Hindawi Publishing Corporation
Disease Markers
Volume 35 (2013), Issue 5, Pages 513–523
http://dx.doi.org/10.1155/2013/613529

Research Article

A Semiautomated Framework for Integrating Expert Knowledge into Disease Marker Identification

Jing Wang,[1] Bobbie-Jo M. Webb-Robertson,[1] Melissa M. Matzke,[1] Susan M. Varnum,[2] Joseph N. Brown,[2] Roderick M. Riensche,[3] Joshua N. Adkins,[2] Jon M. Jacobs,[2] John R. Hoidal,[4] Mary Beth Scholand,[4] Joel G. Pounds,[2] Michael R. Blackburn,[5] Karin D. Rodland,[2] and Jason E. McDermott[1]

[1] *Computational Biology and Bioinformatics, Pacific Northwest National Laboratory, Richland, WA 99352, USA*
[2] *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA 99352, USA*
[3] *Knowledge Discovery and Informatics, Pacific Northwest National Laboratory, Richland, WA 99352, USA*
[4] *Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, UT 84132, USA*
[5] *Department of Biochemistry and Molecular Biology, University of Texas Medical School, Houston, TX 77030, USA*

Correspondence should be addressed to Jason E. McDermott; jason.mcdermott@pnnl.gov

Received 19 March 2013; Accepted 13 August 2013

Academic Editor: Sheng Pan

Background. The availability of large complex data sets generated by high throughput technologies has enabled the recent proliferation of disease biomarker studies. However, a recurring problem in deriving biological information from large data sets is how to best incorporate expert knowledge into the biomarker selection process. *Objective*. To develop a generalizable framework that can incorporate expert knowledge into data-driven processes in a semiautomated way while providing a metric for optimization in a biomarker selection scheme. *Methods*. The framework was implemented as a pipeline consisting of five components for the identification of signatures from integrated clustering (ISIC). Expert knowledge was integrated into the biomarker identification process using the combination of two distinct approaches; a distance-based clustering approach and an expert knowledge-driven functional selection. *Results*. The utility of the developed framework ISIC was demonstrated on proteomics data from a study of chronic obstructive pulmonary disease (COPD). Biomarker candidates were identified in a mouse model using ISIC and validated in a study of a human cohort. *Conclusions*. Expert knowledge can be introduced into a biomarker discovery process in different ways to enhance the robustness of selected marker candidates. Developing strategies for extracting orthogonal and robust features from large data sets increases the chances of success in biomarker identification.

1. Introduction

An unprecedented opportunity for identification of disease biomarker candidates has been provided by the advent of high throughput technologies in the past decade [1, 2]. The explosive growth of large data sets has been overwhelming in terms of the number, size, format, and complexity [3, 4]. While diversified data sets have led to numerous opportunities and studies for discovering new disease marker candidates, the success of those efforts has been largely disappointing in terms of validating the results across populations [4].

Current strategies for biomarker discovery tend to focus on one of two approaches: data-driven [5] or expert knowledge-driven [6]. A data-driven approach makes use of large data sets to unearth the underlying structures embedded in the data to facilitate identification of robust features. The value of this purely statistical approach has been evident in the successful identification of cancer biomarkers, for instance, using an artificial neural network (ANN) model for detecting early stage epithelial ovarian cancer with a panel of five serum markers [7]. In contrast, an expert knowledge-driven approach takes advantage of constantly increased

understanding in pathophysiological mechanisms of diseases at both molecular and systems levels to extract discriminating features of diseases [6]. Despite many intense efforts devoted to the field of biomarker discovery, no robust yet generalizable framework has been widely accepted by the community. Recent studies have suggested that integrating data-driven and knowledge-driven approaches rather than exclusive reliance on either can potentially improve the robustness of selected biomarker candidates and their performances across populations. For instance, an empirical Bayes method has been used to combine the information on pathways and networks into the experimental results of cancer biology [8]. The idea of integrating experimental measurements and existing knowledge is rational and appealing. However, many challenges can be recognized immediately and need to be addressed properly, such as optimal knowledge databases to use, suitable formats of expert knowledge, reasonable ways to integrate these disparate kinds of data, and appropriate selection strategies.

As one of the leading causes of death worldwide, chronic obstructive pulmonary disease (COPD) is a prevalent condition that is characterized by progressive and not fully reversible airflow limitations [9, 10]. No new classes of drugs for COPD treatment have been approved for use in the United States in more than twenty years [11]. Despite associations with multiple pathological components, one hallmark of COPD is a persistent inflammatory state that contributes to a progressive decline in lung function [12]. A mouse model with adenosine deaminase (Ada) deficiency has been established to develop a rapid pulmonary inflammation and progressive destruction of lung tissue that closely mimics many aspects of human COPD and other chronic lung diseases [13, 14]. Adenosine is a molecule routinely generated at sites of inflammation and tissue injury. It is a key signaling molecule involved in multiple intracellular signaling pathways related to the modulation of inflammatory responses [15, 16]. Ada is the purine catabolic enzyme responsible for converting adenosine to inosine, which is frequently induced in response to cell stress or damage and involved in anti-inflammatory, tissue-protective pathways [16, 17].

Accumulating evidence has suggested that elevated adenosine levels in lung are associated with chronic lung diseases in both human and animal models [12, 17]. An in-depth understanding of the biological relevance of Ada in lung will benefit our general understanding of COPD.

In this paper, we describe a semiautomated framework, identification of signatures from integrated clustering (ISIC), for merging data-driven and knowledge-driven approaches into a biomarker selection scheme in an iterative manner, with a defined metric provided for performance evaluation. To demonstrate ISIC, we applied it to proteomics data sets of bronchoalveolar lavage fluid (BALF) and plasma from a mouse model of COPD, the Ada-deficient mouse model [13, 14], to identify marker candidates of COPD. The resulting candidates were subsequently validated in a human plasma data set from a cohort of low body mass index (BMI) smokers with COPD and healthy controls. We believe that ISIC is a novel and powerful tool for integrating data types in the context of biosignature discovery and show that it produces robust results between a model system and human disease.

