Methodological approach from the Best Overall Team in the IMPROVER Diagnostic Signature Challenge

Adi L Tarca,1,2,3,* Nandor Gabor Than,1,4 and Roberto Romero1

1Perinatology Research Branch; The Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD)/National Institutes of Health (NIH)/United States Department of Health and Human Services (DHHS); Bethesda, MD and Detroit, MI USA; 2Department of Computer Science; Wayne State University; Detroit, MI USA; 3Center for Molecular Medicine and Genetics; Wayne State University; Detroit, MI USA; 4Department of Obstetrics and Gynecology; Wayne State University; Detroit, MI USA

Keywords: microarrays, outcome prediction, linear discriminant analysis, systems biology verification, transcriptomics

Introduction

The Industrial Methodology for Process Verification of Research (IMPROVER) Project1 was established as a framework to meet the respective needs of academia and industry, for the assessment of the validity of complex analyses that are now common in systems biology. The Diagnostic Signature Challenge2 was the first initiative of the IMPROVER team aimed at the verification of state-of-the-art approaches to the classification of clinical samples using transcriptomics data.

Since the emergence of the first microarray-based gene expression platforms, class/outcome prediction (e.g., normal vs. disease) has been one of the main applications3,4 (reviewed elsewhere5). Class prediction based on high-dimensional data are usually approached using statistical methods (e.g., logistic regression, discriminant analysis) or machine learning techniques6-8 (e.g., neural networks), as reviewed in the tutorial by Tarca et al.5 The typical challenges that are encountered in class prediction problems concern: (1) how to select the most informative features (see Feature Selection); and (2) which model/algorithm to use for exploring the relationship between feature values and class membership. In addition to these general issues involving class prediction, the Diagnostic Signature Challenge presented other concerns, namely: (3) how to choose the training data sets; (4) how to preprocess these data sets; and (5) how to deal with heterogeneities (e.g., batch, microarray platform) both within the training data sets and between the training and test data sets. Although we do not claim to have the best solution for these issues, we would like to share with the community our straightforward, yet proven reliable, approach to class prediction using microarray data.

The IMPROVER Diagnostic Signature Challenge used crowdsourcing to identify the best methods to classify clinical samples using transcriptomics data. Participating teams used public microarray data sets to develop prediction models in four disease areas, and then made predictions on blinded test data generated by the organizers. Here we describe the approach of the team for the Perinatology Research Branch (Team PRB; AL Tarca, R Romero), that was awarded the best performing entrant prize out of 54 entrants. The key elements of our approach included: (1) selection of training data sets by trial and error; (2) removal of batch effects by pre-processing the test and training data together; (3) the use of statistical significance and magnitude of change to select biomarkers; and (4) optimization of the number of biomarkers via the cross-validated performance of a simple linear discriminant analysis (LDA) model. Not only were our resulting models ranked consistently high, but they also generated parsimonious signatures of as low as two genes, unlike most of the other top-ranked teams that used hundreds of genes for prediction.

*Correspondence to: Adi L Tarca; Email: atarca@med.wayne.edu
Submitted: 01/15/2013; Revised: 04/11/2013; Accepted: 07/31/2013
http://dx.doi.org/10.4161/sysb.25980

Abbreviations: AC, adenocarcinoma; AUPR, Area Under Precision Recall curve; AUROC, Area Under the Receiver Operating Characteristic curve; BCM, Belief Confusion Metric; CCEM, Correct Class Enrichment Metric; COPD, chronic obstructive pulmonary disease; IMPROVER, Industrial Methodology for Process Verification in Research; LDA, Linear Discriminant Analysis; RMA, Robust Multi-array Average; SCC, squamous cell carcinoma; PLIER, Probe Logarithmic Intensity Error; MAS5, Microarray Suite 5; RMA, Robust Multi-array Average; QDA, Quadratic Discriminant Analysis; DLDA, Diagonal LDA; SVM, Support Vector Machines; PRB, Perinatology Research Branch; CDF, chip definition file; LC, lung cancer; MS, multiple sclerosis; PSO, psoriasis

Methodological approach from the Best Overall Team in the IMPROVER Diagnostic Signature Challenge

Adi L Tarca,1,2,3,* Nandor Gabor Than,1,4 and Roberto Romero1

1Perinatology Research Branch; The Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD)/National Institutes of Health (NIH)/United States Department of Health and Human Services (DHHS); Bethesda, MD and Detroit, MI USA; 2Department of Computer Science; Wayne State University; Detroit, MI USA; 3Center for Molecular Medicine and Genetics; Wayne State University; Detroit, MI USA; 4Department of Obstetrics and Gynecology; Wayne State University; Detroit, MI USA

Keywords: microarrays, outcome prediction, linear discriminant analysis, systems biology verification, transcriptomics

Introduction

The Industrial Methodology for Process Verification of Research (IMPROVER) Project1 was established as a framework to meet the respective needs of academia and industry, for the assessment of the validity of complex analyses that are now common in systems biology. The Diagnostic Signature Challenge2 was the first initiative of the IMPROVER team aimed at the verification of state-of-the-art approaches to the classification of clinical samples using transcriptomics data.

