An empirical workflow for genome-wide single nucleotide polymorphism-based predictive modeling

Charalampos S. Floudas, MD, PhD, MS,1 Jeya Balaji Balasubramanian, MS1, Marjorie Romkes PhD2, Vanathi Gopalakrishnan PhD1,2
1Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA; 2University of Pittsburgh Cancer Institute, Pittsburgh, PA

Abstract
Technology is constantly evolving, necessitating the development of workflows for efficient use of high-dimensional data. We develop and test an empirical workflow for predictive modeling based on single nucleotide polymorphisms (SNP) from genome-wide association study (GWAS) datasets. To this aim, we use as a case study SNP-based prediction of survival for non-small cell lung cancer (NSCLC) with a Bayesian rule learner system (BRL+). Lung cancer is a leading cause of mortality. Standard treatment for early stages of NSCLC is surgery. Adjuvant chemotherapy would be beneficial for patients with early recurrence; consequently, we need models capable of such prediction. This workflow outlines the challenges involved in processing GWAS datasets from one popular platform (Affymetrix®), from the results files of the hybridization experiment to the model construction. Our results show that our workflow is feasible and efficient for processing such data while also yielding SNP based models with high predictive accuracy over cross validation.

Introduction
Translational research includes prediction of clinical outcomes, such as patient prognosis and duration of response to treatment, from available evidence. Therefore, improvements in predictive ability can potentially refine the clinical decision making and offer significant benefit to the patients.

The data currently used for prediction are mostly produced from novel high-throughput methodologies. The datasets generated by such methods are high-dimensional, such as microarray gene expression, genome-wide single nucleotide polymorphisms (SNP) and genome-wide methylation profiling data. This trend intensifies the pressure on the bioinformatics community to develop appropriate tools for data manipulation, analysis, and interpretation.

Multiple such tools are being developed and the corresponding publications present the theoretical background and evaluation of each tool, as well as the significance of the results. However, an important but often neglected aspect is a clear description of the workflow of the experiment. Adequately described workflows would facilitate reproduction of the experimental process and enhance comparability of different approaches to the same problem.

This paper describes a workflow applicable to the task of prediction. It is based on our experience with genome-wide SNP based prediction of a clinical outcome in non-small cell lung cancer (NSCLC) and the challenges associated with it. We have utilized the Affymetrix® Genome Wide Human SNP Array 6.0, the Affymetrix® Genotyping Console (AGC) software, the software PLINK for feature selection, the Bayesian rule learning (BRL) system and the Unix command line tools. The workflow can be used with alternate methods for experimental data acquisition, data processing, feature selection and predictive modeling tools. BRL has been used successfully for prediction of disease state based on proteomic and gene expression data, but, to the best of our knowledge, it has not been applied to GWAS data.

Background
Lung cancer is the leading cause of cancer related mortality in both sexes in the US and non-small cell lung cancer (NSCLC) accounts for 85% of cases. The 5-year survival for all stages of NSCLC is 17% but it is better for cases detected when the disease is localized (stage I). For these patients standard treatment is surgery. However in 30-40% the disease will relapse and their disease-free time will be shorter, hence they might benefit from adjuvant chemotherapy (ACT). Selection of patients for ACT is currently based on clinical parameters.

Efforts at identifying consistent prognostic molecular markers for recurrence or survival have used gene expression signatures, SNPs, miRNA and methylation profiling. The standard survival analysis uses Cox regression modeling but typically in high-throughput data, a feature selection step is used for initial filtering of the available
predictors. Analysis using candidate gene SNPs or genome-wide SNPs is of particular interest given that genotypes might be considered more stable in time than gene expression.

Data mining methods are useful techniques for prediction of phenotypes such as disease or health states from high-dimensional “omics” data. A particular method that has been applied successfully is the Bayesian rule learning (BRL) system, which uses a Bayesian score to construct Bayesian networks and to learn probabilistic rule models from them. The models produced are easily interpretable by the biomedical scientist and have been shown to have fewer markers and equivalent or greater classification performance in comparison to models derived from other rule learning methods. In this paper, we develop and apply a novel workflow that permits the application of BRL to a genome-wide SNP dataset generated for early stage lung cancer survival prediction.

**Methods**

Our workflow is presented in Figure 1 and each step is described in detail in the following paragraphs.