2. Methods

2.1. Animal Samples, Patient Samples, and Proteomics Data Collections. Data from the Ada-deficient mice were used for the initial biomarker identification [13]. Bronchial secretions and blood plasma from the Ada −/− and Ada +/− mice were individually collected and processed as described previously [15]. Human plasma samples were selected from a large cohort ($n = 467$) of the Genetics of Addiction program at the University of Utah Medical School [18]. Plasma samples from 7 low BMI smokers with COPD and 7 low BMI never smokers (COPD free) were used for the patient and control samples. All BALF and plasma (mouse and human) samples were processed, tryptic digested, separated, and analyzed using liquid chromatography-mass spectrometry (LC-MS). Detailed information on animal and human sample collections and data collection is provided in Supplemental Information, which includes Supplemental Methods, brief descriptions and rationales of the framework, discussion on COPD data sets used in the current study, Supplemental Figures, Supplemental Tables, and References. Supplemental Information available online at http://dx.doi.org/10.1155/2013/613529.

2.2. Processing of Proteomic Data Sets. The peptides were identified and quantified using a collection of in-house developed tools that are freely available at http://omics.pnl.gov/. For the mouse data, the peak intensity values of the final identified peptides were obtained from the analyses of LC-LTQ-Orbitrap spectral data. The raw peak intensity values were processed in the MatLab environment, including quality control, normalization, protein quantification, and comparative statistical analyses [19–21]. The final peptide abundances were transformed into the $\log_{10}$ scale for the subsequent data analyses. Quality control was a process performed to identify and remove the peptides with an insufficient amount of data across the set of samples [20], as well as to identify and remove the LC-MS data sets that showed significant deviations from the standard behaviors of all LC-MS analyses [22]. The outlier LC-MS data sets were identified at a significance level of 0.0001. The peptides were normalized across all technical replicates to ensure the least amount of bias introduced into the data sets [21]. Specifically, the BALF data were normalized Using a linear combination of order statistics to determine a subset of peptides [23] followed by mean centering, and the plasma data were normalized using a rank invariant peptide subset [21] followed by median centering. The normalized $\log_{10}$ abundance values were averaged across the technical replicates within each biological sample. The subsequent protein quantifications were performed using the most abundant reference peptide through an R-Rollup method [19, 24]. The human plasma data were processed using the same protocols for the BALF data, as described above.

2.3. Significantly Altered Proteins in Mouse BALF and Plasma. The quantified proteins in BALF and plasma were compared

quantitatively and qualitatively between the time-matched Ada +/− and Ada −/− mice (three in each group) at each of the five time points, respectively. Quantitative comparison was performed using a Dunnett adjusted t-test to assess the numeric change in the average abundance of a protein between the two phenotypes at the individual time points. Qualitative comparison (the presence or absence of a protein) was implemented by a G-test, a modified χ^2 test of independence, which assessed the associations between the presence/absence of proteins and the phenotypes of the mice [20]. A significantly altered protein was defined if at least one of the five t-tests or five G-tests was statistically significant ($P < 0.05$) after Bonferroni multiple hypothesis correction. All the protein abundances were compared with their time-matched counterparts. To provide sufficient replicates for the subsequent analyses, we combined the significantly altered proteins from all five time points and grouped them into either disease (Ada −/−) or control (Ada +/−) categories for the BALF and plasma samples, that is, a sample size of 15 mice in each of the groups.

2.4. Distance-Based Hierarchical Clustering. All clustering analyses using the mouse data were performed on the complete data set of significantly altered proteins. Missing values were imputed at the protein abundance level using a regularized expectation-maximization algorithm [25]. The imputation toolbox is freely available at http://www.clidyn.ethz.ch/imputation/index.html. Three different distance matrices were calculated. The first was based on the protein expression profiles and was calculated as Euclidean distance of the protein abundances in the $\log_{10}$ scale. The second was based on the functional relationships between the proteins and was determined by the semantic dissimilarities in the biological process subontology from the Gene Ontology (GO) [26]. The semantic dissimilarity was defined as 1-semantic similarity between protein pairs and calculated using the cross-ontological analysis (XOA) tool [27] or the GOSemSim package [28] in the R statistical language. The third one was the joint distance measure of the previous two. It was calculated as the weighted average of the other two distances with weighting factors of 0.25, 0.50, 0.75, or 1 or the average of their individual logistic functions [29]. A logistic function is a common sigmoidal function with equation:

$$f(x) = \frac{1}{1 + e^{-x}}, \tag{1}$$

where x is associated with a distance matrix, that is, the first or second distance. Specifically, x is expressed as

$$x = \frac{6}{\nu}\left(\delta_{a_i,a_j} - \nu\right), \tag{2}$$

where δ_{a_i,a_j} is a distance or dissimilarity between two proteins, ν is a smooth threshold (chosen as the mean of the distance matrix), and $6/\nu$ is a heuristically chosen parameter of the slope in our calculation [30]. The averaged logistic function of the two distance matrices was used as a candidate of the joint distance measures. The numbers of clusters were empirically determined as 6 and 12 for the BALF and 6 for the plasma data sets based on their sample sizes. Ward's minimum variance linkage was used in all hierarchical clustering [31].

2.5. Expert Knowledge-Driven Disease Model-Related Functional Analysis. A biological function-centric approach was used to determine the functionally enriched biological processes in the BALF and plasma samples of mice [32]. The biological processes are referred to the terms included in the biological process hierarchy in the GO. The significantly changed proteins in the Ada −/− mice (relative to their controls) were mapped to their corresponding genes and compared with a list of all genes in the mouse genome in order to determine the levels of significance for individual biological processes in the data using a hypergeometric test. The GO terms with the enrichment P values smaller than 10^{-8} (in both BALF and plasma) were considered as significantly enriched biological processes in our data sets [32]. For each enriched GO term, we determined its own level (how specific the term is) and its top-level ancestor (broadest category that includes the term) within the biological process subontology. The level of a GO term is the number of steps taken to reach the top-most node when ascending the GO tree starting from the term of interest. The top-level ancestor of a GO term is the ancestor term that is directly below the top-most node ("biological process," GO ID: 8150). The enriched GO terms were then grouped according to their top-level ancestors and resulted in a number of biological process groups that were enriched in the Ada −/− mice.