Since the emergence of the first microarray-based gene expression platforms, class/outcome prediction (e.g., normal vs. disease) has been one of the main applications3,4 (reviewed elsewhere5). Class prediction based on high-dimensional data are usually approached using statistical methods (e.g., logistic regression, discriminant analysis) or machine learning techniques6-8 (e.g., neural networks), as reviewed in the tutorial by Tarca et al.5 The typical challenges that are encountered in class prediction problems concern: (1) how to select the most informative features (see Feature Selection); and (2) which model/algorithm to use for exploring the relationship between feature values and class membership. In addition to these general issues involving class prediction, the Diagnostic Signature Challenge presented other concerns, namely: (3) how to choose the training data sets; (4) how to preprocess these data sets; and (5) how to deal with heterogeneities (e.g., batch, microarray platform) both within the training data sets and between the training and test data sets. Although we do not claim to have the best solution for these issues, we would like to share with the community our straightforward, yet proven reliable, approach to class prediction using microarray data.

*Correspondence to: Adi L Tarca; Email: atarca@med.wayne.edu
Submitted: 01/15/2013; Revised: 04/11/2013; Accepted: 07/31/2013
http://dx.doi.org/10.4161/sysb.25980

Abbreviations: AC, adenocarcinoma; AUPR, Area Under Precision Recall curve; AUROC, Area Under the Receiver Operating Characteristic curve; BCM, Belief Confusion Metric; CCEM, Correct Class Enrichment Metric; COPD, chronic obstructive pulmonary disease; IMPROVER, Industrial Methodology for Process Verification in Research; LDA, Linear Discriminant Analysis; RMA, Robust Multi-array Average; SCC, squamous cell carcinoma; PLIER, Probe Logarithmic Intensity Error; MAS5, Microarray Suite 5; RMA, Robust Multi-array Average; QDA, Quadratic Discriminant Analysis; DLDA, Diagonal LDA; SVM, Support Vector Machines; PRB, Perinatology Research Branch; CDF, chip definition file; LC, lung cancer; MS, multiple sclerosis; PSO, psoriasis

The IMPROVER Diagnostic Signature Challenge used crowdsourcing to identify the best methods to classify clinical samples using transcriptomics data. Participating teams used public microarray data sets to develop prediction models in four disease areas, and then made predictions on blinded test data generated by the organizers. Here we describe the approach of the team for the Perinatology Research Branch (Team PRB; AL Tarca, R Romero), that was awarded the best performing entrant prize out of 54 entrants. The key elements of our approach included: (1) selection of training data sets by trial and error; (2) removal of batch effects by pre-processing the test and training data together; (3) the use of statistical significance and magnitude of change to select biomarkers; and (4) optimization of the number of biomarkers via the cross-validated performance of a simple linear discriminant analysis (LDA) model. Not only were our resulting models ranked consistently high, but they also generated parsimonious signatures of as low as two genes, unlike most of the other top-ranked teams that used hundreds of genes for prediction.
Results

The analytical approach, described in detail in the Materials and Methods section, was applied by Team PRB to compute the class membership probabilities for each sample and phenotype (class) in the test data sets. The Diagnostic Signature Challenge organizers scored the participating teams for each challenge separately using three statistics, the Belief Confusion Metric (BCM), Correct Class Enrichment Metric (CCEM), and Area Under Precision Recall curve (AUPR). The BCM captures the average belief (confidence) that a sample belongs to class $i$ when indeed it belongs to class $i$. In this computation all classes weight the same regardless of the number of samples that are known to belong to each class. CCEM is similar to the accuracy (% of correctly classified samples) and, unlike BCM, is a threshold based metric. For instance in a two class problem when a belief value for a given class is $>0.5$ then that sample is assumed to be predicted in that class. Unlike the accuracy which counts with 1 each sample correctly classified and with 0 each sample misclassified, CCEM accounts for each correctly classified sample with the confidence in that prediction and penalizes each misclassification with their respective confidence. The AUPR is a threshold free metric that summarizes the ability to correctly rank the samples known to belong in a given class when sorted by the decreasing belief values for that class. A precision (P) recall (R) curve is constructed using all possible thresholds on the belief values with the precision being defined as true positives / (true positives + false positives) while the recall as true positives / (true positives + false negatives). All three metrics are transformed as needed to range between 0 (perfect misclassification) and 1 (perfect classification).

The ranking of each team based on each statistic was determined, and then the rank sum across all three statistics was computed for each sub-challenge. The overall rank-sum, covering all four sub-challenges was determined, and based on this figure our team achieved the best overall result with a rank sum of 18 (see Overall Ranking for the Diagnostic Signature Challenge at http://sbvimprover.com/page2). In the following subsections, we briefly describe the performance scores we received in each sub-challenge, disclose the identity of the predictors used and describe the details of the resulting prediction models.

Multiple Sclerosis (MS) Diagnostic sub-challenge

The optimal set of features we found for this sub-challenge involved only two genes, F13A1 (coagulation factor XIII, A1 polypeptide; ENTREZ ID: 2162) and GSTM1 (glutathione S-transferase mu 1; ENTREZ ID: 2944). The training and test data sets, together with the Linear Discriminant Analysis (LDA) model derived from the training data are shown (Fig. 1). The model is a straight line, dividing the feature space into two disjoint areas, composed of either low F13A1 and high GSTM1 expression, associated with a higher risk for MS, or high F13A1 and low GSTM1 expression, associated with a lower risk for MS. Thus, we identified a two-gene model to predict the risk of MS, which was ranked second in this sub-challenge (BCM = 0.629, CCEM = 0.625, and AUPR = 0.819).

