Chen, Mel (2020) *Convergence, connectivity, and continuity: topological perspectives for mining novel biological information from ‘omics data.* PhD thesis.

https://theses.gla.ac.uk/78978/

Convergence, connectivity, and continuity: Topological perspectives for mining novel biological information from ‘omics data

Mel Chen

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Mathematics and Statistics
College of Science and Engineering
University of Glasgow

January 2020

Abstract

In this thesis, we will explore possible applications of topological data analysis to ‘omics data. More specifically, we apply the topologically-based data visualisation technique, Mapper, to gene expression data coming from the fish, Arctic charr (*Salvelinus alpinus*). The fish samples come from the wild, from lakes in Scotland and Russia. Furthermore, the Arctic charr is an interesting study species, since it commonly occurs in two morphs, a bottom/bank-dwelling benthic morph, and an open-water pelagic morph. In general, these morphs share features which are common across lakes, and so provide an opportunity to study a subspecies-level split which is replicated across different populations. This gives an example of parallelism in evolution, and the fact that the split is replicated allows us to test if there are common underlying changes leading to this split, at the level of identical genes, or sets of genes, or genes involved in the same pathways.

We provide an overview of the Mapper algorithm, and also show its application to a breast cancer gene expression dataset, which was the inspiration for our PhD project. When applying Mapper to the Arctic charr, we also investigate the effect of sample size by subsampling the breast cancer data.

As well as applying Mapper, we also use a more mathematical view of the gene expression data to provide a new perspective for looking at the commonly used gene analysis techniques in evolutionary biology, namely, differential gene expression, and gene co-expression analysis.

Finally, we provide an experiment which could be done in the future, assuming the cost of sequencing continues to fall. This experiment incorporates ideas of optimal transport in trying to reconstruct the developmental landscape of Arctic charr. We also discuss other avenues for future work, and current difficulties with applying topological data analysis to gene expression data from wild samples.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my supervisors Drs Liam Watson and Kathryn Elmer for all their support and advice throughout this PhD. I would also like to thank the University of Glasgow, and especially fellow students and department members of the School of Mathematics and Statistics, and the members of the Elmer lab. Special mention to Drs Arne Jacobs and Madeleine Carruthers, who really helped with the evolutionary biology side of this PhD.

A thank you also to the Lord Kelvin Adam Smith PhD Scholarship from the University of Glasgow, which provided the funding for this PhD. I would not have been able to carry out this work without it.

Finally, a thank you to all the friends, family and colleagues who've helped me in one way or another through this degree.

Declaration

I certify that the thesis presented here for examination for a PhD degree of the University of Glasgow is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it) and that the thesis has not been edited by a third party beyond what is permitted by the University's PGR Code of Practice.

The copyright of this thesis rests with the author. No quotation from it is permitted without full acknowledgement.

I declare that the thesis does not include work forming part of a thesis presented successfully for another degree.

I declare that this thesis has been produced in accordance with the University of Glasgow's Code of Good Practice in Research.

I acknowledge that if any issues are raised regarding good research practice based on review of the thesis, the examination may be postponed pending the outcome of any investigation of the issues.

Chapter 1

Introduction

In this thesis, we bring a topological approach to bear on the analysis of biological data. We have two inspirations for this. Firstly, the reduced cost of sequencing has seen an explosion of ‘omics data in the biological sciences. Secondly, topological data analysis (TDA), a new field applying topology to the analysis of data, has seen some recent success analysing ‘omics data from human breast cancer samples [NLC11; Cám16]. The work of this thesis will be to port these ideas from the lab to the wild.

Data from the lab usually come from species such as humans (*Homo sapiens*), mice (*Mus muscus*), zebrafish (*Danio rerio*), etc., which are part of a vast infrastructure, including extensive annotations of gene functions, as well as technical resources such as cell-lines and lab populations of various clones. These allow for more controlled experiments and result in cleaner data.

By contrast, so-called field biologists are interested in organisms from the wild. This invariably involves non-model species, with no control in either the genetic or environmental background. In particular, this means that we are dealing with noisy data. An example could be fish collected from the wild, of which we do not know the rearing or environmental histories. A further restriction when dealing with non-model species is a lack of supporting infrastructure. For example, the annotation of gene functions is not as extensive, if it exists at all, and we lack cell-lines and lab-based populations.

Recent advances in sequencing technology, and the resulting reduction in cost, have led to the widespread use of transcriptomics (mRNA sequencing) not only in model species, but also increasingly in non-model species in the context of evolutionary analysis. This has allowed evolutionary biologists to access a mass of data, such as the expression levels of tens of thousands of genes in their study systems of interest.

RNA-Seq (RNA sequencing) is a technique which was developed in the late 00’s, and allows us to measure the RNA expression levels in tissues [WGS09]. The RNA (gene) expression levels are of interest to us, since they are related to the rate of protein synthesis in cells, and some also have regulatory functions. The quantification of RNA in the cell gives us access to a molecular phenotype, in between the genotype of an organism, determined by its DNA, and its phenotype,[1] determined by its genotype and environment. The RNA expression levels vary between tissues and in different developmental stages and physiological conditions, so having information about this will give us a better idea of how the phenotype arises from an organism's genotype.

Topological Data Analysis (TDA) is a new field of study which has seen recent success in application to noisy datasets [Car09]. A few properties which make topology suited for data analysis is its insensitivity to metric; that is, topology can work with a vague sense of 'nearness' (open sets) rather than with one defined by a particular distance. In particular, it is possible to make choices so that the analysis is robust to noise. Another aspect is that TDA can be used with a notion of distance sensitive to a given parameter, and one can summarise the analysis over different values of this parameter visually. This gives hope for applying it to RNA-Seq datasets, which are noisy and high-dimensional, and finding important features in a low-dimensional representation that accurately summarises the data.

We are inspired by recent applications of TDA to model species, particularly its application to a breast cancer data set which resulted in finding a new subtype of breast cancer [NLC11; Lum+13]. Nicolau *et al.* used Mapper, a data visualisation algorithm based on ideas from topology, on gene expression data in the form of 25,000 gene microarrays coming from 295 breast cancer tumour samples and 13 normal breast tissue samples. After some data processing, the Mapper graph visualisation found a cluster of cancer samples coming from patients with a 100% survival rate, and further statistical analyses showed that this cluster is biologically distinct and was not previously known.

We seek to bring the techniques applied in this context of a model species into the context of evolutionary analysis on a non-model organism. The ultimate goal is to provide a method applicable to other cases of replicated evolution of parallel morphs in non-model species, such as the benthic and limnetic cichlids [Elm+14]. The data we will apply these methods to are 'omics data, including gene expression data, with about the same dimensions (number of genes) as can be found in model species. We hope that TDA will be able to provide new insights into the data, as was the case with the breast cancer example. For example, we may be able to discover more intricate structure, such as an unexpected subgroup, in the Arctic charr data, and isolate a set of genes which is responsible for this structure.

Of course, shifting contexts is not without difficulty. We will be contending with smaller sample sizes, a noisier data set, and a different experimental context. Furthermore, we will be lacking many of the details of the biological function of genes which model species possess. We will write about encountering these challenges, our attempts to deal with them, and how future work could help.

In brief, we first examined the literature on topological data analysis [Car09] in detail, and more specifically the Mapper algorithm [SMC07] and its application to a breast cancer gene expression data set [NLC11]. After this, we investigated our gene

[1] physical, measurable characteristics, such as colour and body shape

expression data set from Artic charr using some standard visualisation techniques, principal component analysis and heatmaps. After this testing from first principles, we reapplied Mapper to a new context with a new non-model organism in an ecological context. This has not been done before with this sort of approach. Unfortunately, unlike the breast cancer example, we find the Mapper algorithm lacking in several regards. In particular, we were unable to find a new subgroup of interest, or even to visualise the benthic and pelagic morphs as subgroups, even though we know the morphs have differing gene expression from previous work [Jac+19].

This set-back caused us to step back and think how else we could better visualise the gene expression data. After exploring a range of techniques to identify their mathematical basis and their potential for relevance to evolutionary context, we finally come up with a method of deforming the gene expression space, which simplifies down to weighting genes. The idea is then to get a weighting of the genes so that, when we visualise the samples under this weighting, the subgroups we are expecting are apparent. We show that this works for a weighting of genes which comes from differential gene expression analysis (DGEA). We also compare this new deformation method with more traditional methods, DGEA, and gene co-expression networks. We find out what parts of this analysis can be reproduced in our new method, and which cannot.

Additionally, we test our method on a larger set of gene expression samples from fruit flies (*Drosophila*) with a larger sample size of 726 compared to our 32 for the Arctic charr. We find that, even with such an increase in sample size, we are still unable to get a clear signal of subgroup compared to the biological noise. We conclude by offering a discussion of a future experiment, which could be done as the cost of sequencing further falls. This synthesises recent work of Schiebinger [Sch+17] with the understanding we have gained of Arctic charr and the development of parallel morphs.

1.1 How to Read this Thesis

This thesis will detail the process of applying Mapper and another data analysis technique inspired by topology to a gene expression data set of salmonid fish (Arctic charr) from lakes in Scotland and Russia.

Since this is an interdisiplinary project, the thesis will be structured around being readable for both biologists and mathematicians. We will begin with two chapters going into the background, one for the maths, and one for the biology. We then move on to two chapters detailing the application of TDA to the Arctic charr data set, and we conclude with a discussion about our findings and potential for future work.

The second chapter will give a literature review of TDA, focusing especially on Mapper. It will also give an example of Mapper applied to a breast cancer data set by Nicolau *et al.* [NLC11]. For mathematicians, there is a section on the workings of Mapper, including some definitions and background to understand the visualisation technique. For the biologists, there is a section describing Mapper with some examples.

The details of Mapper applied to breast cancer will be readable for both audiences.

The third chapter will give a description of the study system, including why Arctic charr are an interesting study species from an evolutionary analysis point of view. We will also include a section for mathematicians, where we treat the study system abstractly, treating the gene expression data from the fish as simply random variables depending on the lake and morph. We will also include some analysis of the data from a standard point of view, using differential gene expression and principal component analysis.

The fourth chapter will show Mapper applied to the Arctic charr data, as well as some investigations of the effect of sample size on Mapper's ability to find novel subgroups. All sections of this chapter are intended to be readable for both mathematicians and biologists.

The fifth chapter will go into the background of our topological perspective for transcriptomics data. The section involving mathematical definitions will primarily be for the mathematicians. We will give an example illustrating the approach for both audiences, by applying it to the Arctic charr data. Additionally, we compare and contrast our approach to differential gene expression analysis and gene co-expression analysis, which are commonly used in evolutionary biology to analyse gene expression data. Finally, we assess the viability of this approach on a Drosophila data set with a larger sample size of about seven hundred.

The sixth chapter and concluding chapter will be a discussion of our attempt to port methods from the lab to the wild, and include a hypothetical experiment which could be conducted to address some of the shortcomings we have found.

Figure 1.1 gives a schematic for the above.

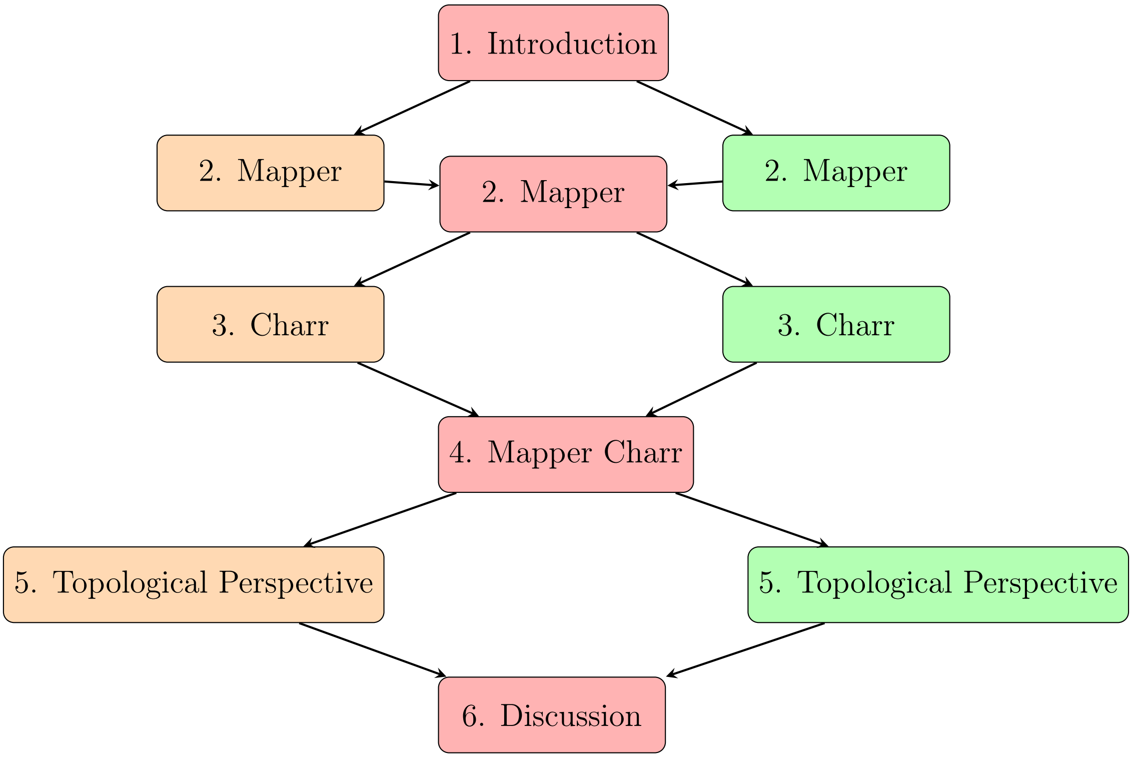


Figure 1.1: How to read this thesis. The chapters are named from the Introduction at the top, to the Discussion at the bottom. Chapters with parts only for mathematicians are to the left in orange, and those with parts only for biologists are to the right in green. Chapters readable for both in their entirety are in the middle, in red. The second chapter, on Mapper, has sections introducing Mapper for both mathematicians and biologists, while the description of its use in [NLC11] finding a new kind of breast cancer aims to be readable for both audiences.

Chapter 2

Mapper and Medical Data

There are two main goals of this chapter. First, we dive into the details of the Mapper algorithm [SMC07]. This will involve writing about its motivation, the mathematical background, and its implementation. Second, we go through an application of Mapper to a breast cancer data set; this application of Mapper motivated this PhD project. The results were new insights into the genetics of breast cancer. In particular, a new subtype of breast cancer with zero percent mortality was discovered. In fact, we will discuss two applications of Mapper to the breast cancer data set, an earlier one by Nicolau *et al.* [NLC11] and a later one featuring in a paper by Lum *et al.* [Lum+13]. Going through this background will set the scene for us to apply Mapper to wild Arctic charr transcriptomic data in Chapter 4.

The biologist may skip §2.1.2 and read from §2.1.4, where we describe the inputs and outputs of the Mapper algorithm and ends with a toy example of Mapper applied to a point cloud sampled from a hand in 3D.

2.1 Mapper

2.1.1 Motivation

Mapper is an algorithm for visualising high dimensional data sets, introduced by Gurjeet Singh, Facundo Mémoli, and Gunnar Carlsson in [SMC07].[1] The motivation for Mapper was to visualise data in low-dimensional spaces, such as the plane ($\mathbb{R}^2$) and 3-space ($\mathbb{R}^3$), in order to make use of people's ability to find patterns by eye in low dimensions.[2] In fact, one can consider Mapper as a combination of two commonly used data visualisation techniques. The first is *projection pursuit* [Hub85], which is a dimension reduction technique that chooses linear projections optimising an objective function, and *principal component analysis* is an instance of this where the objective function is to maximise variance. The second is *clustering*, of which hierarchical clus-

[1] A more mathematically detailed description can be found in §3 of [Car09] , although there Carlsson does not go as much into the implementation of the algorithm or examples.

[2] Gunnar Carlsson mentions this in one of the talks he gave at the 2015 Young Topologists' Meeting. They are available online at: `https://www.epfl.ch/labs/hessbellwald-lab/seminar/ytm2015/`

tering is the most common example. Additionally, a method was desired for finding patterns in the shape of the data, which needed to satisfy the following properties:

Insensitivity to metric. Any such method should work well with different measures of similarity. For instance, we can consider the various measures for distances between gene expression profiles, sequences or more general datasets (e.g. Euclidean, Pearson correlation, etc.). Ideally, we'd want the method to find the same pattern in the data for similar metrics. This means the method should also display "invariance" under "small" deformations and be "coordinate free". This ensures that whatever patterns the method finds will be robust.

Understanding sensitivity to parameter changes. Since many algorithms require an arbitrary choice of parameters before producing a result, it would be useful to have some way to summarise behaviour under all possible choices of parameters. For example, we may want to summarise all parameter choices when using single-linkage clustering, or some other clustering algorithm.

Multiscale representation. To be able to visualise a point cloud at different levels of resolution can be useful for finding structures at various scales, or drawing attention to certain features if they are present over a range of scales, since we can reason that such features are less likely to be artefacts of the data. This also ensures the features we find are more qualitative.

These points can be addressed by adjusting parameters in Mapper. This allows us to ignore the exact distance used.[3] Finally, Mapper provides an easy-to-understand visualisation of the data produced in the form of a graph. We will address all these points in more detail later in the chapter. For now, we move on to the topological inspirations behind Mapper.

2.1.2 Topological Inspirations

In this subsection we will describe the topological notions that Mapper is based on. The interested reader may refer to [Car09, §3.2] for more details. The essential notion is the nerve theorem, Theorem 2.1.13, which gives conditions under which a topological space may be simplified into a simplicial complex. We now proceed to give some relevant background, then describe how these ideas are used in Mapper.

Definitions

Here we will give a few definitions and theorems that will be used in the rest of this subsection. Those who want more background or details can refer to [Mun00] and [Hat02].

[3] However, the extent to which this applies will depend on how robust the features we're looking for are. That is, if we find a feature over more dissimilar distances, then we can consider it a robust one.

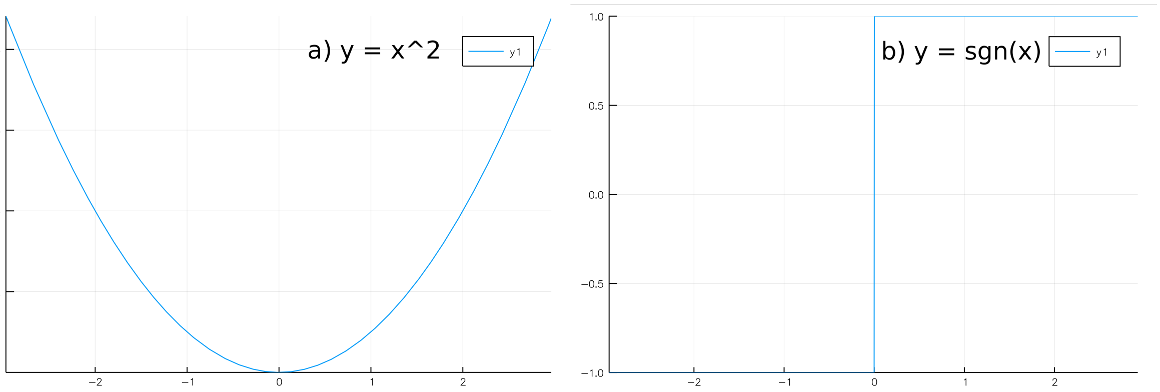


Figure 2.1: Example of functions. **a)** Is a continuous function, $y = x^2$. **b)** Is a discontinuous function, $y = \operatorname{sgn}(x)$, which has a discontinuity at $x = 0$ where it jumps from -1 to 1.

Definition 2.1.1 (Topological Space)**.** A *topological space* is an ordered pair (X, τ) where X is a set and τ is a collection of subsets of X having the following properties:

1. $\emptyset$ and X are in τ.

2. Any union of subcollections of τ is in τ.

3. The intersection of any finite subcollection of τ is in τ.

The subsets in τ are called *open sets*, and τ is called a *topology* on X.

Examples of common scientifically interesting topological spaces are subspaces of $\mathbb{R}^n$, which include point clouds, lines, circles, spheres, and tori.

Definition 2.1.2 (Continuous Function)**.** Let X and Y be topological spaces. A function $f : X \to Y$ is *continuous* if for each open subset $U \subset Y$, $f^{-1}(U)$ is an open subset of X.

In less technical terms, a continuous function $f : X \to Y$ takes points p_1, p_2 which are close in X to points $f(p_1), f(p_2)$ which are close in Y. Figure 2.1 shows examples of functions $\mathbb{R} \to \mathbb{R}$, continuous on the left and discontinuous on the right.

Definition 2.1.3 (Homeomorphism)**.** Let X and Y be topological spaces. We say a function $f : X \to Y$ is a *homeomorphism* if it is a bijection and both it and its inverse f^{-1} are continuous.

If a homeomorphism exists between two topological spaces X and Y, then we say they are *homeomorphic*. In topology, we study spaces up to homeomorphism, that is, properties which remain the same between spaces when they are stretched or deformed without tearing or gluing. A famous example is that a topologist cannot tell the difference between a coffee cup and a donut; the deformation is given by the joke shown in Figure 2.2.

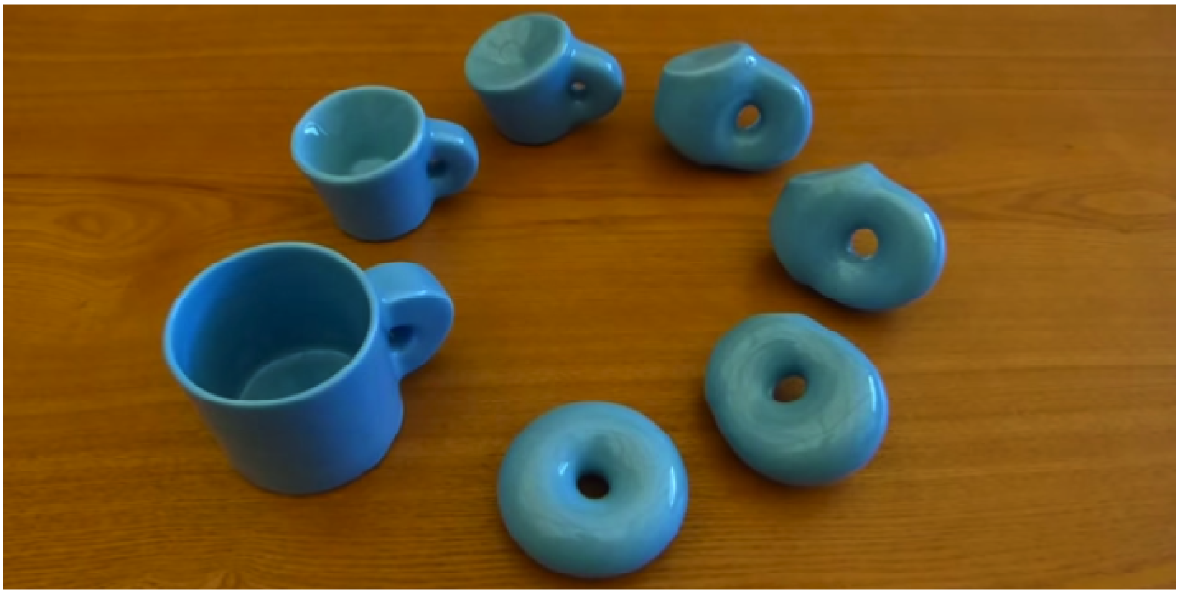


Figure 2.2: This figure suggests how a coffee cup, in the bottom left, can be smoothly deformed, clockwise, into a donut shape, at the bottom. Image taken from a YouTube video by Henry Segerman, and found in a shapeways article: `https://www.shapeways.com/blog/archives/21752-a-3d-printed-topology-joke.html`

Definition 2.1.4 (Homotopy)**.** Let X and Y be topological spaces, and $f_t : X \to Y, t \in [0,1]$ a family of maps. If the function $F : X \times [0,1] \to Y$ given by $F(x,t) = f_t(x)$ is continuous, then we say f_t is a *homotopy.*

If there exists a homotopy connecting two maps $f_0, f_1 : X \to Y$, then we say that f_0 and f_1 are *homotopic*, and we can write $f_0 \simeq f_1$.

Definition 2.1.5 (Homotopy Equivalence)**.** Let X and Y be topological spaces, and $f : X \to Y$ a continuous function. We call f a *homotopy equivalence* if there exists a continuous function $g : Y \to X$ such that $g \circ f \simeq \mathrm{id}_X$ and $f \circ g \simeq \mathrm{id}_Y$.

If there is a homotopy equivalence $f : X \to Y$, then we say that the spaces X and Y are *homotopy equivalent* and we can write $X \simeq Y$. Homotopy equivalence is a less strict condition than homeomorphism, in that if X and Y are homeomorphic, then they are also homotopy equivalent, but not necessarily the other way around. For example, a point $\{*\}$ and a disc D are homotopy equivalent, but not homeomorphic. To speak more illustratively, if homeomorphism allows continuous deformations, then homotopy equivalence also allows us to compress or expand regions to/from a point. This gives rise to the notion of *contractibility*:

Definition 2.1.6 (Contractible)**.** A topological space X is *contractible* if it is homotopy equivalent to a point $\{*\}$.

The vast majority of topological spaces, such as tori, spheres, circles and Euclidean space, consist of infinitely many points. This makes them impossible to represent combinatorially, which we must do in order to run calculations on them with computers.