An expert knowledge-driven disease selection was subsequently implemented on the enriched GO terms selected above. Specifically, a subset of the enriched terms was further selected based on expert knowledge on the Ada-deficient model, the P values of enrichments, and the levels of the GO terms within individual functional groups, that is, the GO terms sharing the same top-level ancestor(s). Those functional clusters with their corresponding proteins were the final results of this expert knowledge-driven disease-model-related annotation selection. Each of the clusters was represented by either all differentially expressed proteins or, alternatively, the top three most differentially altered proteins in the cluster between the two phenotypes for the subsequent analyses in ISIC.

2.6. Bayesian Integration and Classification. A Bayesian integration approach was applied on clusters to derive the optimal probability models for the data sets [33]. Four standard statistical algorithms were applied to the individual clusters (or subsets) generated from the hierarchical clustering or expert knowledge-driven functional annotation to build likelihood probability models: linear discriminant analysis [34], fuzzy k-nearest neighbor [35], multinomial logistic regression [36], and Naïve Bayes [37]. Classification accuracy (CA) was used to evaluate the performances of the individual subsets and the integrations of the multiple subsets of each round of the analyses. CA is a measure of how well predictions match with the actual data. Our approach predicts the disease state of each animal (Ada −/− or Ada +/−), and so true positives (TP; correct predictions of disease) and negatives (TN)

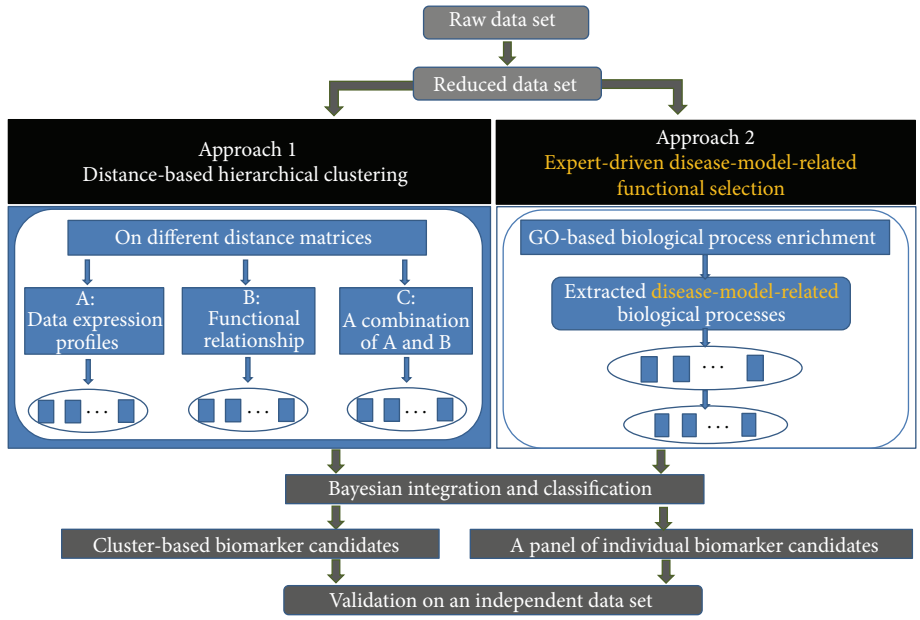

Figure 1: Flowchart of the ISIC framework in a biosignature discovery process.

are compared to incorrect predictions (false positives and negatives; FP and FN). CA is expressed as

$$\mathrm{CA} = \frac{\text{number of correctly classified samples}}{\text{total number of samples}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}. \quad (3)$$

The optimal algorithm for a specific cluster was the one providing the best CA among the four probability models. The posterior probabilities were integrated through a Bayesian approach using the different combinations of the subsets along with their optimal algorithms determined [33, 38]. Within either of the approaches, the CAs from all possible combinations of clusters were calculated, and the highest value was reported as the optimal integrated CA for the data set under that setting. Five-fold cross-validations were performed in all analyses to assure that the probabilities were independent from the training data. The cluster membership and the corresponding proteins of each cluster from the optimal integration were recorded for the comparisons. Once the cluster membership of the optimal integration for a data set was determined, the integration and determination of CA from the selected clusters were repeated 100 times with five-fold cross-validations. The average CA was reported as the final optimal integrated CA of each round of the analyses. Additionally, the integrated CAs were reported from the use of a full data set and a partial data set. The full data set refers to the entire data set that was divided into the numbers of clusters indicated. The partial data set refers to the subset of the clusters that provided the best integrated CA for each combination of parameters (overall number of clusters, distance matrix used, and weight for expert versus data matrix integration.).

2.7. Biomarker Candidate Selection. Biomarker candidate selection was conducted separately in the clustering approach versus the expert-driven functional selection. In the clustering approach, we selected the biomarker candidates as the set of protein clusters that gave the best integrated CAs. In the expert-driven functional selection, the biomarker candidates were selected as the several most differentially altered proteins belonging to the functional clusters providing the best integrated CAs.

2.8. Validation. Validation was performed on a proteomics data set of human plasma at both the cluster (six cluster optimization) and individual protein levels. For the validation on clusters, we used the clusters identified from the mouse BALF and plasma proteins using their joint distance matrices. In each cluster, we filtered to only include proteins that were detected in both human and mouse. The CAs for individual clusters and the integrations from all six clusters were calculated. For the validation on the individual proteins, the biomarker candidates selected from mouse BALF, which were also detected in the plasma data sets, were evaluated in the human plasma data set.

3. Results

3.1. Overview of Approach. The objective of this study was to develop a semiautomated framework for integrating expert knowledge into disease marker selection scheme in an iterative manner guided by the use of a defined metric providing the evaluation of performances. The framework, named ISIC, was designed to serve as a conceptual pipeline rather than a collection of detailed protocols; the basic flow is illustrated in Figure 1. The overall process consists of five components, that is, data reduction, distance-based hierarchical clustering, Bayesian integration and classification, selection of

Table 1: Optimal integrated CAs derived from (A) the distance-based hierarchical clustering and (B) the disease-model-related functional selection approaches.