Psoriasis (PSO) sub-challenge

Again, we identified two features sufficient to predict PSO, comprised of the probesets 232170_at for S100A7A (S100 calcium binding protein A7A; ENTREZ ID: 338324) and 207356_at for DEFB4A (defensin, β 4A; ENTREZ ID: 1673). The training and test samples, together with the LDA model derived from
the training data are shown (Fig. 2). The LDA model defines two areas in the feature space, high S100A7A and DEFB4A expression, associated with PSO, and low S100A7A and DEFB4A expression, associated with the controls. Although our team was ranked 12th in this sub-challenge, our absolute prediction performance was good (BCM = 0.980, CCEM = 0.983, and AUPR = 0.979), indicating that high expression of DEFB4A and S100A7A in the skin are strongly predictive of the risk of PSO.

Chronic Obstructive Pulmonary Disease (COPD) sub-challenge

In addition to our selection of three demographic covariates (smoking status, gender, and age), we also chose 10 microarray probesets (Table 1) as inputs to the LDA model, that we trained to predict COPD risk. Our model was ranked second in this sub-challenge (BCM = 0.701, CCEM = 0.649, and AUPR = 0.937). After unblinding of the test samples, we observed that our 10-gene signature, identified with the training data, was not strongly predictive of COPD on the test data. In fact, a simpler baseline model (involving only the demographic covariates of smoking status, gender, and age) outperformed our initially proposed model, with 10 probesets, and also outperformed the first-ranked team by 10% on the CCEM and approximately 4% on the BCM and AUPR.

Lung Cancer (LC) sub-challenge

Our choice of the top $d = 7$ features from each of the six pairwise comparisons between the four groups (AC1, AC2, SCC1, and SCC2) appeared optimal for the LC sub-challenge, leading to the selection of 25 unique probesets (Table 2). Our risk prediction model was ranked second in this sub-challenge (BCM = 0.459, CCEM = 0.492, and AUPR = 0.454). Note that these performance values are not directly comparable with the same metrics obtained in the previous three, two-class sub-challenges, yet they are used to rank the different teams on the same sub-challenge. The 25-gene signature we identified from the training data (Fig. 3) was useful in distinguishing adenocarcinomas (ACs) and squamous cell carcinomas (SCCs), but not between different stages of the sub-types of disease (Stages 1 and 2).

Discussion

In this article, we have presented the methodological approach we used in the IMPROVER Diagnostic Signature Challenge, to classify clinical samples based on microarray

<table>
<thead>
<tr>
<th>Table 1. The microarray probesets used to predict COPD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probeset ID</td>
</tr>
<tr>
<td>218835_at</td>
</tr>
<tr>
<td>206291_at</td>
</tr>
<tr>
<td>203815_at</td>
</tr>
<tr>
<td>209189_at</td>
</tr>
<tr>
<td>1554899_s_at</td>
</tr>
<tr>
<td>227404_s_at</td>
</tr>
<tr>
<td>205982_x_at</td>
</tr>
<tr>
<td>203645_s_at</td>
</tr>
<tr>
<td>219747_at</td>
</tr>
<tr>
<td>205681_at</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2. The microarray probesets used to predict LC subtypes and stages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probeset ID</td>
</tr>
<tr>
<td>218835_at</td>
</tr>
<tr>
<td>209125_at</td>
</tr>
<tr>
<td>201820_at</td>
</tr>
<tr>
<td>206504_at</td>
</tr>
<tr>
<td>209810_at</td>
</tr>
<tr>
<td>213796_at</td>
</tr>
<tr>
<td>209758_s_at</td>
</tr>
<tr>
<td>228979_at</td>
</tr>
<tr>
<td>213680_at</td>
</tr>
<tr>
<td>206165_s_at</td>
</tr>
<tr>
<td>231240_at</td>
</tr>
<tr>
<td>238017_at</td>
</tr>
<tr>
<td>235075_at</td>
</tr>
<tr>
<td>206204_at</td>
</tr>
<tr>
<td>223806_s_at</td>
</tr>
<tr>
<td>209351_at</td>
</tr>
<tr>
<td>205470_s_at</td>
</tr>
<tr>
<td>206373_at</td>
</tr>
<tr>
<td>205064_at</td>
</tr>
<tr>
<td>232082_x_at</td>
</tr>
<tr>
<td>226960_at</td>
</tr>
<tr>
<td>244056_at</td>
</tr>
<tr>
<td>205595_at</td>
</tr>
<tr>
<td>204455_at</td>
</tr>
<tr>
<td>229271_x_at</td>
</tr>
</tbody>
</table>
data. Although the test and training data sets were obtained by different labs, sometimes on different platforms, the predictions on the blinded test data were encouraging, except for the MS Stage prediction sub-challenge, which was not included in the final ranking of the teams, and hence not discussed in this article. Our strategy was based on (1) RMA preprocessing of the training and test samples together; (2) selection of reduced sets of genes using moderated $t$-test $P$-values and fold-change information; (3) employment of a LDA model; and (4) optimization of the gene selection procedure using a cross-validated AUROC statistic from the LDA model. Interestingly, unlike some of the other best-ranked teams, overall or for each sub-challenge individually, our models were based on a reduced number of predictors, as low as two genes in two of the four sub-challenges. The deliberate and, one can argue, risky strategy for using a low number of genes for outcome prediction was based on previous work, and our own experience that the LDA model works better with fewer, rather than more, features. Thus, we show that the use of a low number of genes, thereby avoiding overfitting, is an optimal approach to outcome prediction using microarray data.

