Guest Editorial—Special Collection Topic: Statistical Systems Theory in Cancer Modeling, Diagnosis, and Therapy

Edward R Dougherty1, Anne-Laure Boulesteix2, Lori A Dalton3 and Michelle Zhang4

1Department of Electrical & Computer Engineering, Texas A&M University, College Station, TX, USA. 2Institute for Medical Information Processing, Biometry and Epidemiology, University of Munich, Munich, Germany. 3Departments of Electrical and Computer Engineering and Biomedical Informatics, The Ohio State University, Columbus, OH, USA. 4Department of Electrical and Computer Engineering, The University of Texas at San Antonio, San Antonio, TX, USA.

AIMS AND SCOPE:

Cancer is a systems disease involving mutations and altered regulation. This supplement treats cancer research as it pertains to 3 systems issues of an inherently statistical nature: regulatory modeling and information processing, diagnostic classification, and therapeutic intervention and control. Topics of interest include (but are not limited to) multiscale modeling, gene/protein transcriptional regulation, dynamical systems, pharmacokinetic/pharmacodynamic modeling, compensatory regulation, feedback, apoptotic and proliferative control, copy number-expression interaction, integration of different feature types, error estimation, and reproducibility. We are especially interested in how the above issues relate to the extremely high-dimensional data sets and small- to moderate-sized data sets typically involved in cancer research, for instance, their effect on statistical power, inference accuracy, and multiple comparisons.

High-Dimensional Statistical Learning: Roots, Justifications, and Potential Machineries by Zollanvari reviews methods that have been developed for high-dimensional data sets. The problem is important in cancer informatics because one is regularly confronted by thousands of features and classical techniques that appeal to asymptotic performance often work poorly when the dimensionality is so great that asymptotic results do not reflect finite-sample performance.

Data Requirements for Model-Based Cancer Prognosis Prediction by Lori Dalton and Mohammadmahdi Yousefi considers cancer prognosis prediction based on the integration of existing pathway knowledge and data. It extends previous work using optimal Bayesian classification based on uncertainty classes of Boolean regulatory networks by considering how population data may be used to estimate network probabilities and by considering optimal Bayesian regression of prognosis metrics.

A New Approach for Identification of Cancer-related Pathways using Protein Networks and Genomic Data by Fonseca et al. addresses the unitization of gene expression data and knowledge of signaling pathways to study the control system in cancer cells. This is accomplished via a statistical computational model. Specifically, a procedure is proposed to recover small protein networks that are differentially expressed in subtypes of breast cancer. These are enriched with specific gene ontologies and new putative cancer genes.

A Bayesian Nonparametric Approach for Functional Data Classification with Application to Hepatic Tissue Characterization by Fronczyk et al. proposes a Bayesian semiparametric model to analyze four interdependent hepatic perfusion computed tomographic characteristics acquired under the administration of contrast using a sequence of repeated scans. The model is applied to measurements from liver regions surrounding malignant and benign tissues, the aim being to cluster the liver regions on the basis of their computed tomography perfusion profiles, which can be used for diagnosing malignant tissue.

Bayesian ABC-MCMC Classification of Liquid Chromatography-Mass Spectrometry Data by Banerjee and Braga-Neto applies the optimal Bayesian classifier using a model of the liquid chromatography-mass spectrometry experiment for proteomic-based classification. Computation of the optimal Bayesian classifier is facilitated by a likelihood-free methodology called approximate Bayesian computation and Markov chain Monte Carlo sampling.

An Application of Sequential Meta-Analysis to Gene Expression Studies by Novianni et al. applies sequential meta-analysis to find gene-expression signatures in acute myeloid leukemia. Sequential meta-analysis combines studies in chronological order while preserving the type I error and prespecifying the statistical power to detect a given effect size. Sequential meta-analysis of seven data sets is used to evaluate whether the accumulated samples show enough evidence or more experiments should be initiated.

Publication Bias in Methodological Computational Research by Boulesteix et al. proposes a new framework to formalize the notion of publication bias in the context of methodological computational research. The authors examine unpublished research with the goal of discovering factors in publication bias to facilitate formalization of the concept.

