Journal of Biomedical Informatics 63 (2016) 120–130

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

An approach for deciphering patient-specific variations with application to breast cancer molecular expression profiles

Radhakrishnan Nagarajan [a,*] (Ph.D), Meenakshi Upreti [b] (Ph.D.)

[a] Division of Biomedical Informatics, College of Medicine, University of Kentucky, KY, USA
[b] Department of Pharmaceutical Sciences, Markey Cancer Center, University of Kentucky, KY, USA

ARTICLE INFO

Article history:
Received 15 March 2016
Revised 6 July 2016
Accepted 27 July 2016
Available online 28 July 2016

Keywords:
Translational bioinformatics
Precision medicine
Data mining
Molecular profiling

ABSTRACT

Several studies have successfully used molecular expression profiling in conjunction with classification techniques for discerning distinct disease groups. However, a majority of these studies do not provide sufficient insights into potential patient-specific variations within the disease groups. Such variations are ubiquitous and manifests across multiple scales with varying resolution. There is an urgent need for novel approaches that falls within the objective of precision medicine and provide novel insights into patient-specific variations and sub-populations within disease groups while discerning the disease groups of interest so as to enable timely and targeted intervention of select subjects. This study presents a selective-voting ensemble classification approach (SVA) for discerning good and poor-prognosis breast cancer samples from their 70-gene molecular expression profile revealing patient-specific variations within the poor-prognosis group. In contrast to traditional classification, SVA adapts the feature sets in a sample-specific manner capturing the proclivity of the samples to each of the disease groups. Correlation between normalized vote counts from SVA and clinical outcomes of the subjects is elucidated. Performance of Support Vector Machine and Naïve Bayes classifier is investigated within the SVA framework and compared to established clinical criteria (Nottingham Prognostic Index, Adjuvant Online, St. Gallen) and Mammaprint approach. Weighted undirected graph abstractions of the ensemble sets of the poor-prognosis test samples is also shown to exhibit markedly different topologies with varying proclivities. These patient-specific networks may reflect inherent variations in underlying signaling mechanisms in the poor-prognosis subjects and reveal potential targets for personalized therapeutic intervention.

1. Introduction

Molecular mechanisms underlying disease phenotypes have been shown to capture the biology of the disease and complement traditional clinical characterization [1–12]. There is an increasing trend in incorporating molecular assays in clinical workflows [13]. While several studies have demonstrated the usefulness of molecular expression profiling for discerning distinct disease groups, their primary goal has been to improve the overall classifier performance with minimal insights into potential patient-specific variations within each of the disease groups. With recent interest in "precision medicine" [14,15] and increasing evidence of marked variations within disease groups, there is an urgent need for developing novel approaches that has the potential to discern the disease groups while revealing patient-specific variations so as to enable personalized treatment regimens in a targeted and timely manner.

Patient-specific variations can manifest across multiple scales. At the molecular scale, studies have clearly attributed variations in DNA sequence composition to variations in disease phenotype (e.g. genotype-phenotype associations) and drug response [16–18]. On a related note, genetic variations that affect a given phenotype should ideally affect intermediate processes such as transcription and translation that are proximal to phenotype as per the central dogma [19]. Variations in transcriptional and translational activities have also been attributed to inherent stochastic mechanisms in molecular systems. More importantly, these stochastic mechanisms have been shown to persist even across isogenic single cells with non-trivial impact on the end phenotype

* Corresponding author at: Division of Biomedical Informatics, College of Medicine, University of Kentucky, 725 Rose Street, Multidisciplinary Science MDS 230F, Lexington, KY 40536-0082, USA.
E-mail address: rnagarajan@uky.edu (R. Nagarajan).

http://dx.doi.org/10.1016/j.jbi.2016.07.022
1532-0464/

[20–23]. At the tissue level, tumor samples have been shown to be heterogeneous comprising of multiple cell types including the tumor cells and those of its microenvironment and characteristic architecture that play a significant role on tumor initiation and progression [24–26]. At the patient level, chronic comorbidities (e.g. Type II Diabetes) have been shown to contribute to marked differences in disease progression [27]. Risk factors (e.g. smoking), lifestyle and demographic factors (e.g. age) can also contribute to patient-specific variations [28,29]. Also, majority of molecular expression profiling are cross-sectional in nature and represent a snapshot in the disease continuum quantized in amplitude, space and time. The disease groups of interest also may not necessarily be well-separated in the disease continuum. For the above reasons and perhaps more, patient-specific variations are to be expected and can be thought of as the cumulative effect of the complex interplay between the various factors across multiple scales. While deciphering the source of variation is a challenging problem in its own merit, accommodating potential variations between subjects as part of the classification process is a critical step in developing patient-tailored treatment regimens and identifying sub-populations with distinct disease trajectories.