The two-gene model that we proposed to predict the risk of MS, was ranked second in this sub-challenge (BCM = 0.629, CCEM = 0.625, and AUPR = 0.819) yet it was substantially worse on all metrics compared with the one of the best performer in this sub-challenge, Team227, that achieved BCM = 0.884, CCEM = 0.883, and AUPR = 0.874. The approach used by Team227 cannot be applied if the entire test data set is not provided in the training phase since it is based on clustering the test samples. Moreover, Team227 used 44 training samples obtained on the same platform as the test samples, while we used 12 training samples obtained on a different microarray platform.

Our performance results on the PSO sub-challenge (BCM = 0.980, CCEM = 0.983, and AUPR = 0.979) were within 2% from the best performer on this sub-challenge (Team294) that made a perfect submission. Team294 used hundreds of genes in their model while we only used two. Results of a post challenge simulation in which we re-run our approach with different random seeds that affect the cross-validation process, and hence the number of features used in the model, provided evidence that our performance on this sub-challenge was the result of a particularly unlucky random seed, with 50% of the trials

Figure 3. The 25-gene signature used to predict LC subtypes (AC or SCC) and stages (1 or 2). The heat map shows expression levels on a color scale of the 25-gene signature in test samples. Genes are clustered based on similarity of their expression profile across all samples using an Euclidean distance. Data are centered and scaled by row.
(typically including about 5 genes) leading to a rank of 2nd or higher (data not shown).

For COPD, our model was ranked second (BCM = 0.701, CCEM = 0.649, and AUPR = 0.937) while the best performer on this challenge, Team122, achieved BCM = 0.683, CCEM = 0.730, and AUPR = 0.942. The fact that both threshold free metrics BCM and AUPR were about the same but Team122 had a notable 0.08 units of improvement on the CCEM metric may be attributed to the fact that their model (bagging of 100 logistic regression models constructed using sequential feature selection) handled better the borderline cases. For instance if predicted probabilities for some of the true COPD samples were 0.49 with LDA but 0.51 with logistic regression BCM and AUPR will be insensitive to such differences but CCEM would count the LDA predictions as wrong because they did not pass the 0.5 threshold.

In the LC sub-challenge, our risk prediction model was ranked second (BCM = 0.459, CCEM = 0.492, and AUPR = 0.454), being only slightly inferior to the one of the best performer in this sub-challenge (Team36), that achieved BCM = 0.479, CCEM = 0.509, AUPR = 0.458. Of note, Team36 did not use an automatic approach to develop their model but picked manually one single marker to distinguish between AC and SCC subclasses being aware that the stages classification (AC1 vs AC2 and SCC1 vs SCC2) were not tractable. In contrast our model included features that appear best to distinguish between cancer stages, and hence contributing noise that could easily explain such minor differences in performance.

Next, we discuss the biological plausibility of the genes we identified as outcome predictors in our models in context, for each sub-challenge separately, although we acknowledge that many other combinations of genes could have led to similar of better prediction performance.

MS Diagnostic

A systematic review of genome-wide expression studies conducted on MS, found little overlap in the gene signatures between studies, showing only 12 genes in common within at least three of the seven studies examined. Unsurprisingly, F13A1 and GSTM1, identified and used in our prediction model, were not among those 12 genes. This can be explained in part by the fact that the seven studies involved in the meta-analysis used older Affymetrix and cDNA microarray platforms not represented in our training data sets. Also, the redundancy in the network structure of gene regulation and the small sample size of these studies can also be a cause for the lack overlap between the lists of differentially expressed genes found in each of the studies. Of note, the training and test data sets included in this sub-challenge were obtained from transcriptomics studies on peripheral blood mononuclear cells, yet in MS, an autoimmune disease, the tissue under attack is the myelin sheath of the brain. As a consequence, the training and test data, mainly reveal non-specific gene signatures associated with T-cell activation and expansion, inflammation and apoptosis (http://sbvimprover.com/). Nevertheless, the low expression of F13A1 and GSTM1 identified in blood mononuclear cells, may suggest a specific association with MS pathophysiology. F13A1 encodes the catalytic A subunit of coagulation factor XIII, the last enzyme activated in the blood coagulation cascade. Coagulation factor XIII catalyzes the formation of crosslinking between fibrin molecules, thereby stabilizing the fibrin clot. The deficiency of this coagulation factor is associated with an increased bleeding tendency, defective wound healing and habitual abortion.22 Of importance, a growing body of evidence has revealed molecular mechanisms linking coagulation and inflammation. The pro-inflammatory role of fibrinogen in MS has also recently been discovered. Moreover, fibrinogen and fibrin degradation products play key roles in determining the extent of local or systemic inflammation.13,14 Therefore, our finding that MS is associated with low expression of F13A1, a catalytic sub-unit of an enzyme critical for the enzymatic cascade in fibrin production and degradation, suggests F13A1 may play a role in local inflammatory processes in MS. GSTM1 encodes the class mu glutathione S-transferase, an enzyme that catalyzes conjugation of reduced glutathione with potentially toxic substrates, including carcinogens, therapeutic drugs, environmental toxins and products of oxidative stress. Oxidative stress has been implicated in MS pathogenesis,15,16 and homozygosity for the GSTM1 null genotype is associated with the long-term prognosis of severe disability for MS patients,17 therefore our finding on the association of low GSTM1 expression with the risk of MS is also of relevance to pathogenesis of this autoimmune disease.