			Optimal Integrated CA						
Clustering based on	Distance matrix		Data expression profiles			Functional relationships		A combinationof the two	
	No. of clusters		1	6	12	6	12	6	12
(A) Different distance matrices	BALF	Full	0.83	0.86	0.79	0.93	0.80	0.81	0.81
		Partial		0.93	0.96	0.96	**1.00**	**1.00**	**0.99**
	Plasma	Full	0.66	0.68		0.54		0.62	
		Partial		0.77		0.79		**0.83**	
	Number of proteins[1]			All		Top 3			
	No. of clusters		1	12		1	12		
(B) Disease model-related functional selection	BALF	Full	0.81	0.90		0.88	0.82		
		Partial		0.93			**0.99**		
	Plasma	Full	0.57	0.56		0.59	0.65		
		Partial		0.73			**0.87**		

[1]This refers to the different number of significantly changed proteins (all proteins or top 3 proteins) used in each cluster.

biomarker candidates, and validation. In addition, an expert knowledge-driven disease model-related functional selection was provided as a parallel approach, the results of which can be compared to those from ISIC using CA as a discriminator between the two approaches. Biomarker candidates were selected based on their CA performances in the individual approaches, which were subsequently validated in a human data set (see Supplemental Information for additional explanation).

3.2. Application on COPD Data Sets. To demonstrate ISIC, we applied it to three proteomics data sets associated with COPD. First, we identified the potential biomarkers in data obtained in the BALF and plasma samples from the Ada-deficient mouse model of COPD. This model system has a clear distinction between diseased and nondiseased samples and therefore is well-suited to developing and testing our classification approach. To validate the candidates identified from the development phase, we chose to examine plasma data from smokers with COPD along with their corresponding controls. This data set, derived from the actual patient samples, allowed us to test whether the signatures identified from mouse would be robust in their ability to classify diseased and nondiseased human samples with complex and varied genetic and environmental backgrounds.

3.2.1. Initial Searching Spaces of BALF and Plasma from the Ada-Deficient Mice. A special effort was focused on how to appropriately handle the missing values in the data sets. Missing values are an inevitable issue in many proteomics studies. It is not uncommon to have 30% or more missing values, that is, measurements for a specific protein that are missing from individual samples but present in others, even from a carefully designed and implemented proteomics data set [39]. In the mouse demonstration data sets, the missing rates were 26% in BALF and 17% in plasma. Dunnett adjusted t-tests were performed on the proteins having adequate abundance values in both types of mice, and G-tests, a modified χ^2 test of independence [20], were implemented on the proteins without adequate abundance values, respectively. The former assesses the quantitative changes, and the latter evaluates the qualitative changes in the individual protein abundances. A quantitative change is self-explanatory, while a qualitative change here refers to a real biological absence or presence of a protein between the two groups of mice. Numbers of changed proteins as well as the direction of change are indicated in Figure S1. The collections of the proteins that are quantitatively (a numerical change in the abundance) and qualitatively (absence/presence) different at individual time points were considered as our initial biomarker searching spaces in BALF (396 out of 532) and plasma (150 out of 351). Heat maps of protein expression from both data sets are depicted in Figure S2, and no distinct patterns are observed in either.

3.2.2. Distance-Based Hierarchical Clustering and Classification Performance. We calculated distances between all proteins based on their abundance levels across all observations or their functional similarity based on their annotations in GO. These sets of distances were then integrated and used for hierarchical clustering. The clusters derived from three different dissimilarity matrices were used in the Bayesian integration and classification step to obtain CA scores for each combination of parameters. No significant differences in the CA scores were observed between different weighted averages or logistic functions (data not shown). The integrated CA scores from using the full and partial data sets are listed in the Table 1(A), and the information from individual clusters is provided in Table S1. The optimal CA scores from the use of the entire data set as a single cluster provided the

Table 2: The list of the enriched general functional groups from the Ada-deficient model of COPD extracted by the expert knowledge-driven functional analysis using the BALF data. This list is based on (A) a GO-based biological process enrichment and (B) the disease-model-related expert selection.

No.	Enriched general functional group	(A) GO-based biological process	(B) The disease model-related GO cluster
1	Immune system process	(1) Immune system process	(13-1)[1] Immune system process
2	Stress/stimulus response	(2) Response to stimulus	(1) Response to stimulus; (2) Response to stress
3	Cellular response to stimulus	(3) Cellular process	(3) Cellular response to stimulus
4	Metabolic process	(4) Metabolic process	(4) Small molecule metabolic process; (5) Oxoacid metabolic process; (6) Oxidoreduction coenzyme metabolic process; (7) Nucleotide metabolic process; (8) Carbohydrate derivative metabolic process
5	Biological regulation	(5) Biological regulation	(9) Regulation of immune system process; (10) Regulation of localization; (11) Regulation of programmed cell death
6	Death	(6) Death	(12) Death
7	Localization	(7) Localization; (8) Establishment of localization	(13-2)[1] Localization; (13-3)[1] Establishment of localization
8	Cellular organization	(9) Cellular component organization or biogenesis	(13-4)[1] Cellular component organization or biogenesis
9	Proliferation	(10) Cell proliferation; (11) Growth; (12) Developmental process	
10	Others	(13) Reproduction and reproductive process; (14) Biological adhesion; (15) Locomotion; (16) Multicellular organismal process; (17) Multiorganism process	(13-1)[1] Immune system process; (13-2)[1] Localization; (13-3)[1] Establishment of localization; (13-4)[1] Cellular component organization or biogenesis

[1]This term belongs to the 13th cluster (others) from the approach B.

baseline performances of the approach. In BALF, using the 396 proteins, the classification performances were approximately 80% CA or higher, and their CA counterparts from plasma (using the 150 proteins) were lower at about 60%. Interestingly, the optimal integrated CA scores in BALF and plasma were both derived from using a subset of the clusters rather than a full data set.

3.2.3. Expert Knowledge-Driven Disease Model-Related Functional Selection and Classification Performance. A total of 303 GO terms (data not shown) were determined as enriched ($P < 10^{-8}$) in the biological process hierarchy from the mouse BALF samples using a hypergeometric test. Intermediate level GO terms were selected based on knowledge of the disease model and then grouped into 13 groups of GO annotations (the rightmost column in Table 2). These groups were summarized into their top-level GO-based biological processes (center column, Table 2) into ten general enriched functional groups for the Ada-deficient mouse model of COPD, which was the final result of the expert knowledge-driven analysis.

The CA performances of the ten functional groups were assessed in a similar way as those in the clustering approach. Specifically, the optimal individual CA scores for functional clusters from using all and the top three differentially expressed proteins within individual clusters were calculated and are summarized in Table S2. The CA results using the entire data set as a single cluster are also provided as the reference points. The integrated CA results are listed in Table 1(B). To our surprise, for both individual and integrated results, the CAs calculated using the top three proteins outperformed the CAs using all proteins in the majority of the cases. In addition, the best integrated CA scores were derived from the partial instead of the full data sets, similar to what we observed in the clustering approach. This similar pattern implies that collecting more data from the same sample source may not guarantee gaining better performances.