**Dataset:** A total of 86 tumor samples from de-identified patients with completely resected stage I NSCLC were genotyped at the University of Pittsburgh Cancer Institute (UPCI) Cancer Biomarkers Facility with the Affymetrix Genome Wide Human SNP Array 6.0 which genotypes 10^6 SNPs. Clinical variables for the patients were made available from the UPCI. Duration of follow-up of patients was until recurrence or at least 1000 days if no recurrence. As endpoint for our modeling the disease free survival (DFS) was chosen, defined as the time until recurrence occurred or until the end of follow-up if no recurrence occurs.

A database can be used instead of a spreadsheet to store patient identification numbers, clinical data, samples identification numbers for any type of sample available for each patient and information pertaining to the samples quality. Such a database can be built on SQLite and facilitates data handling, especially when multiple types of high-throughput data are available for the same patient cohort.

In biomedical data mining, a typical task entails the learning of a mathematical model from gene expression or protein expression data that predicts well the phenotype such as disease or health. Such a task is called classification and the model that is learned is termed a classifier. In data mining, the variable that is predicted is called the target variable (or simply the target), and the features used in the prediction are called the predictor variables (or simply the predictors). The BRL takes continuous or categorical variables as input (potential predictors), but the target variable must be categorical, therefore we classified the lung cancer patients into two categories. We used the 3rd quartile of the distribution of the patients’ DFS as a cut-off point to designated one group as having “poor” survival and the other as having “good” survival. When analyzing survival, transforming continuous values to discrete intervals represents a trade-off; however, discretization based on established clinical criteria provided by the investigator can help ameliorate this. The investigator can also subjectively assess the sample size requirements, the phenotypic parameters and the length of follow-up needed for the model.

**Genotyping:** The image files (.DAT) produced by the SNP chip were processed with the AGC software and transformed to intensity files (.CEL). Automatic quality control (QC) of intensity characterized 19 samples as “out of bounds” and the number of samples for further analysis was reduced from 86 to 67. Genotyping of the “in bounds” samples was performed with the AGC.

**Feature selection:** For the initial feature selection stage we used PLINK, a widely used software for analysis of genome-wide SNP data. Genotypes produced by the AGC are stored in the .CHP file format that is not readable by PLINK. The AGC can export the data to the PLINK format, but this requires that at the time of genotyping the attribute files (.ARR) have been created, containing the phenotypes of interest. However that might not be the case, especially when dealing with cancer survival data instead of a health/disease phenotype, where the collection of samples for genotyping might occur at a fairly short period (e.g. 2 years) while the clinical data to be analyzed with the genotypes might take a longer time (e.g. 3 years). Another case for missing phenotypes at the time of genotyping might be strict blinding, as could be the case in samples procured in the context of a double-blind clinical trial. In this case the genotypes can be exported from AGC as text files (.TXT) and be processed with UNIX scripts to create the format necessary for analysis with PLINK, with a dummy phenotype. Testing can then be performed when the appropriate phenotype is made available. A script for the conversion of Affymetrix genome-wide SNP data to the PLINK format with a dummy phenotype when the attribute files are missing, and a small example dataset are available for download from the Software tab at the website [http://www.dbmi.pitt.edu/probe/](http://www.dbmi.pitt.edu/probe/).
Figure 1. The SNPR workflow for SNP-based Prediction with Bayesian Rule Learning. Column “Process & Platform” refers to the phase and the platform (array/software) used. Column “Data format” refers to the format of the file that contains the data. Rectangular boxes: processing steps; parallelogram boxes: input/output; The dotted horizontal lines create groups of steps occurring within the same platform.

We used PLINK for per sample and per SNP QC. Samples were checked for discordant sex info and genotyping failure rate and we assumed no related individuals or population stratification. We excluded markers with call rate <95%, minor allele frequency <5%, and deviation from Hardy-Weinberg expectations ≤1e-6. After QC, one sample and 281,544 SNPs were removed (66 samples and 628,078 SNPs were retained). PLINK options for association testing include the chi-square for alleles or for genotypes, the Cochran-Armitage trend test and logistic regression. These tests can be used with alternative phenotypes, facilitating the study of multiple aspects of the available data. We used the genotypic test and from the results we selected the 100 SNPs with the smallest p values for the BRL models construction. PLINK can be used to derive a file (.RAW) containing a dataset that will include all samples but only those SNPs that we have selected and the chosen phenotype. In this file, the genotypes for each SNP are coded as categories, where 0 is the major allele homozygous state, 1 is heterozygous, 2 is the minor allele homozygote and 3 is the missing value. At this point a simple transformation (removal of the pedigree columns and moving of the phenotype column at the end) using the UNIX command line tools will get the file ready for BRL.