PSO

We found high expression of two antimicrobial peptide genes found in the skin, S100A7A and DEFB4A, is associated with the development of PSO, a chronic autoimmune disease of the skin involving hyperproliferation of the squamous epithelium. Our results are in accordance with previous findings showing upregulation of a number of antibiotic peptides in psoriatic skin lesions, potentially acting as a “chemical shield” to provide the psoriatic skin with resistance against infection.18 Among those antimicrobial peptide genes, S100A7A encodes psoriasin, a protein implicated in the mediation and regulation of skin differentiation.19,20 Psoriasin acts as a chemoattractant and an alarmin, priming epidermal keratinocytes for cytokine production, to induce the amplification of local inflammation in PSO.20 The gene DEFB4A encodes a defensin originally discovered in psoriatic skin lesions. This defensin (also known as human β-defensin 2) was shown to be produced by epithelial cells and exhibited potent antimicrobial activity against Gram-negative bacteria and Candida albicans.21,22 The association between higher genomic copy numbers for DEFB4A and other β-defensin genes, and the risk of PSO, has previously been established,23 implying a strong genetic component for this disease. Of note, increased expression of β-defensin 2 in the skin leads to high systemic concentrations of the peptide in patients with PSO, with β-defensin 2 blood concentrations strongly correlated with disease activity, suggesting its utility as a surrogate biomarker for PSO disease activity.24 In this regards, it will be important to ensure that DEFB4A expression is reduced when psoriatic lesions improve in response to treatment.25 All these findings are in agreement with our model (Fig. 2).
COPD

COPD is a major cause of morbidity and mortality among smokers worldwide. It is characterized by chronic bronchitis, inflammation of the large airways, obstruction of small airways, and destruction of the alveoli. The specific challenge of this sub-challenge was that the majority of the training data sets were from small airway transcriptomics, while the test data sets were from large airway transcriptomics. Of importance, it has recently been shown there is a non-uniform distribution of leukocytes in the small and large airways in COPD, with increased macrophages in large airways, and increased CD8+ T lymphocytes and neutrophils in small airways. Moreover, histone deacetylase 2 expression also differs between these two airway segments in COPD patients in small airways. Moreover, histone deacetylase 2 expression also differs between these two airway segments in COPD patients that smoke, suggesting epigenetic and inflammatory differences between small and large airways in COPD. Overall, these results suggest that the transcriptomic signature of COPD in small and large airways is likely divergent. Indeed, we did not regard the 10 genes we used as predictors in our COPD model, based on small airway transcriptomics data (Table 1), as highly predictive for this condition in the large airway transcriptome, as they did not contribute much to the overall prediction performance of the model. The baseline covariates (gender, age, and smoking status) were more important for prediction than the gene expression values, corresponding well with the fact that 90% of COPD patients are male smokers. As we described in the Results section, the use of the baseline covariates alone, results in a better model (as assessed by prediction performance on the test data set), than the model we submitted for this challenge. Although one can argue that this simpler model was chosen after unblinding of the test data results and therefore is optimistically biased, its simplicity prevents it from overfitting. Indeed, the relatively small contribution of the gene expression levels was apparent to us from the training data, by the slight improvement of the AUROC from 80% (baseline covariates alone) to only 86% (gene expression with baseline covariates). However, as it did not appear to be detrimental, the expression data were used in the model.

LC

LC is the leading cause of cancer deaths in men and women, and non-small cell LC, which includes squamous cell carcinomas and ACs, is the most common type (http://sbvimprover.com/). We found 25 probesets, corresponding to 24 genes, predictive of these two LC subtypes. The list of genes included four encoding for keratins (KRT6A, KRT5, KRT6B, and KRT14), which are assigned by the Gene Ontology consortium to the structural constituent of cytoskeleton molecular function. Not only has dysregulation of the cytoskeleton previously shown to be involved in LC, but also differential expression of keratin proteins in the proteome of these two subtypes of non-small cell LC was recently observed. Remarkably, three small proline-rich protein encoding genes (SPRR1A, SPRR1B, and SPRR3) and four surfactant or surfactant-associated protein encoding genes (SFTPA2, SFTPB, SFTA2, and SFTA3) were also on our selected list, in accordance with the former trio being highly expressed in squamous epithelial cells and carcinomas, and the latter being characteristic for secretory alveolar cells and ACs. Another gene from our list, DSG3, encodes for desmoglein 3, a component of the intercellular desmosomal junction, which is highly upregulated in squamous cell carcinoma of the lung, compared with AC, and has been suggested as a good marker of the former.

The strengths of our approach reside in the fact that it proved to be reliable on several data sets with relatively minimal user interaction in the process and tuning from one data set to the other. Also, this approach produced gene signatures that were by far more parsimonious than most of the top ranked teams. A reduced biomarker set is desired in an industrial context as it reduces the cost of eventual commercial kits that could be developed for disease screening. In addition the genes we identified show strong biological plausibility of disease pathogenesis. Some of the aspects that we would do differently in a similar setup are:

1. Use a larger number of cross-validation repetitions, since the value of 5 that we used can still lead to unstable results when determining the optimal number of features;
2. Automate the process of finding best cut-offs on P-values and fold-changes so that more solutions can be explored;
3. Instead of relying only on the AUC values in the model optimization process we could have used the three metrics used for team ranking;
4. The selection of training data sets, especially for the MS Diagnostic sub-challenge could have been better. The best team in this sub-challenge performed much better by using a more than three times larger training set.

