Current Developments in Machine Learning Techniques in Biological Data Mining

Bioinformatics and Biology Insights
Volume 11: 1–4
© The Author(s) 2017
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1177932216687545
SAGE

Gerard G Dumancas[1], Indra Adrianto[2], Ghalib Bello[3] and Mikhail Dozmorov[4]

[1]Department of Mathematics and Physical Sciences, Louisiana State University, Alexandria, LA, USA. [2]Quantitative Analysis Core, Arthritis & Clinical Immunology Research Program, Oklahoma Medical Research Foundation, Oklahoma City, OK, USA. [3]Department of Environmental Medicine & Public Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [4]Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, USA.

AIMS AND SCOPE

This supplement is intended to focus on the use of machine learning techniques to generate meaningful information on biological data.

This supplement under *Bioinformatics and Biology Insights* aims to provide scientists and researchers working in this rapid and evolving field with online, open-access articles authored by leading international experts in this field. Advances in the field of biology have generated massive opportunities to allow the implementation of modern computational and statistical techniques. Machine learning methods in particular, a subfield of computer science, have evolved as an indispensable tool applied to a wide spectrum of bioinformatics applications. Thus, it is broadly used to investigate the underlying mechanisms leading to a specific disease, as well as the biomarker discovery process.

With a growth in this specific area of science comes the need to access up-to-date, high-quality scholarly articles that will leverage the knowledge of scientists and researchers in the various applications of machine learning techniques in mining biological data.

The area of biology has evolved in a manner that encompasses mining meaningful information to answer the next biological question. Machine learning techniques are computational methods that use "experience" to improve performance or to make accurate predictions. Within the context of this supplement, experience refers to past information such as electronic data available to the learners, which are then consequently used for analyses.[1]

Over the years, the collection of biological information has been rapidly increasing due to the developments and improvements of existing technologies and facilities. An excellent example would be the Human Genome Project, founded in 1990 by the US Department of Energy and the US National Institutes of Health and was eventually completed in 2003. The rapid growth of these massive data eventually led to the need for the use of computational and statistical methods to organize, maintain, and interpret biological results.[2]

There is a strong motivation in the use of machine learning methods in knowledge discovery and data mining to generate models of biological implications. The history of the relationship between machine learning and biology is considered long and complex. One of the early techniques used in machine learning called perceptron constituted an attempt to mimic and model the the behavior of a biological neuron. Consequently, the area of artificial neural network (ANN) emerged from this attempt.[3]

This supplement covers a wide array of topics involving the use of machine learning techniques to extract meaningful information from genetic and clinical data with the primary objective of answering biological questions. Various applications of machine learning techniques in different areas are covered in this supplement, including its use for predicting human leukocyte antigen (HLA)-peptide binding activity, integrating disparate short-read alignment algorithms for mapping next-generation sequencing reads to a reference genome, identifying similar diseases by semantic and genomic similarity, as well as development of risk assessment tools for prediction of life expectancy using genetic algorithm (GA) and weighted quantile sum (WQS) regression. The network analysis, as well as other machine learning techniques, is reviewed by Luo et al.[4] This review summarizes different methods and tools for predicting binding properties of the HLAs. The HLA system produces a variety of peptides that play a critical role in immune system regulation by recognizing foreign antigens and presenting them to different types of immune cells. The review by Luo et al[4] covers a wide variety of machine learning methods for predicting HLA binding, from regression-based techniques and decision trees through support vector machines (SVMs), hidden Markov models, and ANNs. The authors discuss the strengths and limitations of each method and propose a combination approach that uses multiple types of prediction methods to address some of the limitations.

The complexity of immune system was also addressed in the paper by Dozmorov et al.[5] The authors compare 2 methods, csSAM and DSection, for detection of cell type–specific differential expression by analyzing RNA-seq data obtained from a heterogeneous mixture of immune cells of healthy individuals and patients diagnosed with an autoimmune disease systemic lupus erythematosus. Both methods use linear regression to estimate cell type–specific gene expression differences, whereas the DSection method also uses Bayesian approach to estimate cell proportions. Dozmorov et al[5] compared the results of cell type–specific differential expression analysis with genes differentially expressed in heterogeneous mixture of cells. In addition to evaluating csSAM and DSection methods applied to the cell type–specific differential expression analysis of RNA-seq data, the authors provide a brief methodologic overview of gold standard tools for differential expression analysis of RNA-seq data.