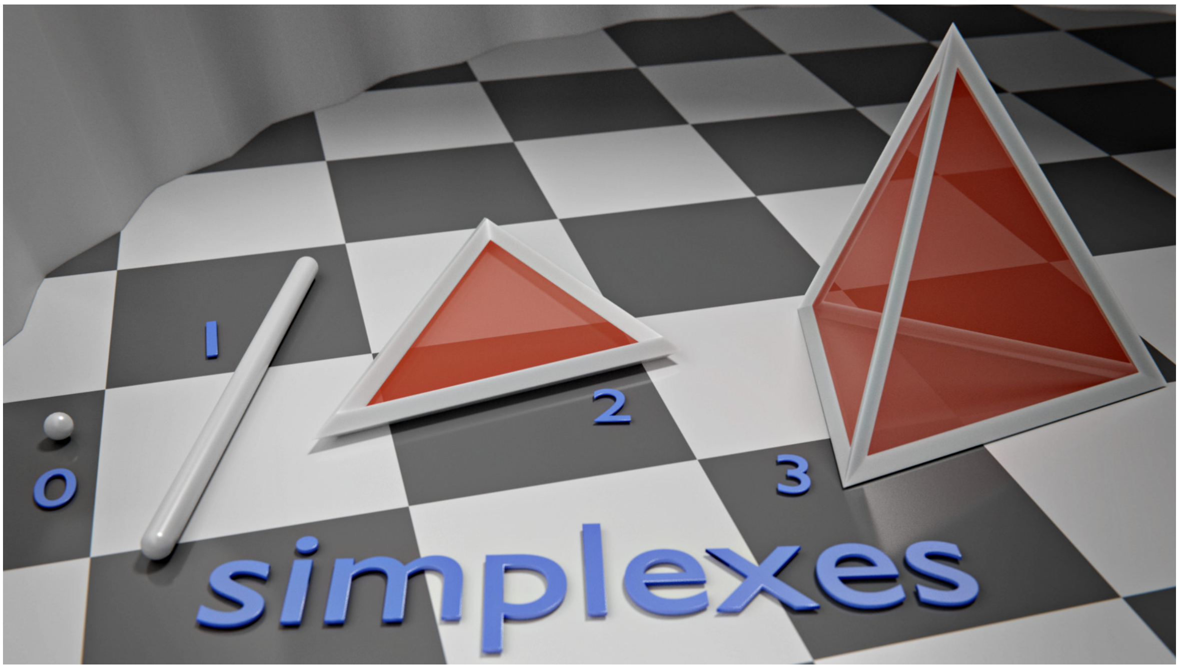


Figure 2.3: The standard n-simplices which can be represented in 3 dimensions. Namely, the $0, 1, 2, 3$-simplices, corresponding to a point, line, triangle and tetrahedron. Image from: `https://commons.wikimedia.org/wiki/File:Simplexes.jpg`

We can solve this by introducing discrete representations of spaces, called simplicial complexes. We will now give a few definitions regarding these objects.

Definition 2.1.7 (Abstract Simplicial Complex)**.** Let S be a set. An *abstract simplicial complex* is a collection K of non-empty finite subsets of S such that, if $X \in K$ and $Y \subset X$, then $Y \in K$.

On computers, S will be a finite set, and the collection of subsets allows us to give a purely combinatorial description a topological space. There is a construction called the *geometric realisation*, which allows us to associate a topological space $|K|$ to an abstract simplicial complex K. However, we will not give details of it here, and instead use the following examples to illustrate the relation between abstract simplicial complexes and topological spaces.

Example 2.1.8 (Standard n-Simplex)**.** Let S be a set with $n + 1$ elements, where $n \in \mathbb{Z}_{\geq 0}$, and K the collection containing every subset of S. Then standard n-simplex is the geometric realisation $|K|$ in $\mathbb{R}^{n+1}$ given by the convex hull of the points $\{(1, 0, 0 \ldots, 0), (0, 1, \ldots, 0), \ldots, (0, 0, \ldots, 0, 1)\}$. The four standard n-simplices which can be represented in 3 dimensional space are shown in Figure 2.3.

Example 2.1.9 (Triangle)**.** Consider the set $\{0, 1, 2\}$ and let K be the collection of all its subsets. Its geometric realisation in $\mathbb{R}^3$ is given by the convex hull of the points $\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$. See Figure 2.4 for a plot of this particular triangle.

One can consider abstract simplicial complexes as a generalisation of triangulation. For example, just as we can triangulate a hexagon in 2D, by using six triangles which meet at the centre, we can divide a solid sphere into tetrahedra in 3D.

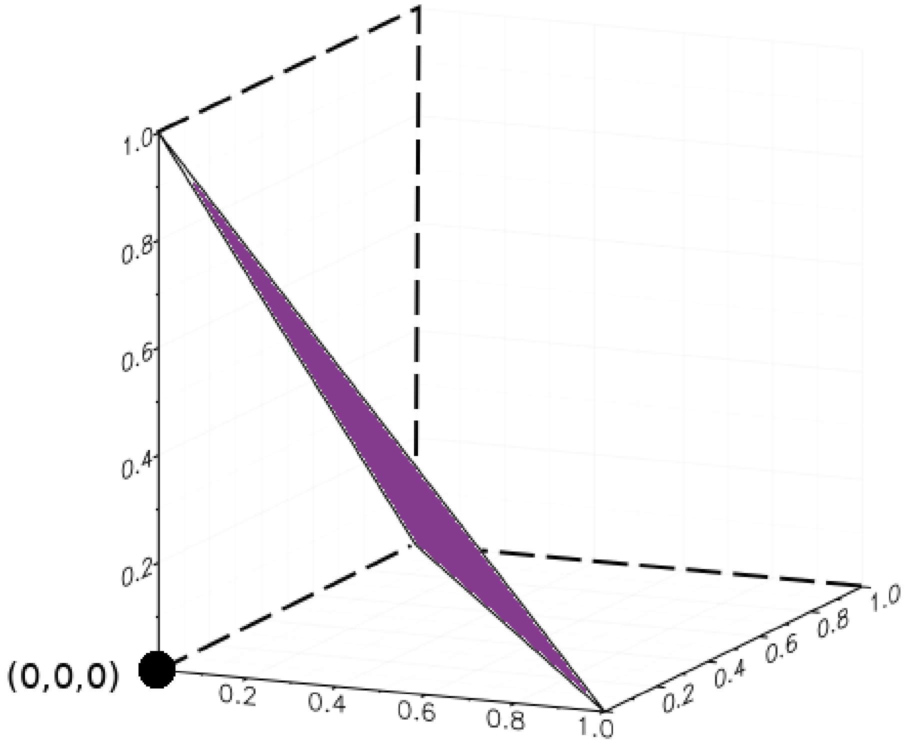


Figure 2.4: The geometric realisation of the standard 2-simplex, namely, the triangle as a subset of $\mathbb{R}^3$.

We can now finally give definitions to describe the nerve, which is a method of associating a simplicial complex to a topological space.

Definition 2.1.10 (Cover). Let X be a topological space and $\mathcal{U}$ a collection of subsets of X. We say that $\mathcal{U}$ *covers*, or is a *covering* for, X if $\bigcup_{U \in \mathcal{U}} U = X$. Furthermore, $\mathcal{U}$ is an *open cover* if all its elements are open.

Definition 2.1.11 (Paracompact). Let X be a topological space. We say that X is a *paracompact* space if every open cover $\mathcal{U}$ has a locally finite open refinement.

The paracompact condition on X guarantees the presence of certain 'nice' properties we would like to have. In particular, it guarantees that any open cover of X admits a partition of unity subordinate to it, and the proof of the Nerve Theorem (Theorem 2.1.13) relies on this. It will never be an issue for us in practice, since all metric spaces (spaces with a distance function $d(-,-)$) are paracompact.

Definition 2.1.12 (Nerve). Let X be a topological space and $\mathcal{U}$ an open cover of X. Then the *nerve* of $\mathcal{U}$ is the simplicial complex $N\mathcal{U}$ which corresponds to a set which has one element v_α for each open set of the open cover $\mathcal{U}_\alpha$, and contains the finite subsets corresponding to open sets of the open cover with non-empty intersections.

The nerve has the topology of its geometric realisation.

Theorem 2.1.13 (Nerve Theorem). *[Hat02, Corollary 4G.3, p. 459] If $\mathcal{U}$ is an open cover of a paracompact space X such that every nonempty intersection of finitely many sets in $\mathcal{U}$ is contractible, then X is homotopy equivalent to the nerve $N\mathcal{U}$.*

The nerve theorem captures the essential topological inspiration behind Mapper. The result is that, if we have an open cover where the intersections are simple enough, i.e. contractible, then the space and its nerve are similar, i.e. homotopy equivalent, topologically speaking.

The Idea of Mapper

The motivation behind Mapper is the desire to find a map from a topological space X to a simplified discrete network model we can calculate and visualise on computers. The primary topological inspiration is the nerve, which takes us from X to a simplicial complex $\Delta(A)$, where A is the indexing set of a cover. As mentioned above, we can think of simplicial complexes as spaces built from points, edges, triangles, tetrahedra, and their higher-dimensional analogues. The nerve theorem (Theorem 2.1.13) gives conditions under which an open cover $\mathcal{U}$ of X will give a homotopy equivalent simplicial complex. This is useful since it guarantees us a discrete representation of a space which preserves topological properties up to homotopy. In particular, any loops, spheres, or other higher-dimensional holes which exist in the original topological space X will still exist in a discrete representation of it, if we take care in how we build such a representation.

Mapper now consists of trying to construct a suitable open cover of a space which, when we take the nerve, will give a simplicial complex simple enough to visualise while also keeping as much of the topology of the original space. This will take place in three steps. Firstly, we introduce a reference map $\rho : X \to Z$ from our topological space X to some metric space Z, where Z is appropriately simple.[4] Then from an open cover $\mathcal{V}$ of Z, we produce an open cover of X by taking the pullback (pre-images) of the sets under ρ, which we can write as $\rho^*(\mathcal{V})$.[5]

Secondly, once we have this open cover of X, Mapper introduces a refinement step by splitting the open sets of $\rho^*(\mathcal{V})$ into their path-connected components. This step results in an open cover of X which is closer to satisfying the requirements of the nerve theorem.

Finally, all that remains is for us to take the nerve. We can then visualise it, if the resulting simplicial complex is low-dimensional enough. In particular, if we end up with a 1-dimensional simplicial complex, we can visualise it as a graph, as in Figure 2.6.

2.1.3 Discrete Spaces

In practice we will be dealing with *point clouds*, not infinite topological spaces.

[4] Typically, Z is one or two dimensional.

[5] Note: when $Z = \mathbb{R}$ and the cover is by overlapping intervals then we get something which can be thought of as a discrete approximation to the Reeb graph [Ree46], and indeed Mapper's convergence to the Reeb graph in this case was proved in 2015 [MW15; CO18].

Definition 2.1.14 (Point Cloud)**.** A *point cloud*, X, is a finite set of points, with the discrete topology. The metric we define on this set will depend on our use case. For example, in the case where X is a subspace of Euclidean space, $\mathbb{R}^n$ for some $n > 0$, then we can take the induced Euclidean metric.

The notion of covers, and pullbacks of covers from a reference metric space, can be reused without difficulty. But what is analogous to path-connected components? The Mapper algorithm uses here the idea of *clustering*, where the clusters take the role of the path-connected components.

Definition 2.1.15 (Clustering Algorithm)**.** A *clustering algorithm* is a function on a finite number of points with a measure of distance, which outputs a partition of the points.

In Mapper single linkage clustering is implemented, with a chosen parameter ϵ. The result of this clustering corresponds to the path-connected components of the space consisting of balls of radius ϵ at each of the points in the point cloud. To summarise, the Mapper pipeline, starting from a point cloud X, is:

1. Define a reference map $\rho : X \to Z$ from your point cloud X to a reference metric space Z. This ρ is called the *filter* function.

2. Select a covering $\mathcal{U}$ of Z. For example, if $Z = \mathbb{R}$ we can take a covering by overlapping intervals.

3. If we have the cover $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$, where A is the index set for $\mathcal{U}$, we construct subsets of X by pulling back $\mathcal{U}$ over ρ, that is we get subsets $X_\alpha = \rho^* U_\alpha$.

4. Select a value of ϵ to input into a single linkage clustering algorithm, which we apply to each set X_α, giving us clusters of each set. At this point, we have a covering of X parametrised by (α, c), where $\alpha \in A$ and c is one of the clusters of X_α.

5. Construct the simplicial complex with elements all possible pairs of (α, c), and with subsets $\{(\alpha_0, c_0), (\alpha_1, c_1), \ldots, (\alpha_k, c_k)\}$ where the corresponding clusters have non-empty intersection.

We demonstrate this pipeline with the following example of Mapper applied to a point cloud approximating a circle:

Example 2.1.16 (Mapper Applied to a Noisy Circle)**.** Figure 2.5 gives an example of Mapper applied to a point cloud sampled from a circle, where we add Gaussian noise. Here the reference map is the height function $\rho : \mathbb{R}^2 \to \mathbb{R}$, given by $\rho(x, y) = y$. We take a covering $\mathcal{U}$ of the height by three intervals, which results in a covering $\rho^*(\mathcal{U})$ by three open sets on our original circle, given by colours in the figure. We note that the middle (purple) interval can be separated into two clusters. The bottom part of the

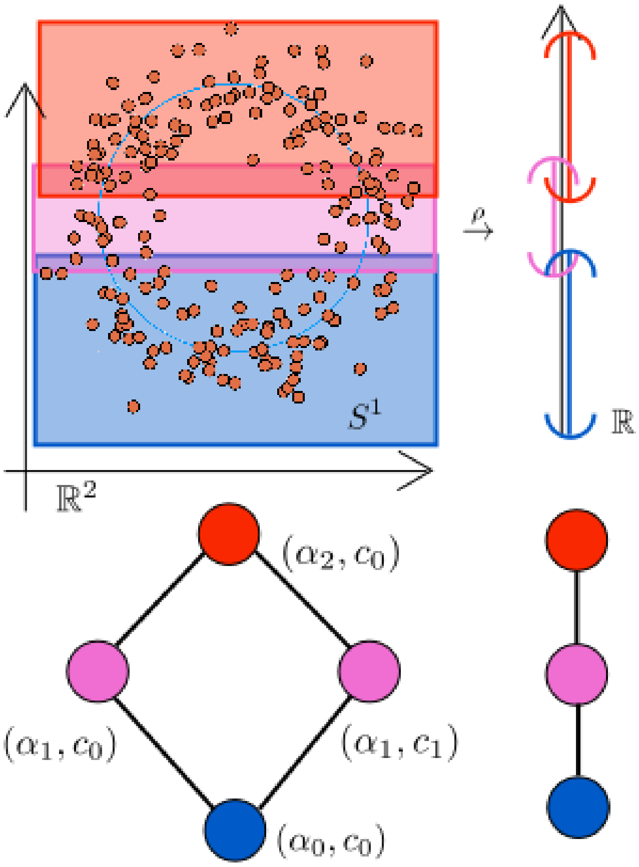


Figure 2.5: In the top left, we have a point cloud coming from a noisy sampling of the circle S^1 in $\mathbb{R}^2$. The reference map ρ is the height function, mapping to the line $\mathbb{R}$ in the top right. Note that we have a covering of $\mathbb{R}$ by three open intervals, represented by the three bracketed lines. The coloured rectangles in the top left represent the pullback of the covering on $\mathbb{R}$. Note that the middle rectangle contains two clusters, by inspection. The bottom right graph represents the nerve of the covering without breaking the middle into two clusters, note how it appears like a line. The bottom left graph represents the nerve of the covering, with the middle broken into two clusters, note how this has the same homology as the circle. This example is modelled on one in [Car09, §3.2].

figure shows the resulting simplicial complexes we get when we take the covering $\rho^*(\mathcal{U})$ of S^1 versus the one with the middle split into two colours. The graph on the right is a line, while the one on the left more closely approximates the circle.

2.1.4 Inputs and Outputs

We have given an overview of the steps in applying the Mapper algorithm. Now let us summarise its outputs and inputs, along with a few comments. For inputs, Mapper requires $\{X, \rho, Z, \mathcal{U}\}$, that is, a point cloud X, a reference map $\rho : X \to Z$, and a covering $\mathcal{U}$ of the reference space Z, along with a clustering algorithm (usually single linkage clustering with parameter ϵ). The output is a simplicial complex with dimension bounded by the covering dimension of $\mathcal{U}$. The covering dimension of $\mathcal{U}$ is given by a positive integer, d, so that any collection of more that $d+1$ sets in $\mathcal{U}$ has empty intersection. This means, by our construction, we will not have any simplices of dimension greater than d.

Note that we can get a multiresolution structure by varying the cover on Z. In particular, we can consider sequences of covers that include one into the other (for example, take a collection of coverings of $\mathbb{R}$ by steadily larger intervals). However, this requires that our clustering algorithm does not split clusters, if more points are added to a space (or else we will not get functions between the resulting simplicial complexes). Single linkage clustering has this property, since adding more points to a point cloud we're clustering can only cause clusters to merge. In more technical terms, single linkage clustering is *functorial* under inclusions. So two points assigned to the same cluster by single linkage with a certain parameter ε will still be assigned to the same cluster when more points are added, namely, under inclusion into a larger set of points. Complete linkage clustering does not have this property.

Finally, let us discuss the choice of ϵ for the (single linkage) clustering step.[6] The idea is to choose either a single ϵ, or otherwise an ϵ_α for each $\alpha \in A$, that is, for each set of the cover, such that if two sets of the cover, U_α and $U_{\alpha'}$ have a non-empty intersection, then ϵ_α and $\epsilon_{\alpha'}$ produce the same set of clusters when applied to the pre-image of $U_\alpha \cap U_{\alpha'}$. Practically speaking ϵ_α is chosen by looking at the dendrogram for single linkage clustering applied to the pre-image of U_α, and picking a value in a place where a large gap appears. The reasoning being that shorter edges are required to merge points which "should" belong in a cluster, while relatively longer edges are required for merging clusters themselves.[7]

An additional point to consider is Mapper's sensitivity to parameter changes, as discussed by Carrière and Oudot in [CO18].

So, in summary, Mapper requires a point cloud X, a reference map $\rho : X \to Z$ (also called a filter function) which gives a value in some reference space Z to each data point, a cover of Z, and finally some clustering function which is applied to the pulled back cover on the data set. The nerve of this covering is the simplicial complex produced by Mapper. Unusually for data analysis, we are given choices when applying the Mapper algorithm. However, this allows us to find properties that persist for any "generic" choice, which leads us to suspect that such properties are significant. See also Table 2.1 for a summary with examples.

[6] Gunnar Carlsson's presentation about Mapper at the 2015 Young Topologists' Meeting refers to this as the "magic fudge".

[7] We have described how Mapper produces a simplicial complex. Visualising it, so we can see the results in 2D, or at most 3D, is another matter. Most applications use $\mathbb{R}$ as the reference space, so the result is a 1-simplex, that is, a graph. We can then draw these graphs using certain software applications. Typically, the node sizes correspond to the number of points in its cluster, and the colour is by filter function, but there are other options here. The nodes are also usually positioned in order of increasing filter function, in some direction.

Inputs (or Choices)	
Point Cloud X	Gene expression profiles
Reference Map $\rho : X \to Z$	Height function
Cover $\mathcal{U}$ of Z	Open intervals
Clustering Function	Single linkage clustering
Output	
Simplicial Complex	Visualised as a graph

Table 2.1: Mapper inputs and output summary.

Example

For illustrative purposes, we offer another example here taken from [Lum+13], which we show in Figure 2.6 corresponding to Figure 1 from the Lum *et al.* paper.

(A) In this example, the point cloud is sampled from a hand in 3D.

(B) The reference map is illustrated using a colour gradient, which is by x-coordinate from right-to-left (blue to red), one can also think of this as 'height' from the base of the hand.

(C) We see here the cover and clustering. The cover is shown by the hand being split up into overlapping portions along the left-right axis, and the result of the clustering in this case corresponds precisely to the connected components of the hand, that is, each isolated portion of the split up hand is its own cluster.

(D) We see here the output of Mapper, which is a graph with nodes corresponding to each cluster in part **(C)**, and edges between nodes with clusters which overlap.

2.2 Application to Medical Data

This section will detail the application of Mapper to a breast cancer gene expression data set from a paper of Nicolau *et al.* [NLC11]. The data set is high dimensional, consisting of expression values from thousands of genes, with each gene corresponding to one dimension. In particular, we will outline the reasoning behind the choices made in the application of Mapper, going from the initial pre-processing of the data, to the filter function (reference map) chosen, and the clustering algorithm used. Lastly, we describe how, with aid from the output of Mapper, Nicolau *et al.* determined that the cluster containing the newly discovered *c-MYB*$^+$ breast cancer subgroup was of interest.

2.2.1 The Data

We begin with a brief description of the data and its sources. There were two sources of data, firstly the 295 tumour samples from the *Nederlands Kanker Instituut* (*NKI*)

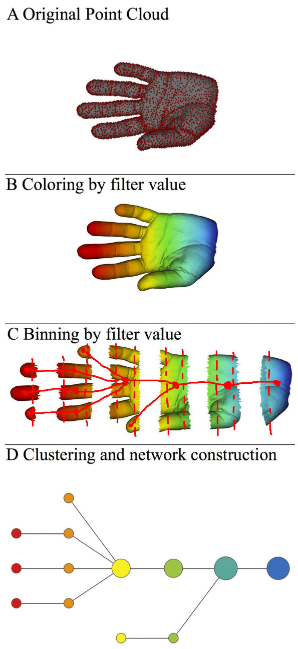

A Original Point Cloud
B Coloring by filter value
C Binning by filter value
D Clustering and network construction

Figure 2.6: A figure showing the Mapper algorithm applied to a point cloud arising from sampling a hand in 3D. We have edited **C** from the original [Lum+13, Fig. 1] by showing the clusters in the pullback of the cover more clearly. **A**) shows the point cloud in 3D, sampled from the hand. **B**) shows the filter function using colour. In this case, the filter function is the x-coordinate, from right-to-left. **C**) shows the cover, which has split up the hand into overlapping section, and it additionally shows the clusters, which are just the connected components of the split up hand. **D**) shows the Mapper graph output.

[Vij+02], and secondly 13 normal breast tissue samples, called the *Breast Cancer Normal* (*BCN*) data.[8] The data come from 25k microarrays.[9]

2.2.2 Pre-processing

The first step is to pre-process the gene expression data. We will give an outline of this process in this paragraph, and go over the steps in more detail in the rest of this subsection. Pre-processing involves checking the quality of the samples, and keeping only samples of high quality (i.e. at least 70% of the sample's genes have recorded expression values). Next missing data is imputed (that is, inferred from the data we have), and the actual genes are identified by mapping the microarray probes to an appropriate database.[10] Once we have the expression values and their corresponding genes, we take advantage of the fact that we have a source of 'diseased' (breast cancer)

[8] Note: I've had difficulty finding the source of these. There are dead links to some of the raw data in the older papers. The closest I've come is Nicolau *et al.*'s DSGA paper [Nic+07] which cites their breast cancer data as 63 primary tumour samples and 13 normal tissue samples coming from [Zha+04], although that paper itself states that they have only three normal breast tissue samples. I later received a copy of the data after contacting Nicolau.

[9] For mathematicians, it's sufficient to know that microarrays are a technology which can be used to measure gene expression.

[10] Nicoleau *et al.* [NLC11] use a *knn* algorithm with $k = 10$ to impute their missing data, and use the UniGene cluster ID build 219 to determine their genes. The exact algorithm and database of genes used will, of course, depend on the application.

samples and 'normal' samples. We normalise using DSGA [Nic+07] (see a later part of §2.2.2 for more details) between the two data sets, then use the expression of the normal samples to work out a subspace of normal expression. This is then subtracted from all the samples, leaving us with only the 'diseased' component of the expression. Finally, we filter out a certain number of genes, by taking those which have the highest diseased expression.

Extract Values

To get the gene expression values, the authors take the microarray data (from 25,000-gene arrays) and keep only the samples with at least 70% high quality data. Next they use a *knn* algorithm [Tro+01] with $k = 10$ to impute the missing data, that is they use the mean value of the 10 nearest samples to predict the value of a gene in a sample which is missing its value.[11] and finally they use UniGene cluster ID (build 219) to cluster the samples into UniGene clusters[12].

The result is that 18,790 UniGene clusters are found in the *NKI* data, and 18,791 in the *BCN* data. Of these, 12,237 UniGene IDs matched between the two data sets. So, each sample now corresponds to a vector $\vec{T} \in \mathbb{R}^{12,237}$ of gene expression values.

Subsequently, the vectors of gene expression level corresponding to the 12,237 UniGenes for each tissue sample were normalised, to have the same magnitude as the mean of the 13 normal tissue vectors. This accounts for any systematic differences in the way the 295 tumours vs. the 13 normals were measured, since they come from two distinct sources.

DSGA

Disease-specific genome analysis (*DSGA*) [Nic+07] is a method which decomposes 'omic data into two terms, a *normal component*, which is the part of the data best mimicking healthy tissue, and a *disease component*, which is the rest, and one can think of this component as a measure of how aberrant the tissue is. In equation form, if we have a vector $\vec{T}$ of 'omic data, we get a decomposition into a normal component, $Nc.\vec{T}$, and a disease component $Dc.\vec{T}$.

$$\vec{T} = Nc.\vec{T} + Dc.\vec{T}$$

The normal component $Nc.\vec{T}$ is calculated by fitting the tissue data $\vec{T}$ onto a linear model, called the *Healthy State Model* (*HSM*), which is calculated from normal tissue data using a *FLAT* construction, as detailed in [Nic+07, Computational Details Supplement], which we will now summarise.

We begin with our 13 normal tissue sample data $\{\vec{N}_1, \ldots, \vec{N}_{13}\}$. To reduce the impact of expression which is unique to individual tissue samples, we replace each of

[11] The authors give no particular reason for using 10.

[12] UniGene clusters correspond to actual genes in humans

the normal tissue sample data vectors by their approximation with all the other vectors. This gives us the 13 vectors $\{\hat{N}_1, \ldots, \hat{N}_{13}\}$ where

$$\hat{N}_i = \sum_{j \neq i} \beta^i_j \vec{N}_j$$

and the β^i_j are chosen to minimise the distance between $\vec{N}_i$ and $\hat{N}_i$.[13] We will refer to this set as the *FLAT* normal vectors.