Materials and Methods

The participants in the Diagnostic Signature Challenge were asked to make predictions about unlabeled samples in four disease areas: PSO, MS, COPD, and LC. Full details on the significance of these diseases and other relevant details are available at http://www.sbvimprover.com/. Our approach to the Diagnostic Signature Challenge is described for simplicity as a series of four separate steps at which appropriate decisions had to be made. These four steps were: selecting training data sets, data preprocessing, classifier selection, and feature selection. Although our goal was to implement a strategy that was common to all four sub-challenges, slight variations were needed to accommodate the following specifics:

Number of classes

For all sub-challenges, the number of classes (possible outcomes) was two (control, disease), except for the LC sub-challenge, where it was necessary to distinguish four phenotypes. Therefore, the feature selection step involved differences between two- vs. four-class setups.

Training and test data set platforms

The training data of our choice, and the corresponding test data sets generated by the organizers, were obtained on the same platform for the COPD, LC, and PSO sub-challenges, but on a different platform for the MS Diagnostic sub-challenge. The data preprocessing step was adjusted accordingly based on this dichotomization.
Our overall strategy can be seen as a cascade of choices, among available alternatives, for every major aspect of the classifier development (Fig. 4).

**Selecting training data sets**

The different data sets suggested for each sub-challenge were obtained using different microarray platforms by different labs in different batches. Such non-biological differences in the training data can obscure real biological differences that are predictive for the phenotype. Our approach was to use only some of the suggested training data sets, ensuring those chosen were obtained on the same platform as the test data set. This was performed for the COPD, LC, and PSO sub-challenges, but not for the MS Diagnostic sub-challenge.

For the COPD sub-challenge, among all data sets profiled on the HG-U133 Plus2.0 platform, the platform used to generate the test data, we selected 66 control and 22 disease samples included in three of the largest data sets (GSE10006, GSE8545, and GSE5058).

For the COPD sub-challenge, among all data sets profiled on the HG-U133 Plus2.0 platform, the platform used to generate the test data, we selected 66 control and 22 disease samples included in three of the largest data sets (GSE10006, GSE8545, and GSE5058). For the PSO sub-challenge, all suggested training data sets and the test data were profiled using the HG-U133 Plus2.0 platform. Therefore, we included both test sets, GSE13355 (64 control and 58 disease samples) and GSE14905 (21 control and 33 disease samples), with a total of 85 control and 91 disease samples. Note, that for GSE14905, our tally included an additional 5 samples in the disease group that were not shown at http://www.sbvimprover.com/.

For the LC sub-challenge, we used all three provided data sets (GSE10245, GSE18842, and GSE2109), as they were profiled using the same platform as the test data set (HG-U133 Plus 2.0). One sample in each group was removed because of either bad hybridization or their appearance as outliers in Principal Component Analysis (PCA) plots. The outliers removal was based on a visual inspection of PCA plots, yet a more rigorous and consistent approach could have been used instead. The final number of samples in each of the groups designated as AC1, AC2, SCC1, and SCC2 were 50, 19, 56, and 16, respectively.

For the MS Diagnostic sub-challenge, we attempted to use the same strategy for training data selection. The test data were generated using the HG-U133 Plus2.0 platform, therefore we selected the only two data sets profiled on the same platform (GSE21942 and E-MTAB-69). An additional Illumina data set, GSE19224, was also initially considered as we had used it in the MS Stage sub-challenge, that was subsequently removed from the scoring process. However, using the cross-validated Area Under the Receiver Operating Characteristic (AUROC) of the classifier (described in the following sections), we concluded that the prediction quality was too poor, likely for a number of reasons, including: (1) GSE19224 included MS samples but no controls; (2) GSE21942 included controls but no MS samples; and (3) E-MTAB-69 included both control and MS samples but the controls included cases with neurological disorders of a non-inflammatory nature. Due to these reasons, we used only one training data set (GSE23832) in the final trial, which was profiled on a different platform (Human Gene 1.0 ST) and was small, but did offer the advantage of including both MS and appropriate control samples. This data set contained 4 controls and 8 disease samples, although only 4 MS samples were shown on the IMPROVER website. An overall summary of the training and test data sets used in all four sub-challenges is provided (Table 3).

**Data preprocessing**

The typical approach to manage batch effects in class prediction applications is to preprocess the data within a batch...
and then correct for biases in each batch separately.\textsuperscript{56} For the Affymetrix microarray platforms involved in our training and test data sets, several established methods exist for preprocessing (background correction, normalization and categorizing probe intensities into probeset/gene expression summaries), including MAS 5.0,\textsuperscript{37} RMA,\textsuperscript{38} GCRMA,\textsuperscript{39} and PLIER.\textsuperscript{40} Batch effect methods that can be used include mean-centering, standardization, a modified version of COMBAT,\textsuperscript{41} and ratio-based methods.\textsuperscript{36} For instance the mean-centering approach works by setting to 0 the mean expression levels (on a log scale) of a given gene in each of the batches available. Instead of relying on the assumption that the test data will contain a reasonable fraction of both disease and control samples, as most of these methods assume, we chose to preprocess the training and test samples together, when the same platform was involved. Alternatively, when more than one platform was involved, each batch was preprocessed separately and then quantile normalization was applied on gene level expression summaries.

