Breast Cancer Digital Pathology using Unsupervised Feature Learning

Research Objectives

- To automatically identify tumour/normal cell populations in H&E stained breast cancer pathology images, using a range of unsupervised feature learning methods
- To correlate these cell populations with important clinical outcomes such as survival time and metastasis, to show that we can make useful prognostic predictions for individual patients
- (PhD follow-up) Use gene expression data to relate these populations to the underlying biology
- (PhD follow-up) To build predictive models using the learned features, to predict survival time and likelihood of metastasis in breast cancer patients

Why is the project interesting?

Huge numbers of digital images are now being generated routinely in hospital pathology laboratories. Unsupervised machine learning methods do not require label information (which would require detailed attention from an expert pathologist), which means we can learn structure from large numbers of these images. These learned features can then be used to develop predictive biomarkers for the clinical outcome of breast cancer patients, such as survival time, and even effectiveness of different treatments for individual patients. This approach can ultimately be used to combine pathology image information with molecular data such as gene expression and genomic sequences, to give a fully personalised treatment regime for each cancer patient.

Data

The METABRIC study is the largest genomic study of a single epithelial cancer to date. In total 2,000 breast cancer tumours have been molecularly profiled and their pathological images generated. These data are expected to provide important clue to the understanding of breast cancer and patient outcome prediction.

Methods

Unsupervised feature learning methods such as Sparse Principal Component Analysis (PCA), Independent Component Analysis (ICA), K-means clustering, sparse filtering, deep learning.

Deliverables

- Image feature set that compactly distinguishes between the different cell types present in breast cancer H+E stained histopathology images
- A set of distinct cell types, as identified by clustering analysis. These cell types will then be linked back to the known cell types, as identified by expert pathologists and supervised classifiers trained with known cell class labels.
- Correlation of identified cell types with known clinical outcomes such as survival time

Who will benefit from this research?

Cancer patients.
Cancer researchers and clinicians including pathologists and oncologists.

Avenues for a follow-up PhD project

There is significant scope for this work to develop into a full PhD project (and beyond), using the full METABRIC data set, including both pathology images and also extensive genomic and gene expression measurements. There are a range of important scientific and medical goals that can be addressed, including large-scale unsupervised identification of cell types, investigations of organisation of cells in image, and progostic prediction of important clinical outcomes such as survival time and metastasis. There is also significant importance to the integration of these image data with the METABRIC molecular data such as gene expression and copy number variation (genomic) data, both in terms of bioinformatics methodology development and advances in the scientific understanding of breast cancer. Similar methodologies can be applied to other cancer types that are readily available from The Cancer Genome Atlas project.

We anticipate that the methods developed on this work will be turned into an R software package, so that the wider community can benefit from this work. For relevant references, please contact Dr Richard Savage (r.s.savage@warwick.ac.uk)