Computational Toxicology xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Computational Toxicology

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

The sbv IMPROVER Systems Toxicology computational challenge: Identification of human and species-independent blood response markers as predictors of smoking exposure and cessation status

Vincenzo Belcastro [a,*,1], Carine Poussin [a,*,1], Yang Xiang [a], Maurizio Giordano [o], Kumar Parijat Tripathi [o], Akash Boda [a], Ali Tugrul Balci [l,3], Ismail Bilgen [l,3], Sandeep Kumar Dhanda [n,3], Zhongqu Duan [i,k,3], Xiaofeng Gong [i,k,3], Rahul Kumar [m,3], Roberto Romero [d,e,f,g,h,3], Omer Sinan Sarac [l,3], Adi L. Tarca [b,c,3], Peixuan Wang [i,k,3], Hao Yang [i,j,3], Wenxin Yang [i,j,3], Chenfang Zhang [i,k,3], Stéphanie Boué [a], Mario Rosario Guarracino [o], Florian Martin [a], Manuel C. Peitsch [a], Julia Hoeng [a]

[a] PMI R&D, Philip Morris Products S.A., Quai Jeanrenaud 5, 2000 Neuchatel, Switzerland[2]
[b] Department of Obstetrics and Gynecology, Wayne State University School of Medicine, Detroit, MI, USA
[c] Department of Computer Science, Wayne State University College of Engineering, Detroit, MI, USA
[d] Perinatology Research Branch, NICHD/NIH/DHHS, Bethesda, MD, USA
[e] Perinatology Research Branch, NICHD/NIH/DHHS, 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] SJTU-Yale Joint Center for Biostatistics, Shanghai Jiao Tong University, Shanghai, China
[j] School of Mathematics Sciences, Shanghai Jiao Tong University, Shanghai, China
[k] Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China
[l] Istanbul Technical University, Istanbul, Turkey
[m] Institute of Microbial Technology, Sector 39A, Chandigarh 160036, India
[n] La Jolla Institute for Allergy and Immunology, 9420, Athena Circle, La Jolla, CA 92037, USA
[o] Istituto di Calcolo e Reti ad Alte Prestazioni CNR, Via P. Castellino, 111 80131 Napoli, Italy

ARTICLE INFO

Article history:
Received 20 April 2017
Received in revised form 23 June 2017
Accepted 12 July 2017
Available online xxxx

Keywords:
Systems toxicology
Computational challenge
Gene signature
Smoking biomarker
Blood biomarkers

ABSTRACT

Cigarette smoking entails chronic exposure to a mixture of harmful chemicals that trigger molecular changes over time, and is known to increase the risk of developing diseases. Risk assessment in the context of 21st century toxicology relies on the elucidation of mechanisms of toxicity and the identification of exposure response markers, usually from high-throughput data, using advanced computational methodologies.

The sbv IMPROVER Systems Toxicology computational challenge (Fall 2015-Spring 2016) aimed to evaluate whether robust and sparse (≤40 genes) human (sub-challenge 1, SC1) and species-independent (sub-challenge 2, SC2) exposure response markers (so called gene signatures) could be extracted from human and mouse blood transcriptomics data of current (S), former (FS) and never (NS) smoke-exposed subjects as predictors of smoking and cessation status. Best-performing computational methods were identified by scoring anonymized participants' predictions.

Worldwide participation resulted in 12 (SC1) and six (SC2) final submissions qualified for scoring. The results showed that blood gene expression data were informative to predict smoking exposure (i.e. discriminating smoker versus never or former smokers) status in human and across species with a high

* Corresponding authors.
E-mail addresses: vincenzo.belcastro@contracted.pmi.com (V. Belcastro), carine.poussin@pmi.com (C. Poussin), yang.xiang@pmi.com (Y. Xiang), maurizio.giordano@cnr.it (M. Giordano), parijat24@gmail.com (K.P. Tripathi), akash.boda@uqconnect.edu.au (A. Boda), alitugrulbalci@gmail.com (A.T. Balci), ibilgen@itu.edu.tr (I. Bilgen), mdusdhanda@gmail.com (S.K. Dhanda), zhqduan@sjtu.edu.cn (Z. Duan), ikergong@gmail.com (X. Gong), rahulk.aiims@gmail.com (R. Kumar), roberto.romero@wayne.edu (R. Romero), ssarac@itu.edu.tr (O.S. Sarac), adi@wayne.edu (A.L. Tarca), peixuan_wang@qq.com (P. Wang), eric_yang09@126.com (H. Yang), ywxsjtu@sjtu.edu.cn (W. Yang), cfzhang2015@sjtu.edu.cn (C. Zhang), stephanie.boue@pmi.com (S. Boué), mario.guarracino@na.icar.cnr.it (M.R. Guarracino), florian.martin@pmi.com (F. Martin), manuel.peitsch@pmi.com (M.C. Peitsch), julia.hoeng@pmi.com (J. Hoeng).

[1] Equal contribution.
[2] Part of Philip Morris International group of companies.
[3] Challenge best performers.

http://dx.doi.org/10.1016/j.comtox.2017.07.004

level of accuracy. By contrast, the prediction of cessation status (i.e. distinguishing FS from NS) remained challenging, as reflected by lower classification performances. Participants successfully developed inductive predictive models and extracted human and species-independent gene signatures, including genes with high consensus across teams. Post-challenge analyses highlighted "feature selection" as a key step in the process of building a classifier and confirmed the importance of testing a gene signature in independent cohorts to ensure the generalized applicability of a predictive model at a population-based level.