Principal Component Analysis (*PCA*) is now applied to the *FLAT* normal vectors to identify an appropriate subspace, which we will consider to be our *HSM*, that is, all expression falling in this subspace will be considered "normal". The choice of dimension reduction is determined by the Wold invariant [Wol78]:

$$W(l) \approx \left(\frac{\lambda_l^2}{\lambda_{l+1}^2 + \cdots + \lambda_R^2} \right) \frac{(n - l - 1)(R - l)}{(n + R - 2l)}$$

where λ_i is the ith singular value of the *PCA*, R is the number of normal samples (13 in this case) and n is the number of genes (12,237 in this case). Roughly speaking, λ_i gives a measure of the amount of data[14] in the ith direction, so Wold's invariant $W(l)$ is proportional to the smallest signal (λ_l) divided by all the noise ($\lambda_{l+1}, \ldots, \lambda_R$), if we were to pick l as our number of dimensions. As an approximation to a signal-to-noise ratio, we desire a value of l for which $W(l)$ spikes up, or shows an abrupt drop.[15]

We now choose the appropriate dimension l subspace from our *PCA* to be our *HSM*. Now we can properly define our decomposition of a data vector $\vec{T}$ as:

$$\vec{T} = Nc.\vec{T} + Dc.\vec{T}$$

where $Nc.\vec{T}$ is the projection of $\vec{T}$ onto the *HSM* and $Dc.\vec{T}$ is the complement, given by their difference.

Gene Thresholding

Now that we have disease vectors $Dc.\vec{T}$ for each of our tissue samples, we will threshold the data so that only genes that show a significant deviation from the healthy state are considered. The goal here is to find genes which have unusually high expression in the disease components. For each of the 12,237 genes, Nicolau *et al.* recorded the 5th and 95th percentiles of the disease components of the 295 tumours, that is, they

[13] i.e. $\hat{N}_i$ is a projection of $\vec{N}_i$ into the subspace spanned by $\{\vec{N}_1, \ldots, \vec{N}_{i-1}, \vec{N}_{i+1}, \ldots, \vec{N}_{13}\}$.

[14] which we think of as variation

[15] Apart from [Wol78], and this paper, I've been unable to find the Wold invariant online. The statisticians I've talked to also have no idea about it. Monica Nicolau has indicated that the Wold invariant is from the aforementioned paper. However, the method used in [Wol78] does not match what she has used. In any case, as mentioned in the DSGA paper's supplementary information [Nic+07], the usual method of choosing an appropriate dimension for the PCA is to go by proportion of variance explained (depends on the application, but > 0.9 is typical) and to pick a dimension after which the singular value drops suddenly, the idea being you only care about variation above a certain threshold.

take about the 15th smallest value and the 15th greatest value of the 295.[16] Now they take the higher absolute value of the two and call it $MaxAbs595$, and we have this value for each of the 12,237 genes. They then took the 85th and 98th percentiles of these $MaxAbs595$ values, that is, for the 12,237 $MaxAbs595$ values, they take about the 1,836th and 245th highest. The genes chosen for analysis were those which had values above the 85th percentile, and were also highly correlated ($r > 0.6$, Pearson's correlation coefficient) to at least three genes having values above the 98th percentile. This ensured that the retained genes deviated significantly from the *HSM* in highly correlated groups. Nicolau *et al.* ended up with 262 genes retained.

2.2.3 Applying Mapper

Mapper is now applied. This is the step where topology plays a role. The input was a point cloud of 295 tumour points plus 13 normal points, with distances given by the Pearson correlation distance, $d(i,j) = 1 - \operatorname{cor}(i,j)$, where $\operatorname{cor}(i,j)$ is the Pearson correlation, defined as follows:

Definition 2.2.1 (Pearson correlation)**.** The Pearson correlation between two samples i and j is given by

$$\operatorname{cor}(i,j) = \frac{\sum_{k=1}^{m}(\mathbf{y}_{k,i} - \bar{\mathbf{y}}_i)(\mathbf{y}_{k,j} - \bar{\mathbf{y}}_j)}{\sqrt{\sum_{k=1}^{m}(\mathbf{y}_{k,i} - \bar{\mathbf{y}}_i)^2}\sqrt{\sum_{k=1}^{m}(\mathbf{y}_{k,j} - \bar{\mathbf{y}}_j)^2}}$$

where $\mathbf{y}_{k,i}$ is the expression of gene k in the i^{th} sample, and $\bar{\mathbf{y}}_i$ is the mean expression of sample i.

The correlations were calculated in $\mathbb{R}^{262}$ on the retained genes using the disease component values. A family of functions, $f_{p,k}$ were used as filter functions (reference maps), taking the data points to a value in $\mathbb{R}$, given by the L^p norm taken to the k^{th} power. That is, if we have a disease component associated with a data point $Dc.\vec{T} = (g_1, g_2, \ldots, g_{262})$, then we get:

$$f_{p,k}(Dc.\vec{T}) = \left(\left(\sum_{r=1}^{262} |g_r|^p\right)^{\frac{1}{p}}\right)^k$$

Nicolau *et al.* used values of $p = 1, \ldots, 5$ and $k = 1, \ldots, 10$, to change the relative importance genes with high variance in the disease components, and also to change the scaling. These affect the ability to pick out features in a visualisation. The cover chosen for $\mathbb{R}$ was to divide the image of $f_{p,k}$ into 15 equal intervals, with 80% overlap.

The resulting output were one dimensional simplices, which they visualised as Mapper graphs. They modified these by excluding all vertices corresponding to clusters with only one data point in them, as we see in Figure 2.7 ($p = 2, k = 4$). The authors now

[16] Keep in mind that, after normalisation, we can have negative expression values.

find several groups of tumours that stand out by looking at the arms of the Y-shape of the Mapper graph. One arm corresponds to the normal tissue and cancers with normal-like expression. Two arms display high levels of deviation from normal. One of these correspond to Basal tumours, the other to ER^+ tumours (ones which are not normal-like). Of note is the new subgroup they discovered, the *c-MYB*$^+$ tumours, which stood out as being the most dense segment of the ER^+ sequence, and exists even when vertices containing only one data point are removed.

Comparing against Clustering

Nicolau *et al.* now compared the Mapper approach with clustering. To do this, they applied a clustering algorithm called average linkage clustering to the same data as Mapper and found it unable to distinguish the *c-MYB*$^+$ subgroup, as shown in Figure 2.8. Rather, clustering scatters the tumours in the *c-MYB*$^+$ subgroup, and even those in the ER^+ arm of the Mapper output. This step shows that it's possible to use topology, as incorporated in Mapper, to find something new in a gene expression dataset.

2.2.4 Distinguishing *c-MYB*$^+$

After finding an interesting subgroup, Nicolau *et al.* now argue that it deserves its own new classification. This is due to:

- Its uniformity in molecular signature
- Its clinical and survival properties
- It's validated in other breast cancer data sets
- It does not fit into previously identified breast cancer types

Note that this argument is statistical, and largely biological, work, not topological.

For survival and clinical outcomes, it's found that the patients in the *c-MYB*$^+$ group in the study had a 100% overall survival rate, with no recurrence or death from disease, with a median time to follow-up of 8.5 years.

For classification,the authors used the molecular subtypes found in [Sør+03] (*Basal*, *ERBB2*, *Luminal A*, *Luminal B*, and *Normal-like*) to classify the 22 *c-MYB*$^+$ tumour samples. The result was that six of the tumours had correlation > 0.1 to one of the five centroids, and the rest were left unclassified. This indicates that this subgroup does not fit neatly into any of the existing classifications, implying that it has not been discovered before.

There are two techniques the authors use to find a molecular signature. Firstly, the authors use *Prediction Analysis of Microarrays* (*PAM*) [Tib+02], which utilises a method of "nearest shrunken centroids", to tell if, using the disease component of all the 12,237 genes, the *c-MYB*$^+$ subgroup is distinct from normal tissue, and shows uniform

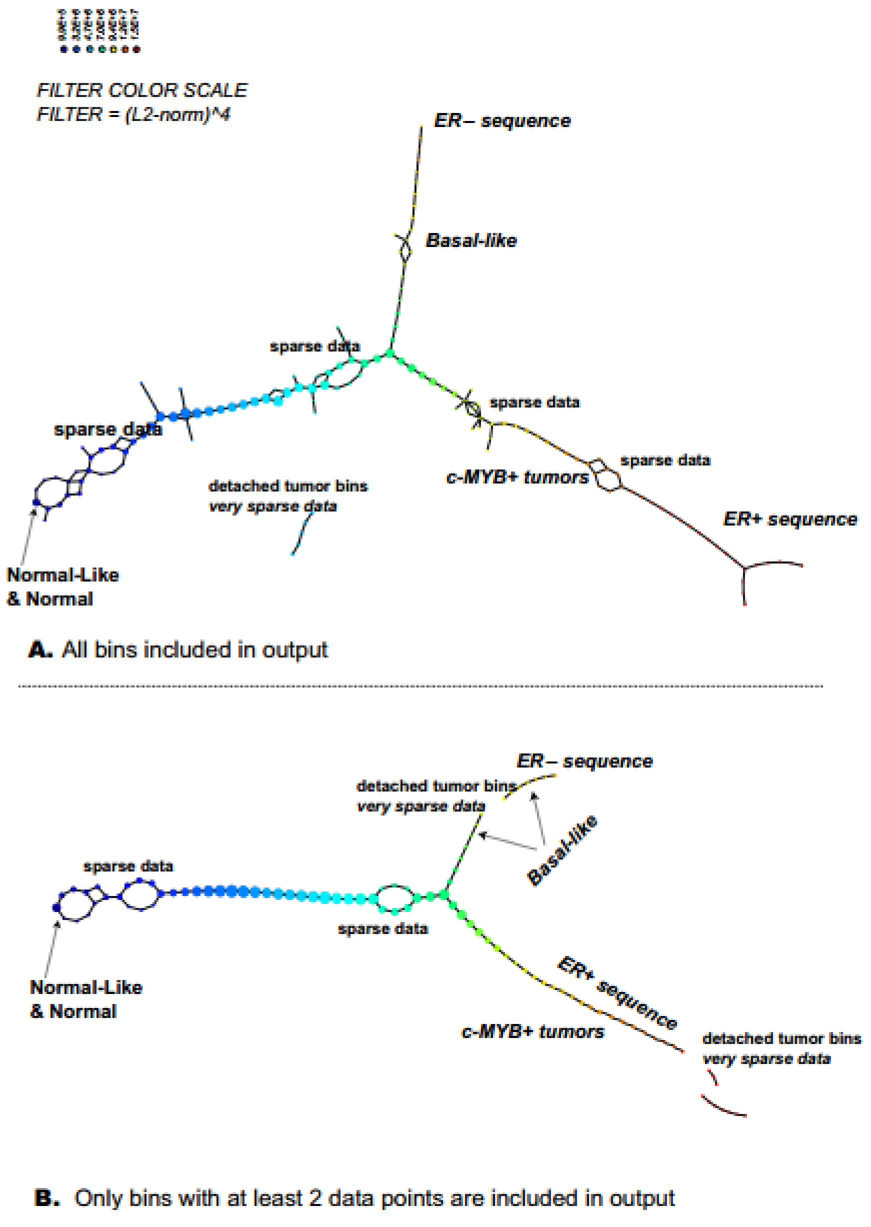


Figure 2.7: Figure S3 from [NLC11]. (A) shows the Mapper output with all vertices included. (B) Shows the output excluding vertices with only one data point. The size of a vertex is correlated to the number of data points in them, and the colour is the value of the filter function, which in this case one can think of as a measure of disease, i.e. away from “normal”, with blue being low and red being high.

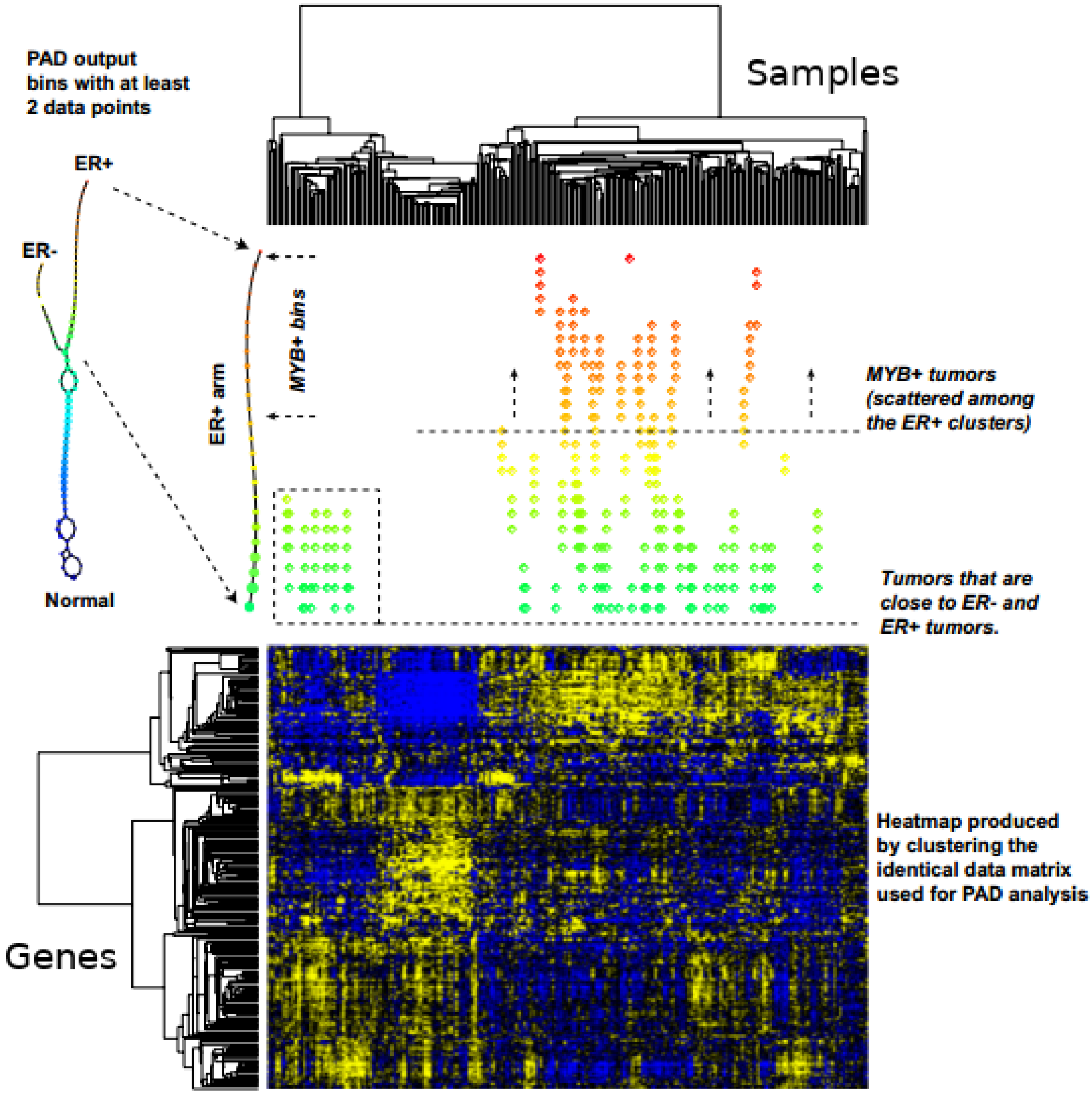


Figure 2.8: Figure S4 from [NLC11]. A comparison between Mapper and average linkage clustering. The data used in both cases was the *DSGA*-transformed data, taking the 262 genes which passed the thresholding. The top half shows the ER^+ arm of the Mapper output magnified, with the top half being the *c-MYB*$^+$ tumours. The position of the tumours in the vertices in the clustering dendrogram are given to the left of the Mapper output, from which we see that tumours which are close in the Mapper output, even coming from the same vertex in some cases, are scattered throughout the clustering dendrogram. In particular, the *c-MYB*$^+$ tumours are scattered in the dendrogram, while being close in the Mapper output. The bottom half of the picture shows the heatmap produced when clustering the data. In the heatmap, the columns are samples, and the rows are genes. The expression level is yellow for high and blue for low. See §3.4.2 for more details on heatmaps.

characteristics. Of particular note, *PAM* analysis finds two predictor genes which distinguished between the *c-MYB*$^+$ subgroup and normal tissue with 0 error. These genes were *TSH-releasing hormone* and *proprotein convertase subtilisin/kexin type 1.* The authors indicate that being able to use only two genes to separate *c-MYB*$^+$ from normal suggests that the subgroup is both distinct and homogeneous.

Secondly, the authors use *Significance of the Analysis of Microarrays* (*SAM*) [TTC01] to detect genes with statistically significant changes in expression.[17] *c-MYB* was, as expected, one of the top significant overexpressing genes.

Validation

Nicolau *et al.* then check if the *c-MYB*$^+$ subgroup was present only in the *NKI* data set, or if it is also present elsewhere. This validation was done on two other breast cancer sets, a *Ullevål University Hospital* (*ULL*) data set [Lan+07] of 80 breast cancers, and *HERSCH* [Her+08], a set of 232 tumours. The authors first selected 46 tumours of ductal histological type that had been in the study for longer than 10 mo from the *ULL* data, and 188 primary breast tumours with good-quality RNA. *DSGA* was applied to the selected tumours, using the UniGene clusters in common with the normal *BCN* data set, there were 17,441 Unigenes in common for the *ULL* data, and 18,898 for the *HERSCH* data.

The set of genes showing significant differences in expression in the *c-MYB*$^+$, identified by *SAM*, were used to extract a set of four tumours in the *ULL* set and 37 tumours in the *HERSCH* set. Considering also tumours highly correlated to that set, the authors find six of 46 tumours in the *ULL* and 19 of 188 in the *HERSCH* to be of the *c-MYB*$^+$ subgroup. Lastly, looking at this subgroup of tumours, they found the patients had 100% survival and no recurrence, just like in the *NKI* data set.

2.2.5 Discussion

We have seen an example of Mapper applied to a gene expression data set. Briefly, we first need to preprocess the data. This involves cleaning up missing data, as well as determining which parts of the gene expression data are normal expression versus disease expression, and extracting the disease component as being of interest. We then input the preprocessed data into Mapper, these being the point cloud of samples and the distances between them (calculated from the disease components), the filter function(s) (magnitude of diseased expression), and coverings for the reference space (intervals for the magnitude of diseased expression). We have also seen how the simplicial complex (graph in our one-dimensional case) was cleaned up by removing the vertices corresponding to small (single data point) clusters.

This visualisation allowed us to find an interesting subgroup (*c-MYB*$^+$ breast cancers) which had not been found before, and is not apparent using a standard clustering

[17] They are listed in the supplementary data of [NLC11].

method. It occurs as a distinct group in the Mapper output, but is split up by hierarchical clustering. The statistical validity of this subgroup is then checked, based on testing properties of the subgroup found by visual inspection through Mapper. These properties include: consistent gene expression in the *c-MYB*$^+$ subgroup, consistent clinical properties, and existence in other breast cancer datasets.

In the next section, we will go through another application of Mapper to the same data set. This application differs in that the preprocessing step and filter function chosen are much simpler, yet Lum *et al.* still manage to recover the same result.

2.3 Reapplying Mapper

A paper of Lum *et al.* [Lum+13], showcasing Mapper applied to several datasets, contains an example of applying Mapper to the *NKI* data [Vee+02], which is the same source of data that Nicolau *et al.* use, whose paper [NLC11] we went through in the previous section. In this section, we will go through this application. In this case, different preprocessing and filter functions are used, however, we still get the same result. So, this will show how Mapper can be applied to a data set which has been preprocessed in a different fashion, and yet recover the same essential structure. This highlights flexibility as a key feature.

They also run Mapper on another breast cancer microarray data set, *GSE2034* [Wan+05]. In both cases, the Mapper output is similar, and this gives us the ability to compare between different data sets, even if they come from different experiments.

2.3.1 Pre-processing

Lum *et al.* decided to use the 1,500 top varying genes in their analysis, after trying various values (24K, 11K, 7K, 3K, 1.5K) and seeing which one gave the most distinct branching of the Mapper graph output, as shown in Figure 2.9.[18]

2.3.2 A Simpler Filter

The data points are tumour samples, using the Pearson correlation distance[19] across the 1,500 chosen genes.

The filter function they used for Mapper was something the authors call *L-infinity centrality*, which assigns to each data point the value of the maximum distance that point is from any other point in the data set. That is, if X is our data set, and d, a distance function on it, then our filter function, f, is given by

$$f(x) = \max_{y \in X} d(x, y)$$

[18] Strictly speaking, this is not pre-processing, since the authors picked the number of genes they used by looking at the output. To truly be pre-processing, the figure 1,500 should have been decided before looking at the output.

[19] In geometric terms, this is $1 - \cos(\theta)$, where θ is the angle between the two gene expression vectors.

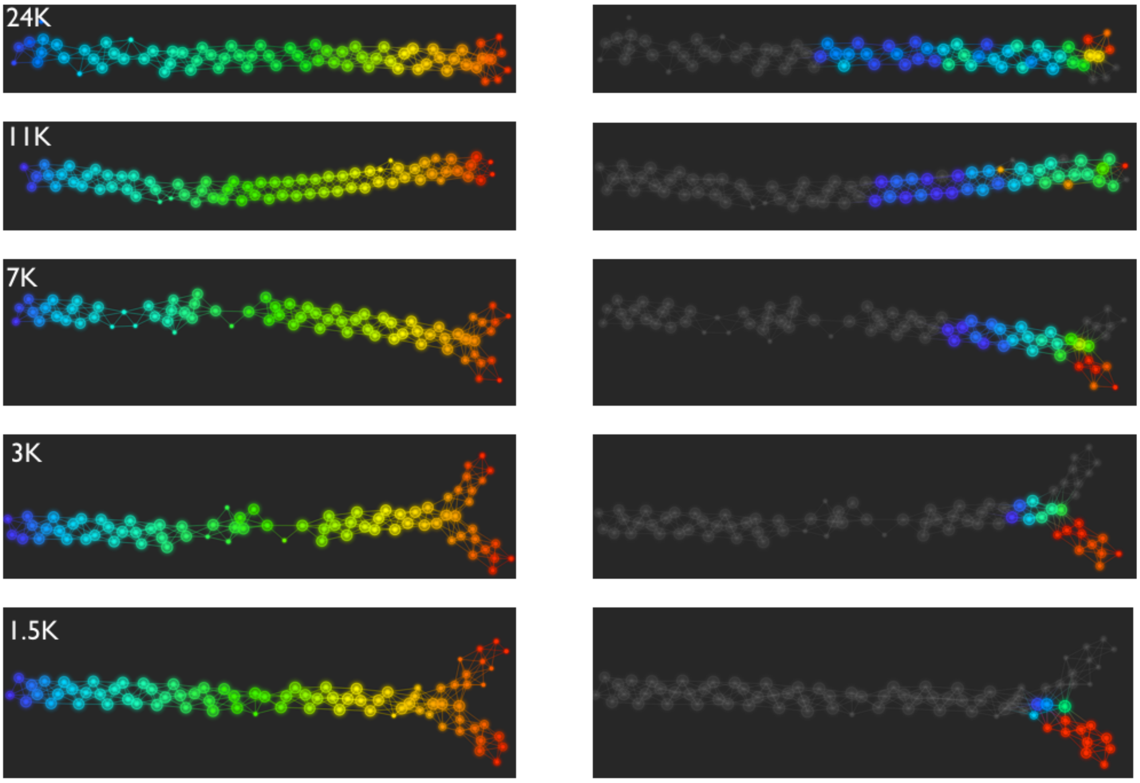


Figure 2.9: Figure S1 from [Lum+13]. The Mapper graph of the *NKI* data, using only the patients who survived, so corresponding to the bottom of panels A or B in Figure 2.10. The rows represent the analysis done with a different number of the top most varying genes, from 24,000 in the top, to 1,500 at the bottom, which is the same as Figure 2.10. The panels in the left column have been coloured by L-infinity centrality values, with blue low and red high. The panels in the right column have the nodes coloured by the percentage of the data points in them belonging to the bottom-right flare of the 1,500 graph. With grey being 0% and then a gradient from blue to red, red being 100%. The idea is to see where the data points corresponding to the *c-MYB*$^+$ subgroup end up as we increase the number of genes used in the analysis. We see that as the number of genes used in the analysis increases, the *c-MYB*$^+$ subgroup in the bottom-right flare spreads out and mixes with non-flare points. This indicates that the inclusion of genes which do not vary obscures the presence of the *c-MYB*$^+$ subgroup in the data.

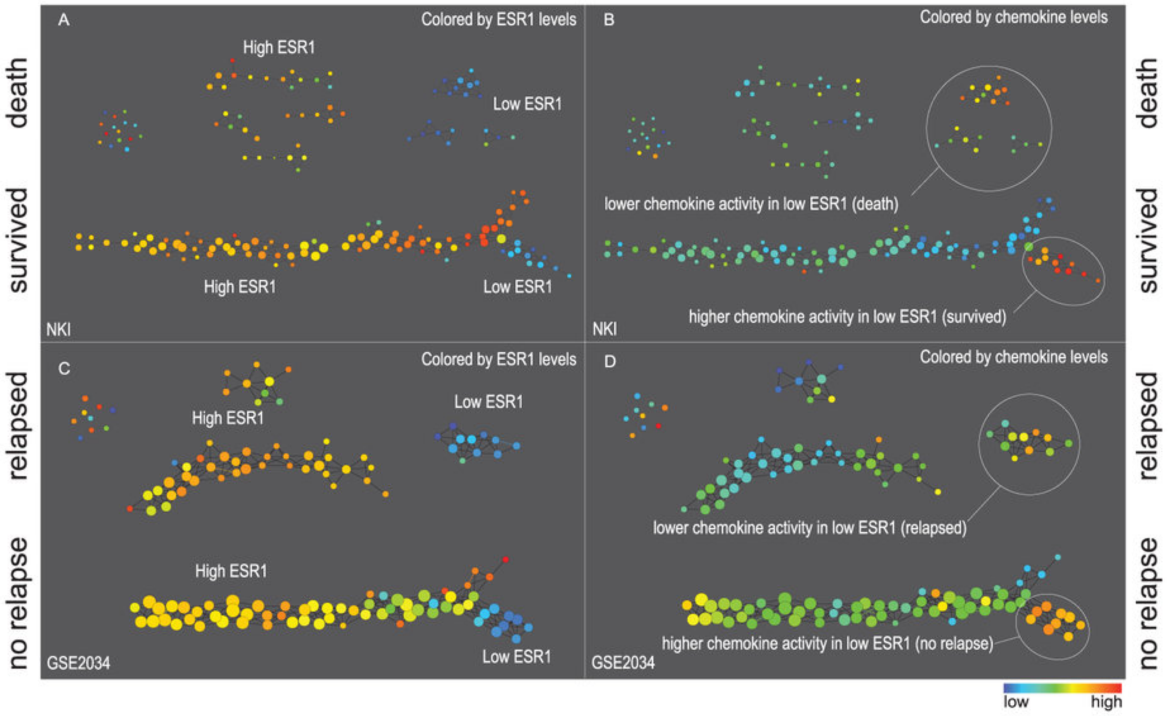

A
Colored by ESR1 levels
High ESR1
Low ESR1
High ESR1
Low ESR1
NKI
B
Colored by chemokine levels
lower chemokine activity in low ESR1 (death)
higher chemokine activity in low ESR1 (survived)
NKI
C
Colored by ESR1 levels
High ESR1
Low ESR1
High ESR1
Low ESR1
GSE2034
D
Colored by chemokine levels
lower chemokine activity in low ESR1 (relapsed)
higher chemokine activity in low ESR1 (no relapse)
GSE2034
death
survived
relapsed
no relapse
low
high

Figure 2.10: Figure 2 from [Lum+13]. The top panels, A and B, are from the *NKI* data, and the bottom panels, C and D, are from the *GSE2034*. In both cases, a binary filter has been applied in the top/down direction, being death/survival, and relapse/no relapse, respectively. From left to right the nodes of the graphs are positioned by their L-infinity centrality values. Finally, colouring has been done, the left two panels, A and C, are coloured by the average expression of the ESR1 gene, and the right two panels, B and D, are coloured by the average expression of the genes in the KEGG chemokine pathway. In both cases blue is low and red is high.

