Getting the Most From Molecular Cancer Data using Unsupervised Feature Learning

Research Objectives

- To use unsupervised feature learning algorithms to extract prognostic signatures from molecular cancer data (gene expression, copy-number variation)
- To use the signatures to train supervised learning algorithms to predict patient survival time
- To perform cross-validation and test set validation to establish the best strategies for the extraction of prognostic features in breast cancer molecular data

Why is the project interesting?

Huge volumes of molecular data are now being generated for the study of cancer. These data measure the underlying molecular machinery in tumours and hold the potential to revolutionise both treatment and our understanding of the disease.

Key to the effective use of these data is the development of methods for getting the most information from them. This is often highly challenging, as molecular data are typically noisy, high-dimensional measurements of complex underlying biological processes. We therefore need to identify the best cutting-edge tools with which to handle these data.

Unsupervised feature learning is ideal in this context. Molecular data types typically contain gene- or genomic-locus-level information about biological networks/pathways. Extracting a relative small number of key biological signatures (features) is therefore likely to be an effective general strategy. These features can then be used for many scientific and medical analyses, for example the prediction of prognostically useful quantities such as the likely survival time for a given patient.

Methods

- Unsupervised learning algorithms such as Sparse Principal Component Analysis, Sparse Filtering, K-means, Independent Component Analysis, deep learning methods
- Supervised learning methods for predicting survival time, such as the Cox proportional hazards model, GLMNET, Random Survival Forests, and Generalised Boosting Models
- Cross-validation and test-set validation
- Metrics will be concordance index, root-mean-square-error (uncensored cases only)
- The data will be gene expression and copy-number variation data sets for 2000 patients from the DREAM/Sage Bionetworks Breast Cancer Prognosis Challenge (http://www.the-dream-project.org/challenges/sage-bionetworks-dream-breast-cancer-prognosis-challenge)

Deliverables

- Quantitative cross-validation comparison of the effectiveness of different unsupervised feature learning algorithms in extracting prognostic signatures from molecular breast cancer data
- Comparison to an existing gene set, generated by extensive mining of the cancer literature, and known to perform well on these data
- Validation of this on a separate, held-out test data set

Who will benefit from this research?

Cancer patients
Clinical oncologists
Cancer researchers working with molecular data

Avenues for a follow-up PhD project

There is huge scope for this work to develop into a full PhD project. This work would focus on the following principal areas, each of which will produce strong, novel research outcomes.

- Development of novel unsupervised learning algorithms, for example using Bayesian nonparametrics (the Indian Buffet process)
- Further investigation of deep learning methods as ways to better extract prognostic signatures from molecular cancer data
- Development of an R software package to apply a range of these methods to molecular data
- Application of these methods to other major genomic cancer data sets, such as those from The Cancer Genome Atlas (TCGA), or the STAMPEDE prostate cancer trial
- Follow-up bioinformatics analysis of the learned features, to identify the likely biological function/s represented by each feature