In conclusion, the Systems Toxicology challenge demonstrated the feasibility of extracting a consistent blood-based smoke exposure response gene signature and further stressed the importance of independent and unbiased data and method evaluations to provide confidence in systems toxicology-based scientific conclusions.

Introduction

Holy grail in systems toxicology: can specific markers of exposure response to chemical(s) be identified in blood?

Humans are constantly exposed to individual or mixtures of chemicals (e.g., cigarette smoke, pollutants, pesticides, and other chemicals) that may have effects on their cells. If chemicals are harmful and/or cumulative doses of chemicals exceed a threshold limit value, exposure can lead to cellular/tissue damage and dysfunction, which in turn can increase risk of disease development. Hence, identification of specific markers elicited in response to (specific) chemicals is important to assess the exposure status of subjects and to associate exposure with toxicity outcomes. The appropriate combination of identified markers constitutes a specific exposure response fingerprint or signature discriminating exposed and non-exposed subjects. Such a signature may also distinguish formerly-exposed and never-exposed subjects. Exogenous chemicals (e.g., lead), chemical-derived metabolites, and endogenous molecules produced by primarily exposed organs (e.g., lung, gut) can pass into the blood stream and may induce molecular changes in blood cells [1]. Therefore, investigating whether specific markers in response to chemical exposure can be identified in blood cells may be highly valuable for monitoring chemical exposure [2,3]. Interestingly, new 'omics' technologies (e.g., genomics, transcriptomics, proteomics, metabolomics, lipidomics) can be applied to toxicity testing in order to increase efficiency and provide a more data- and system-driven approach to exposure response assessment [3,4].

Gene signature-based classification models for biological/clinical status prediction

Transcriptomics-based technologies enable biological mechanistic insights to be gained by measuring whole-genome gene expression levels. Transcriptomics data have also been used extensively to develop classification models predictive of disease diagnosis or prognosis, tumor subtyping, adverse drug response, and therapeutic outcome [5–8]. Gene signatures are generally derived from disease-relevant tissues such as liver, lung, and tumors. However, blood can be collected easily for diagnostics (minimally invasive) and only small quantities are necessary for transcriptomics profiling. Therefore, more and more investigations have used blood samples to identify gene signatures that may be leveraged for the development of tests such as In Vitro Diagnostic Multivariate Index Assays [9,10]. The real-world application of gene signature-based classification models as reliable tools for predictive medicine is still limited [11]. This is mainly because (i) it is difficult to identify robust and sufficiently sparse signatures for the development of ready-to-use diagnostic tools and (ii) the way models are built often leads to poor predictive performance when applied to new individual samples (e.g., lack of validation in independent cohorts to test robustness and generalized applicability in populations, or the use of transductive method-based models [12]).

The systems biology verification Industrial Methodology for PROcess VErification in Research (sbv IMPROVER [13]; https://sbvimprover.com) project aims to verify methods and data in systems biology/toxicology using double-blind performance assessment. Over the past six years, sbv IMPROVER organized crowd-sourced challenges covering a broad range of scientific questions [14–18]. The first one titled Diagnostic Signature Challenge in 2012 was designed to assess to what extent models trained on transcriptomics data available in public repositories could predict the diagnosis of individual subjects in unrelated datasets for four disease types [18]. Many of the classification models proposed were transductive (i.e., training and test sets are processed together and prediction model solely applies to this specific test set) rather than inductive (i.e., the signature model is applied to a single new sample without retraining), which may lead to poor classification on a new single patient sample and may be impracticable for real-world application. These limitations were considered in the design and constraint of new classification problems in our latest computational challenge open to the scientific community and described below.

Application of omics-based classification to toxicogenomics using blood as surrogate tissue: prediction of tobacco smoke exposure and cessation

In liver and pulmonary toxicity studies, gene signatures have been identified successfully in blood showing (i) capability to predict exposure and toxicity to chemicals such as acetaminophen in liver (drug-induced liver injury) or crystalline silica in lung; (ii) superior sensitivity as predictors of toxicity compared with the classical toxicity markers in rats; and (iii) to some extent, similarities in pathways and functions that are perturbed in primary tissue and blood [3]. These findings, in addition to its easy access, make blood highly relevant as a surrogate to identify gene expression-based signatures as specific markers for toxicological evaluation and risk assessment. Smoking is a major risk factor for the development of various diseases (e.g., cardiovascular and lung diseases) [19]. Smokers are exposed to a mixture of thousands of chemical constituents when cigarette smoke is inhaled. Among them, some constituents or their metabolites that pass into the blood circulation elicit systemic effects distal from the lungs, the primary site of exposure. For example, changes in gene expression in circulating peripheral blood cells are associated with several systemic immune and inflammatory-related disorders [20,21]. Smoking cessation has been shown to revert some cigarette smoke-induced functional and molecular changes back to non-smoker levels or intermediate levels depending on the subject's smoking history (e.g., smoking duration, consumption) and

Please cite this article in press as: V. Belcastro et al., The sbv IMPROVER Systems Toxicology computational challenge: Identification of human and species-independent blood response markers as predictors of smoking exposure and cessation status, Comput. Toxicol. (2017), http://dx.doi.org/10.1016/j.comtox.2017.07.004