This is in contrast to the filter function used in [NLC11], where the authors first used *DSGA* to get a measure of how diseased the gene expressions were. It is interesting that they obtain a similar result, despite using a simpler filter function in this analysis. However, both functions act as a proxy measure of disease, since more unusual (and hence, far from the centre) gene expression is indicative of disease.

They also use a binary filter to split their graph into two, which is survival/death in the *NKI* data, and no relapse/relapse in the *GSE2034* [Wan+05] data. They find a graph output by Mapper in both cases, as we can see in Figure 2.10, and mention that here we see Mapper's great utility in being able to compare data sets without transforming coordinates to account for how the two data sets measured gene expression. That is, the authors have used Mapper on two different gene expression datasets to produce a similar visualisation, without normalising the two datasets between each other. In particular, samples from one Mapper output are similar to samples occupying the same area in the other Mapper output.

2.3.3 Validation Again

Lum *et al.* validate the significance of the features of the Mapper output by simulating random data. They look at two interesting structures, flares, which are long linear segments, and groups, which are subsets of neighbouring vertices. The idea is to look at how often a certain structure appears in random data. So, if we get something in the Mapper output of our actual data that only appears rarely, then we know we have found something which may be of further interest.

Validating Flares

Flares are long linear segments, which may indicate an interesting subgroup in the data, like the $c\text{-}MYB^+$ subgroup in this example. To test their significance, Lum *et al.* generated 1000 data sets of the same dimensionality as the original data. The entries in each column were given by a Gaussian distribution with zero mean and constant variance across all columns. Mapper is then applied to each generated data set, and a flare detection algorithm is applied to the resulting graph. The number of flares is counted, and compared to the number found in the graph of the original data.

The flare detection algorithm is as follows:

1. Associate an eccentricity value $e(n)$ to each node of the graph, given by:

   $$e(n) = \sum_{m \in V(G)} d(n, m)$$

   where $V(G)$ is the vertex set of the graph, and d is the graph distance.[20]

2. For each connected component of the graph G, compute the zero-dimensional persistent homology as follows. We start at the maximum eccentricity value for each connected component, and decrease the value from there, adding nodes as we pass their eccentricity value (so nodes are added from high to low eccentricity). Meanwhile, we keep track of the number of components, and the eccentricity value at which they were born and at which they died. We will end up with a set of components, parametrised by their birth and death eccentricity values.

3. We keep track of the range of each component, and normalise it by dividing by the range of the eccentricity values, so we get values in $[0, 1]$ assigned to each component. Higher values indicate longer components, which were likely started by high eccentricity nodes, and so correspond to flares.

4. For each $x \in [0, 1]$, we look at the number of components with an associated value greater than x. This gives us a non-decreasing function of x. We look for the longest interval over which this function stays constant and greater than one, and

[20] The idea is to differentiate between nodes in more central regions of the graph and nodes at the ends of flares.

take the corresponding components as flares. The algorithm outputs the number of flares.

Note that the algorithm is only applied to connected components which make up more than 10% of the overall graph. The authors state that the above algorithm, applied to randomly generated data as described above, did not produce more than one flare except once in 1000 Monte Carlo simulations. This suggests that the presence of flares has a significance level of about 0.001, and will tend to be due to structure in the data, rather than noise or happenstance.

Validating Groups

The purpose of this section is to validate the significance of groups in the Mapper output displaying a high proportion of a certain trait. For instance, Lum *et al.* state that in previous analyses of the breast cancer microarray data they had found connected families of adjacent vertices where the survival is perfect (that is, a subgroup where everyone survived). The question now arises, what is the significance of such a group occurring, and is it just an artefact of applying Mapper? To answer this question, the authors consider relative density measures of survival on the breast cancer microarray data, as follows.

The microarray data set can be considered as a *finite metric space* X, once we have chosen a distance on it.[21] We can then consider a proxy function for density on X, for instance:

$$\rho_\sigma(x) = \frac{1}{|X|} \sum_{x' \in X} k_\sigma(d(x, x'))$$

where $k_\sigma(t)$ is a probability distribution, for instance, the normal distribution, with mean zero and standard deviation σ.

Now, we can similarly come up with a proxy function for density considering only a subset of the patients, for instance, the ones who survived. Let us call this set L, and the resulting proxy function:

$$\rho_\sigma^L(x) = \frac{1}{|L|} \sum_{x' \in L} k_\sigma(d(x, x'))$$

We are interested in the value $q(x) = \frac{\rho_\sigma^L(x)}{\rho_\sigma(x)}$, a proxy for the relative density of live patients, compared to the density of all patients in the metric space.

To study how significant a certain value of $q(x)$ is, we will assume a null hypothesis that the set L of living patients occurred randomly. We now select sets L' of the same number of points as L, chosen uniformly at random, and look at the corresponding maximum value of the quotient $q^{L'}(x) = \frac{\rho_\sigma^{L'}(x)}{\rho_\sigma(x)}$.

To do this, Lum *et al.* performed a Monte Carlo simulation by repeated selection of sets L', and recorded the values $\mu^{L'} = \max_{x \in X} q^{L'}(x)$. That is, they selected sets

[21] For gene expression data, this will rarely be the Euclidean distance of $\mathbb{R}^n$.

of points of size $|L|$ at random from the breast cancer data set, and calculated the maximum density of the 'living' patients relative to the density of all patients. The distribution of these values were then used to determine how significant a given value of $q(x)$ is. In particular, the probability of a value of $q(x)$ at least as extreme as the $q(x)$ associated with the high survival flare in the breast cancer data is less than 10^{-4}.

2.3.4 Discussion

We have seen a second application of Mapper to the same dataset. The pre-processing step and filter function chosen were different, but we still ended up with a similar result to the first application. This serves to confirm the existence of the *c-MYB*$^+$ breast cancer subgroup. Furthermore, we see that applying Mapper to another dataset also leads to a similar-shaped Mapper graph. This shows that we can use Mapper to make a visual comparison between datasets which come from different experiments. Finally, we have seen another method less dependent on discipline-specific statistics for determining the statistical significance of structures (flares and groups) in Mapper graph output.

2.4 Discussion

We have described the Mapper algorithm, and seen its application to breast cancer gene expression (microarray) data to detect a novel subgroup of breast cancer.

In applying Mapper, there are four choices that must be made:

1. We must have a choice of distance function on our data set.
2. We must choose a reference space and a map to it. (Our filter function.)
3. We must choose a cover of the reference space.
4. We must cluster the pre-image of our cover of the reference space.

In cases where the choice is unclear, we vary the choices made and see what interesting features persist in the outputted Mapper graphs. For instance, in Section 2.2 (the main breast cancer example), distance was given by Pearson correlation, and the pre-image clustered by single-linkage, with the parameter chosen in a suitably large gap between the clusters changing. However, the filter function, a measure of the diseased state, was changed by considering different L^p norms raised to different powers, which represents giving differing weights to either all genes, or just the most varying ones, and the cover of the reference space was changed by adjusting the number of intervals and their overlap, which affected the resolution of the resulting Mapper graph. The removal of bins with only one data point is another way the Mapper graph can be cleaned up. It corresponds to removing outliers, so is another way to test the robustness of the features encountered. Ultimately, structures which persist across parameter changes are considered robust, and so undergo further testing. This is the key principle of applying Mapper.

When interesting structure is encountered, it must be validated. In Subsection 2.3.3 we see the statistical validation of flares and groups by Monte Carlo simulation, as described in the Methods section of [Lum+13]. However, if we are to find out why the structure we encountered is interesting, we must dig deeper into the data, as described in §2.2.4, where Nicolau *et al.* tested a subgroup found by Mapper, by looking at the survival data of the patients, and testing for genes which had significant changes in expression, as described in [NLC11].

Now some remarks to conclude. In order to use Mapper, a metric is *essential*. In both approaches, the Pearson correlation distance was used. *c-MYB*$^+$ could not have been found without a metric, at least not using Mapper. Now, the validation of *c-MYB*$^+$ can take place without a metric, but this is mere comparing between the *c-MYB*$^+$ and the other tumours, once they have already been separated into subgroups.

Mapper is dimension reduction along with clustering applied locally. The dimension reduction aspect is given by the reference map ρ to some reference space. There are even cases where the reference map gives values of the first two principal component values, from a PCA. The clustering is applied locally, in the preimages of the open cover chosen for the reference map. This allows Mapper to discriminate more between points than applying a clustering function alone, since points separated by the reference map will never be clustered together.

This results in Mapper picking up structure that traditional clustering methods cannot. In particular, when the reference space is $\mathbb{R}$, Mapper is an approximation of the Reeb graph [MW15; CO18]. This makes it possible for Mapper to detect the number of holes in a surface, assuming a fine enough sampling of points [CM+04]. As another contrast with clustering, we can consider the circle or the 'Y'-shape (as in the breast cancer example). Both consist of a single path-connected component, so it would be reasonable to expect that a clustering algorithm would cluster a point cloud from them into a single cluster.

Mapper is a visualisation tool which helps in extracting interesting features from point cloud data. Such features are usually seen from its graph output as groups, flares, or loops. While Mapper provides a rich avenue for the initial exploration of data, making further sense of the output usually requires finer analysis and knowledge of the system studied. In particular, studying the groups or structures of interest in the Mapper output will involve looking at differences between the group of interest and the rest of the data. For instance, in the discovery of the *c-MYB*$^+$ subgroup, further analysis involving the gene expression was required to validate the existence of the subgroup. Mapper may tell you that there is interesting structure in your data, but it will not tell you why, or exactly what it arises from. That requires further familiarity with the data, or system being analysed, as well as other empirical and analytical approaches. On that note, the next chapter will introduce the problems currently occurring in ecological transcriptomics, as well as our main study species, the Arctic charr (*Salvenlinus alpinus*), found all over the Northern hemisphere.

Chapter 3

Transcriptomics in the Wild

This chapter introduces the current problems in ecological transcriptomics that we are seeking to address, and also gives the biological background on our study species *Salvelinus alpinus*, the Arctic charr. We shall show what motivates us to study it, and also give some results of standard differential gene expression techniques.

For mathematicians, we have a small subsection defining the space we are working in and some concepts in the data section (§3.3.3). If desired, mathematicians can read from this section and skip the background on the biology.

3.1 Ecological Transcriptomics

The transcriptome consists of all RNA molecules in a cell or population of cells; transcriptomics is a field studying the transcriptome. Typically, we analyse the mRNA (messenger RNA) which gives us the gene expression of the cells. mRNA is transcribed from the DNA sequence of genes in the nucleus, and exported to the cytoplasm where it is translated into a protein sequence by ribosomes. It therefore gives us a way to measure the rate of protein synthesis associated to given genes. The gene expression is a dynamic source of information that is intermediate between the phenotype (physical characteristics) and the genotype (the code of the cell contained in the DNA). In the context of ecology, mRNA molecules are extracted from tissues of interest from individuals of species of interest. For example, in the next section we will describe our interest in the Arctic charr, from which we take white muscle tissue samples for transcriptomic analysis.

Recent advances in high throughput DNA sequencing technology have led to an explosion of data in the field of transcriptomics. We have advanced from microarrays, which give the expression of $\sim 10,000$ genes, to RNA-Seq, capable of giving the expression of $\geq 100,000$ genes [Wan+17]. Furthermore, falling costs have given us the means to apply RNA-Seq to 10s of samples from wild non-model species.

The older microarray technology was constrained to target only transcripts with a know sequence, and has difficulty detecting rarely expressed transcripts. RNA-Seq, on the other hand, can capture novel transcripts, not just those with a known sequence, and also has a greater range of sensitivity then microarrays. However, RNA-Seq is still considerably more expensive then microarrays, and also generates a higher volume of data [KWG11].[1]

The tools for analysing RNA-Seq data have not kept pace. In particular, differential gene expression is analysed in a gene-by-gene manner, which doesn't take into account the co-dependence of gene expression, while gene network approaches, accounting for the relationships between genes, are specialised for model species in the lab, and cannot take into account novel information from non-model systems.

A key aim of this project is to develop a method capable of taking into account the co-dependencies in gene expression in the form of a gene network, that works on non-model species in the wild. Essentially, such a method should be capable of telling which group of genes expressed in what manner results in which phenotype. It should also be powerful enough across the noise of different sites and lineages. This is ambitious, however, and we will settle, in the first instance, for being able to group genes into sets of co-expressed genes, and then tell which sets are important for the differences between certain phenotypes.

An additional challenge of studying wild species is that there will always be an effect of environment, since we do not have the ability to do controlled experiments, as in the laboratory. In particular, if the samples we gather are from different locations, there will be the effect of local adaptation and genetic drift to take into account. This is important, since any data analysis method for tackling gene expression in an ecological context will have to deal with these issues.

In the following section, we will describe a species of fish, the Arctic charr, which we will study. This will be a test for whatever model we come up with, since it is a non-model species in the wild, with well-characterised variation depending on the part of the lake the fish inhabits.

3.2 A Natural Model

We want an ecological and evolutionary model which will help answer a major outstanding question in the field of ecology, which is the molecular basis of how diversification occurs rapidly. With this consideration, we pick the Arctic charr (*Salvelinus alpinus*), a species belonging to a family of fish called the Salmonids. These are famous for their ecological and evolutionary diversity, and their rapid evolution. This is hypothesised to be the result of a whole genome duplication in the common ancestor of all Salmonids [Rob+17]. Whatever the reason, this diversity provides us good test species for finding patterns in 'omics data relevant to the processes of ecological divergence and speciation. In particular, there are fish species in the Salmonid family which occur in benthic and pelagic morphs in the same lake, with Arctic charr being an example of such [Kle10].

[1] There are also still issues with the short ($\sim$ 100 bp) lengths of the reads and errors.

3.2.1 Morphs

Morphs are a sub-species distinction based on phenotype. An example is the black and white morphs of the peppered moth (*Biston betularia*), with the black morph appearing after the Industrial Revolution in England in the 19$^{\text{th}}$ century [CS13]. *Ecomorphs* are morphs based on correlation with specific habitats. In the text, we will use morph and ecomorph interchangeably. In the context of Arctic charr, there are two morphs of particular interest. These are the benthic and pelagic morphs, as mentioned above. Commonly, there will be Arctic charr of both morphs in the same lake. In fact, these morphs also occur amongst some other lake-dwelling Salmonids, and even more generally in other fish species, like cichlids. This gives the hope of generalising our results to those fish which also possess this axis of diversification. We will now proceed to a description of these morphs.

The benthic morph dwells near the lake bottom and feeds on benthos (organisms which live near or on the lake bottom such as bloodworms) while the pelagic morph dwells in open water and feeds on plankton. Differences between the morphs depend on lake, but we can describe some typical traits. The benthic morphs are usually smaller than the pelagics, and have a less streamlined body shape, since they are not as adapted to swimming in open water. Jaw and head shape differ too, mostly due to adaptation to diet and foraging. For example, the pelagic morphs have more and finer gill-rakers than the benthics, since they catch plankton by filter-feeding, while the benthics dig around at the lake bottom. Figure 3.1 shows the benthic and pelagic morphs in Loch Tay.

The occurrence of ecomorphological parallelism across independent lakes may indicate the beginnings of the process of speciation, at the very least, it shows that local adaptation is taking place. This is an ongoing area of research, and its mechanisms are still unclear [Elm+10].

3.3 Data

In this section, we will give a description of the biological and transcriptomic data, as well as a brief outline of the transcriptomic pipeline.

3.3.1 Samples

The biological samples we have are from 36 wild caught individuals of Arctic charr, comprised mainly of benthic and pelagic morphs from four lakes. A list of the morphs and lakes is shown in Table 3.1. We have four individuals for each lake/morph combination. Figure 3.2 shows the locations of the lakes, and Figure 3.1 shows exemplars of the two morphs in Loch Tay. These lakes were chosen since Arctic charr of both benthic and pelagic morphs are present in each lake, which gives us the opportunity to find parallelism between the four different lakes along this axis of diversification [Jac+19].

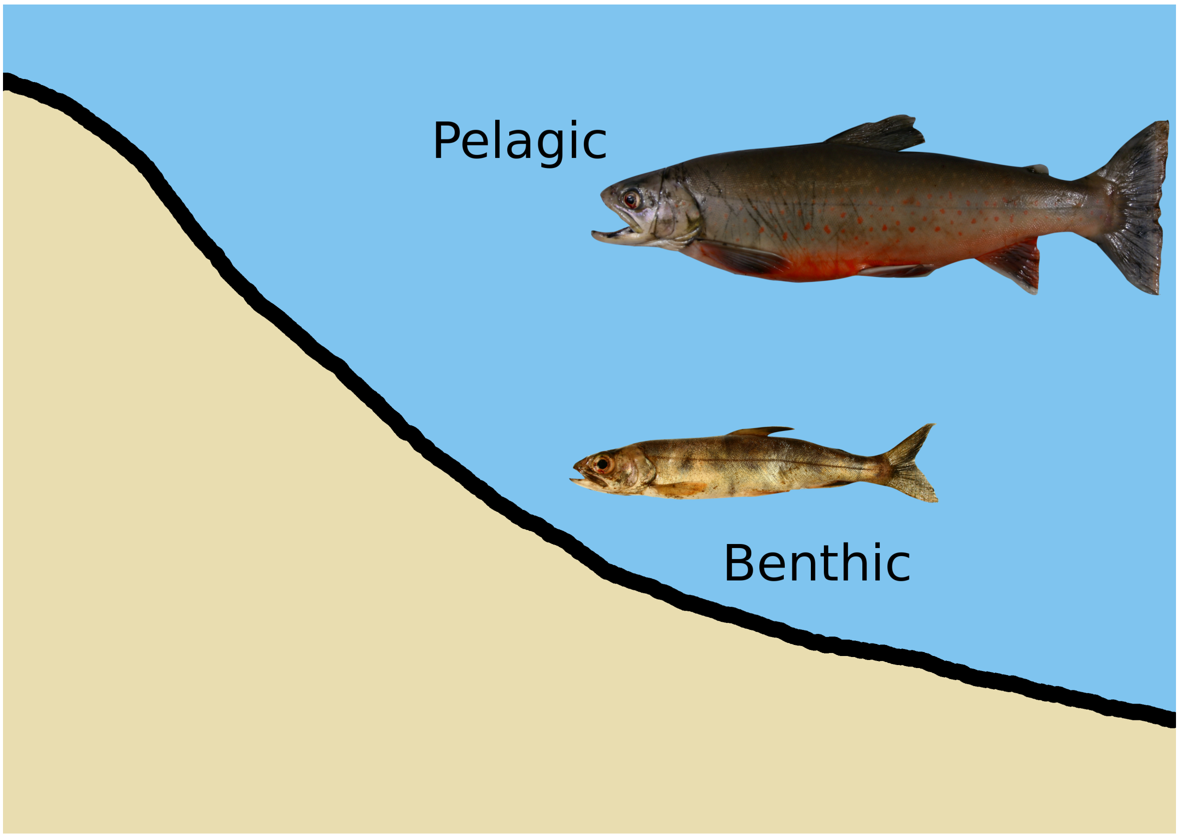


Figure 3.1: This figure shows the two morphs in Loch Tay. The pelagic is the larger one dwelling in open water, while the benthic is smaller dwelling near the bottom. Note differences in head and body shape, as well as size have been examined in other studies [Jac+19].

Lake/Loch	Name	Ecomorph
Awe	Autumn	Pelagic
	Spring	Benthic
Dughaill	Benthic	Benthic
	Pelagic	Pelagic
Tay	Large	Pelagic
	Small	Benthic
Kamkanda	Dwarf	Benthic
	Large	Piscivorous
	Small	Pelagic

Table 3.1: A table giving the lakes and lochs the Arctic charr have been sampled from, as well as their name and corresponding ecomorph. Note that we refer to the fish as benthic or pelagic morphs throughout the text, rather than by their lake-specific name.

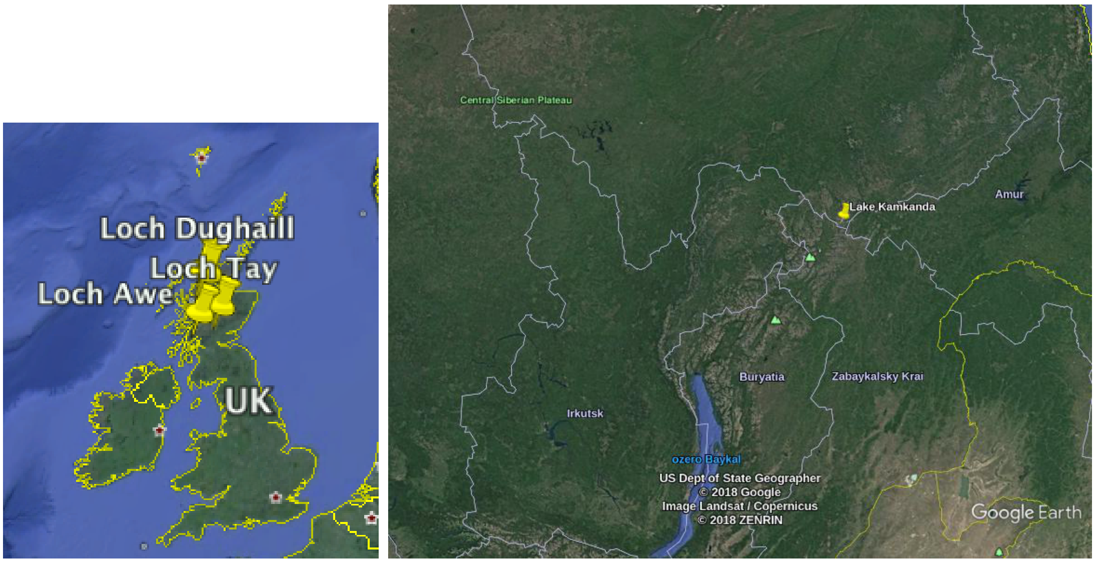


Figure 3.2: This figure shows the locations of the lochs and lakes the Arctic charr have been sampled from. We are using samples from the three lochs in Scotland, and Lake Kamkanda in Russia.

The key point here is that, if we do find parallelism, then the parallel aspects of these morphs will indicate genetic regions and mechanisms behind the emergence of these morphs in general. Furthermore, these same regions and mechanisms could be involved in the beginnings of ecological speciation, at least along the benthic/pelagic axis.

3.3.2 Transcriptomics

In this section, we will give a brief outline of the methods used for obtaining transcriptomic data from the Arctic charr.

The data are derived from RNA extracted from white muscle tissue in 36 Arctic charr, four individuals from each category in Table 3.1. The RNA was prepared and sequenced by Glasgow Polyomics in an Illumina NextSeq 500 machine, outputting about 20–30 million paired-end 2 × 75bp reads for each sample. The *.fastq* files were aligned to an Arctic charr *de novo* transcriptome with 33,126 transcripts prepared by

Madeleine Carruthers and Andre Yurchenko from the Elmer lab [Car+18]. Alignment was carried out in January 2017 using Bowtie2 (Galaxy Version 2.2.6.2) [LS12] on Glasgow University's Galaxy server [Afg+16] with options *paired-end alignment, local, no-mixed, no-discordant, -a (no ceiling)* and the rest defaults. Please see [Car+18] for more details on the approach followed for the RNA sequencing.

Transcript cluster counts were calculated with Corset (Version 1.06) [DO14], defining nine groups of four individuals each (one group for each category in Table 3.1). This resulted in a count table, with a total of 19,015 different transcript clusters with at least 10 reads over all 36 individuals. The count table was input into R (version 3.3.2) [R C17] and normalised using the regularised logarithm transformation function in the DESeq2 (version 1.14.1) [LHA14] R package, with one group for each morph/lake.

The rationale behind transforming the raw counts is to obtain expression values which are approximately normally distributed. In raw count data, highly expressed and highly variable genes will end up dominating the results. The log transformation takes this into account by considering variation relative to expression [BC64].

3.3.3 Mathematics

In this section, we show how we can consider the gene expression dataset mathematically as vectors/points in a gene expression space. We furthermore give a description of how to incorporate categorical information as part of our dataset.[2]

Suppose that we have processed the output of our transcriptomics pipeline into a gene expression dataset as a set of n vectors, $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, in $\mathbb{R}^m$, where n is the number of samples and m the number of genes. We may form this into a data matrix as follows:

Definition 3.3.1 (Data Matrix)**.** Let A be an $n \times m$ matrix, with the rows given by vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ in $\mathbb{R}^m$, corresponding to gene expression values of n samples given over m genes.

This also allows us to find relations between genes by considering the m column vectors $\{\mathbf{g}_1, \ldots, \mathbf{g}_m\}$ in $\mathbb{R}^n$.