There has been recent interest in understanding patient-specific variations using high-throughput molecular assays and integrated data sources in conjunction with novel algorithms. With the recent explosion in genome sequencing technologies a number of studies have investigated variations in genetic markers in a patient-specific manner [30–32]. Studies (e.g. PARADIGM, Pathifier) [33,34] have also proposed integrated approaches to decipher patient-specific variations in pathways that are curated from existing pathway databases. The present study takes a different tack to this problem and investigates patient-specific variations from molecular expression profiles using a classification framework without relying on prior knowledge regarding any interaction between the molecules of interest. While it focuses on transcriptional expression profiles, the framework as such is generic and can be extended to accommodate mixed data types as well as integrated data sets. Classification techniques can be broadly classified into single classifiers and multiple or ensemble classifiers. (a) *Single Classifiers*: Single classifiers have been successfully used to discern distinct disease groups from their molecular expression profiles [1,35]. However, such an approach traditionally uses all the molecular markers simultaneously as features in discerning the samples between the disease groups. On a related note, using all the features simultaneously in the classification may also render the dimensionality of the feature space comparable to that of the sample-size resulting in a sparse representation of the samples in a high-dimensional space and potential overfitting. Missing values are not uncommon in molecular assays and a missing value even across a single feature can affect the overall performance of single classifiers that use all the features simultaneously in the classification process. While statistically motivated approaches such as imputation can be used to alleviate these issues, such approaches are based on implicit assumptions [36]. (b) *Multiple Classifiers/Ensemble Classifiers*: Unlike traditional single classifiers, ensemble classifiers [37–40] rely on the "wisdom of crowds" where the classification label is predicted based on the collective decision of a team of *base classifiers* [37]. The merits of ensemble classifiers from statistical, computational and representational standpoints under certain implicit assumptions are discussed elsewhere [37]. Recent studies have successfully demonstrated the usefulness of ensemble approaches for disease classification and identifying novel transcriptional targets of critical canonical pathways [41–44]. Traditional ensemble classification techniques fall under two broad categories, namely bagging [45] and boosting [46–49]. *Bagging* (*B*ootstrap *agg*regat*ing*), falls under parallel ensemble methods where bootstrapped realizations are generated with replacement from the given empirical training sample. Each bootstrapped realization in some sense represents a unique example for the classifier to learn from, although a proportion (~63%) of the empirical training sample are retained in each of these realizations. Subsequently, the class label is predicted by combining or aggregating the individual predictions across the bootstrapped realizations using voting strategies (e.g. majority voting) [40]. Bagging framework is implicitly incorporated in popular ensemble classification techniques such as random forest (random decision forests) [50–52] that combines the predictions across multiple unstable classifiers such as decision trees. Such an approach has been shown to outperform single decision tree classifiers. However, unlike traditional bagging, the feature sets across each of the decision trees in a random forest are randomly generated [50,51]. Such a choice has been shown to enhance the overall diversity of the ensemble. On the other hand, sequential ensemble methods such as boosting [46–49] have also been proposed as a suitable alternative to bagging. Unlike bagging, boosting combines the predictions across multiple weak learners whose accuracy is better than that of random guess to form a strong learner from weighted training samples. The weights of the training samples are adjusted in an iterative manner with more preference (boost) given to misclassified samples in the classification process. Boosting techniques have been successfully used for classification of disease groups from molecular expression profiles [42–44]. As with any classification technique, prudent choice of hyperparameters is critical for optimal performance of ensemble classifiers. While single and ensemble classifiers discussed above have been successfully used across a number of applications, they do not necessarily provide sufficient insights into potential variations between the samples within the groups in their out-of-the-box or native form. The present study addresses this critical aspect with emphasis on retrieving patient-specific feature sets while discerning the disease groups of interest. The proposed **S**elective **V**oting Ensemble Classification **A**pproach (**SVA**) is essentially an ensemble classification framework and relies on our recent efforts on investigating inherent heterogeneity within disease groups [53]. Patient-specific variations in SVA are captured by the normalized vote-counts and patient-specific ensemble sets as a result of majority voting. SVA does share some similarities to random forest, in the sense it implicitly uses bootstrapping with replacement to generate training samples as in random forest and determines the labels of the samples using the collective decision across base classifiers. However, the base-classifiers in SVA are accompanied by pairs of features implicitly restricting the dimensionality of the feature space unlike that of the random forest. Using pairs of features is not uncommon and have been shown to yield superior performance in discerning disease groups by other independent studies [54]. As noted earlier, SVA is essentially a classification framework and ideally can accommodate any classification technique without constraining the base classifiers to be decision trees as in random forest. The primary contributions of the present study are as follows:

- Present an ensemble classification framework (SVA) that provides insight into potential patient-specific variations by adapting the feature sets selectively across the samples within and between the disease groups in the classification process.
- Generate network abstractions of the patient-specific ensemble sets to identify dominant nodes and edges with high-confidences that may reveal potential targets for therapeutic interventions and variations in signaling patterns in a patient-specific manner.