From Transcriptomics to Predictive Toxicology

The Systems Toxicology Computational Challenge

Carine Poussin, Ph.D., Vincenzo Belcastro, Ph.D. and Julia Hoeng, Ph.D.

Exposure to external toxicants (cigarette smoke, pollutants, pesticides etc.) can induce significant molecular changes in human blood. Given that blood is easily accessible, it would be advantageous to identify specific markers in blood cells that could predict whether an individual had been exposed to a given toxicant. Such knowledge would have valuable implications for the toxicological risk-assessment of chemicals, drugs and consumer products, as well as for diagnostics.

However, blood is a complex tissue to analyze, primarily due to the many different cell sub-populations it contains. Molecular changes brought about by exposure to a toxicant may involve a complex interplay of a sub-set of the chemicals present in the toxicant itself, molecules produced by the exposed organ (e.g., the lungs or the gut), and chemical-derived metabolites.

Furthermore, the real-world application of models based on blood markers for predictive classification of individuals is uniquely challenging. The difficulty resides in the identification of relevant markers in blood after chemical exposure, the low success of correct classification when predictive models are applied on new individual blood samples, and the translation of these techniques into practical ready-to-use tools. In addition, most pre-clinical toxicological in vivo studies are conducted in rodents, adding a degree of complexity when applying the results to humans.

The Systems Toxicology Computational Challenge

The sbv IMPROVER Systems Toxicology Computational Challenge was designed to explore these issues and to help increase scientists' understanding of what is necessary to reach higher levels of predictability and robustness in predictive toxicology. Specifically, the scientific questions raised in the challenge concerned the identification of blood response markers and models that can predict smoking exposure or cessation status.

The challenge was open to anyone working in computational sciences who develops predictive modeling techniques. Provided with blood transcriptomics datasets,

Carine Poussin, Ph.D, is Senior scientist, Computational Biology at Philip Morris International; **Vincenzo Belcastro, Ph.D.,** is Scientist, Systems Biology at Philip Morris International; **Julia Hoeng, Ph.D.,** is Director of Systems Toxicology, Biological Systems Research, Philip Morris International. (Sbvimprover.RD@pmi.com)

Inductive vs Transductive Approaches to Prediction

A previous sbv IMPROVER challenge—the Diagnostic Signature Challenge—assessed the extent to which markers for four distinct diseases could be extracted from transcriptomics data held in public repositories (training datasets), and then used to make diagnostic predictions on completely unrelated datasets (test datasets). An interesting aspect of this earlier challenge was that most of the models developed were transductive, i.e., they relied on processing both training and test datasets within the same model, class predictions then being made for subjects from the test datasets.

When it comes to real-world application, transductive approaches to prediction are limited since they cannot reliably be generalized and independently applied to new datasets. The Systems Toxicology Computational Challenge was structured to address this problem, stipulating that the models proposed must be inductive, i.e., capable of being applied to new, individual blood samples without the need for any adjustments, thus making them potentially suitable for ready-to-use diagnostic tools.

"We were driven by the desire to create a model that can both lead to valuable biological insight, and be implemented in practice at the lowest possible cost. The Systems Toxicology Computational Challenge has allowed us to test the quality of our research and I'm delighted that our approach has proved to be robust."

— Adi L. Tarca, Associate Professor
Wayne State University, School of Medicine
First place best-performer sub-challenge one, joint second place best-performer sub-challenge two

Matias Castello / EyeEm / Getty Images

participants were asked to solve two tasks. First, they were asked to derive predictive classification models that would distinguish current tobacco smokers from non current smokers (prediction of smoking exposure status). Second, they were asked to discriminate non current smokers as former smokers and never smokers (prediction of cessation status). Anonymized participants' submissions were then scored against a gold-standard dataset, with final results and rankings approved by an independent expert scoring review panel.

The challenge included two independent sub-challenges each aiming to address both tasks using human blood data only (sub-challenge 1) and both human and mouse blood data together (sub-challenge 2). The first sub-challenge explored whether gene expression changes in human blood are sufficiently informative to predict smoking exposure or cessation status. The second investigated the issue of species translatability, with the identification of species-independent blood markers applicable from in vivo rodent studies to clinical blood samples to assess exposure status in humans.

Challenge participants used their own computational techniques to make their predictions, with best-performers achieving accuracy of up to 95 percent in distinguishing current tobacco smokers from non-smokers. Predicting whether non-smokers were former smokers or never smokers was more challenging, suggesting that these two groups are likely to have similar gene expression profiles.

(continued on next page)

BEST PERFORMING TEAMS	
Sub-Challenge 1	***Team Members***
1st	Adi L. Tarca, Associate Professor, Wayne State University, School of Medicine, USA Prof Roberto Romero, Wayne State University, School of Medicine, USA
2nd	Anonymous Entry
3rd	Xiaofeng Gong, Shanghai Jiao Tong University-Yale Joint Center for Biostatistics, China Wenxin Yang, Shanghai Jiao Tong University-Yale Joint Center for Biostatistics, China Zhongqu Duan, Shanghai Jiao Tong University-Yale Joint Center for Biostatistics, China Peixuan Wang, Shanghai Jiao Tong University-Yale Joint Center for Biostatistics, China Hao Yang, Shanghai Jiao Tong University-Yale Joint Center for Biostatistics, China Chenfang Zhang, Shanghai Jiao Tong University-Yale Joint Center for Biostatistics, China
Sub-Challenge 2	***Team Members***
1st	Ömer Sinan Saraç Associate Professor, Istanbul Technical University, Turkey İsmail Bilgen, Istanbul Technical University, Turkey Ali Tuğrul Balcı, Istanbul Technical University, Turkey
2nd (joint)	Adi L. Tarca, Associate Professor, Wayne State University, School of Medicine, USA Prof Roberto Romero, Wayne State University, School of Medicine, USA
2nd (joint)	Dr Sandeep Kumar Dhanda, La Jolla Institute for Allergy & Immunology, USA Dr Rahul Kumar, London, UK

A total of 135 scientists from around the world registered for the challenge, the primary incentive being the opportunity to vigorously and objectively test their methodologies on high-quality, large scale datasets. The results were presented in July at the Intelligent Systems for Molecular Biology conference in Orlando, Fla. The first placed best-performing teams of each sub-challenge presented their methods during the official technology track of the conference, while other high ranking teams presented their methods at a dedicated sbv IMPROVER symposium that ran alongside the conference.

sbv IMPROVER

The Systems Toxicology Computational Challenge is the latest to be run under the sbv IMPROVER umbrella, a crowdsourcing initiative led and funded by Philip Morris International which is designed to test and verify scientific methods and results. The three previous challenges were the Diagnostic Signature Challenge, which asked participants to identify robust diagnostic signatures across four disease areas, the Species Translation Challenge, which sought to refine understanding of the limits of rodent models as predictors of human biology, and the Network Verification Challenge, designed to review biological network models for use in toxicological risk assessment. Further information about the sbv IMPROVER project is available on the project website: www.sbvimprover.com