Adams et al[6] provided a novel examination on the use of GA for determining variables predictive of mortality. Their manuscript offers a novel method involving the use of a GA approach for selection of predictive variables from health questionnaire data. The selected variables are then used to construct predictive models of 5-year mortality using various machine learning techniques. Parametric and nonparametric machine learning algorithms are emerging computational methods that have increasing applications in the area of bioinformatics and computational biology. Results obtained from this study will provide novel insights for computational biologists and bioinformaticians to use GA in conjunction with machine learning techniques to efficiently select important variables and also determine their collective prediction accuracy. The various machine learning techniques used in the study included gradient boosting, ANN, elastic net, SVM, ridge regression, logistic regression, random forest, least absolute shrinkage and selection operator (LASSO), partial least squares-discriminant analysis, and decision trees. The optimization of variable selection for questionnaire data and the construction of predictive models using selected variables are areas of interest for researchers and clinicians alike. The study demonstrates the feasibility of various machine learning techniques for developing both prognostic and explanatory models using data collected via surveys or questionnaires.[6]

Bello et al[7] used a statistical technique known as weighted quantile sum (WQS) regression to develop a model that condenses the information from a variety of health markers into a composite index (referred to as the health status metric [HSM]). The HSM can be used as a holistic measure of overall health and also as a risk score for predicting all-cause mortality. Indeed, results of their study indicated that the index was highly predictive of life expectancy and long-term health-related outcomes such as hospital utilization. Weighted quantile sum regression is a novel, penalized regression method that was developed to handle multicollinear data, the situation whereby complex correlation patterns exist among multiple variables. Weighted quantile sum controls the variance inflation arising from multicollinearity by imposing nonnegative and unit-sum penalties on model coefficients. It is a powerful alternative to other penalized regression techniques, such as LASSO and the elastic net, which are commonly used in machine learning. The study demonstrates the utility of WQS in predictive modeling and development of risk scores.

REFERENCES

1. Mohri M, Rostamizadeh A, Talwalkar A. *Foundations of Machine Learning*. Cambridge, MA: The MIT Press; 2012:427 pp.
2. Machine learning and data mining in bioinformatics. http://lpis.csd.auth.gr/publications/Tzanis_Handbook09.pdf. Accessed September 26, 2016.
3. Tarca AL, Carey VJ, Chen X, Romero R, Drăghici S. Machine learning and its applications to biology. *PLoS Comput Biol*. 2007;3:e116.
4. Luo H, Ye H, Ng HW, et al. Machine learning methods for predicting HLA-peptide binding activity. *Bioinform Biol Insights*. 2015;9:21–29.
5. Dozmorov M, Dominguez N, Bean K, et al. B-cell and monocyte contribution to systemic lupus erythematosus identified by cell-type-specific differential expression analysis in RNA-Seq data. *Bioinform Biol Insights*. 2015;9:11–19.
6. Adams LJ, Bello G, Dumancas GG. Development and application of a genetic algorithm for variable optimization and predictive modeling of five-year mortality using questionnaire data. *Bioinform Biol Insights*. 2015;9:31–41.
7. Bello G, Dumancas GG, Gennings C. Development and validation of a clinical risk-assessment tool predictive of all-cause mortality. *Bioinform Biol Insights*. 2015;9:1–10.

Lead Guest Editor **DR. GERARD G DUMANCAS**

Professor is presently an Assistant Professor of Chemistry at Louisiana State University of Alexandria. He is also presently the Vice President for Research and Development at LipidTech, LLC, a start-up company he cofounded in January 2015. He completed his Postdoctoral Fellowship (2012-2014) at the Oklahoma Medical Research Foundation specializing in the areas of bioinformatics, chemometrics, and molecular genetics. He finished his PhD in Chemistry at Oklahoma State University in 2012 specializing in analytical chemistry (chemometrics). His present research encompasses several of science, including bioanalytical chemistry, environmental chemistry, chemometrics, bioinformatics, and machine learning techniques. Dr Dumancas is the author or coauthor of 18 research articles, 17 book chapters, 1 book, and an intellectual property disclosure. He has accumulated a total of 32 oral and poster research presentations. He also presently holds editorial appointments in the Research Journal of Applied Sciences, Engineering, and Technology, American Journal of Epidemiology and Infectious Disease, American Oil Chemists' Society (AOCS) Inform (2011-2013), AOCS USA Section Leadership Team, and the International Journal of Stem Cell Research and Transplantation (webpage: http://dumancas.weebly.com; LinkedIn: https://www.linkedin.com/pub/gerard-dumancas-phd/35/497/25; researchgate: http://www.researchgate.net/profile/Gerard_Dumancas/publications).

