Arran Schlosberg

Mike and Carole Ralston 2015 Recipient Final Report Summary

With the generous aid of the RCPA Foundation Mike and Carole Ralston Travelling Fellowship, I was able to undertake a year of bioinformatics research at the Wellcome Trust-MRC Institute of Metabolic Science, leading to the award of MPhil from the University of Cambridge. Building on existing investigations of obesity and neurogenetic control of appetite, my work concentrated on a concept known as *dimensionality reduction* with a focus on transcriptomic profiles.

Considering a dummy example of linear regression measuring two attributes for only two specimens, it is trivial to perfectly describe a line through these data points. Extending this to the third dimension, a plane such as a sheet of paper will be defined by three specimens and three measured attributes. This pattern continues such that any model with more measured attributes than there are data points—most certainly the case with a transcriptome of approximately 20,000 genes and only tens to hundreds of specimens—will always have a solution that fully conforms to the empirical data.

Whilst this may appear to be of benefit, such models suffer from overfitting and too closely focus on nuances specific to the particular data set. As such, these models do not perform well in more general cases. Dimensionality reduction aims to cull measured attributes such that only salient features remain for the creation of statistical models. Two broad approaches exist—feature *selection* and feature *extraction*. Across various disciplines, *features*, *dimensions*, and *attributes* are synonymous.

In the domain of feature selection, whereby distinct attributes are chosen, my work demonstrated that an apparently unrelated algorithm in the domain of information theory (Fayyad and Irani; 1993) can be repurposed to rank genes by their ability to discriminate tissues of origin when quantifying respective mRNA expression. Analysing transcriptomes from neuronal nuclei implicated in feeding behaviour (Yeo and Heisler; 2012) and performing an in-depth literature review, I qualitatively demonstrated a favourable positive predictive value with highly ranked genes routinely described in existing literature. Furthermore, the approach exhibited high sensitivity as assessed by its ability to recall key genes (Yeo and Heisler; 2012) in the highest rankings. Computationally efficient, this method finds potential application in biomarker discovery and can similarly be applied to any ordinal -omics or biochemical data.

Feature extraction, a common example being principal component analysis (PCA), aims to discover weighted combinations of attributes that capture characteristics of empirical data most relevant to the statistical analyses at hand. My attempts to impute missing expression data with a form of artificial neural networks known as denoising autoencoders (Vincent et al; 2008, 2010) were unsuccessful. However, my results support the ability to numerically summarise transcriptomes, an approach that has previously been demonstrated to detect clinical correlates in breast cancer (Tan et al; 2014).

Although I will no longer pursue a Fellowship of the College, I will continue as a member of the RCPA Informatics Committee. My decision to leave traditional medical training was based on a desire to undertake cutting-edge research in computational pathology and, as such, I have recently commenced employment with Google where I plan to work at the intersection of clinical medicine and so-called big data.