns of dependence not limited to linearity. In this situation, the generalized measures of association more adequate than the Pearson's correlation and capable of capturing both linear and nonlinear correlations are needed. Recently, generalized correlation coefficients have been frequently discussed [4] and their application to large scale genomic data illustrated through microarray gene expression time-course analysis [5].

Currently, the generalized measures of dependency mainly refer to the concepts of rank correlation and information theory based measures. The rank based correlation is well represented by Hoeffding's D [6] which measures the difference between the joint ranks of two random variables (X, Y) and the product of their marginal ranks. The information theory based approaches include mutual information (MI) [7] and maximal information coefficient (MIC) [5,8]. By providing the amount of information one variable reveals about another, MI measures the dependency between two variables of any type. In the middle of last century, Linfoot [9] proposed the information coefficient of correlation which is a monotone increasing function of mutual information with attractive properties for measuring dependency. Using binning as a means to apply MI on continuous random variables, the MIC [5] can be seen as a continuous variable counterpart to MI. MIC searches over various possible grids through binning to achieve maximal mutual information between two variables. A general overview of the main methods used to identify dependency between random variables has been provided and applications illustrated using microarray gene expression data [4].

In a very recent paper published in Scientific Reports [2], we reported a signature of 82 plasma proteins that predicted the increase of inflammation marker C-reactive protein from index day to next-day using proteome analysis in 91 incident kidney transplant recipients. C-reactive protein is an acute-phase-reactant and is an early nonspecific indicator of infectious or inflammatory situations. Although important, current methods cannot determine the day-to-day development of C-reactive protein at the time of its measurement in plasma. The paper showed that it is possible to define a plasma protein signature to predict the increase of next-day C-reactive protein. The predictive proteins were selected from 359 quantified plasma proteins by correlating plasma protein concentrations of each protein with changes of next-day C-reactive protein using the Pearson's correlation coefficient. Feature selection was done by recursively shrinking correlation smaller than a predefined threshold to zero and using the remaining subset of proteins for prediction model building using support vector machines. Leave-one-out cross validation estimated a sensitivity of 81%, and a specificity of 69%, and an overall accuracy of 77%.

Taking the same dataset, we explored prognostic protein signature selection using Hoeffding's measure of dependence, which is a nonparametric measure of association that detects more general departures from independence [6]. Following the same procedure as in Tepel et al. [2] but replacing Pearson's correlation in the feature selection step with Hoeffding's D measure, a 62-protein signature was selected for prediction model building. Our new list of proteins performed about equally well as the 82-protein signature with a sensitivity of 79%, a specificity of 70% and a mean accuracy of 76%. Noticeably, among the 62 proteins selected, 48 overlapped with the published 82-plex signature with 14 new proteins. Our novel application of generalized association measure in feature selection in prediction analysis of high dimensional data shows that, by relaxing the linear relationship assumption, the non-traditional method of association could help with more efficient feature selection while maintaining high prediction accuracy.

The capability of handling both linear and nonlinear associations promotes the use of the generalized correlation measures in analysing massive and complex omics data with aim at ultimately disentangling and interpreting the complex patterns of relationships between omics data concepts in an integrative manner. Taking the relationship between gene expression and DNA methylation for example, multiple studies have been conducted in analysis their correlation using Spearman's correlation coefficient and reported predominantly low or even poor correlation patterns [10,11]. Here, we think that the more adequate generalized correlation methods should help to characterize the biological relationship more adequately and precisely. Moreover, the generalized correlation can also be a useful tool for investigating the functional dependency between sets of attributes in omics data.

Recently, De Siqueira Santos et al. [4] reviewed and evaluated the main methods for identifying dependency between random variables and provided a suggestive list of methods for use in different types of datasets. The main methods can be easily implemented using free R packages such as matie (https://cran.r-project.org/web/packages/matie/), FNN (https://cran.r-project.org/web/packages/FNN/), minerva (https://cran.r-project.org/web/packages/minerva), and Hmisc (https://cran.r-project.org/web/packages/Hmisc/).

***Corresponding author:** Qihua Tan, MD, PhD, Professor, Epidemiology, Biostatistics, and Biodemography, Dept of Public Health, University of Southern Denmark, J. B. Winslows Vej 9B, DK-5000, Odense C Denmark, Tel: 0045 65503536; E-mail: qtan@health.sdu.dk

Received October 20, 2015; **Accepted** November 10, 2015; **Published** November 17, 2015

Citation: Tan Q, Tepel M, Beck HC, Rasmussen LM, Hjelmborg JVB (2015) Generalized Measure of Dependency for Analysis of Omics Data. J Data Mining Genomics Proteomics 7: 183. doi:10.4172/2153-0602.1000183

References

1. Van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415: 530-536.
2. Tepel M, Beck HC, Tan Q, Borst C, Rasmussen LM (2015) The 82-plex plasma protein signature that predicts increasing inflammation. Sci Rep 5: 14882.
3. Royston P, Altman DG (1994) Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. Applied Statistics 43: 429-467.
4. De Siqueira Santos S, Takahashi DY, Nakata A and Fujita A (2013) A comparative study of statistical methods used to identify dependencies between gene expression signals. Brief Bioinform.
5. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, et al. (2011) Detecting novel associations in large data sets. Science 334: 518-524.
6. Hoeffding W (1948) A non-parametric test of independence. Am Math Stat 19: 546-557.
7. Shannon CE, Weaver W (1949) The mathematical theory of communication. Urbana, IL: University of Illinois Press.
8. Speed T (2011) A correlation for the 21st century. Science 334: 1502-1503.
9. Linfoot EH (1957) An informational measure of correlation. Information and Control 1: 85-89.
10. Wagner JR, Busche S, Ge B, Kwan T, Pastinen T, et al. (2014) The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. Genome Biol 15: R37.
11. Kulis M, Heath S, Bibikova M, Queirós AC, Navarro A, et al. (2012) Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. Nat Genet 44: 1236-1242.