3.2.4. Selection of Biomarker Candidates from Different Approaches. In the distance-based clustering approach, the biomarker candidates were the protein clusters. Specifically, 215 (out of 396) proteins in two clusters (with the total number of clusters set as six) or 129 proteins in two clusters (with the total number of clusters set as twelve) in BALF from the best performing combination of clusters were considered to be the biosignatures. Similarly, a group of 13 proteins from the best performing cluster were selected as a narrowed set of biomarker candidates in plasma. Because our approach combines patterns in abundance with functional relationships, we hypothesized that these signatures would

Table 3: The validation results (in CA) on the cluster-based biomarker candidates using a human plasma data set.

Functional group no. in Table 2	CA from mouse plasma-defined clusters in		CA from mouse BALF-defined clusters in	
	Mouse plasma	Human plasma	Mouse BALF	Human plasma
1	0.54	0.93	0.79	0.93
2	0.58	0.86	0.93	0.79
3	0.56	0.71	0.72	0.71
4[1]	0.83	0.71	0.79	0.79
5	0.63	0.79	0.83	0.64
6	0.53	0.79	0.62	0.64
1, 4	0.80	0.86		
2, 5[1]			1.00	0.71
1–6	0.83	**0.93**	0.81	**0.86**

[1]The best performing clusters from the indicated mouse data set.

be more robust relative to the individual candidates with top performances.

In the disease model-related functional selection driven by expert knowledge, CAs from the top three most differential proteins of each cluster generally outperformed the CAs from all differential proteins of the cluster. Therefore, we examined the biomarker candidates in the top three proteins from each GO term instead of all proteins under it. The information on the five best CA performances is listed in Table S3, including the CA scores, the cluster names, and the protein lists. A single functional cluster, carbohydrate derivative metabolic process, yielded the best CA score of 0.99, which includes complement C3 (CO3/C3, Uniprot/gene symbol), prothrombin (THRB/F2), vitamin D-binding protein (VTDB/GC), and v-type proton ATPase 16 kDa proteolipid subunit (VATL/Atp6v0c). These proteins were also present in some of the top-performing clusters from our cluster-based approach. Note that most proteins have multiple annotations and the six proteins that consistently recurred in the top-performing functional clusters were considered as a list of biomarker candidates of COPD (using the BALF data set) from the expert knowledge-driven approach. The panel of six biomarker candidates includes THRB, VTDB, CO3, VATL, adiponectin (ADIPO/ADIPOQ), and liver fatty acid-binding protein (FABPL/FABP1).

3.2.5. Validation. To compare the robustness of signatures derived using different approaches and to validate our findings, we chose to use a data set of human plasma samples. These samples were taken from subjects with low BMI (<25) who smoke and have been diagnosed with COPD and their corresponding healthy controls. A total of 44 proteins in human data were differentially expressed in the mouse plasma, which was used in validation. The optimal CAs from using the 44 common proteins in the six clusters that were defined by the mouse plasma are listed in Table 3. The integrated CA from using all six clusters was 0.93 in the human and 0.83 in the mouse plasma. Though the best-performing cluster (the fourth cluster) in mouse did not provide a better performance in human, the top integrated CA (using the first and fourth clusters) gave comparable classification result in human plasma. Validation using the six clusters defined by the mouse BALF showed similar outcomes (Table 3). We also calculated the individual CAs of the 44 common plasma proteins in mouse and human data sets, respectively. The results show that the top-performing individual proteins in mouse do not provide consistent classification performances using human data and only marginally discriminate the patients from their controls (Table S4). This is in contrast to our findings in the cluster-based biomarker candidates in the mouse plasma, which could also classify human patients quite well. The receiver operating characteristic (ROC) curves and the areas under the curves (AUC) were also performed for validation, which obtained comparable results relative to the CAs (Figure S3). The ROC curves provide the estimates for sensitivity and specificity, which is a commonly used evaluation metric in clinical studies, of the tested biomarker candidates.

At the level of individual proteins, four out of the six candidates selected by the expert-driven functional selection were also identified in human plasma samples. The CA scores from individual marker candidates and several top integrations of both mice and human samples are summarized in Table 4. This panel of BALF-based marker candidates, that is, THRB, VTDB, CO3, and ADIPO, consistently showed better performances in human plasma relative to those in mouse plasma. Reasonable results were observed in all three specimens: the CA score of 0.93 in the mouse BALF, 0.70 in the mouse plasma, and 0.93 in the human plasma. Detailed information on the four candidates is illustrated in Figure 2, including the average fold changes and their regulation directions in the Ada −/− group relative to the controls. The significance levels of the protein abundance changes are also provided, which were determined by the P values from the corresponding t-tests or G-tests.

3.2.6. Comparison of Classification Performances at the Different Time Points. With the data sets available for the time-matched diseased and controlled animals, we were able to compare the individual and integrated optimal CAs at the different time points during the developmental course of COPD from the Ada-deficient mouse model. In particular, we compared the optimal CAs derived from the proteins that were individually and cumulatively significantly changed at

Table 4: The validation results in CA on the individual biomarker candidates of COPD in the human data and the CAs from the mouse data.

Individual protein or a panel of proteins	Optimal CA in			Belong to the general functional group[1]
	Mouse BALF	Mouse plasma	Human plasma	
Prothrombin, THRB	0.86	0.50	0.93	1–10
Vitamin D-binding protein, VTDB	0.69	0.63	0.86	2–10
Complement C3, CO3	0.69	0.67	0.79	1–10
Adiponectin, ADIPO	0.66	0.53	0.64	1–9
THRB; VTDB	0.97	0.57	0.93	
THRB; CO3	0.86	0.70	0.93	
VTDB; CO3	0.83	0.67	0.79	
THRB; CO3; ADIPO	0.86	0.70	1.00	
VTDB; CO3; ADIPO	0.83	0.70	0.93	
THRB; VTDB; CO3; ADIPO	**0.93**	**0.70**	**0.93**	

[1]The enriched functional clusters refer to the general functional groups listed in Table 2: 1: immune system process; 2: stress/stimulus response; 3: cellular response to stimulus; 4: metabolic process; 5: biological regulation; 6: death; 7: localization; 8: cellular organization; 9: proliferation; 10: others.