In all cases, our samples will have derived from some experiment, so they will possess phenotypic, and other, properties. It is also the case that we may have properties, such as annotations, for our genes. Since we are interested in the replicate populations of non-model organisms which have undergone the same phenotypic diversification, that is, we see the same morphs (which can be thought of as subpopulations) in different locations, we will restrict our phenotypic properties to two nested levels. The top level will be the *locations*, and the second level will be the *morphs*, which would ideally be the same across all locations. In both cases, we will be dealing with categorical variables. An example of this phenotypic information is in Table 3.1, where the lakes are the

[2] Note: We use *categorical* in contrast to *numerical* in statistical language, which means incorporate properties which are described by a finite number of distinct labels/categories.

locations, and the ecotypes are the morphs, with the exception that Lake Kamkanda has an additional piscivorous ecotype. Furthermore, we will assume that we have no reliable gene annotations, since we are dealing with a non-model species.

Definition 3.3.2 (Nested Subsets)**.** We assume that we have groupings associated to our n samples. Firstly, we assume we have a partition of our samples into *locations*. Secondly, we assume we have partitions of the samples at each location into *morphs*.

Note that all these ways of partitioning our samples are done via their properties, and this is what we mean by categorical data.

3.4 Analysis

In this section, we will use our data, in the form of vectors in Euclidean space (36 vectors in $\mathbb{R}^{19,015}$) to proceed with some standard analyses. We do some processing, and visualise the data with PCA and heatmaps. This lets us visualise the data, and is a first pass, to see if we have any unusual outliers or unexpected groupings. Furthermore, we can check if there is parallelism apparent between the benthic and pelagic morphs in the different lakes. This will turn up on our PCA plots as the difference between the benthic and pelagic morphs being in approximately the same direction in each lake. This has been apparent in previous cases where parallel evolution has been found, such as in Midas cichlids of Nicaragua [Elm+14, Fig. 3].

3.4.1 Lake Effect

First, we note that the data can be grouped by lake, as in Table 3.1. We can see in Figure 3.4, and we know from previous research, that there is a large effect of lake on the gene expression of the Arctic charr. We must take this into account in order to focus on the pervasive signal of the differences between the morphs. We can do this by taking the barycentre (centroid) of the gene expression of all the fish in that lake as an estimate of the expression of the fish in that lake. For example, if we have the expression vectors of the eight fish in a lake $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_8\}$, then their barycentre is $\bar{\mathbf{x}} = \frac{\sum_{i=1}^{8} \mathbf{x}_i}{8}$. A plot of the centroid for the Arctic charr is given in Figure 3.3.

Once we have the lake effect, we can subtract it from the gene expression of all our fish samples, and continue our analysis. If we consider the example with eight fish above, this gives us the vectors $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_8\}$, where $\mathbf{y}_i = \mathbf{x}_i - \bar{\mathbf{x}}$. This procedure can be repeated for any number of mutually exclusive subgroups.

3.4.2 Visualisation

Visualising data allows us to inspect it for any obvious groupings or patterns. This serves two purposes. First, we can check if, in our rough 2D visualisation, we see the grouping of data that we expect, for example, into different conditions. Second, we can

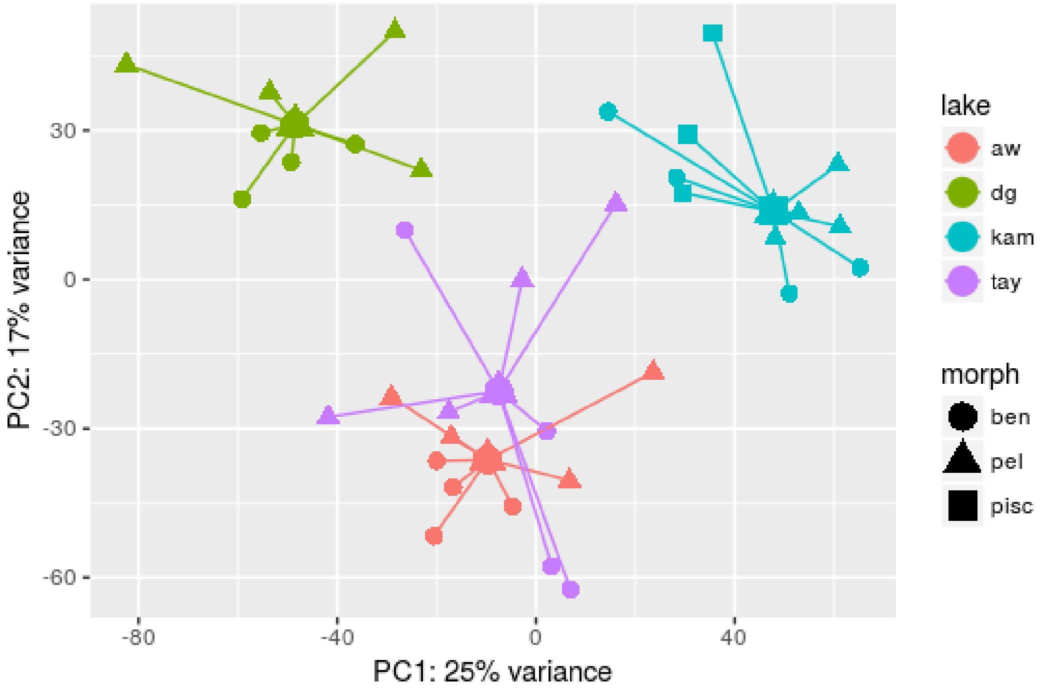


Figure 3.3: An example of the finding the centroid for the Arctic charr. Here we have only plotted their first two principal components. The coordinates of their centroid are given by the average over each coordinate. In this case, it's the large centre point joined to all the samples from the lake by a line.

see if there are other groups in the data which we have not taken into consideration, for example, a subgroup in one of our conditions, or unusual outliers.

There are a few standard ways to visualise data. We first go through two of these methods, principal component analysis (PCA) and heatmaps, and show their results on our Arctic charr gene expression data. In the next chapter, we will follow up by using the Mapper method.

Principal Component Analysis

Principal Component Analysis (PCA) is a method of dimension reduction. PCA uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. This transformation is defined in such a way that the first principal component has the largest possible variance, and each succeeding component has the highest possible variance under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set. In more technical terms, we can consider PCA as a change of basis to a new coordinate system, where the first coordinate accounts for the greatest variance, the second coordinate for the second greatest, and so on [BCV13, §5.1].

To illustrate this, let X be a matrix where the columns are samples, and the rows are shifted to have mean zero.[3] In PCA, we wish to transform X to a new coordinate

[3] This simplifies matters, since it allows us to use the sum of squares to calculate the variance,

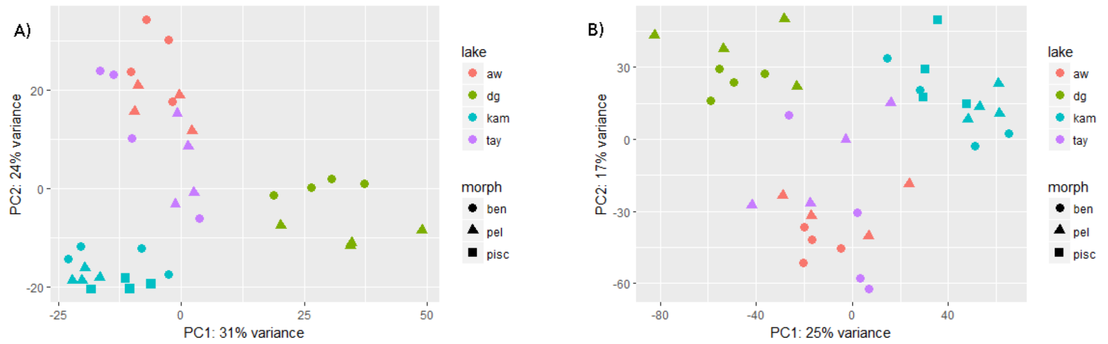


Figure 3.4: **A)** Showing the first two PCs of PCA applied to the 500 genes with greatest variance of the regularised log-transformed Arctic charr transcriptomic data. Notice the broad division into Kamkanda (Russian), Dughaill (North Scottish), and Tay and Awe (more Southerly Scottish). **B)** Showing the first two PCs of PCA applied to all genes of the regularised log-transformed Arctic charr transcriptomic data. Note that the axes are different between the two pictures, however, the overall groupings are maintained.

system $T = WX$ where W is the transition matrix and the columns of T correspond to sample coordinates in the new system.

The rows of W are given by unit vectors which maximise the variance in each coordinate (row) of T, so the first row of W maximises the variance in the first coordinate, the second row maximises the remaining variance in the second coordinate, and so on. In fact, the rows of W are given by eigenvectors of the matrix XX^T, the covariance matrix of X, ordered by their eigenvalue.

Figure 3.4 gives a visualisation of the regularised logarithm transformed Arctic charr transcriptomic data. It is a PCA plot where the left diagram is given on the top 500 varying genes only, while the right uses all 19,015 genes. Figure 3.5 has the lake centroids subtracted. Note that the lake effect (§3.4.1) has been lost, but there is still some indication of separation by morph.

Heatmaps

A heatmap (or shading matrix) gives a visualisation of the values in a matrix. In our case, the heatmap is produced from a matrix of Euclidean distances between the Arctic charr, calculated on the basis of the gene expression values of either the 500 genes with the most variance, or all 19,015 genes.

Some heatmaps have the dendrogram of a hierarchical clustering algorithm on their axes. This is the case with our heatmaps, which have the dendrograms from the single-linkage clustering algorithm.

Figure 3.6 gives a visualisation of the regularised logarithm transformed Arctic charr transcriptomic data. It is a heatmap of the distances, with single-linkage clustering. The left diagram is given on the 500 genes with the most variance, while the right

rather than having to shift every coordinate by its mean first. Note, we can also normalise each row to have variance one, in which case XX^T is the correlation, rather than the covariance, matrix in the following paragraph.

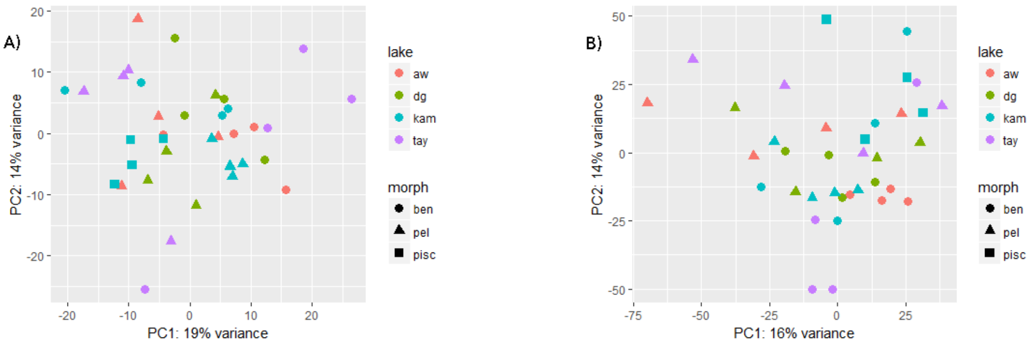


Figure 3.5: **A)** Showing the first two PCs of PCA applied to the top 500 varying genes of the regularised log-transformed Arctic charr transcriptomic data, with lake centroids subtracted. Note that now most of the benthic fish are to the top right, compared to their lake's respective pelagic morphs. **B)** Showing the first two PCs of PCA applied to all genes of the regularised log-transformed Arctic charr transcriptomic data, with lake centroids subtracted. Note that now most of the benthic fish are to the bottom right, compared to their lake's respective pelagic morphs.

uses all genes. Figure 3.7 gives the same visualisation, only with the lake centroids subtracted.

3.5 Discussion

We have provided the background to our Arctic charr gene expression data, as well as some analysis methods, primarily focussing on accounting for the effect of lake, and visualisation. The PCA plots (Figure 3.5) suggest some parallelism in morphs between lakes, but heatmap outputs show us groupings based on lake, with a few distinguished morphs when we subtract the lake effect. No unexpected groupings are apparent. In the next chapter, we will take a look at Mapper applied to this Arctic charr gene expression data.

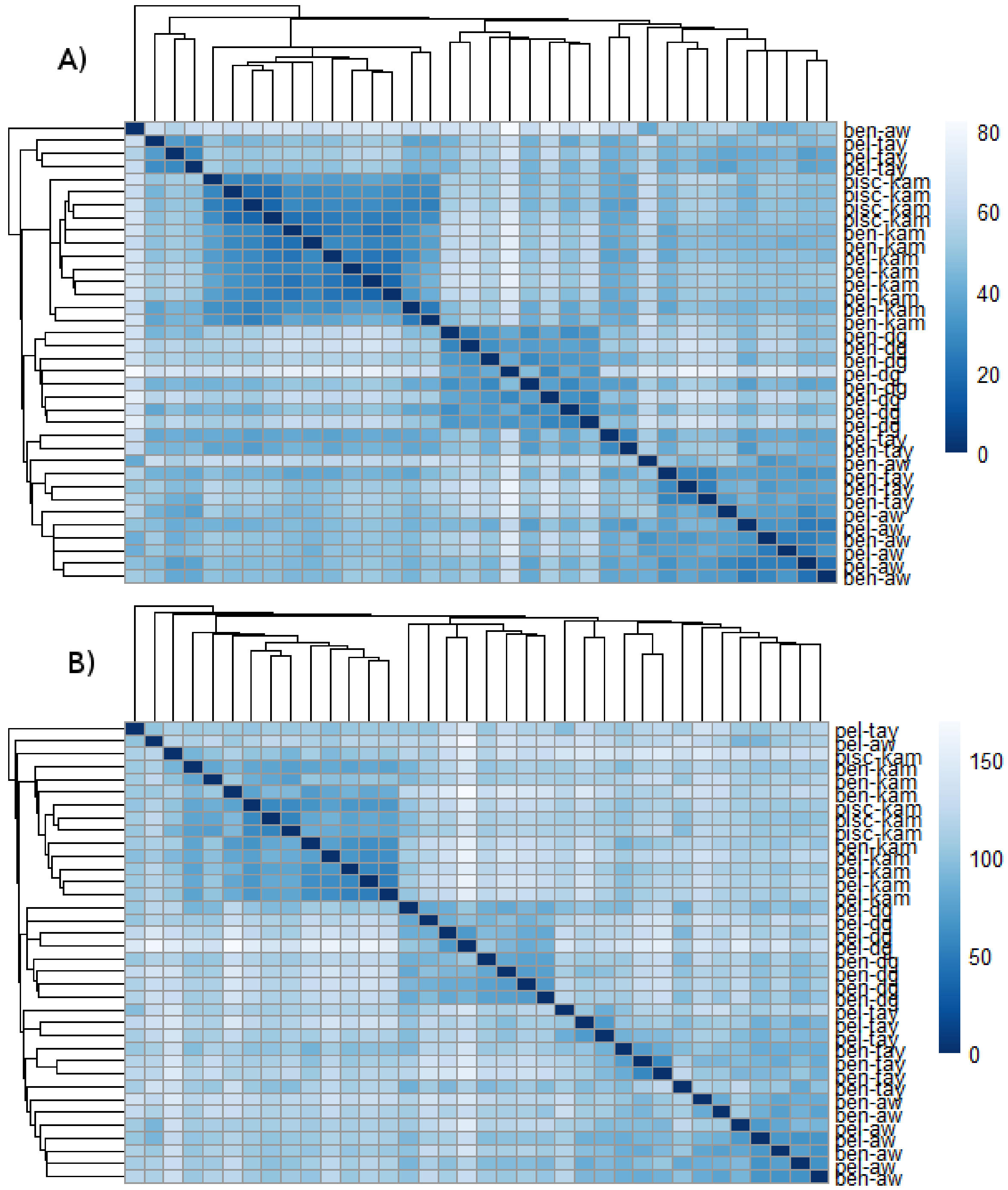


Figure 3.6: **A)** Showing a heatmap of Euclidean distances, along with single-linkage clustering, on the 500 genes with the most variance of the regularised log-transformed Arctic charr transcriptomic data. Notice the broad division into Kamkanda (Russian), Dughaill (North Scottish), and Tay and Awe (more Southerly Scottish). (The groups are apparent as squares of darker coloured blocks in the heatmap, indicating that all members of that square have a shorter distance to each other than other members.) There are some outliers from Tay and Awe located near the top of the diagram. **B)** Showing a heatmap of Euclidean distances, along with single-linkage clustering, on all genes of the regularised log-transformed Arctic charr transcriptomic data. Again we see grouping into Kamkanda (Russian) and Dughaill (North Scottish), with the Tay and Awe (more Southerly Scottish) group being less distinct. Again, there are outliers from Tay and Awe near the top of the diagram, though they're different from (**A**).

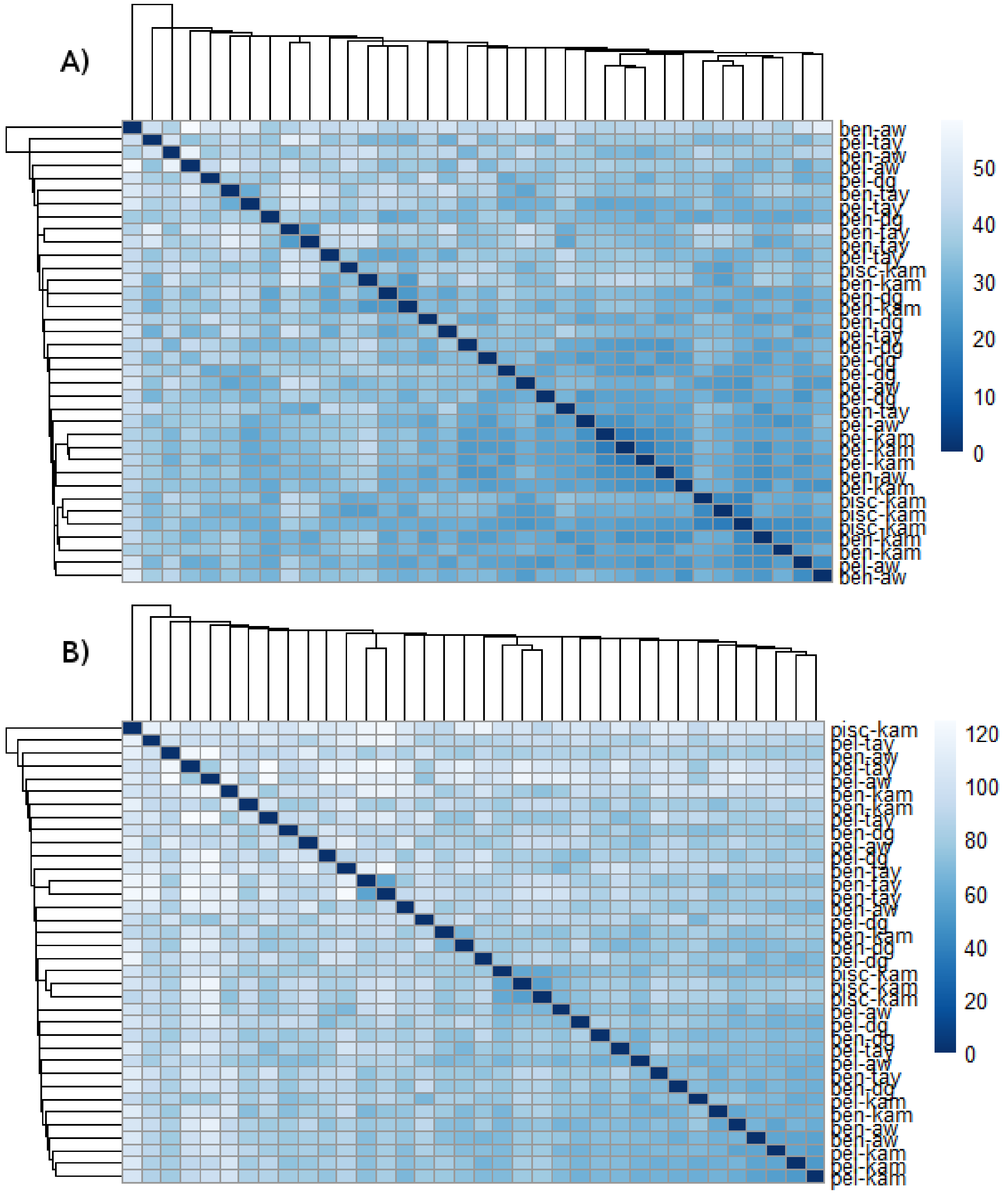


Figure 3.7: **A)** Showing a heatmap of Euclidean distances, along with single-linkage clustering, on the top 500 varying genes of the regularised log-transformed Arctic charr transcriptomic data, with lake centroids subtracted. Much of the lake structure has been lost, while groupings by morph are not apparent. There are still some notable groupings by lake and morph, such as the Kamkanda piscivorous (large morph) and the Kamkanda pelagics (small morph). Note the change in scale from Figure 3.6. **B)** Showing a heatmap of Euclidean distances, along with single-linkage clustering, on all genes of the regularised log-transformed Arctic charr transcriptomic data, with lake centroids subtracted. Again, the groupings are no longer distinct. Like (**A**), a couple of notable groups still remaining are the Kamkanda piscivorous (large morph) and the Kamkanda pelagics (small morph).

Chapter 4

Mapper Applied to Arctic Charr

We have seen from the previous chapters of Mapper applied to breast cancer gene expression data (§2.2), that it is possible to use it to find a new unexpected grouping in gene expression data, and that further investigation of the group showed how it is clinically and biochemically distinct. Inspired by this example, in this chapter, we now port the method to gene expression data from a non-model species.

In the first section, we will describe how we applied Mapper to the Arctic charr gene expression data, and the resulting visualisations. We show the effect of changing the preprocessing, and the Mapper parameters. Finally, we settle on a simpler version of preprocessing the data, using only the Euclidean distance, rather than the more advanced Pearson correlation based on DSGA. In this case, we expect to see at least two subgroups, showing the benthic and pelagic morphs. However, we still produce no clear subgroups in our Mapper output. We suspect that the small sample size is to blame for us not finding any significant properties in our Arctic charr gene expression data. So, in the final section, we make an investigation of the sample size required to show the new group in the breast cancer expression data set of [NLC11; Lum+13], which we went through in §2.2. Subsampling this reproduction suggests that we need about two hundred samples before we can find a subgroup of interest.

4.1 Varying Mapper

We varied the Mapper parameters over many trials to find the optimal way of displaying the data. This involved changing both methods of preprocessing the Arctic charr gene expression data, as well as the parameters used by the Mapper algorithm. The main goal here is to find a visualisation which will highlight a novel subgroup, or some other interesting feature, of the Arctic charr data.

4.1.1 Pre-processing

We initially applied a procedure similar to Disease-specific genome analysis (DSGA) [Nic+07] (see §2.2.2). Instead of trying to account for the variation in the normal tissue (as the distinction was between normal and breast cancer tissues), we used DSGA to account for the difference among the pelagic morphs or the benthic morphs, respectively. The idea was that, since each morph was found over all four lakes, if we accounted for the variation in a single morph, we would also account for the variation between the lakes.

Figure 4.1 shows the Mapper output for the Arctic charr gene expression data, with DSGA applied with the benthics as the 'normal' group. One of the benthic morphs from Dughaill has been removed as an outlier, since it would have great influence on what is considered 'normal' benthic expression. The colour is by the filter function, which can be considered as a distance from 'benthic-ness', similar to how the original use of DSGA in breast cancer used a distance from normality. Blue is low and red is high. As expected, we see most of the benthic morphs grouped in the large blue node to the right, since we have accounted for most of their variation using DSGA. On the other hand, no real pattern is found in the pelagic morphs, they are scattered all throughout the rest of the Mapper output. They do not even concentrate by lake (except for Kamkanda, whose pelagics are found to the left of the output).

Figure 4.2 gives Mapper output with the pelagics as the 'normal' group. This single change leads to quite a difference in the outputs, since this one seems to lack any structure whatsoever. However, this is not stable with regard to parameters, since a small change (not shown here) causes the output to change into a single cluster with individual points jutting out. We now move on to looking at the effect of parameter changes.

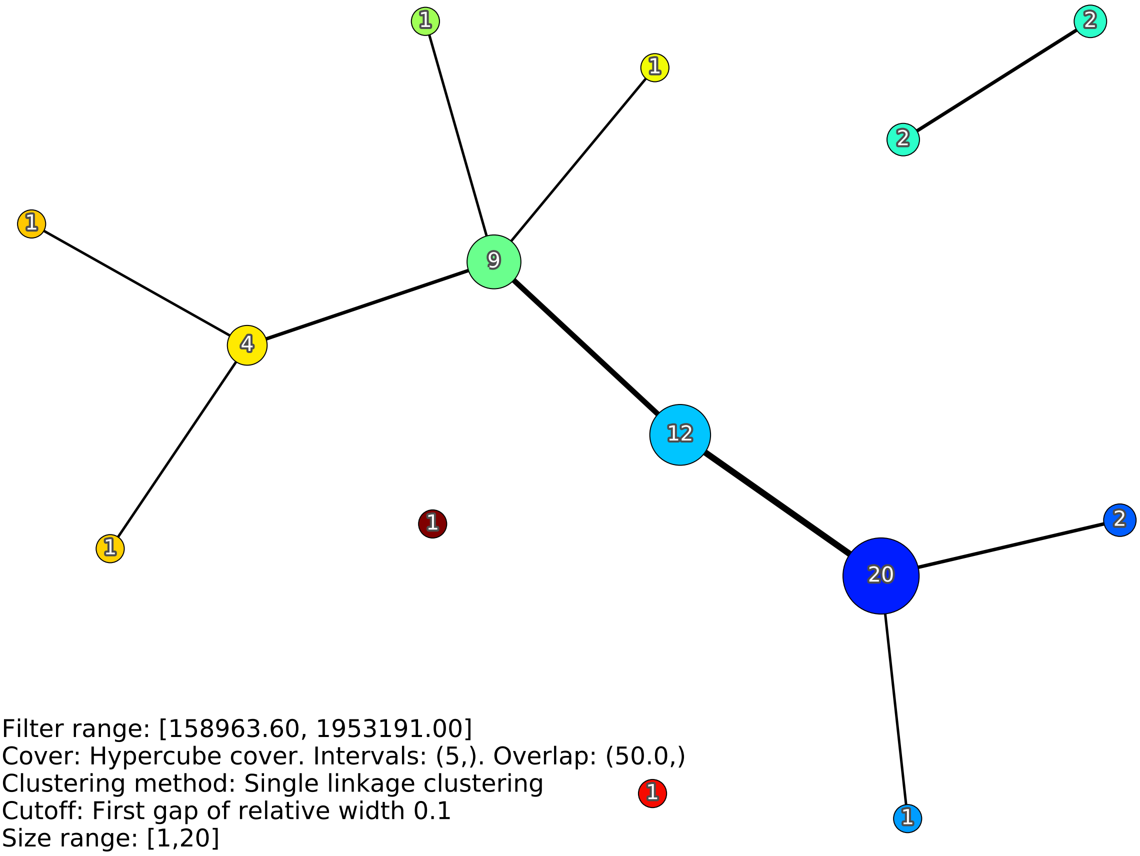
We see that preprocessing to make the benthics 'normal' shows up no patterns in the pelagics, while making the pelagics 'normal' gives lots of isolated small clusters. We now proceed to testing how changing Mapper parameters changes the output.