The two microarray platforms we used to generate the training data sets in the four sub-challenges were Affymetrix arrays HG-U133 Plus 2.0 and Human Gene 1.0 ST. The first platform is a 3’ end biased array containing probes grouped into probesets (one or a few per gene), while the second platform has multiple probes on different exons of a given gene. Our data preprocessing approach was tailored appropriately depending on whether or not the training and test data sets were produced on the same platform.

The following steps were used to summarize the expression level at the probeset level for the COPD, LC, and PSO sub-challenges for which all training and test data were obtained using the HG-U133 Plus 2.0 platform:

1. Run RMA on training samples separately, using the chip definition file (CDF) HuGene-1.0-st-v1 to summarize expression at the transcript level.

2. Average expression of all transcripts that can be mapped to the same gene using annotation from the Bioconductor\textsuperscript{44} package hugene10stranscriptcluster.db.

3. Test data sets:
   1. Run RMA on test samples.
   2. Discard probesets from the test data not present in at least 5 samples.
   3. Average expression of all probesets that can be mapped to the same ENTREZ ID.

After these steps, both training and test data sets are summarized at the gene level (ENTREZ ID). Genes not common between the training and test data sets were discarded, and quantile normalization was applied to both training and test samples, with a batch effect removal method, using the preprocessCore package.\textsuperscript{45}

**Classifier selection**

Class prediction applications rely on a particular model or algorithm (classifier) that learns to associate the values of the attributes (features) to the probability of each possible outcome (class). Different types of classifiers produce different decision boundaries between classes, and therefore perform differently when applied to the test data. The classifier we chose and used consistently in this challenge was an LDA. Assuming that the training data are organized in a $n \times p$ matrix $X = (x_{ij})$, where $x_{ij}$ represents the measured value of the feature $j$ in the sample $i$, every row of the matrix $X$ is a vector $x_j$ with $p$ features to which a class label $y_j$ is associated, $y = 1,2,...,c,...,K$. The LDA works by assuming that $x$ follows a multivariate normal distribution, $N(\mu, \Sigma)$, with a different mean vector in each class $\mu$ and a common covariance matrix $\Sigma$. Plugging in the multivariate normal probability density function formula, the sample derived estimates of the covariance matrix and class mean vectors, and the class-conditional probabilities $p(x|y = c)$, were obtained.\textsuperscript{46} The posterior probabilities requested by the challenge organizers for each sample were derived from $p(x|y = c)$ after multiplying them with the prior

### Table 3. Summary of the training and test data used in the sub-challenges

<table>
<thead>
<tr>
<th>Sub-challenge</th>
<th>Platform</th>
<th>Sample Size</th>
<th>Platform</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSO</td>
<td>HG-U133_Plus_2</td>
<td>66/22</td>
<td>HG-U133_Plus_2</td>
<td>35/27</td>
</tr>
<tr>
<td>COPD</td>
<td>HG-U133_Plus_2</td>
<td>85/91</td>
<td>HG-U133_Plus_2</td>
<td>16/24</td>
</tr>
<tr>
<td>LC</td>
<td>HG-U133_Plus_2</td>
<td>50/19/56/16</td>
<td>HG-U133_Plus_2</td>
<td>41/36/34/39</td>
</tr>
<tr>
<td>MS Diagnostic</td>
<td>HuGene-1.0-st</td>
<td>4/8</td>
<td>HG-U133_Plus_2</td>
<td>32/28</td>
</tr>
</tbody>
</table>

Sample size indicates the number of controls samples and (/) number of disease samples. For LC the groups order is AC1, AC2, SCC1, and SCC2. The covariance matrix $\Sigma$ is square of dimension $p \times p$. The element $\Sigma_{ij}$ of this matrix is the covariance between the variables $i$ and $j$. Discarding undetected probesets ensures that low expression features are not selected as markers at subsequent steps. Assuming equally sized test and training sets, each containing a balanced number of cases and controls, keeping all probesets detected present in at least 25% of all samples would ensure that a gene detected in at least half (50%) of the samples of any group would be retained. This type of filtering is discussed in detail elsewhere.\textsuperscript{43} The last step, although considered optional, ensures that the probesets found as markers correspond to known genes.

For the MS Diagnostic sub-challenge, the procedure above was slightly modified to account for the difference in the array platform between the training (Human Gene 1.0 ST) and test (HG-U133 Plus 2.0) data sets.
probabilities of each class \( p(y = c) \) and standardizing these products to sum up to 1.0:

\[
p(y = c \mid x_i) = \frac{p(x_i \mid y = c) \cdot p(y = c)}{\sum_{c=1}^{K} p(x_i \mid y = c) \cdot p(y = c)}
\]

Because we had no a priori expectations about the prevalence of each class in the test data set, we assumed that \( p(y = c) = 1/K, \) that is, all classes have the same prior probability. The linear discriminant was previously found to perform better when the number of features is low,\(^{10}\) and hence, our next step was to search for a reduced set of inputs into the LDA model. When the number of features \( p \) in the LDA model increases, the number of parameters to estimate in the variance-covariance matrix increases even faster \((p^2 + p)/2)\) and this is leads to unreliable estimates given the typically low number of samples that were available for training.

