Computational Toxicology xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Computational Toxicology

journal homepage: www.elsevier.com/locate/comtox

Human blood gene signature as a marker for smoking exposure: Computational approaches of the top ranked teams in the sbv IMPROVER Systems Toxicology challenge

Adi L. Tarca [a,b,*], Xiaofeng Gong [c,d], Roberto Romero [e,f,g,h], Wenxin Yang [c,i], Zhongqu Duan [c,d], Hao Yang [c,i], Chengfang Zhang [c,d], Peixuan Wang [c,d]

[a] Department of Obstetrics and Gynecology, Wayne State University School of Medicine, Detroit, MI 48202, USA
[b] Department of Computer Science, Wayne State University College of Engineering, Detroit, MI, USA
[c] SJTU-Yale Joint Centre for Biostatistics, Shanghai Jiao Tong University, Shanghai, China
[d] Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China
[e] Perinatology Research Branch, NICHD/NIH/DHHS, Bethesda, MD, and Detroit, MI 48201, USA
[f] Department of Obstetrics and Gynecology, University of Michigan, Ann Arbor, MI 48109, USA
[g] Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI 48825, USA
[h] Center for Molecular Medicine and Genetics, Wayne State University, Detroit, MI 48201, USA
[i] School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, China

ARTICLE INFO

Article history:
Received 17 April 2017
Received in revised form 12 June 2017
Accepted 12 July 2017
Available online xxxx

Keywords:
Systems toxicology
Computational challenge
Gene signature
Smoking biomarker
Predictive modeling

ABSTRACT

Crowdsourcing has emerged as a framework to address methodological challenges in omics data analysis and assess the extent to which omics data are predictive of phenotypes of interest. The sbv IMPROVER Systems Toxicology challenge was designed to leverage crowdsourcing to determine whether human blood gene expression levels are informative of current and past smoking. Participating teams were invited to use a training gene expression dataset to derive parsimonious models (up to 40 genes) that can accurately classify subjects into exposure groups: smokers, former smokers that quit for at least one year, and never-smokers. Teams were ranked based on two classification performance metrics evaluated on a blinded test dataset. The analytical approaches of the first- and third-ranked teams, that are presented in detail in this article, involved feature selection by moderated *t*-test or LASSO regression and linear discriminant analysis (LDA) and logistic regression classifiers, respectively. While the 12-gene signature of the top team allowed the classification of current smokers with 100% sensitivity at 93% specificity, discriminating former smokers from never-smokers was much more challenging (65% sensitivity at 57% specificity). Gene ontology molecular functions and KEGG pathways associated with current smoking included *G protein-coupled receptor activity*, *signaling receptor activity*, *calcium ion binding*, and the *Neuroactive ligand-receptor interaction* pathway. Selection of marker genes by either moderated *t*-test or multivariate LASSO regression followed by LDA or logistic regression, are robust approaches to classification with omics data, confirming in part findings of previous sbv IMPROVER challenges. While current smoking is accurately identified based on blood mRNA levels, smoking cessation for more than one year is accompanied by a "normalization" of the expression of certain mRNAs, making it difficult to distinguish former smokers from never-smokers.

Introduction

Omics technologies are widespread in life science research because of their ability to measure the activity of tens of thousands of molecules (e.g., mRNAs), providing a systems-level snapshot of molecular activity in a biological sample [1]. Omics data are used to identify novel disease subtypes and markers of disease [2–5], understand the mechanisms of disease and normal physiologic

Abbreviations: KEGG, Kyoto Encyclopedia of Genes and Genomes; IMPROVER, Industrial Methodology for PROcess VErification in Research; LDA, Linear Discriminant Analysis; LASSO, Least Absolute Shrinkage and Selection Operator; sbv, systems biology verification.

* Corresponding author at: 3990 John R., Detroit, MI 48201, USA.
E-mail address: adi@wayne.edu (A.L. Tarca).

http://dx.doi.org/10.1016/j.comtox.2017.07.003
2468-1113/

processes [6,7], and more recently, to evaluate the impact of exposure to different toxicants, termed *systems toxicology* [8–10]. Given the high-dimensionality of these data and hence the ease with which data overfitting can occur [11], an important concern in the field is how to develop reliable models that relate gene expression patterns to an outcome of interest, such as exposure to a toxicant. Crowdsourcing the analysis of omics data has emerged as a framework to identify robust methodologies that address this concern [12]. The Industrial Methodology for Process Verification of Research (IMPROVER) [13] was introduced as an industry-driven initiative to determine best practices for predictive modeling with omics data, similar to the MicroArray Quality Control (MAQC)-II initiative [14] sponsored by the U.S. Food and Drug Administration. The first two sbv IMPROVER initiatives addressed the task of human disease classification from gene expression data (Diagnostic Signature Challenge) [15,16], and prediction of protein phosphorylation and pathway perturbation as responses to stimuli within and across species (Species Translation Challenge) [17–19]. The Systems Toxicology challenge, which is the subject of this article, was designed to determine whether mRNA transcription levels in blood hold information on smoking status (sub-challenge 1, SC1), and whether there is a molecular signature which is translatable from animal models to humans (sub-challenge 2, SC2). In SC1, microarray gene expression data generated from blood samples of current and former smokers, as well as from never-smokers, was provided to participating teams to develop supervised classification models [20]. The organizers provided two test datasets in which the exposure status was concealed from participants, while the identity of the teams was concealed from the organizers, who ranked the teams based on their prediction performance on the test dataset.