**Prediction:** The prediction phase consists of training and testing of the model. During training the rules are learned on the complete dataset using rule-based classifiers and in the next phase they are tested with 5-fold cross validation. Various settings for classifier-parameters can be used and the cross validation can suggest which parameter settings for the classifier would result in more general models.

Rule Learner (RL), a knowledge-based rule induction tool is one such classifier used to learn a rule-model. The Global (GBRL) and Local Bayesian Rule Learner (LBRL), the other tools used in our analysis, induce a Bayesian
network from a given dataset. Bayesian rules that predict a categorical target are inferred from this network. The Global Rule Learner generates a parsimonious rule model, which examines all combinatory values of the conditional probability table of the Bayesian network. The Local Rule Learner used here is a generalization of the Global Rule Learner that finds a decision tree representation of the Bayesian network.

Results

The results from the 5-fold cross validation from three types of rule-based classifiers in the BRL+ system are presented in Table 1. RL generates 16 rules comprising of 14 variables representing SNPs. GBRL generates 48 rules selecting 3 variables. Out of the 3 variables, GBRL has only one variable in common with RL. LBRL generates 30 rules with 4 variables. LBRL has only one variable in common with RL and no variable in common with GBRL.

Table 1. Performance measures across 5-fold cross-validation by the three rule-based classifiers in the BRL+ module - Rule Learner (RL), Global Bayesian Rule Learner and Local Bayesian Rule Learner - Decision Tree (DT). The average and standard error values are computed over the 5-fold cross-validation

| Algorithm                | Accuracy† (%) | Balanced Accuracy‡ (%) | Sensitivity□ (%) | Specificity□ (%) | No. of variables selected§ |
|--------------------------|---------------|------------------------|-----------------|-----------------|-----------------------------|
| Rule Learner (RL)        | 90.91 (incl. abstentions*) | 86.67 +/- 9.71 | 100 +/- 0.0 | 73.33 +/- 19.44 | 14                          |
| Global Bayesian RL       | 80.30         | 70 +/- 6.45           | 90 +/- 0.0      | 50 +/- 12.91    | 3                           |
| Local Bayesian RL- DT    | 83.33         | 80.67 +/- 9.11        | 88 +/- 3.74     | 73.33 +/- 19.44 | 4                           |

Accuracy†: determined over confusion-matrix obtained at the end of the 5-fold cross-validation; balanced accuracy‡: average sensitivity and specificity with standard error; sensitivity□: average sensitivity and standard error for predicting the “poor survival” class correctly; specificity□: average specificity and standard error for predicting the “good survival” class correctly; number of variables selected§: total number of attributes that compose the rule model; *abstentions: Rule Learner abstains from classifying an instance when it fails to match any of the learnt rules. Rule Learner treats this as incorrect classification.

Discussion

For the downstream analysis, out of the 100 SNPs with the lowest p-values from the genotypic test for association in PLINK, we mapped the 44 intragenic SNPs to 33 genes. Functional analysis of this gene list was performed through the use of Ingenuity Pathway Analysis (IPA, Ingenuity® Systems, www.ingenuity.com) and indicated cancer as the most significantly associated disease for 9 of the 33 genes and cell-to-cell signaling and interaction as the most significantly associated biological function for 8 of the 33 genes. Specific examples include the CHODL (chondrolectin) gene, associated with shorter survival of patients with NSCLC, the CDH13 (cadherin 13) gene, which is hypermethylated in NSCLC and the CHST11 (carbohydrate (chondroitin 4) sulfotransferase 11) gene, associated with lung colonization from breast cancer.