**Feature selection**

The data preprocessing step produced a training expression matrix \( X_i^j \) and a test expression matrix \( Z_i^j \), where \( i \) corresponds to samples and \( j \) to features. The classification model described above uses \( X_i^j \), and the corresponding class label \( y_j \), to fit the LDA model. A reduced number of features, predictive of the outcome, needed to be identified before the classifier could be trained. The LDA model in particular requires that the number of features is lower than the number of training samples. This feature selection step is described below for each sub-challenge.

**MS Diagnostic**

This sub-challenge differed from the others by the size of the training data set, which was small (4 controls and 8 disease samples). The following steps were used to select a reduced set of biomarkers as inputs to the LDA model:

1. Apply a moderated \( t \)-test, implemented in the *limma* package,\(^47\) to compute differential expression statistics at the probeset level between disease and control groups.
2. Retain genes with \( P < 0.005 \) and sort them by decreasing fold change.
3. Begin with the top \( d = 2 \) features and move sequentially through the list, one gene at a time, up to 5 genes.
4. The best value of \( d \) is chosen via a 4-fold cross-validation (repeated 10 times) maximizing the AUROC of the LDA model using the top \( d \) features. Here \( t_{AUROC} = \text{mean}(AUC)/\text{sd}(AUC) \) across all 25 point estimates.

**COPD**

The same steps and parameters used above for the PSO sub-challenge were also used for the COPD sub-challenge, with the following changes: (1) at step 1, smoking status, gender, and age were included as covariates in the linear model; (2) at step 3, probesets with \( P < 0.01 \) were retained and then sorted by decreasing fold change; and (3) the LDA model included also smoking status, gender, and age covariates.

**LC**

The feature selection step for the LC sub-challenge included several modifications, due to the four-class discrimination problem:

1. Apply a moderated \( t \)-test at the probeset level for each of the six possible between-group comparisons (SCC1-SCC2, SCC1-AC1, SCC1-AC2, SCC2-AC1, SCC2-AC2, and AC1-AC2).
2. Map probesets to ENTREZ IDs, and retain only the probeset with the smallest \( P \)-value per gene in each comparison separately.

Note, that for each complete 4-fold cross-validation, one ROC curve was constructed from the 12 samples. What differs from one repetition of the cross-validation to the next is how the 12 samples are split into 4 folds. The choice of the \( P \)-value threshold at step 2 was made by trial and error, maximizing the AUROC on the training data. The trial and error procedure involved testing a few possible \( P \)-values thresholds such as 0.05, 0.01 and 0.005 and choosing the one with the best AUROC value. The most stringent cut-off was limited by the need to have at least 5 genes meeting the threshold at every single trial in the repeated cross-validation process, where 5, as described above, is the maximum number of features we have considered to include in the model for this sub-challenge.

**PSO**

Unlike the MS Diagnostic subchallenge, the PSO training sample size was larger, and the features were probesets instead of genes. Therefore, feature selection included the following steps:

1. Apply a moderated \( t \)-test to compute differential expression statistics at the probeset level between disease and control groups.
2. Map probesets to ENTREZ IDs and keep only the probeset with the smallest \( P \)-value for each gene.
3. Retain probesets with \( FC > 8 \) and sort them by increasing \( P \)-value.
4. Start with the top \( d = 2 \) features and move sequentially through the list, two genes at a time, up to 20.
5. The best value of \( d \) is chosen via a 5-fold cross-validation (repeated 5 times) maximizing \( t_{AUROC} \) of a LDA model using the top \( d \) features. Here \( t_{AUROC} = \text{mean}(AUC)/\text{sd}(AUC) \) across all 25 point estimates.

The choice of the fold change threshold at step 2 was made by trial and error maximizing \( t_{AUROC} \) on the training data. The trial and error procedure involved testing a few possible fold-change thresholds such as 2, 4, and 8 fold and choosing the one with the best AUROC value. The most stringent cut-off was limited by the need to have at least 20 genes meeting the threshold at every single trial in the repeated cross-validation process, where 20, as described above, is the maximum number of features we have considered to include in the model for this challenge. The choice of 5 folds in the cross validation was made to allow that several samples from each class appear in the test set so that an ROC curve can be constructed for each test set. In contrast a lower number of folds would remove too many samples from the model training process which could be detrimental.
3. Retain probesets with $P < 0.01$ and sort them by decreasing fold change.

4. Start with the top $d = 1$ feature from each comparison and move sequentially through the list, one gene at a time, up to $7$.

5. The best value of $d$ is chosen via a 5-fold cross-validation (repeated 5 times) maximizing $t_{\text{AUROC}}$ of a LDA model using the top $d$ features. Here, $t_{\text{AUROC}} = \frac{\text{mean}(\text{avgAUC})}{\text{std}(\text{avgAUC})}$ across all 25 point estimates. The $\text{avgAUC}$ is the average AUC for the four possible ROC curves that were constructed by contrasting each class against the remaining three.

The $P$-value cutoff at step 3 was determined via trial and error by maximizing $t_{\text{AUROC}}$.

All analyses were performed using the R statistical language and environment\textsuperscript{49} version 2.15 and other infrastructure packages available in Bioconductor version 2.16.

References


5. The best value of $d$ feature from each comparison and $P < 0.01$ and sort them by decreasing $P$-value.

Acknowledgments

This research was supported, in part, by the Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health and U.S. Department of Health and Human Services. We thank Maureen McGerty for her careful revision of the manuscript.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.
45. Bolstad, BM. preprocessCore: A collection of preprocessing functions. (R package version 1.18.0). 2012