4.1.2 Parameter Changes

In this subsection, we will look at the effect of changing parameters on the Mapper output. We will do this with the Arctic charr gene expression data, with DSGA applied with the benthics as the 'normal' group. This is because we observe output with more structure than having pelagics as the 'normal' group, so the effect of changing parameters will be more noticeable.

First we vary the filter function. Instead of using the Euclidean distance (the L^2-norm) from 'benthic-ness', as in Figure 4.1, we instead use the L^p-norm raised to some power m, where $p, m \in \mathbb{N}$. In Figure 4.3, we have used the L^5-norm on the left, and the L^2-norm squared on the right. The effect of these different norms is to weight the dimensions differently when calculating distances. Higher values of p give more weight to dimensions with greater differences, while raising to some power changes how the points are covered in the reference space.

Next, we look at varying the number of intervals used to cover the reference space. The effect of varying the number of intervals is to increase or decrease the number of



Filter range: [158963.60, 1953191.00]
Cover: Hypercube cover. Intervals: (5,). Overlap: (50.0,)
Clustering method: Single linkage clustering
Cutoff: First gap of relative width 0.1
Size range: [1,20]

Figure 4.1: Mapper output on data pre-processed with DSGA, where the benthics were treated as the ‘normal’ group. Covering was by five intervals with 50% overlap, and clustering was single-linkage with a cut-off where there is a gap of 10% of the range. One of the benthic morphs from Dughaill has been removed as an outlier. The colour is by the filter function, which can be considered as a distance from ‘benthic-ness’, similar to how the original use of DSGA in breast cancer used a distance from normality. Blue is low and red is high. Most of the benthic samples are in the large blue-coloured node, and the rest of the pelagic samples are scattered throughout. No grouping by lake is apparent.

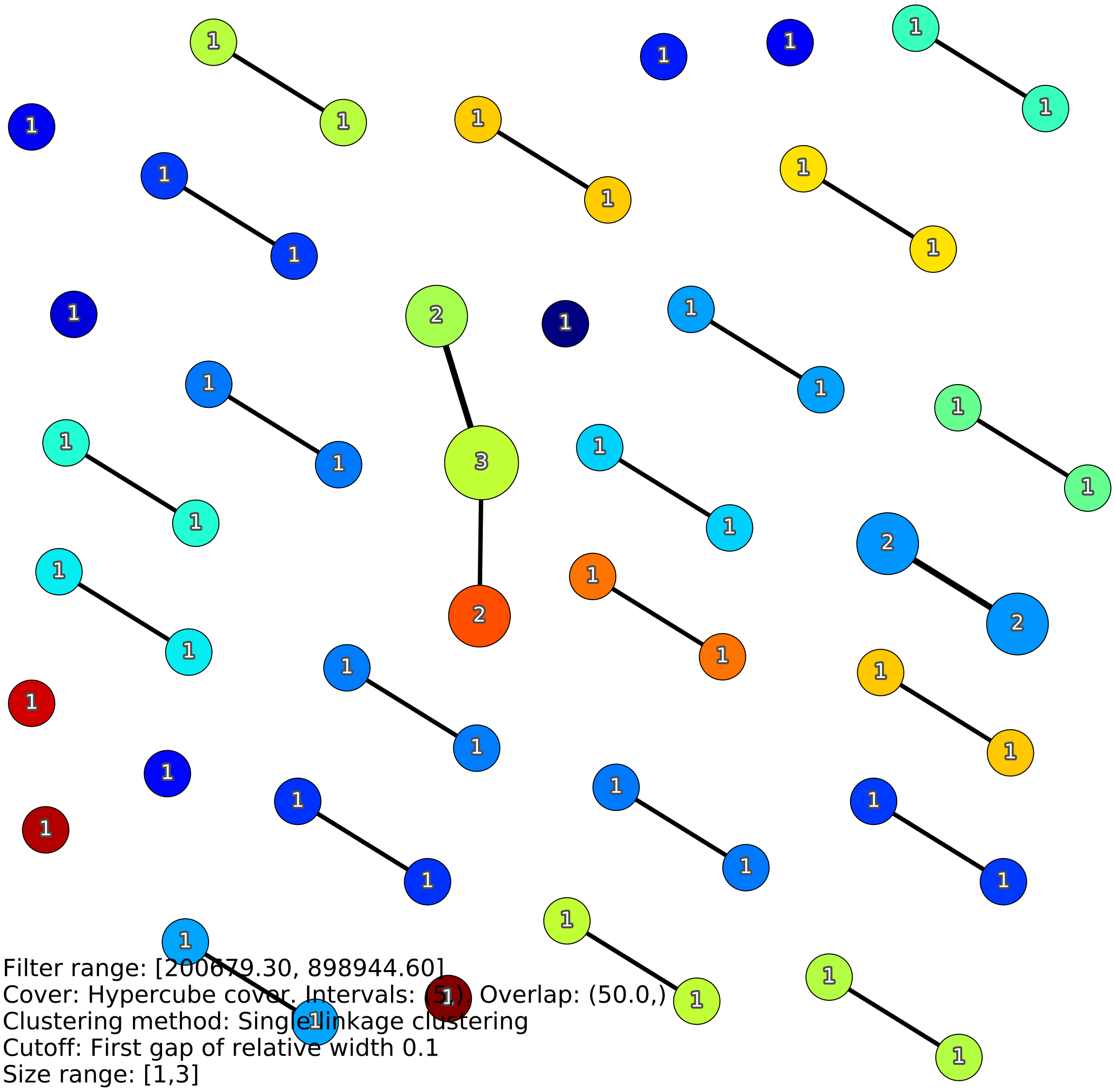

Filter range: [200679.30, 898944.60]
Cover: Hypercube cover. Intervals: (5,) Overlap: (50.0,)
Clustering method: Single linkage clustering
Cutoff: First gap of relative width 0.1
Size range: [1,3]
1
2
3

Figure 4.2: Mapper output on data pre-processed with DSGA, where the pelagics were treated as the 'normal' group. Covering was by five intervals with 50% overlap, and clustering was single-linkage with a cut-off where there is a gap of 10% of the range. One of the benthic morphs from Dughaill has been removed as an outlier. The colour is by the filter function, which can be considered as a distance from 'pelagic-ness', similar to how the original use of DSGA in breast cancer used a distance from normality. Blue is low and red is high. We see that the output is incredibly scattered. However, this is not stable with regard to parameters, since a small change (not shown here) causes the output to change into a single cluster with individual points jutting out.

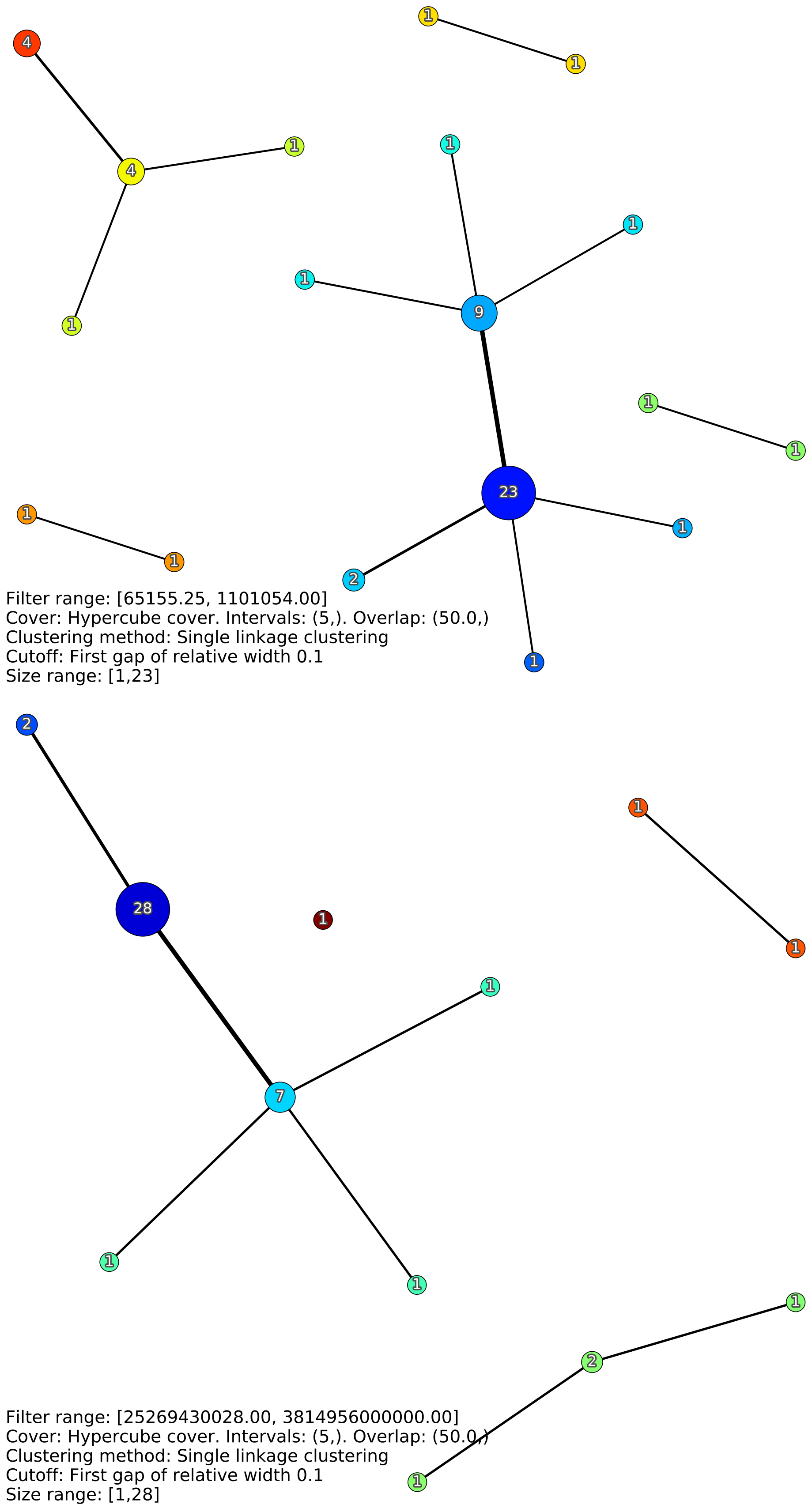

Filter range: [65155.25, 1101054.00]
Cover: Hypercube cover. Intervals: (5,). Overlap: (50.0,)
Clustering method: Single linkage clustering
Cutoff: First gap of relative width 0.1
Size range: [1,23]
Filter range: [25269430028.00, 3814956000000.00]
Cover: Hypercube cover. Intervals: (5,). Overlap: (50.0,)
Clustering method: Single linkage clustering
Cutoff: First gap of relative width 0.1
Size range: [1,28]

Figure 4.3: Varying filter function. We have used the L^5-norm on the top, and the L^2-norm squared on the bottom. In both cases, the Mapper output graph is more disconnected when compared to Figure 4.1. The bottom is not as scattered as the top. In this case, the filter function affects the spread of the samples, so the preimages of the intervals contain different samples, and these cluster differently.

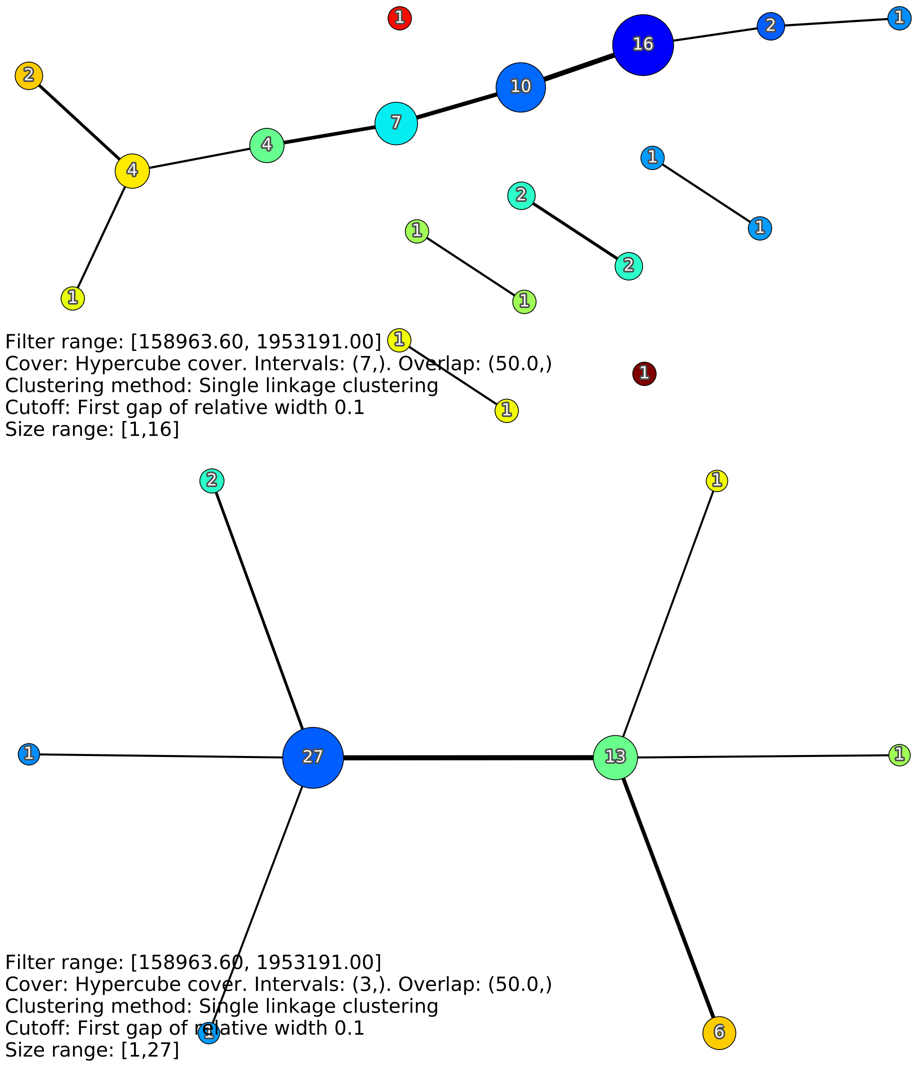

1
2
16
2
1
10
7
4
4
1
2
1
2
1
1
1
Filter range: [158963.60, 1953191.00]
Cover: Hypercube cover. Intervals: (7,). Overlap: (50.0,)
Clustering method: Single linkage clustering
Cutoff: First gap of relative width 0.1
Size range: [1,16]
1
1
1
2
1
27
13
1
1
6
Filter range: [158963.60, 1953191.00]
Cover: Hypercube cover. Intervals: (3,). Overlap: (50.0,)
Clustering method: Single linkage clustering
Cutoff: First gap of relative width 0.1
Size range: [1,27]

Figure 4.4: Varying intervals. Compared to Figure 4.1, with 5 intervals, we see more nodes in the Mapper output on the top, with 7 intervals, and fewer on the bottom, with 3 intervals. Note that with more intervals, the graph also becomes more disconnected, and we have a higher number of isolated points.

nodes in the Mapper output, as can be seen in Figure 4.4.

Finally, we look at changing the clustering function used by Mapper. The clustering function influences how the clusters (nodes of the Mapper output) are produced from the preimage of the cover on the reference space. At one extreme, we get one node for each preimage (where all points are assigned one cluster), at the other, we get one node for each point. Figure 4.5 shows how varying the clustering function varies the number, size and connectivity of the nodes.

So, we have seen that changing the parameters and pre-processing can have quite a large effect on the Mapper output. The pre-processing in particular can lead to a total change in the output, while varying the filter function, covering, and clustering function have more subtle effects. As a reminder, the goal was to find a combination of parameters which produces interesting and informative output, while remaining robust to parameter changes.

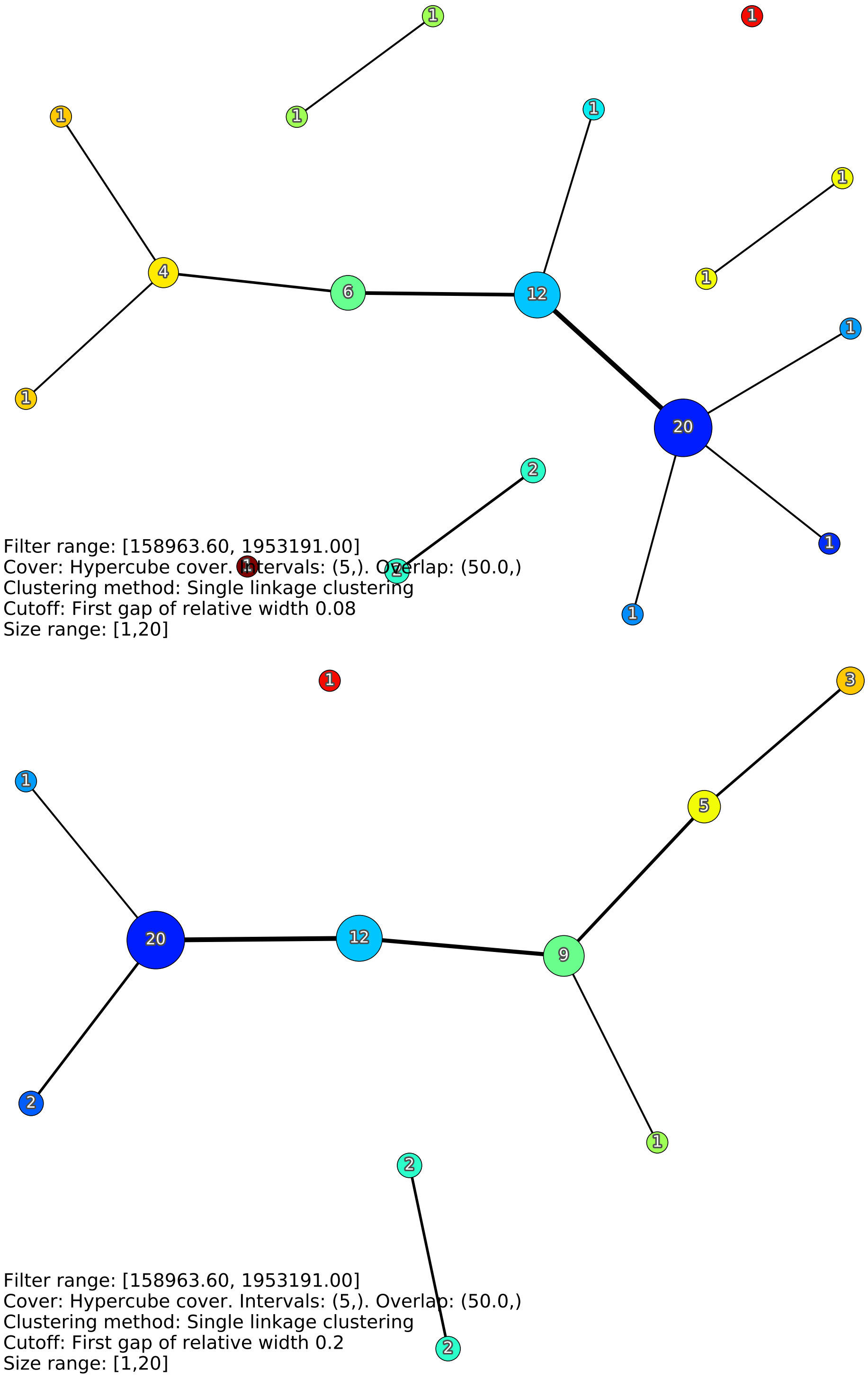

1
1
1
1
1
4
6
12
1
1
1
20
2
2
1
1
1
Filter range: [158963.60, 1953191.00]
Cover: Hypercube cover. Intervals: (5,). Overlap: (50.0,)
Clustering method: Single linkage clustering
Cutoff: First gap of relative width 0.08
Size range: [1,20]
1
3
1
5
20
12
9
2
1
2
2
Filter range: [158963.60, 1953191.00]
Cover: Hypercube cover. Intervals: (5,). Overlap: (50.0,)
Clustering method: Single linkage clustering
Cutoff: First gap of relative width 0.2
Size range: [1,20]

Figure 4.5: Varying clustering function. Compared to Figure 4.1, with a cut-off at 10% of the range, we see more nodes in the Mapper output on the top, with a cut-off at 8%, and fewer on the bottom, with a cut-off at 20%.

Finally, we have seen that preprocessing to make the benthics 'normal' shows up no patterns in the pelagics, while making the pelagics 'normal' gives lots of isolated small clusters. However, this seems to be very sensitive to changing the parameters, i.e. covers and clustering function, used to produce the Mapper output, and consequently we cannot say that this difference in preprocessing is significant.

Subsequently, we decided to use simpler distances and filter functions (Euclidean distances and norms) to visualise our data Arctic charr gene expression data. Partly, this is because when Lum *et al.* reapplied Mapper to the breast cancer data set [Lum+13], they found success using a simpler distance and filter function (§2.3.2).

4.2 Euclidean Distance

In this section we will show a couple of visualisations of our Arctic charr gene expression data using the Mapper algorithm [SMC07], as described in §2.1. Since, in the previous section, we were unable to identify any significant patterns, we will now aim instead to reproduce structure in the Mapper output which we know should already be in the data. Namely, we want to get output where the split between the benthic and pelagic morphs, which we see in the wild, is apparent.

4.2.1 Data

The input data to Mapper is a point cloud (see Def. 2.1.14) of 36 points, $\mathbf{y}_i \in \mathbb{R}^N$, where $N = 19,050$ genes and the distances are given by the Euclidean metric on the lake centroid centred gene expression vectors. §3.3.2 describes the process for getting the coordinates of these points, from extracting the RNA molecules, to finally getting normalised read counts. These are the same points used for the PCAs and heatmaps in the previous chapter.

Filter Function

The filter function is given by the Euclidean norm of the vectors, i.e. a radial Morse function

$$\begin{aligned} f : \mathbb{R}^N &\to [0, \infty) \\ \mathbf{x} &\mapsto |\mathbf{x}| \end{aligned}$$

The motivation for choosing this filter function was the observation that if the fish are sufficiently concentrated at their lake and morph barycentres and the direction and magnitude of change are similar enough, implying parallelism, then the Mapper output would consist of two nodes, one for each morph. In the case where the direction and magnitude of change are sufficiently different, suggesting a different biological basis of the differentiation between the lakes, we should get more nodes, possibly as many as eight, one for each benthic/pelagic and lake combination.

Parameters

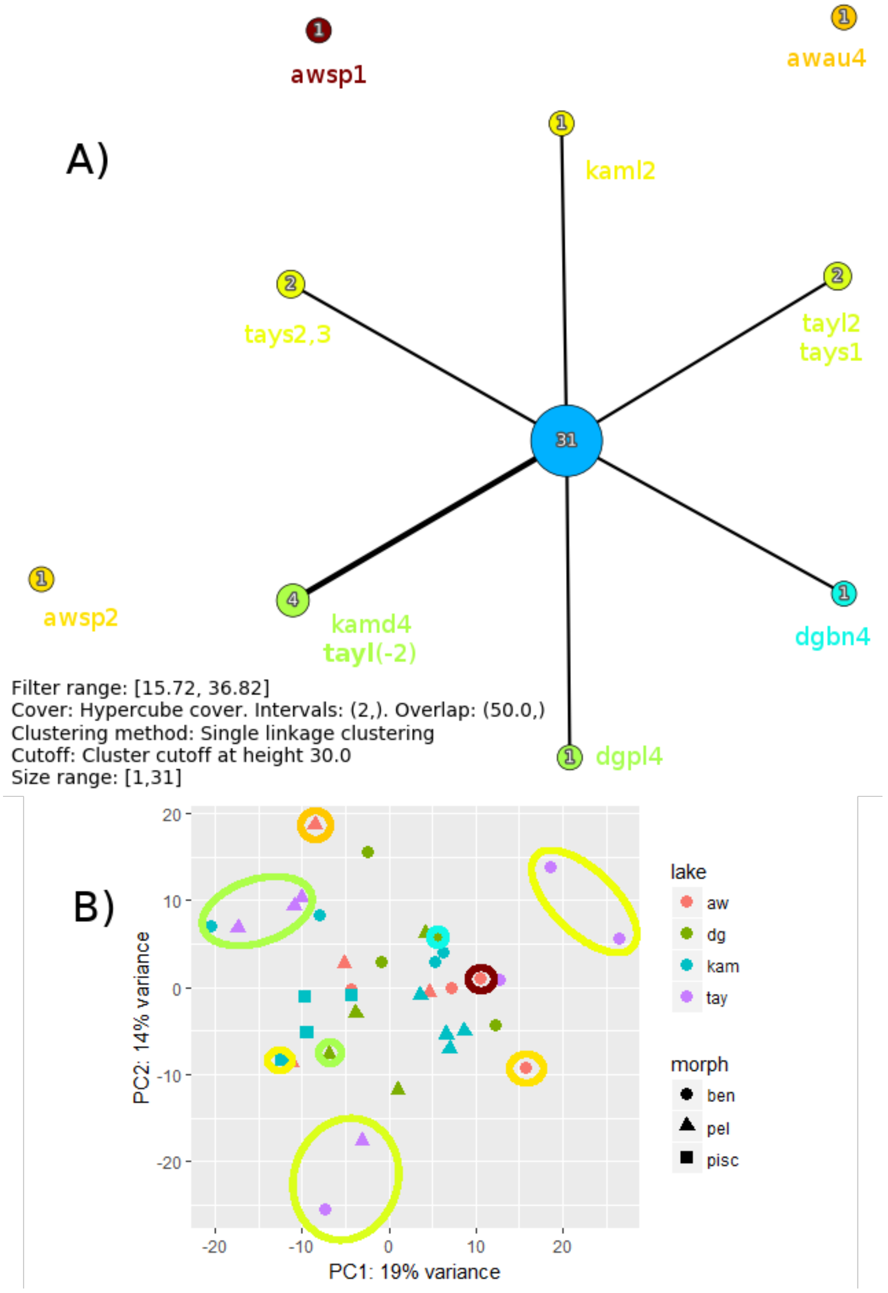
We utilise an open source Mapper algorithm called Python Mapper [MB13]. There are a few parameters which affect the Mapper output. The ones which we will change are, firstly, the number of intervals for the covering of the filter function, and, secondly, the cut-off point of the single-linkage clustering function. These are the last two items in Table 2.1, since the data points and filter function/reference map are input into the software. These are given by text files of pairwise distances, and filter function values, respectively.[1]

4.2.2 Visualisation

In the Mapper outputs, we find that the data are scattered away from the barycentres to such an extent that we see only one component, with branches representing points located particularly far from their lake centroid, and no particular groupings. Figures 4.6 and 4.7 show the Mapper output on the regularised logarithm transformed Arctic charr transcriptomic data, with lake centroid subtracted, using the top 500 varying genes and all genes, respectively. The Mapper outputs have the same overall shape, just with different samples in the outlying nodes, indicating that samples with unusual gene expression in only the top 500 varying genes, are not the same as those with unusual gene expression when considering all genes. In particular, there is no splitting into benthic and pelagic groups that we would expect. This lack is what will drive us to consider a new topological approach to the Arctic charr data in the next chapter.