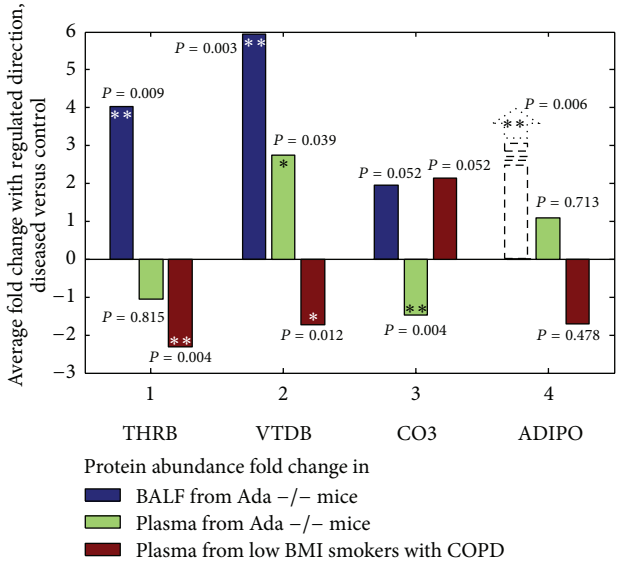

Figure 2: The bar graph of the average fold changes of the protein abundances in diseased group relative to their controls of four potential biomarkers identified in mouse BALF. The positive fold changes indicate the observed upregulation in the diseased group, the Ada −/− mice, and the negative fold changes indicate the observed downregulation. The significances of these changes are indicated with two (P value is between 0.01 and 0.001), one (P value is between 0.01 and 0.05), or no asterisk (P value is greater than 0.05). The arrow in dashed line of ADIPO shows that this protein was present in the BALF of the Ada −/− mice but absent in the Ada +/− group. The mouse data from the last two time points (on days 38 and 42) were used for this analysis.

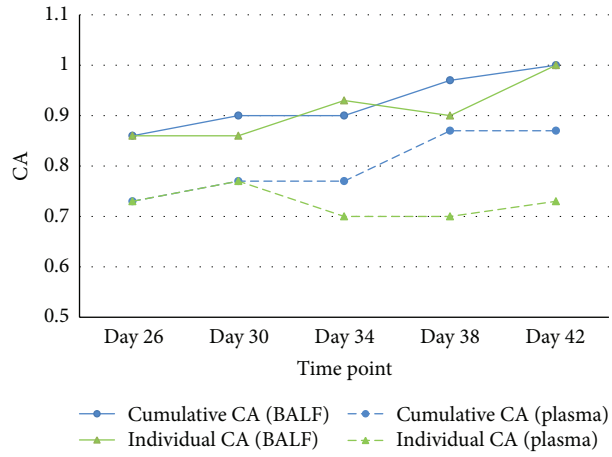

Figure 3: The comparative results in the optimal CAs between the BALF (solid lines) and plasma (dashed lines) from the Ada-deficient mice during the disease developmental course. The CAs were derived from the cumulatively (blue) and individually (green) significantly changed proteins at different time points. The CAs were obtained from the resulting clusters using the joint distance matrix of protein expression patterns and their functional relationships (XOA).

the five time points in both fluids (Figure 3). The individually changed proteins at a specific time point refer to the proteins that showed significant alterations in their abundances at this single time point, while the cumulatively changed proteins at the same time point refer to a collection of individually changed proteins from day 26 up to this time point. Both the individual and cumulative CAs from BALF (solid lines) consistently outperformed their counterparts from plasma (dashed lines) in terms of discriminating between diseased and nondiseased animals.

Not surprisingly, the ability of this discrimination shows an increasing trend as the Ada −/− mice get sicker in both plasma and BALF. It is also interesting that these four candidates are able to classify mice fairly well at very early time points, even before outward manifestations of disease.

4. Discussion

In the field of biomedical science, the primary challenge has been shifting from data generation to data interpretation. The explosive growth of high dimensional data sets has demanded the development of semiautomated or automated tools as a necessity for knowledge discovery [4, 40]. In this study,

we introduced the ISIC as a framework that is designed for integrating experimental measurements and expert knowledge into disease marker identification in a semiautomated manner. One main merit of the ISIC lies in the manner of integration, which simultaneously combines data- and knowledge-driven information into a quantitative format. A pipeline was assembled accordingly and demonstrated on several proteomics data sets of COPD to identify biomarker candidates of COPD.

The focus of the semiautomated clustering approach is to separate the initial marker searching candidates, that is, the differentially expressed proteins in the data sets, into several different groups that contain features with similar expression patterns and functionalities within groups. In contrast, the expert-driven functional selection may be somewhat subjective; however, it can be an efficient way to extract a handful of biomarker candidates with the incorporation of proper knowledge. The classification performances of this demonstration of COPD data sets on both approaches obtained comparable results that were both quite good. It is also worth mentioning that our intention here is to illustrate the individual merits and weaknesses of both approaches in the biomarker selection schemes in order to gain insight on how to comprehensively and efficiently extract valuable information from data sets.

Biomarker identification is a process to select a limited number of biomolecules that convey the essential biological information distinguishing a disease state from a nondisease state. In the clustering approach, our results show that our cluster-based biomarkers are much more robust in their ability to classify human patients than the individual proteins. A possible explanation is that features in clusters may capture more consistent and comprehensive information from data relative to the individual proteins. We are currently working to include an extra step of feature selection, which will rigorously identify subsets of proteins with optimal CAs, to focus smaller biomarker sets.

We found that small sets of proteins could be selected with good performance using our expert knowledge-driven approach. The biomarker candidates selected in this way have a subjective component but also can potentially filter out the low quality markers identified from pure statistical processes. Another limitation of expert-driven strategy is that not all gene or gene products have annotations, which eliminates the possibilities for exploring the functional relationships among them in the currently available knowledge databases.

One noticeable consistency of the two approaches is that all the optimal classification performances, indicated by the optimal integrated CAs, resulted from using partial instead of the full data sets (Table 2). This trend was particularly striking when using the top-three most differentially changed proteins to represent the individual GO terms in clusters defined by the expert-knowledge driven functional selection (Table S2). This indicates that simply using more data collected from the same data source does not guarantee improved performance and that redundant information most likely is included in the additional data.

