Background

I am an Australian PhD student in my final year of study working on an interdisciplinary project based between the School of Mathematics and Statistics and the Institute of Biodiversity, Animal Health and Comparative Medicine at the University of Glasgow. My research is on applying a new kind of data analysis, Topological Data Analysis, to the analysis of noisy high-throughput sequencing data coming from fish sampled in the wild. In 2018 I was awarded a grant of £4,250 from the Jim Gatheral Travel Scholarship to fund a four month visit to the University of British Columbia in Vancouver, Canada.

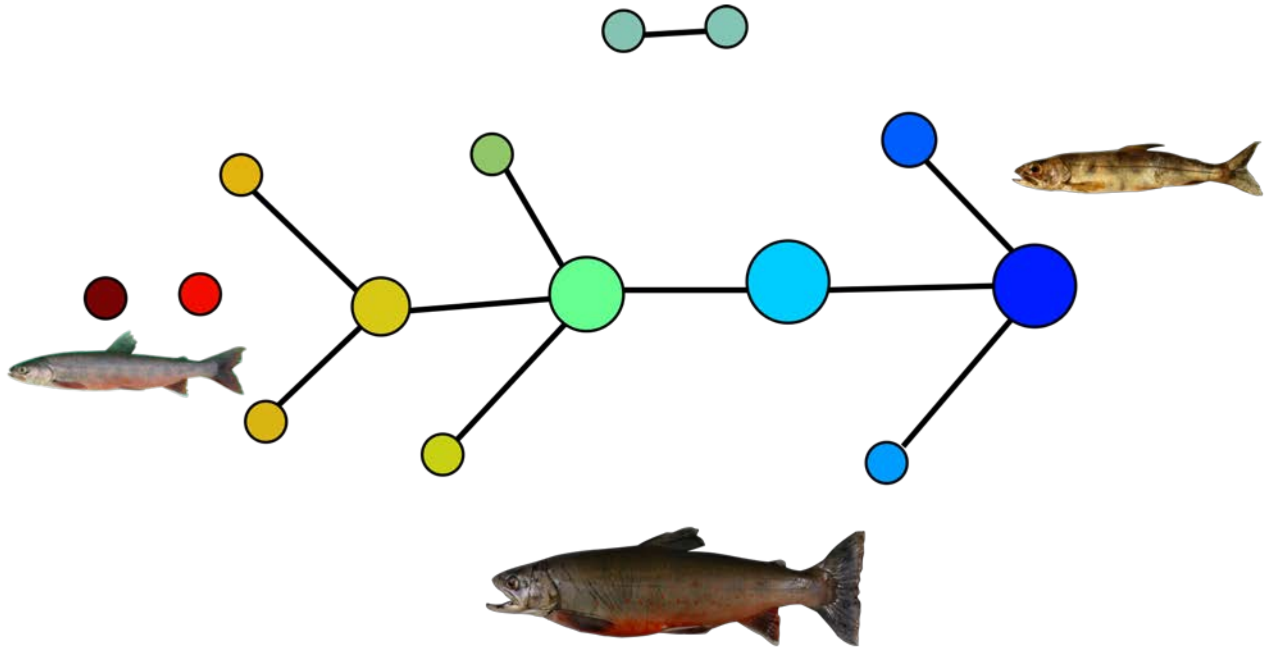
On a walk in Lynne Valley, just northwest of Vancouver

Reasons for applying for the scholarship

My research involves applying ideas from topology – a branch of mathematics that is specifically adapted to treat qualitative properties such as connectivity – to make topological models of gene expression, in order to better study speciation. The motivation is to better analyse the shear volumes of data from new genomics techniques. In particular, with the advent of new generation sequencing, a great deal of gene expression data has become available to work with. My aim is to use topology to model gene expression, visualise the associated data sets, and give new insight into speciation at the genetic level.

My project involves taking ideas from topology which have been applied to humans, for example in finding a new subtype of breast cancer and adapting them for use in other species. The model we are using considers gene expression data sets as points in some high dimensional vector space. (With number of points equal to the number of samples, and number of dimensions equal to the number of genes.) From there, we can visualise the data by adapting a method of Nicolau and Carlsson, called Mapper.

In particular, I have been working on gene expression data from Arctic charr. This is a species of fish found in lakes all around the Arctic circle. It is of interest to us, since it occurs in two different forms (an open-water and a shore) in many lakes. I have been engaged in making a topological model, which can account for the two morphs in the different lakes, as well as deal with the noisiness of the gene expression data we have acquired from the fish.



Using ideas from topology to visualise gene expression data from Arctic charr

A secondary concern has been investigating how we can use topology to find gene sets which are related in gene expression data sets. This can be done by taking into consideration the correlation between the genes, then using a topological method to visualise and find gene sets.

The travel scholarship allowed me to visit Assistant Professor Liam Watson, an expert in topology. Spending time at UBC offered me the chance to attend their seminars, which included a talk by Dr Geoff Schebinger who studies ontogenetic stochastic processes in developmental biology and cellular reprogramming. This allowed me to maximise my exposure to mathematics I would not have encountered at the University of Glasgow.

Details of my visit

I was at UBC from early January to late April 2019. During that time, I wished to work out in greater detail the mathematics underlying my topological model, and how it relates to and is distinct from current techniques used in gene expression analyses (such as weighted gene correlation network analysis (WGCNA)).

Being at UBC allowed me to attend seminars in the supportive and world-renowned environment of UBC Maths. Particularly useful was the fact that UBC maths was hosting job talks at this time. The candidates would each give one seminar and a more general colloquium talk. A number of the candidates gave talks about mathematical and statistical methods they had come up with to analyse gene expression data. One of these individuals was Dr Geoff Schebinger, who gave a talk using optimal transport to model the developmental landscape of stem cells in gene expression space. The results of his high-powered modelling experiment (315,000 single cell RNA-seq expression profiles at 40 time points over 18 days of stem cell reprogramming) and the resulting interactions and discussions with Liam were crucial for coming to a new understanding of the possibilities and limitations of mathematical modelling of gene expression data.

As a result, I managed to come up with a theoretical outline for how to port the idea of optimal transport from a cell developmental context to a developmental biology context. This involved

Other grad students: (L-R) Kim, Sebastian, and Mihai

reworking Schebinger's idea of learning gene regulatory networks by using a cell's expression level at a certain time point to predict the expression level of the cell's descendants in the next time point and applying it instead to predicting the change in morphology of an organism, Arctic charr in my case, as it develops.

A further result of these discussions was more clarity on the role of sample size in the statistical distinctions we could make. More specifically, we proved mathematically that the number of distinct categories which can be found in a dataset with a certain statistical confidence is inversely related to the sample size.

Impact of the travel scholarship

This trip, made possible by the Jim Gatheral Travel Scholarship, has provided the circumstances and ideas for the final chapter of my PhD thesis, allowing me to pull all the strands of my research together.

Personally, I found the opportunity to experience research at another environment invaluable for broadening my horizons and seeing how research is conducted at different universities.

Finally, I would like to thank the University of Glasgow and the Jim Gatheral Travel Scholarship for making this research trip possible, and the Mathematics Department at UBC for hosting me.

Snow on the UBC campus