Note that in the bottom-left of the Mapper output graphs, we see some text which gives the parameters used for the visualisation. The filter range gives the minimum and maximum values of the filter function, the cover type is always the ‘Hypercube cover’, which means that we cover by intervals with an identical range, the number of these intervals is given by the ‘Intervals:’ number, and the overlap by ‘Overlap:’ in percent. The clustering method is usually single-linkage clustering, and is what is used to cluster the points in the pre-images of the covers of the filter function, the cut-off is a description of how the clusters are chosen from the clustering function, either a cut-off at a certain height, or at a gap of relative width some fraction to the total height of the clustering tree. The size range gives the minimum and maximum number of points/samples in any one node.

[1] We should note here that it is currently quite difficult to get the Python Mapper software to run. It requires older versions of some dependencies, such as WX python. For software implementing Mapper, I would recommend, as of 18/06/2019 the `knotter` software by rosinality `https://github.com/rosinality/knotter` and the `Kepler Mapper` software [VS19].



Filter range: [15.72, 36.82]
Cover: Hypercube cover. Intervals: (2,). Overlap: (50.0,)
Clustering method: Single linkage clustering
Cutoff: Cluster cutoff at height 30.0
Size range: [1,31]

Figure 4.6: **A)** Mapper applied to the top 500 varying genes of the regularised log-transformed Arctic charr transcriptomic data, with lake centroids subtracted. The cover used was two intervals with 50% overlap, and the clustering was single-linkage with a cut-off at 30. The members of the outlying nodes of the graph have been listed. There appear to be no strong groupings by lake or morph. In fact, there are nodes with both morphs, such as the size 4 one to the bottom left, with most of the Tay pelagic morph, and a member of the Kamkanda benthic morph. **B)** shows the corresponding two dimensional PCA plot for the Arctic charr transcriptomic data, with node membership indicated by outlines in corresponding colours. Note, the large node with 31 samples has not been outlined in (**B**) to simplify the diagram.

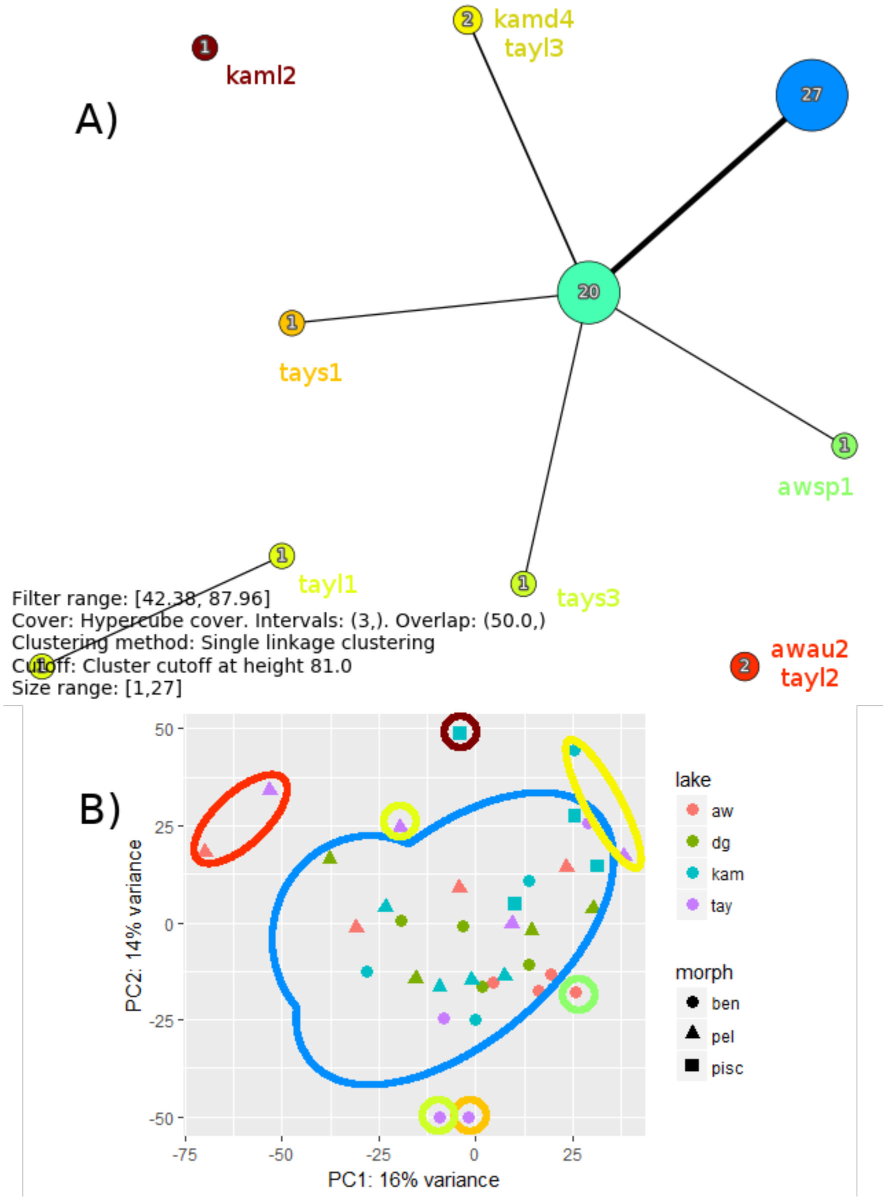


Figure 4.7: **A)** Mapper applied to all genes of the regularised log-transformed Arctic charr transcriptomic data, with lake centroids subtracted. The cover used was three intervals with 50% overlap, and the clustering was single-linkage with a cut-off at 81. The members of the outlying nodes of the graph have been listed. There appear to be no strong groupings by lake or morph. Note the presence of different fish in the outlying nodes. **B)** shows the corresponding two dimensional PCA plot for the Arctic charr transcriptomic data, with node membership indicated by outlines in corresponding colours. Note, the node with 20 samples has not been outlined in (**B**) to simplify the diagram.

4.3 Reproducing the Example

Since we did not succeed in finding any significant features in Mapper output, we will undertake an investigation into this matter. We suspect that the low sample size (36 samples) of the Arctic charr gene expression data set could be an issue. This is contrast to the dataset used in the breast cancer example, as found in [Lum+13] and §2.2 of this thesis, which had a sample size of 295. We proceed by first reproducing the Mapper output used by Lum *et al.* in their paper. Afterwards, in §4.3.2, we subsample to determine the sample size required so that the *c-MYB*$^+$ is apparent in the output. We use the open-source Python Mapper program [MB13] for our reproduction with data from the Netherlands Cancer Institute [Vee+02].[2]

4.3.1 Reproduction

We follow Figure 2.9 from the supplementary data of [Lum+13], where the authors take the surviving patients, and look at the Mapper output for taking their filters and distances on the top varying genes.

Methods

We take the 5% of genes with the highest variance, and use these to calculate Pearson correlation distances between the breast cancer samples. See Definition 2.2.1 for the definition of Pearson correlation, and the distance based on it is described in the paragraph preceding the definition. The filter function used is L-infinity centrality which assigns to a data point the value of the maximal distance to any other data point in the set. We use Ward clustering [War63], which minimises the total within-cluster variance, in Python Mapper [MB13] to produce something similar to the desired output.[3]

Output

Figure 4.8 shows the reproduction, where the lower branch on the right seems to be our group of interest, since it contains 21 samples, about the same as the 22 found in [NLC11].

When we take the top 50%, 25%, 10% and 5% varying genes, we usually get something like Figure 2.9 from [Lum+13], but not quite as clean looking, and occasionally with a loop at the end, rather than a branch.

We were unable to do a more stringent test of these Mapper outputs, since neither paper actually explicitly lists which samples in the breast cancer data set are actually *c-MYB*$^+$.[4]

[2] Since the online resource where the gene expression data was kept has now gone offline, we thank Monica Nicolau [NLC11] for providing the data to us.

[3] This was chosen since the clustering algorithm produced Mapper outputs most similar to those of Lum *et al.*, and furthermore, the authors have not specified which clustering algorithm they used.

[4] There have also been problems following this up with the authors, since the contact details on the

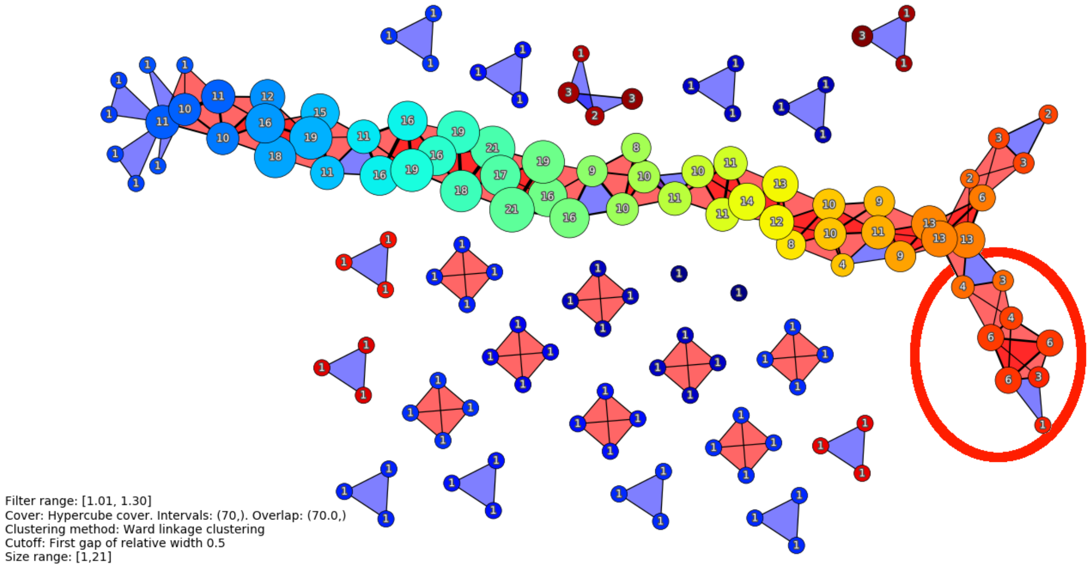

Filter range: [1.01, 1.30]
Cover: Hypercube cover. Intervals: (70,). Overlap: (70.0,)
Clustering method: Ward linkage clustering
Cutoff: First gap of relative width 0.5
Size range: [1,21]

Figure 4.8: Mapper output taking the 5% top varying genes ($\sim$ 1100). 70 intervals, with 70% overlap. Filter function was the L-infinity centrality, as used in [Lum+13]. Clustering was Ward clustering, with a cut-off at 50% of the diameter. The purported $c\text{-}MYB^+$ samples on the lower branch on the right have been circled.

4.3.2 Subsampling

We took subsamples of the 295 samples, by taking 250, 200, 150, and 100 at random, by producing a vector of random non-repeating numbers between 1 and 295 in R, and then extracting those rows from the data matrix. We then take the top 5% varying genes and visualise the subsample using Mapper, as in the previous subsection. We decreased the number of intervals used at each subsampling to 60, 50, 40 and 30 to keep the output connected, otherwise, the other inputs to Mapper were the same as in Figure 4.8. Output can be found in Figures 4.9–4.12. Note how the branch becomes less distinct as we take smaller samples, eventually disappearing when we take 100. We do not show it here, but if we colour the nodes by their proportion of purported $c\text{-}MYB^+$ samples, we can see that these start to disperse further out along the main line as we take smaller and smaller subsamples.

4.4 Discussion

When applying Mapper to the Arctic charr gene expression data, we see that pre-processing to make the benthics 'normal' shows up no patterns in the pelagics, while making the pelagics 'normal' gives lots of isolated small clusters. We have also demonstrated how changing the pre-processing, cover, and clustering function can change the Mapper output, which in turn shows that the difference between the two pre-processings is not significant. Subsequently, we try to visualise the benthic/pelagic split using Mapper, but this also fails.

paper are no longer valid, and it is unclear where the authors are now.

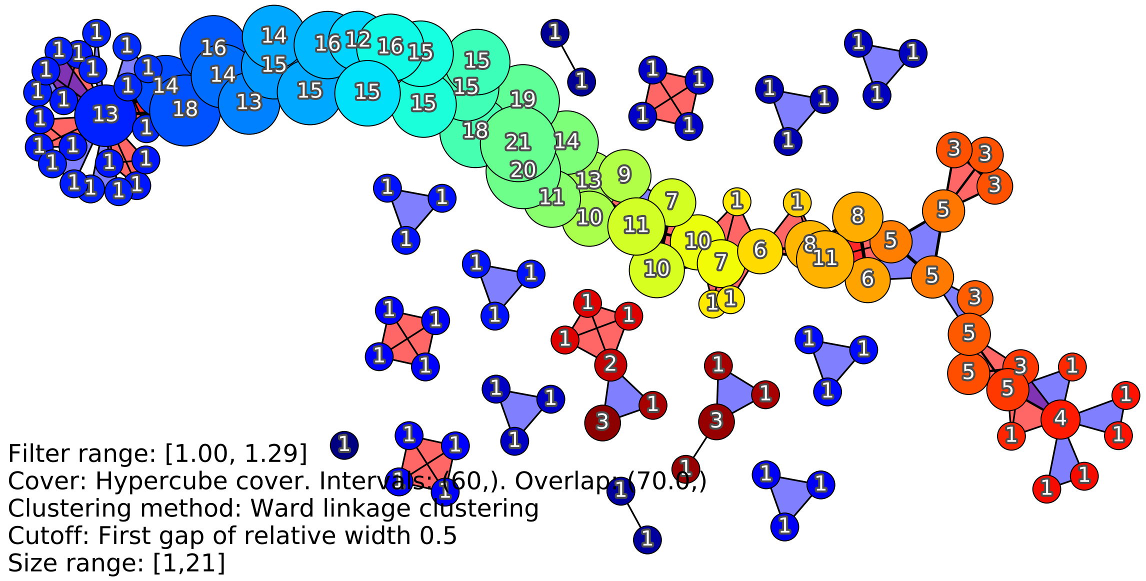

Filter range: [1.00, 1.29]
Cover: Hypercube cover. Intervals: (60,). Overlap: (70.0,)
Clustering method: Ward linkage clustering
Cutoff: First gap of relative width 0.5
Size range: [1,21]

Figure 4.9: Mapper output on a subsample of 250 (184 survivors) taking the 5% top varying genes (∼ 1100). 60 intervals, with 70% overlap. Filter function was the L-infinity centrality, as used in [Lum+13]. Clustering was Ward clustering, with a cut-off at 50% of the diameter.

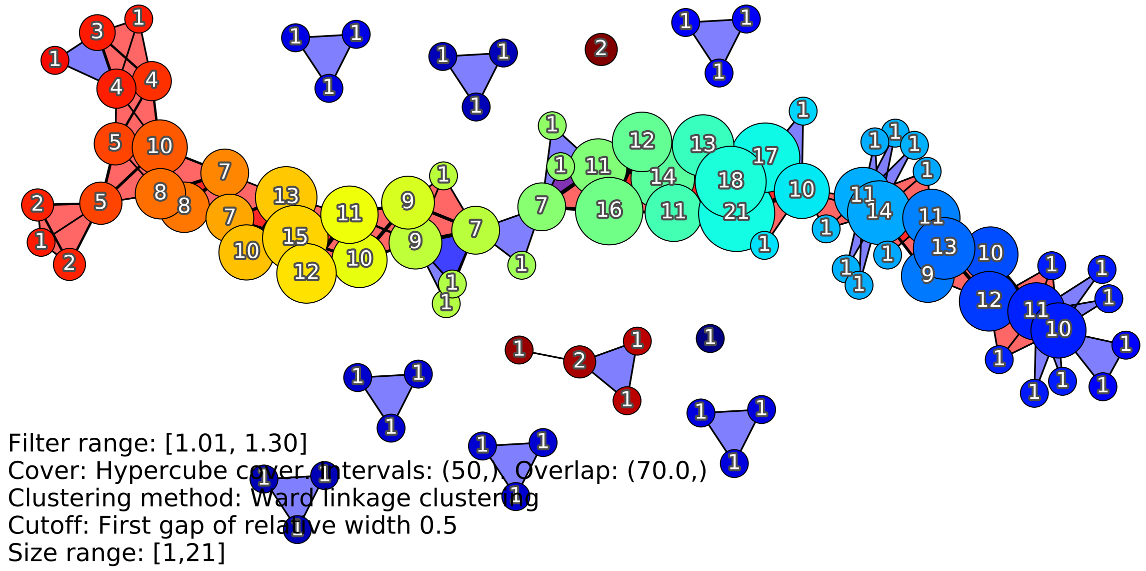

Filter range: [1.01, 1.30]
Cover: Hypercube cover. Intervals: (50,). Overlap: (70.0,)
Clustering method: Ward linkage clustering
Cutoff: First gap of relative width 0.5
Size range: [1,21]

Figure 4.10: Mapper output on a subsample of 200 (144 survivors) taking the 5% top varying genes (∼ 1100). 50 intervals, with 70% overlap. Filter function was the L-infinity centrality, as used in [Lum+13]. Clustering was Ward clustering, with a cut-off at 50% of the diameter.

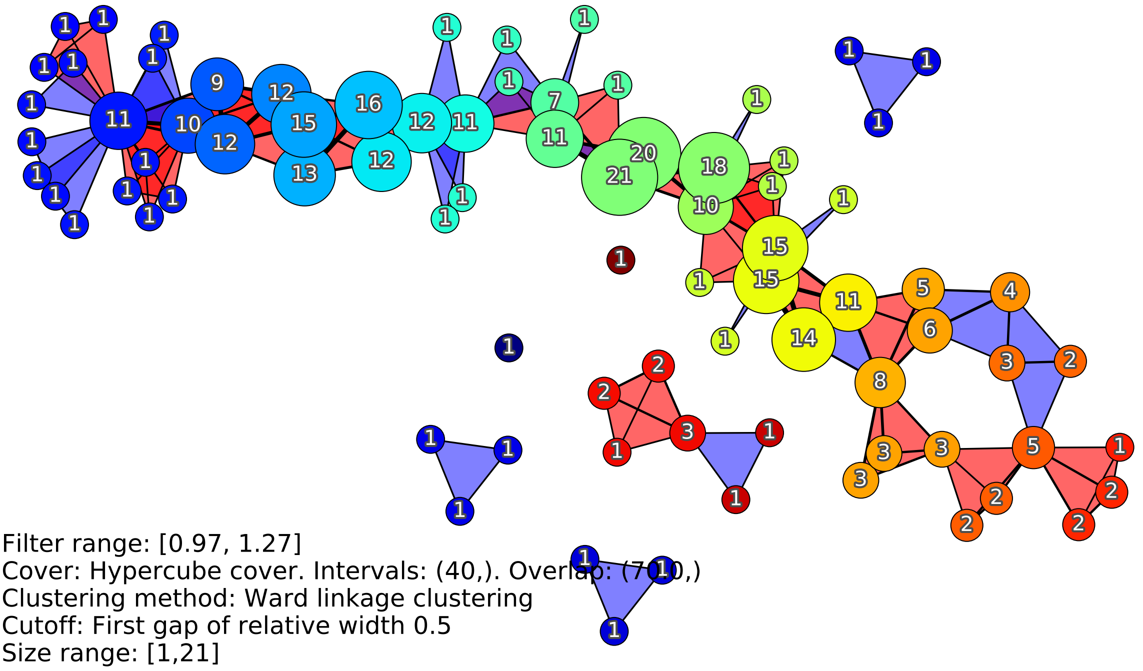

Filter range: [0.97, 1.27]
Cover: Hypercube cover. Intervals: (40,). Overlap: (70.0,)
Clustering method: Ward linkage clustering
Cutoff: First gap of relative width 0.5
Size range: [1,21]

Figure 4.11: Mapper output on a subsample of 150 (113 survivors) taking the 5% top varying genes ($\sim$ 1100). 40 intervals, with 70% overlap. Filter function was the L-infinity centrality, as used in [Lum+13]. Clustering was Ward clustering, with a cut-off at 50% of the diameter.

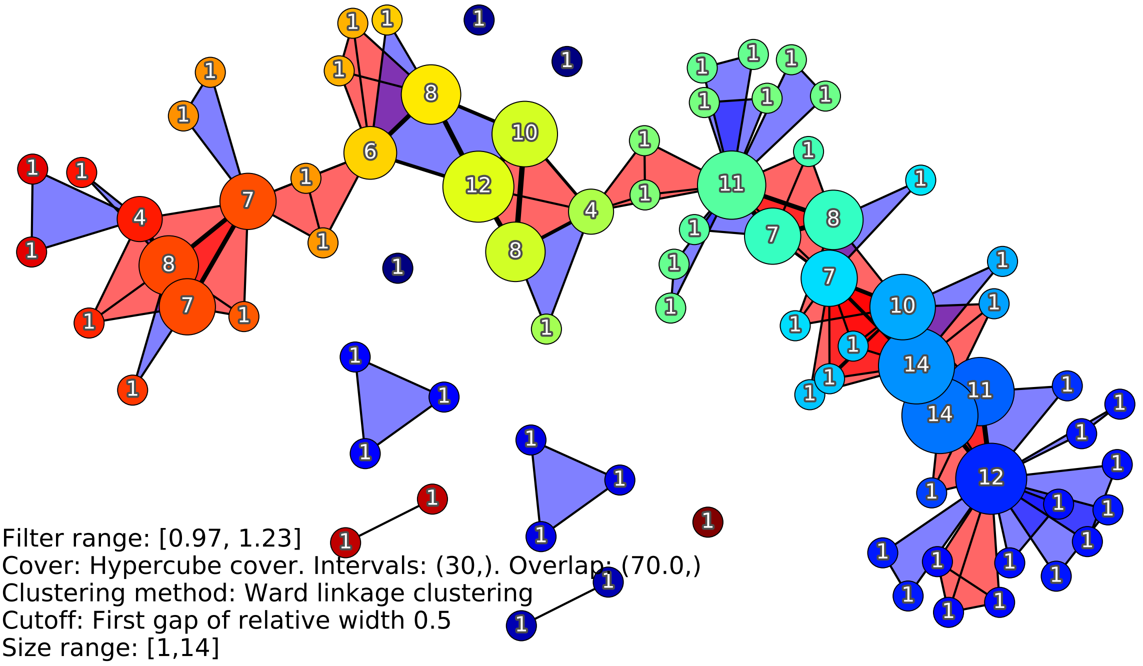

Filter range: [0.97, 1.23]
Cover: Hypercube cover. Intervals: (30,). Overlap: (70.0,)
Clustering method: Ward linkage clustering
Cutoff: First gap of relative width 0.5
Size range: [1,14]

Figure 4.12: Mapper output on a subsample of 100 (74 survivors) taking the 5% top varying genes ($\sim$ 1100). 60 intervals, with 70% overlap. Filter function was the L-infinity centrality, as used in [Lum+13]. Clustering was Ward clustering, with a cut-off at 50% of the diameter.

We suspect the lack of anything significant in the Mapper output could be explained by our small sample size. To investigate this, in §4.3.2 we go back to our original motivation, the breast cancer study of Lum *et al.* [Lum+13] and reproduce their Mapper output. Subsequently subsampling this reproduction suggests that we need about two hundred samples before we can find a subgroup of interest.

Sample Size We see that sample size plays a big role in being able to distinguish the *c-MYB*$^+$ samples. In particular, when dropping from 200 to 150 samples we see that the branch where the *c-MYB*$^+$ samples are found is merged into the main line, as shown in Figs. 4.10 and 4.11, so we can no longer distinguish it in the Mapper output. This occurs since there are no longer enough *c-MYB*$^+$ samples for the Mapper algorithm to distinguish them and cluster them on their own.

In conclusion, we have seen that Mapper allows us to visualise the data in another manner, and from §2.2 we know that it can let us see additional structure, not visible to PCA or clustering. In this case, we have found no additional structure using Mapper. The inability to observe the difference between benthic and pelagic morphs is the motivation for us to develop a topologically-inspired perspective involving deformation. In the next chapter, we will describe this different approach to applying ideas from topology to gene expression data, focusing more on the genes, their correlations, and their expression between the ecomorphs.

Chapter 5

A Topological Perspective for Ecological Transcriptomics

In this chapter we give a topological approach for analysing gene expression data, and show its application to our Arctic charr data. We will also look at how our topological perspective works on a large *Drosophila* dataset [Lin+15], consisting of 726 samples. Finally, we will also compare and contrast our approach with a couple of standard methods currently used to analyse gene expression data in evolutionary biology. The first, and most commonly used, is a standard differential expression analysis, as implemented in the R package DESeq2 [LHA14]. The second involves looking at gene co-expression, and we will look at its implementation in the WGCNA R package [LH08].

Our approach was inspired by the goal of visualising the split between the benthic and pelagic morphs in the Mapper output. This split can be seen in Fig. 3.5, where we note that the benthic fish tend to be in one direction, and the pelagic in the other. In the previous chapter, we naïvely tried to find this split in Mapper output, using the Euclidean distance, and the Euclidean norm as a filter function. The results are shown in Figs. 4.6 and 4.7, where we do not see such a split.