As previously emphasized, ISIC is intended to be conceptual as well as flexible. The five components function independently and collaboratively. Each component serves its distinct functionality and is implemented at a different stage in the biomarker discovery process. The independence between them makes it easy to tailor individual compartments for specific data sets or to substitute using other methods with similar functions. The data reduction largely is a data-dependent process. As a means to group similar data into clusters based on a similarity criterion, the distance-based hierarchical clustering can also be achieved by other grouping mechanisms, such as k-means, self-organized maps, and fuzzy clustering [41]. In the model integration portion, Bayesian integration can be replaced with a support vector machine [42, 43], decision trees and random forests [44, 45], and artificial neural networks [2], which have been applied in many types of data integration [4]. In terms of performance evaluation, CA was chosen mainly due to its suitability in cases with more than two categories. An ROC curve and the measurement of AUC [46] as well as recall and precision [47] are both reasonable substitutions for the CA but mostly limited to cases with binary responses.

In conclusion, we describe a generalizable framework for integrating expert knowledge into processes of disease biomarker discovery. Our framework, ISIC, consists of several independent and collaborative components and is flexible enough to accommodate addition, subtraction, and modification of analyses. The integration of data-driven and knowledge-driven information is used in a distance-based clustering approach in a semiautomated manner. An expert-driven functional selection approach was also performed to select individual proteins for comparison to our automated approach. We identified signatures in a mouse model of COPD and subsequently validated them in a human cohort, where they demonstrated a comparable accuracy in discriminating patients with COPD from those without COPD. This was in contrast to standard approaches to identify biomarkers in the mouse model, which were not robust in the human cohort. We believe that ISIC represents a generalizable platform for identification of robust biosignatures from integrated data sources.

Acknowledgments

The authors would like to thank Dr. Harish Shankaran for his help in the functional enrichment analysis. This study was supported by the Signatures Discovery Initiative, a component of the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory, a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy under Contract DE-AC05-76RL01830. The proteomics measurements were conducted in the Proteomics Facility at the Environmental Molecular Sciences Laboratory, a DOE BER national scientific user facility at Pacific Northwest National Laboratory (PNNL) in Richland, WA. Support was also received from the Pulmonary Systems Biology Initiative of the Battelle Memorial Institute. The work on the patient samples was supported by funds from National Institutes of Health (NIEHS) (Grant U54-ES016015).

References

[1] D. Ghosh and L. M. Poisson, "'Omics' data and levels of evidence for biomarker discovery," *Genomics*, vol. 93, no. 1, pp. 13–16, 2009.

[2] B. P. Bradley, "Finding biomarkers is getting easier," *Ecotoxicology*, vol. 21, no. 3, pp. 631–636.

[3] Z. Feng, R. Prentice, and S. Srivastava, "Research issues and strategies for genomic and proteomic biomarker discovery and validation: a statistical perspective," *Pharmacogenomics*, vol. 5, no. 6, pp. 709–719, 2004.

[4] J. E. McDermott, J. Wang, H. D. Mitchell et al., "Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data," *Expert Opinion on Medical Diagnostics*, vol. 7, no. 1, pp. 37–51, 2013.

[5] T. Wei, B. Liao, L. Ackermann et al., "Data-driven analysis approach for biomarker discovery using molecular-profiling technologies," *Biomarkers*, vol. 10, no. 2-3, pp. 153–172, 2005.

[6] L. Chen, C. Wang, I.-M. Shih et al., "Biomarker identification by knowledge-driven multi-level ICA and motif analysis," in *Proceedings of the 6th International Conference on Machine Learning and Applications (ICMLA '07)*, pp. 560–566, December 2007.

[7] Z. Zhang, Y. Yu, F. Xu et al., "Combining multiple serum tumor markers improves detection of stage I epithelial ovarian cancer," *Gynecologic Oncology*, vol. 107, no. 3, pp. 526–531, 2007.

[8] S. M. Hill, R. M. Neve, N. Bayani et al., "Integrating biological knowledge into variable selection: an empirical Bayes approach with an application in cancer biology," *BMC Bioinformatics*, vol. 13, article 94, pp. 94–109, 2012.

[9] D. L. Hoyert, K. D. Kochanek, and S. L. Murphy, "Deaths: final data for 1997," *National Vital Statistics Reports*, vol. 47, no. 19, pp. 1–104, 1999.

[10] A. D. Lopez and C. C. Murray, "The global burden of disease, 1990–2020," *Nature Medicine*, vol. 4, no. 11, pp. 1241–1243, 1998.

[11] S. R. Rosenberg and R. Kalhan, "Biomarkers in chronic obstructive pulmonary disease," *Translational Research*, vol. 159, no. 4, pp. 228–237, 2012.

[12] Y. Zhou, D. J. Schneider, and M. R. Blackburn, "Adenosine signaling and the regulation of chronic lung disease," *Pharmacology and Therapeutics*, vol. 123, no. 1, pp. 105–116, 2009.

[13] M. R. Blackburn, S. K. Datta, and R. E. Kellems, "Adenosine deaminase-deficient mice generated using a two-stage genetic engineering strategy exhibit a combined immunodeficiency," *Journal of Biological Chemistry*, vol. 273, no. 9, pp. 5093–5100, 1998.

[14] H. Zhong, J. L. Chunn, J. B. Volmer, J. R. Fozard, and M. R. Blackburn, "Adenosine-mediated mast cell degranulation in adenosine deaminase-deficient mice," *Journal of Pharmacology and Experimental Therapeutics*, vol. 298, no. 2, pp. 433–440, 2001.

[15] M. R. Blackburn, J. B. Volmer, J. L. Thrasher et al., "Metabolic consequences of adenosine deaminase deficiency in mice are associated with defects in alveogenesis, pulmonary inflammation, and airway obstruction," *Journal of Experimental Medicine*, vol. 192, no. 2, pp. 159–170, 2000.

[16] M. R. Blackburn, C. G. Lee, H. W. Young et al., "Adenosine mediates IL-13-induced inflammation and remodeling in the lung and interacts in an IL-13-adenosine amplification pathway," *Journal of Clinical Investigation*, vol. 112, no. 3, pp. 332–344, 2003.