This article describes the approaches and results of the first- and third-ranked teams (the second-ranked team was invited to participate in the article, but declined owing to a conflict of interest) that participated in SC1, focusing on the key aspects of the methodology that led to their successful models, as well as on the similarity and differences of the resulting cigarette smoke exposure classification models.

Methods

Challenge organization

The human training dataset that was made available to participating teams was based on the Queen Anne Street Medical Centre (QASMC) clinical case-control study conducted at the Heart and Lung Centre (London, UK) according to good clinical practices (see ClinicalTrials.gov, identifier NCT01780298). The dataset included gene expression data from 109 smokers, 57 former smokers, and 58 never-smokers. The test dataset was obtained by profiling blood samples from a banked repository (BioServe Biotechnologies Ltd., Beltsville, MD, USA), and included samples from 27 smokers, 26 former smokers, and 28 never-smokers. After total RNA extraction, hybridization was performed on Affymetrix GeneChip® Human Genome U133 Plus 2.0 arrays (Affymetrix, Santa Clara, CA, USA). CEL files of raw gene expression data were background corrected, normalized, and summarized into one expression value per ENTREZ Gene ID using frozen robust microarray analysis (fRMA) [21]. For the exact versions of software used, refer to the technical document from the challenge organizers [22].

Participating teams were invited to develop a prediction rule based on the training data and apply it to the test data. The test data were released in two batches to ensure that they were not used in the training processes in any way, as had happened with the second-best overall team in the sbv IMPROVER Diagnostic Signature Challenge [16]. For each test sample, the teams provided a confidence value (probability ranging from 0 to 1) that the sample was taken from a smoker (p_s) as opposed to a non-current smoker (former smoker or never-smoker) ($1 - p_s$). For samples assigned to the non-current smokers group ($p_s \leq 0.5$), a second classification was requested to determine whether the sample came from a former smoker (p_{fs}) or from a never-smoker ($1 - p_{fs}$). Of note, teams that wrongly classified non-current smokers as smokers were penalized by imputing the missing confidence values of these samples in the former smoker vs. never-smoker classification with the worst-case scenario (e.g., p_{fs} for a former smoker misclassified as current smoker was set to 0.0 instead of the ideal 1.0). An alternative evaluation of the prediction performance for this classification task of the first and third ranked teams was based on actual confidence values (requested post-challenge) to compute performance estimates that were comparable and independent between the two classification tasks (smoker vs. non-current smoker and former smoker vs. never-smoker).

The submissions from the participating teams were ranked by computing the area under the precision recall curve (AUPR) [23] and the Matthews correlation coefficient (MCC) [24] for both classification tasks on the test dataset. While the AUPR statistic monitors the precision (positive predictive value) as a function of recall (sensitivity) determined at all cut-off points on the confidence values, the calculation of the MCC statistic involves using a single threshold on the confidence values (i.e., 0.5), hence giving a snapshot of the prediction performance at a single (fixed) sensitivity value. Each team's sum of ranks, based on the rank from each of the four statistics (two metrics × two classification tasks) was used to rank the teams. To determine whether or not the performance metric (AUPR or MCC) of a given team was better than expected by chance, the empirical distribution of the AUPR and MCC statistics was determined as described by *Belcastro* et al. in this issue. The top three teams were invited to contribute to this manuscript; the team ranked second, however, could not join owing to a conflict of interest.

Approach of the first-ranked team (Team 264)

The approach of the first-ranked team (ALT and RR, Team 264) was similar to the one used in previous sbv IMPROVER challenges [17,25], and was applied in the same manner to both classification tasks. In essence, with this approach genes are first ranked by moderated *t*-test [26] which is known to be more robust than ordinary *t*-test in selection of truly differentially expressed genes. With moderated t-tests gene level expression variance is shrunk toward a common value derived using information from all genes on the array. This ensures that genes are not selected as best predictors solely because they have small within group variance, and that, genes that appear to have large variance in the training data are not completely overlooked as long as they have a sizable difference in means between groups. Further, genes are filtered out regardless their significance p-value from the moderated *t*-test if their magnitude of change between groups does not exceed a minimum threshold, since larger differences are more likely to be reproducible in the test set. These steps are designed to remove noisy genes from the pool of candidate genes out of which an optimal number of predictors will be selected as follows:

It involved the following steps:

1) Rank genes by moderated *t*-test [26] *p*-values and select the top N_F genes (optimal value of N_F to be determined), as the genes with nominal $p < 0.05$ and fold change in expression between groups greater than a given Fold Change Threshold (*FCT*). If there are no N_F genes meeting these criteria, use (or complement with) top genes ranked solely by *p*-values. Of

Please cite this article in press as: A.L. Tarca et al., Human blood gene signature as a marker for smoking exposure: Computational approaches of the top ranked teams in the sbv IMPROVER Systems Toxicology challenge, Comput. Toxicol. (2017), http://dx.doi.org/10.1016/j.comtox.2017.07.003