Email: gerard.dumancas@okstate.edu

Guest Editors

DR. INDRA ADRIANTO

Professor is a Research Assistant Member and Director of Operations of the Quantitative Analysis Core of the Arthritis & Clinical Immunology Research Program at Oklahoma Medical Research Foundation. He completed his PhD at the University of Oklahoma. He has extensive research experience in both traditional and advanced statistics, statistical genetics and genomics, biostatistics, bioinformatics, genetic epidemiology, admixture mapping, next-generation sequencing data analysis (DNA and RNA sequencing), network modeling, data mining, and machine learning. He now works primarily in the area of autoimmune and inflammatory disease genetics and genomics. He is the a lead or contributing author on multiple manuscripts published in major journals such as *Nature Genetics, Arthritis & Rheumatology, and American Journal of Human Genetics*, among others. Learn more about Dr Adrianto by visiting his institutional webpage: http://omrf.org/research-faculty/scientist/adrianto-indra/.

Email: Indra-Adrianto@omrf.org

Guest Editors

DR. GHALIB BELLO

Professor is an Assistant Professor in the Division of Biostatistics at the Department of Environmental Medicine and Public Health in the Icahn School of Medicine at Mount Sinai, New York. He earned his PhD in Biostatistics at Virginia Commonwealth University and completed a postdoctoral fellowship in Statistical Genetics at the Oklahoma Medical Research Foundation. Dr. Bello's research interests are in environmental health, machine learning, genetic epidemiology and risk prediction for personalized medicine. His current research focus is on the development of statistical methodology for rigorous quantitative assessment of the health effects of exposure to complex mixtures of environmental pollutants. His work in this area involves the use of machine learning tools and network analysis for evaluation of mixture effects and synergistic interactions among environmental chemicals. He is also involved in efforts to develop and improve data driven approaches to healthcare. He has used machine learning and data mining techniques to build risk models for the prediction of life expectancy and onset of various chronic diseases, and is exploring novel approaches for the quantification of allostatic load, cumulative disease burden and aging-related physiological dysregulation. Dr. Bello is the author of multiple research articles and book chapters. Learn more about him by visiting his institutional webpage: http://icahn.mssm.edu/profiles/ghalib-bello.

Email: ghalib.bello@mssm.edu

Guest Editors

DR. MIKHAIL DOZMOROV

Professor is an Assistant Professor of Biostatistics at Virginia Commonwealth University. His primary area of expertise is computational genomics and epigenomics. His research involves developing computational and statistical methods for the integrative analysis of “omics” data and focuses on precision medicine approaches. He has developed a bioinformatics program and a biostatistics approach, GenomeRunner (http://www.integrativegenomics.org/), to automate genome and epigenome exploration. He uses functional epigenomic data from the ENCODE and Roadmap Epigenomics projects to understand the biology of complex disorders, such as cancer and autoimmune diseases. In his research, he works with data from TCGA, dbGAP, GEO, and PGC2 databases and collaborates with scientists interested in an in-depth understanding of the molecular mechanisms from a global perspective. He has published more than 40 peer-reviewed publications, actively participates in presenting on scientific conferences, serves on the Board of Directors of the MCBIOS bioinformatics conference, and is a statistical advisor of *PLoS ONE* family of journals. Learn more about Dr Dozmorov by visiting his institutional webpage: http://www.biostatistics.vcu.edu/mikhail-g-dozmorov/.

Email: mikhail.dozmorov@vcuhealth.org

TYPE: Editorial

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARAT ION OF CONFL ICT ING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPON DING AUTHOR: Gerard G Dumancas, Louisiana State University, 8100 Hwy 71 South Alexandria, Louisiana, 71302, USA. Email: gerard.dumancas@okstate.edu