[17] A. V. Sauer, I. Brigida, N. Carriglio et al., "Autoimmune dysregulation and purine metabolism in adenosine deaminase deficiency," *Frontiers in Immunology*, vol. 3, pp. 1–19, 2012.

[18] H. Jin, B.-J. Webb-Robertson, E. S. Peterson et al., "Smoking, COPD, and 3-nitrotyrosine levels of plasma proteins," *Environmental Health Perspectives*, vol. 119, no. 9, pp. 1314–1320, 2011.

[19] M. M. Matzke, J. N. Brown, M. A. Gritsenko et al., "A comparative analysis of computational approaches to relative protein quantification using peptide peak intensities in label-free LC-MS proteomics experiments," *Proteomics*, vol. 13, no. 3-4, pp. 493–503, 2013.

[20] B.-J. M. Webb-Robertson, L. A. McCue, K. M. Waters et al., "Combined statistical analyses of peptide intensities and peptide occurrences improves identification of significant peptides from MS-based proteomics data," *Journal of Proteome Research*, vol. 9, no. 11, pp. 5748–5756, 2010.

[21] B.-J. M. Webb-Robertson, M. M. Matzke, J. M. Jacobs, J. G. Pounds, and K. M. Waters, "A statistical selection strategy for normalization procedures in LC-MS proteomics experiments through dataset-dependent ranking of normalization scaling factors," *Proteomics*, vol. 11, no. 24, pp. 4736–4741, 2011.

[22] M. M. Matzke, K. M. Waters, T. O. Metz et al., "Improved quality control processing of peptide-centric LC-MS proteomics data," *Bioinformatics*, vol. 27, no. 20, Article ID btr479, pp. 2866–2872, 2011.

[23] P. Wang, H. Tang, H. Zhang, J. Whiteaker, A. G. Paulovich, and M. Mcintosh, "Normalization regarding non-random missing values in high-throughput mass spectrometry data," *Pacific Symposium on Biocomputing*, pp. 315–326, 2006.

[24] A. D. Polpitiya, W.-J. Qian, N. Jaitly et al., "DAnTE: A statistical tool for quantitative analysis of -omics data," *Bioinformatics*, vol. 24, no. 13, pp. 1556–1558, 2008.

[25] T. Schneider, "Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values," *Journal of Climate*, vol. 14, no. 5, pp. 853–871, 2001.

[26] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[27] C. Posse, A. Sanfilippo, B. Gopalan et al., "Cross-ontological analytics: combining associative and hierarchical relations in the gene ontologies to assess gene product similarity," in *Computational Science*, Lecture Notes in Computer Science, pp. 871–878, 2006.

[28] G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang, "GOSemSim: an R package for measuring semantic similarity among GO terms and gene products," *Bioinformatics*, vol. 26, no. 7, Article ID btq064, pp. 976–978, 2010.

[29] D. H. von Seggern, *CRC Standard Curves and Surfaces With Mathematica*, Applied Mathematics & Nonlinear Science, Chapman and Hall/CRC, London, UK, 2nd edition, 2006.

[30] D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer, "Co-clustering of biological networks and gene expression data," *Bioinformatics*, vol. 18, supplement 1, pp. S145–S154, 2002.

[31] J. H. Ward, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[32] J. E. McDermott, H. Shankaran, A. J. Eisfeld et al., "Conserved host response to highly pathogenic avian influenza virus infection in human cell culture, mouse and macaque model systems," *BMC Systems Biology*, vol. 5, article 190, pp. 190–212, 2011.

[33] B.-J. M. Webb-Robertson, L. A. McCue, N. Beagley et al., "A Bayesian integration model of high-throughput proteomics and metabolomics data for improved early detection of microbial infections," *Pacific Symposium on Biocomputing*, pp. 451–463, 2009.

[34] M. Ahdesmäki and K. Strimmer, "Feature selection in omics prediction problems using cat scores and false nondiscovery rate control," *Annals of Applied Statistics*, vol. 4, no. 1, pp. 503–519, 2010.

[35] A. F. Atiya, "Estimating the posterior probabilities using the K-nearest neighbor rule," *Neural Computation*, vol. 17, no. 3, pp. 731–740, 2005.

[36] P. MacCullagh and J. A. Nelder, *Generalized Linear Models*, Monographs on Statistics and Applied Probability, Chapman and Hall/CRC, London, UK, 1989.

[37] T. Mitchell, B. Buchanan, G. Dejong et al., "Machine learning," *Annual Review of Computer Science*, vol. 4, pp. 417–433, 1989.

[38] N. Beagley, K. G. Stratton, and B.-J. M. Webb-Robertson, "VIBE 2.0: visual integration for bayesian evaluation," *Bioinformatics*, vol. 26, no. 2, pp. 280–282, 2010.

[39] S. Oh, D. D. Kang, G. N. Brock, and G. C. Tseng, "Biological impact of missing-value imputation on downstream analyses of gene expression profiles," *Bioinformatics*, vol. 27, no. 1, Article ID btq613, pp. 78–86, 2011.

[40] E. Younesi, L. Toldo, B. Muller et al., "Mining biomarker information in biomedical literature," *BMC Medical Informatics and Decision Making*, vol. 12, article 148, pp. 148–160, 2012.

[41] R. Nugent and M. Meila, "An overview of clustering applied to molecular biology," *Methods in Molecular Biology*, vol. 620, pp. 369–404, 2010.

[42] W. S. Noble, "What is a support vector machine?" *Nature Biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[43] M. G. Schrauder, R. Strick, R. Schulz-Wendtland et al., "Circulating micro-RNAs as potential blood-based markers for early stage breast cancer detection," *PLoS ONE*, vol. 7, no. 1, Article ID e29770, 2012.

[44] C. Kingsford and S. L. Salzberg, "What are decision trees?" *Nature Biotechnology*, vol. 26, no. 9, pp. 1011–1013, 2008.

[45] R. Díaz-Uriarte and S. Alvarez de Andrés, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, article 3, 2006.

[46] M. H. Zweig and G. Campbell, "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine," *Clinical Chemistry*, vol. 39, no. 4, pp. 561–577, 1993.

[47] D. M. V. Powers, "Evaluation: from precision, recall and Fmeasure to ROC, informedness, markedness & correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.