Vertical data integration for melanoma prognosis

Kaushala Jayawardana*
School of Mathematics and Statistics, University of Sydney, Australia
kaushala@maths.usyd,edu.au

Samuel Mueller
School of Mathematics and Statistics, University of Sydney, Australia
samuel.mueller@sydney.edu.au

Sarah-Jane Schramm
The University of Sydney at Westmead Millennium Institute, Australia
Melanoma Institute Australia, Australia
ssch2971@uni.sydney.edu.au

Graham J. Mann
The University of Sydney at Westmead Millennium Institute, Australia
Melanoma Institute Australia, Australia
gmann@usyd.edu.au

Jean Yang
School of Mathematics and Statistics, University of Sydney, Australia
jean.yang@sydney.edu.au

In this paper we outline the integration of clinical and omics data, to improve the prognosis capabilities of a predictive model. Traditionally, clinical data alone has been used to predict a disease outcome. However, with the generation of complex datasets from high-throughput biotechnologies, the interest of researchers has been focused on utilizing these data and the vast level of information they provide, to improve the prognosis of disease outcome. Integrating the components from different platforms has become a crucial step to better understand the relationships between clinical and omics data and the information they provide to explain/predict some response. It is an open question how to best combine different types of variables, as the large dimension of omics data can completely dominate the modelling procedure. We use clinical data from stage III melanoma patients, in a framework which combines bootstrap sampling to account for stability and multiple imputation to account for missingness in clinical data (B-MI), to produce a model with good predictive properties in unraveling the biomarkers in stage III melanoma. We exploit the availability of other high-throughput omics data on the same set of patients, more specifically, gene expression data, protein data and microRNA data, to explore methods in integrating omics data, with special focus on lasso based methods. Such an integration aims to add another dimension in understanding the predictive power of biomarkers in critical diseases like melanoma.

Keywords: omics data, prognostic model, lasso