Integrating Multiscale Modeling with Drug Effects for Cancer Treatment by Li et al. reviews multiscale modeling for cancer treatment. Although systems biology focuses on multifactorial controls over biological processes, systems pharmacology...
studies drugs regarding the pharmacokinetic and pharmacodynamic relations accompanying drug interactions. Multiscale methods are required to integrate models from molecular levels to cellular, tissue, and organism levels.

Assessing Combinational Drug Efficacy in Cancer Cells by Using Image-based Dynamic Response Analysis by Sima et al. proposes a new measure to assess combinational drug efficacy. Although traditional efficacy measures focus on the extent of drug inhibition, this paper measures the speed of killing, based on live cell imaging. This dynamic response trajectory approach takes both extent and speed into account, thereby revealing synergisms that would otherwise be missed.

Optimized Prediction of Extreme Treatment Outcomes in Ovarian Cancer by Misganaw et al. uses the lone star algorithm to identify 25 genes with which to classify patients with ovarian cancer into one of the three groups relative to platinum-based chemotherapy: super responders, medium responders, and nonresponders. The paper also proposes a discriminant function to divide the patient population into two survival classes.

Design of Probabilistic Random Forests with Applications to Anticancer Drug Sensitivity Prediction by Rahman et al. generalizes random forests by representing the regression trees as probabilistic trees and analyzing the nature of heteroscedasticity. This representation facilitates analytical computation of confidence intervals and tree weight optimization provides stricter confidence intervals with comparable performance in mean error. The method is applied to drug sensitivity prediction.

Flow Cytometry-Based Classification in Cancer Research: A View on Feature Selection by Hassan et al. examines feature selection in cancer-related machine learning regarding accuracy and stability in the context of very small samples. The paper also considers model selection among the 11 regularization paths of logistic regression classifiers and compares cross-validation with a recently proposed Bayesian error estimator.

An NGS Workflow Blueprint for DNA Sequencing Data and Its Application in Individualized Molecular Oncology by Li et al. discusses issues pertaining to next-generation sequencing (NGS): reduction in sequencing cost, enhancement of sequencing quality, improvement of technical simplicity and reliability, and development of semi-automated and integrated analysis workflow. The authors conduct a literature search and summarize a 4-stage NGS workflow to provide a systematic review on NGS-based analysis.

Quantitative Proteomic Approach for MicroRNA Target Prediction Based on 18O/16O Labeling by Ma et al. develops a quantitative proteomic approach based on 18O/16O labeling and applies it on Kaposi sarcoma-associated herpesvirus (KSHV) microRNA (miR) target prediction. A method is proposed whereby several 18O/16O data processing algorithms are integrated to identify the messenger RNAs of downregulated proteins as potential targets in KSHV miR-transfected human embryonic kidney 293T cells.
Guest Editors

DR LORI A DALTON

Dr Lori A Dalton is an Assistant Professor of Electrical and Computer Engineering and an Assistant Professor of Biomedical Informatics at The Ohio State University. She completed her PhD at Texas A&M University. She now works primarily in Genomic Signal Processing. Dr Dalton is the author or coauthor of 18 published journal papers and has presented at 19 conferences and workshops. Learn more about Dr Dalton by visiting her institutional Web page: http://www2.ece.ohio-state.edu/~dalton/.

Email: dalton@ece.osu.edu

Guest Editors

DR MICHELLE ZHANG

Dr Michelle Zhang is an Assistant Professor of Electrical & Computer Engineering at The University of Texas at San Antonio. She completed her PhD at State University of New York at Stony Brook and has previously worked at University of New Hampshire and Greehey Children’s Cancer Institute. She now works primarily in proteomic mass spectrometry, Bayesian statistical signal processing methods, bioinformatics, biomarker discovery, and classifications. Learn more about Dr Zhang by visiting her institutional Web page: http://compgenomics.utsa.edu/jzhang/indexroot.htm.

Email: michelle.zhang@utsa.edu

TYPE: Editorial: Special Collection: Statistical Systems Theory in Cancer Modeling, Diagnosis, and Therapy

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

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

CORRESPONDING AUTHOR: Edward R Dougherty, Department of Electrical & Computer Engineering, Texas A&M University, College Station, TX, USA. Email: edward@ece.tamu.edu