Classification and Regression Analysis of Lung Tumors from Multi-level Gene Expression Data

Pratheeba Jeyananthan
School of Electronics and Computer Science,
University of Southampton,
Southampton, UK
pj2u16@soton.ac.uk

Mahesan Niranjan
School of Electronics and Computer Science,
University of Southampton,
Southampton, UK
mn@ecs.soton.ac.uk

***Abstract*—We study classification and regression problems in lung tumours where high throughput gene expression is measured at multiple levels: epi-genetics, trancription and protein. We uncover the correlates of smoking and gender-specificity in lung tumors. Different genes are indicative of smoking levels, gender and survival rates at these different levels. We also carry out an integrative anaysis, by feature selection from the pool of all three levels of features. Our results show that the epi-genetic information in DNA methylation is a better marker for smoking status than gene expression either at the transcript or protein levels. Further, surprisingly, integrative anlysis using multi-level gene expression offers no significant advantage over the individual levels in the classification and survival prediction problems considered.**

***Index Terms*—Lung cancer, Survival prediction, Smoking mutated genes, Multi-omic, Gender specific genes, TCGA data repository**

I. Introduction

Cancer is a complex disease formed by heritable and environmental factors. Understanding of it causes and dynamics at the molecular level is crucial for early diagnostics and treatment planning because it is known that drug response, for example, widely varies across sub-populations of cancer sufferers. Hence molecular level stratification of patients is vital for knowing which drug will have good response from which group of patients. Advances in high throughput functional genomics, in which gene sequences, their epi-genetic modifications, expressions, and relative abundance of their products (i.e. proteins) are measured at a genomic level (i.e. throughout the genome), coupled with machine learning methods, is seen as an appealing avenue to pursue in the development of a better understanding of the disease and its effective treatment. Among the various cancers, lung cancer is a highly prevalent one with a specific environmental factor – smoking – being its main cause. Over 85% of incidence of lung cancer relate to smoking, and is the subject of this study.

High throughput measurement of gene expression at various levels (gene, epi-genetic, protein etc.) produce representations of a tumor sample in very high dimensions (methylation at $27,000$ sites, $20,000$ transcripts, $7,000$ proteins etc.). Two important issues arise when we attempt to apply statistical inference or machine learning methods on such data. Firstly, the number of patients on whom such measurements can be made is often much smaller, of the order of a few tens or hundreds. This leads to the problem often known a $n << p$ problem (*i.e.* number of samples far fewer than dimensionality). In this setting, any inference method, those that include density estimation in particular, suffer the effects of the *curse of dimensionality*. Techniques such as feature selection, feature reduction by subspace projections or regularisation have to be carefully applied to deal with this issue. Secondly, information contained in measurements taken at different levels of gene expression is often not the same, due to various types of regulatory mechanisms actin in cells. For example, genes that are transcribed need not all be translated into protein. Hence we will observe their expression at the level of the transcriptome, though they could have very different cellular function at the protein level. Thus, integrative analysis of measurements taken at different levels is of importance.

Several transcriptome-based studies have been reported in the literature, including the irreversible effect of tobacco [1], identifying differentially expressed genes between smokers and non-smokers [6], [7], [22] and differentiating smokers from non-smokers [20] or current smokers from others [24]. Methylation data has been used in studies to show how smoking could be identified using single [3], [5] or multiple methylation sites [35], and to infer the effects of maternal smoking on new-borns [27]. Methylation data was analysed between former and current smokers to see how methylation changes by smoking [36] and its residuals has also been suggested as a good marker of smoking in the past [30].

Relating lung cancer with smoking shows that smoking related methylations identified by [3], [5] have most association with lung cancer [14], [38]. This property of methylation was identified in transcriptome data as well [31].

Another area that has attracted attention is prediction of survival from high throughput genomic data. A wide range of algorithms including Bayesian ensemble method [4], PCA [10] and Wavelet based gene selection [13] for gene selection from trnscriptome data. Similarly survival prediction has been attempted from trancriptome and proteome data [8], [17], [25], [29], [32], [37].Interestingly, [33] report a study that links RNA degradation to survival in non-small cell lung cancer patients.

Gender specificity of gene expression in response to smok-