Table 2. Functional analysis of the 33 genes mapped from the 100 SNPs of the feature selection stage

| Diseases and Disorders               | p-value                  | No. of Genes |
|-------------------------------------|--------------------------|--------------|
| Cancer                              | 6.75x10^-4 - 4.27x10^-2  | 9            |
| Inflammatory Response               | 6.75x10^-4 - 4.43x10^-2  | 6            |
| Cardiovascular Disease              | 1.61x10^-3 - 4.74x10^-2  | 5            |
| Molecular and Cellular Functions    | p-value                  | No. of Genes |
| Cell-To-Cell Signaling and Interaction | 2.49E-05 - 4.74x10^-2   | 8            |
| Cell Morphology                     | 1.36x10^-4 - 4.58x10^-2  | 10           |
| Cellular Movement                   | 3.59x10^-4 - 4.89x10^-2  | 10           |

The good performance observed over cross-validation of the rule-based classifiers suggests that under the limitations of the dataset and our assumptions, the model should generalize well. The rules with the high coverage of the training dataset can be examined further for biomarker discovery.
Conclusion

We have applied BRL to predict early stage lung cancer survival from genome-wide SNP data. We presented our empirical workflow, SNPR, which efficiently overcomes the challenges associated with the predictive modeling of such large-dimensional SNP datasets. Limitations of our work include the non-standard treatment of survival and not including as covariates clinical variables (e.g., age) known to significantly affect survival. This case study presents a specific workflow to a commonly present challenge in various laboratories researching biomarkers to assist in early detection, monitoring and prognosis of disease. However, our workflow can be easily generalized and adapted to other experimental platforms and data mining tasks needed for biomarker discovery applications. Particularly, our results show that it is a feasible and efficient workflow while also yielding SNP-models with high predictive accuracy for lung cancer survival.

Acknowledgments: The authors wish to thank the Cancer Biomarkers Facility of the University of Pittsburgh Cancer Institute for sample processing that are supported in part by award P30CA047904.

Grant support: The research reported in this publication was supported by the following grants from the National Institutes of Health: National Cancer Institute Award Number P50CA090440, National Library of Medicine Award Number R01LM010950, and National Institute of General Medical Sciences Award Number R01GM100387. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Purcell S, Neale B, Todd-Brown K et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559-75.
2. Gopalakrishnan V, Lustgarten J, Visweswaran S, Cooper GF. Bayesian rule learning for biomedical data mining. Bioinformatics. 2010;26:668–75.
3. Bigbee WL, Gopalakrishnan V, Weissfeld JL et al. A Multiplexed Serum Biomarker Immunoassay Panel Discriminates Clinical Lung Cancer Patients from High-Risk Individuals Found to be Cancer-Free by CT Screening. J Thorac Oncol 2012, 7:698-708.
4. Ryberg H, An J, Darko S et al. Discovery and verification of amyotrophic lateral sclerosis biomarkers by proteomics. Muscle Nerve. 2010;42:104-111.
5. American Cancer Society. Cancer Facts and Figures 2012. Atlanta, GA: American Cancer Society; 2012.
6. Zhu C, Ding K, Strumpf D et al. Prognostic and predictive gene signature for adjuvant chemotherapy in resected non-small-cell lung cancer. J Clin Oncol. 2010;28:4417-24.
7. Huang Y, Heist RS, Chirieac LR et al. Genome-wide analysis of survival in early-stage non-small-cell lung cancer. J Clin Oncol. 2009;27:2660-67.
8. Lu Y, Govindan R, Wang L et al. MicroRNA profiling and prediction of recurrence/relapse-free survival in stage I lung cancer. Carcinogenesis. 2012;33:1046-54.
9. de Fraipont F, Levallet G, Creveuil C et al. An apoptosis methylation prognostic signature for early lung cancer in the IFCT-0002 trial. Clin Cancer Res. 2012;18:2976-86.
10. Clearwater S, Provost F. RL4: A tool for knowledge-based induction. In Proceedings of the Second International IEEE Conference on Tools for Artificial Intelligence, IEEE CS Press, 1990:24–30.
11. Chickering DM, Heckerman D, Meek C: A Bayesian Approach to Learning Bayesian Networks with Local Structure. In: UAI ’97. 1997:80-9.
12. Lustgarten, JL. A Bayesian rule generation framework for ‘Omic’ biomedical data analysis. PhD Dissertation, Pittsburgh: University of Pittsburgh; 2009 Available from: http://etd.library.pitt.edu/ETD/available/etd-03062009-175216/