To make this split apparent in the Mapper output, in this chapter, we introduce the idea of *perturbation* of the gene expression space. The idea is to find a perturbation which makes the difference between the benthic and pelagic morphs apparent in the Mapper output. We subsequently simplify this perturbation to weighting the genes (dimensions), since this is tractable on computers, being only an n-dimensional problem, and we do not have to take into account an $n \times n$ rotation matrix.

5.1 Mathematics

In this section, we will outline the mathematical underpinnings of the gene expression data, and how they motivate our chosen method. We will begin by looking at the input data, and the assumptions we have made about it. We expand on the definitions of §3.3.3, so we recommend readers to refer back to that section before carrying on.

5.1.1 Replicate Effect

First, show that the *lake effect*, mentioned in §3.4.1 is an example of a more general *replicate effect*. We can use this more general idea for population variation which we wish to take into account when combining across replicates. Once we have removed this effect, the difference we actually care about in our replicated experiments will be more obvious, although we will still have individual variation. In the Arctic charr, this corresponds to removing the lake effect, and being left with the effect of morph.

In general, this is difficult to define, but since our analysis will involve looking at the samples as if they were points in gene expression space, we can take into account the replicate effect by an appropriate centring of the data.

Definition 5.1.1 (Replicate Effect)**.** The *replicate effect* is a vector in $\mathbb{R}^m$, where m is the number of genes under consideration, associated with one of the experimental replications in our dataset of interest.

We can consider the replicate effect as the gene expression of a generic sample from a given replication.

For our purposes, we will use a simple method of finding the replicate effect. As used for the lake effect, we use the barycentre (centroid). Additionally, if we have an uneven number of samples from each group of interest, then we may want to take this into account when finding the replicate effect. This can be done by taking the barycentre of each group in an experimental replication separately, and then taking the barycentre of all the groups.

5.1.2 Perturbation

Our goal is to perturb the Euclidean distance function in such a way that groups of interest become apparent when visualised, for example in Mapper output. In our case, this is the grouping by morphs. We are assuming that the difference between groups can be found at the transcriptomic level, and we would like to know which genes are involved. The idea is that a perturbation which finds some difference between the morphs will give us this information.

We can perturb the Euclidean distance in $\mathbb{R}^m$ by picking an inner product. This is equivalent to transforming the points by a symmetric matrix, which can be decomposed as a combination of a rotation (orthogonal matrix) and a scaling on each dimension (diagonal matrix). Practically speaking, the rotation matrix is too large for computers to deal with (it's usually about a $30,000 \times 30,000$ matrix), so we just use a weighting vector $\mathbf{w} = (w_1, w_2, \ldots, w_m)$ where m is the number of genes under consideration and $w_i \in \mathbb{R}_{\geq 0}$. Additionally, using a rotation makes it more difficult to interpret the results, since we are no longer dealing with weights on genes, but weights on combinations of genes.

Definition 5.1.2 (Weight Vector)**.** Let $\mathbf{w} = (w_1, w_2, \dots, w_m)$, where m is the number of genes under consideration and $w_i \in \mathbb{R}_{\geq 0}$, be a *weight vector.*

Furthermore, note that our results will be the same up to scaling, so we could also have $w_i \in [0, 1]$, where at least one $w_j = 1$.

A key assumption is that the genes which are given high weight in the perturbation are important for distinguishing between groups. Additionally, if the groups exist under sufficiently generic perturbations, then this indicates some robustness about the groups.

Searching Gene Space We can use perturbations of a method for searching gene space. There are a couple of ways to go about this. One is to use a gene weighting which is given to us, for example, a set of genes where we weight genes in the set with a weight of 1, and genes outside with a weight of 0. Then we can test how effective this gene weight is at separating known groups, either by visualising, for example with Mapper, or using some other statistical measure of separation between groups, like an F-statistic (Def. 5.4.25). Another is to fix some measure for separation between groups, and then search for weights of genes which maximise that measure. This method does not work with having to visually inspect outputs for evaluating the effectiveness of a gene weight.

In the next section, we will show that there are perturbations which separate the Arctic charr by morphs. We do this by using information we already know about genes which separate out groups.

5.2 Proof of Concept

We now give an example of a perturbation such that groups of interest become apparent in the Mapper output. This is to show that a method based around finding such perturbations can work in practice. We will pick a 'perturbation' which considers only a certain subset of genes (so these genes have weight 1, and the rest have weight 0). The choice of genes will be determined by the Welch's t-test [Wel47]. For each gene, we apply the t-test where our two populations are given by the 16 benthic fish and the 16 pelagic fish. The test is based on the t statistic, which is given by the difference between the means of the two populations, standardised by their sample size and variances. A probability (p-value) is associated to the t-statistic, which measures the probability that we get more extreme values in a t-distribution with a certain number of degrees of freedom. You can think of genes with low p-values as fitting a certain pattern, in this case, the gene will have different average expression between the benthic and pelagic fish, and also show low variance within morphs.

Once we have p-values associated to each of our genes, we apply an adjustment for multiple testing, in this case the Benjamini–Hochberg (BH) method [BH95]. We then

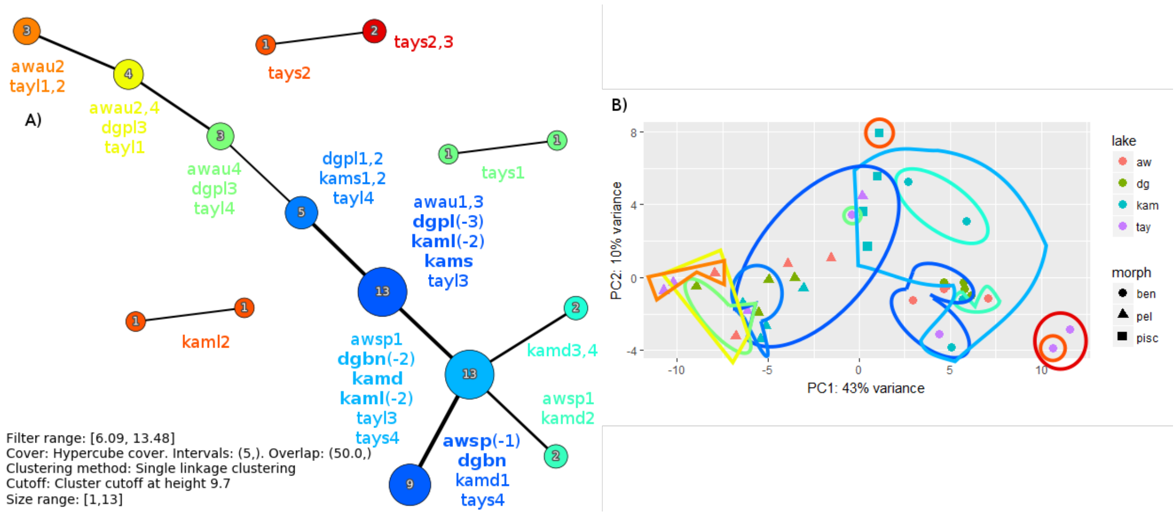


Figure 5.1: **A)** Showing the output of Mapper applied to the 284 genes with significantly different means between the benthic and pelagic morphs, as calculated on the regularised log-transformed Arctic charr transcriptomic data with lake centroids subtracted. Filter function was the Euclidean magnitude. Note the separation into benthic and pelagic morphs. We have a long flare consisting of pelagic morphs, while the benthic are mostly concentrated in the bottom blue size 9 nodes, with most of the Tay benthic (small) morph in their own small components. The fish belonging to each of the nodes are circled in the PCA plot in (**B**).**B)** Showing the first two PCs of PCA applied to the 284 genes with significantly different means of between the benthic and pelagic morphs, as calculated on the regularised log-transformed Arctic charr transcriptomic data with lake centroids subtracted. Note the expected separation of the fish into pelagic and benthic morphs. The unusual lone Tay benthic (small) morph seen near the centre has been detected before, e.g. it shows up opposite of all the other Tay benthics in 3.4. The Tay pelagic (large) morph near the centre hasn't seemed unusual from the figures in §4.

pick a cut-off, say, our adjusted p-value has to be less than 0.05.[1] This results in 284 genes where we find a significant difference between the benthic and pelagic fish.

Figures 5.1 and 5.2 show the results of applying Mapper, PCA, and clustering to this data. Note that we have the expected separation into benthic and pelagic morphs in each case.

5.3 Experiments

5.3.1 Residual Replicate Effect

In this subsection, we will describe work on weighting a set of differentially expressed genes. The idea here is that, even though we have removed a first-order approximation of the replicate effect, as given in Definition 5.1.1, there might still be some differences between replicate experiments when we look at each group separately. We want to find genes for which this remaining difference is low, while this difference between the

[1] Adjusting p-values for multiple comparisons and then picking a cut-off is functionally the same as picking a more strict cut-off on the original p-values.

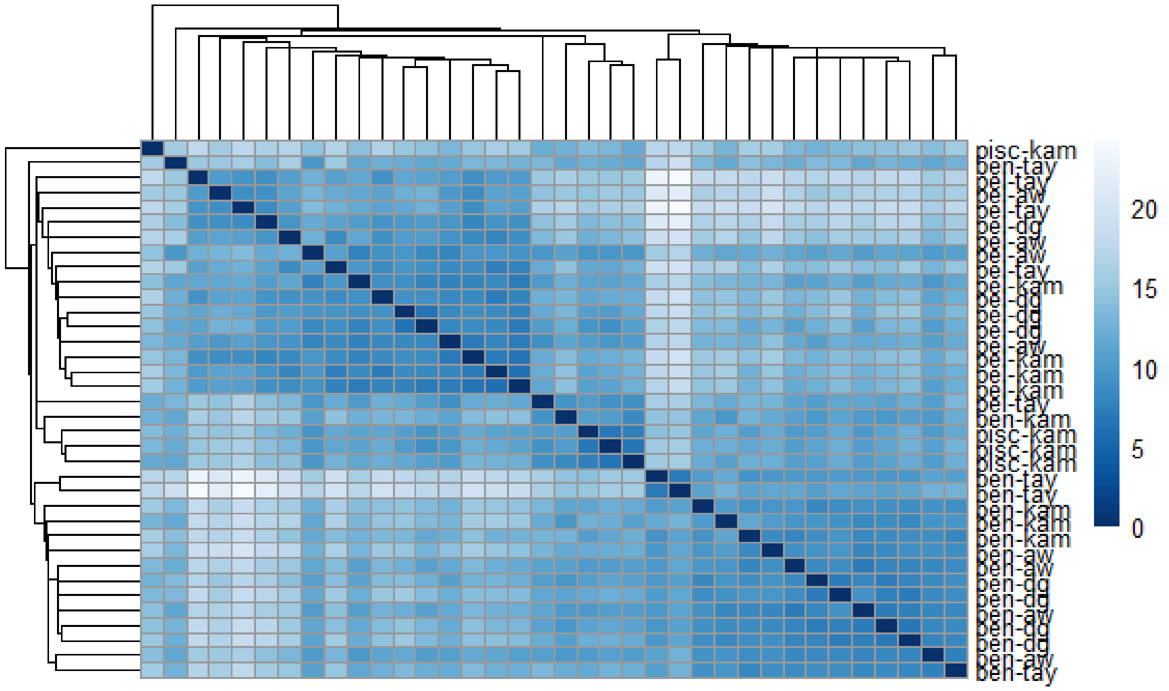


Figure 5.2: Showing a heatmap of Euclidean distances, along with single-linkage clustering, on the 284 genes with significantly different means of between the benthic and pelagic morphs, as calculated on the regularised log-transformed Arctic charr transcriptomic data with lake centroids subtracted. Note again the expected clusters pelagic and benthic fish. We see unusual Tay benthic (small) morph (as in Figure 5.1 (**B**)) as an outlier, it's the second row.

groups is high. Biologically, this corresponds to genes which have the similar expression in each group, over all the replicates. These are genes which are behaving in the same way across replicates, and so are showing a parallelism. To do this, we will focus on a set of differentially expressed genes, which we know have a high difference between the groups, and then weight these genes to get out a subset which has substantially smaller differences between replicates, when we look within groups.

First, suppose we have the results of some statistical test for differential expression and we're given gene expression data, with replicate effect subtracted, in the form of an $n \times m$ data matrix A, with entries $a_{ij} \in \mathbb{R}$ the expression of the j^{th} gene in the i^{th} sample, accounting for location. Suppose we have k differentially expressed genes, indexed by $d_1, d_2, \ldots, d_k$, where d_1 is the most differentially expressed gene, and d_k the least.

We will investigate a couple of different methods for testing if there is any residual replicate effect. One will be plotting the expression of genes with different weightings on the genes, and seeing if this can make the difference between locations disappear. We will do this with both Mapper and PCA. The other will be to come up with some sort of test statistic in the case where we look to see the differences between the locations within groups.

In theory, we should see a reduced replicate effect if we take the most differentially expressed genes, say, the top ten. We want to extend this to look for gene sets which are not just the top most differentially expressed one, for which this also occurs.

For the statistical test, we will use the analysis of variance (ANOVA) test. The goal is, for every group, to test if the replicate effect is still present in it or not. So, suppose we have r groups and l replications. Then for every group, we look at the samples belonging to that group, and take the centroid for each replication in that group.[2] The ANOVA F-test statistic is given in Definition 5.4.25. In the case where we expect no difference between replications, then the F-statistic is known to have an F-distribution $F \sim F(k-1, N-k)$. Note that the F-test assumes that every replication has the same variance, and that is it normally distributed.

When conducting the F-test using all genes, we find no statistically significant residual lake effect for either the benthic or pelagic morphs. The p-values are around 0.6 in both cases.

We will now look at figures and the p-values where we have taken the top ten genes which are best at separating the lakes in a particular morph by this measure, as well as the ten genes which are the worst (i.e. they don't see the difference between the lakes) and plot these out on PCA and Mapper graphs, to see if this effect is visible when we visualise. Figures 5.3–5.6 are PCA plots of the top and bottom ten differentially expressed genes at separating out the lakes in each morph. The p-values in the cases of benthic, best and worst separation, and pelagic, best and worst separation, are respectively, approximately, 0.0002, 0.96, 0.0006, and 0.99. So we see that selecting individual genes which are known to have a large residual lake effect gives us a large lake effect overall, while selecting those with none still leaves us with none.

In any case, we do see that there are specific genes which have a significant residual lake effect in the dataset, but overall the effect is not significant, unless we zoom in and select genes for which we see this effect.

5.3.2 Drosophila

In this section, we will test the ability for differential gene expression techniques to increase the group signal relative to the replication signal (noise) by looking at our Arctic charr data set, and also a *Drosophila* data set from [Lin+15], which has a much larger sample size of 726 compared to our 32. The goal is to see if increasing the sample size will make it easier to distinguish between groups, visually, or if there is always too much noise inherent in biological organisms. For the *Drosophila* data set, we consider the environment as the signal, and sex as the noise, in analogy to the Arctic charr.[3]

[2] Note: This is different to taking the centroid overall, which we did originally to find and subtract the replicate effect.

[3] Since for the Arctic charr, the lakes are the large 'signal' which we do not care about. Similarly, in the *Drosophila*, sex provides the largest signal.

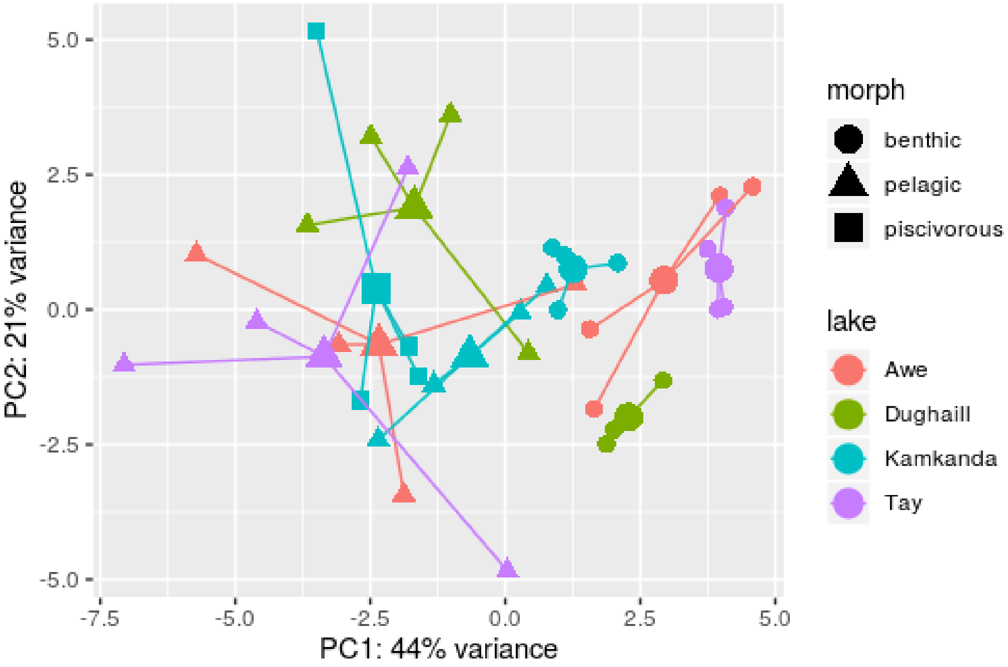


Figure 5.3: A PCA of the gene expression values of the ten differentially expressed genes with the lowest p-value under an ANOVA, testing if the groupings by lake are statistically significant in the benthic morph. That is, these ten genes are the 'best' under this metric at separating out the lakes in the benthic morph.

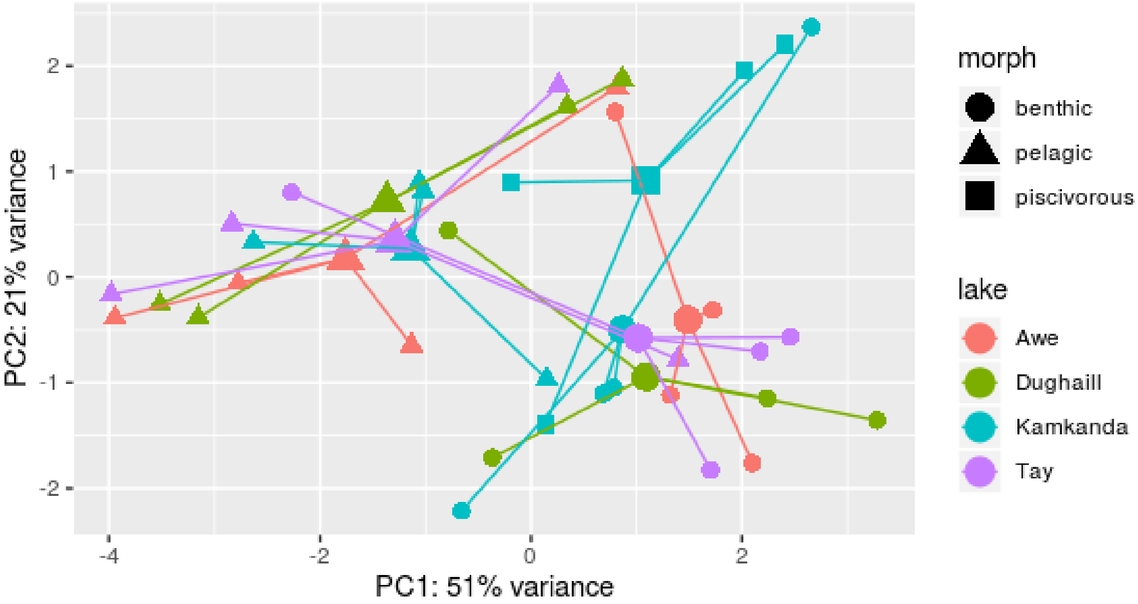


Figure 5.4: A PCA of the gene expression values of the differentially expressed ten genes with the highest p-value under an ANOVA, testing if the groupings by lake are statistically significant in the benthic morph. That is, these ten genes are the 'worst' under this metric at separating out the lakes in the benthic morph.

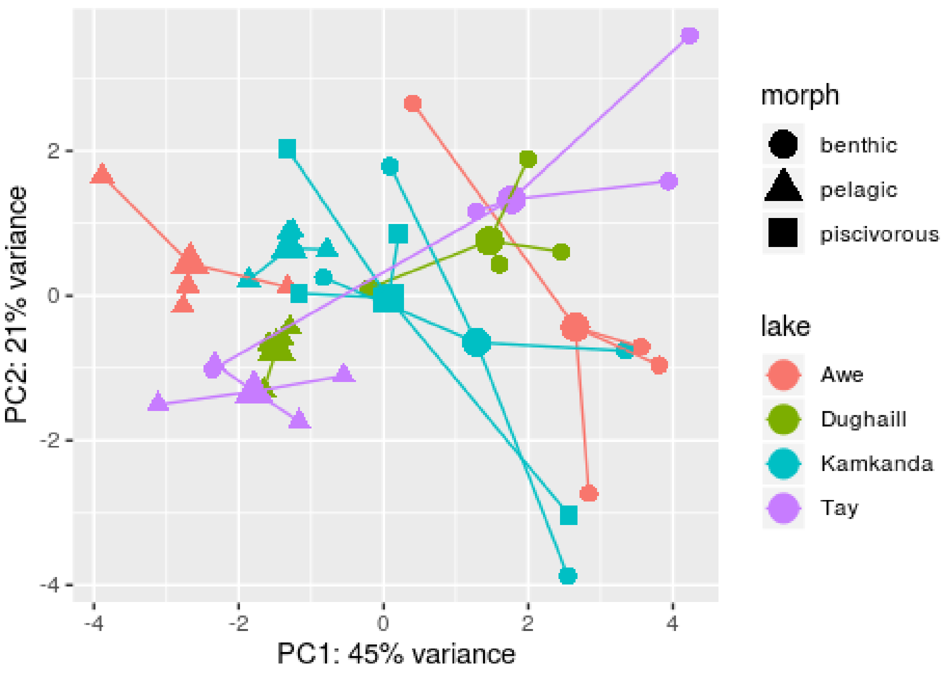


Figure 5.5: A PCA of the gene expression values of the ten differentially expressed genes with the lowest p-value under an ANOVA, testing if the groupings by lake are statistically significant in the pelagic morph. That is, these ten genes are the 'best' under this metric at separating out the lakes in the pelagic morph.

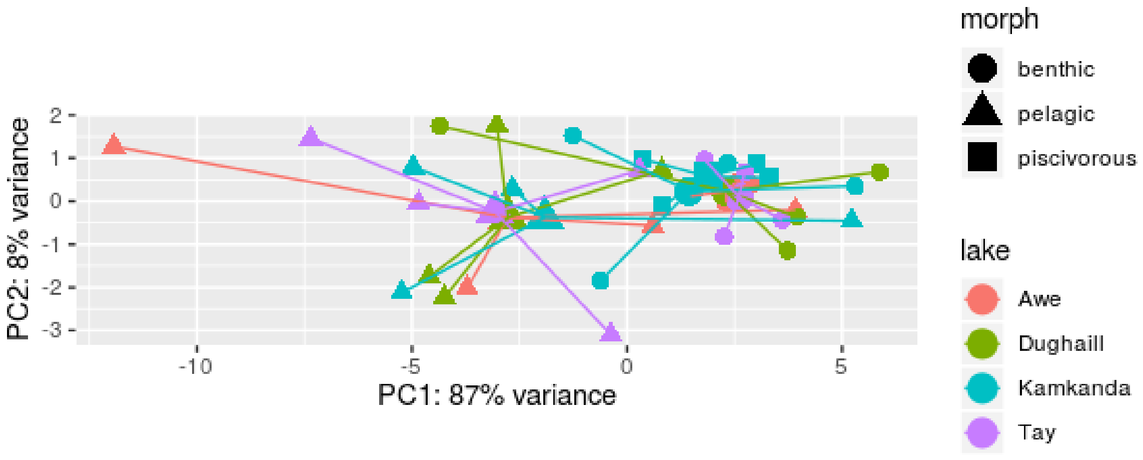


Figure 5.6: A PCA of the gene expression values of the ten differentially expressed genes with the highest p-value under an ANOVA, testing if the groupings by lake are statistically significant in the pelagic morph. That is, these ten genes are the 'worst' under this metric at separating out the lakes in the pelagic morph.

Data

The 726 samples come from eight flies of each sex, taken from 16 *Drosophila* Genetic Reference Panel (DGRP) genotypes. The *Drosophila* were raised by Lin *et al.* at three different calendar times, with all other environmental variables controlled. These included parental culture density, food, temperature, light/dark cycle, mating status, social exposure, and the circadian time of RNA extraction. We downloaded the read counts from GEO with accession number GSE60314.

Procedure

The procedure we use here is as follows:

1. We take the normalised Arctic charr data, and do not take into account the effect of location. Similarly, we take normalised *Drosophila* data, and do not take into account any effect of sex, environment, or genotype.

2. For the Arctic charr, we estimate the size of the morph signal by taking the average of Euclidean distance from the benthic and pelagic morph centroids in each lake to the centroid of the whole lake. To estimate the size of the lake signal (noise) we look at the average distance from the lake centroids to the overall centroid.

3. We repeat for the *Drosophila*, using environment and sex in place of lake and morph.

4. We use a test for differential expression of genes between morphs for the Arctic charr, and environment for the *Drosophila*. We use the program sleuth [Pim+17] for the Arctic charr, and DESeq2 [LHA14] for the *Drosophila*, and look for genes with an adjusted p-value of < 0.05.

5. We take the set of differentially expressed genes and repeat the above analysis from steps 2 and 3, looking at the morph signal versus the lake noise. We want to see if taking only the differentially expressed genes(/transcripts) reduces the signal-to-noise ratio by any significant amount.

Versus Charr

When taking all 109,584 transcripts of the Arctic charr, we get a morph signal (average Euclidean distance from morph/lake centroid to the respective lake centroid) of 121.3241 and a lake signal (average Euclidean distance of lake centroid to overall centroid) of 147.9235. When taking only the 574 differentially expressed transcripts, we get a morph signal of 18.07353 and a lake signal of 17.70094. We can see that the morph signal has increased relative to the lake 'signal', but there has been a decrease of signals (i.e. distances) overall, since we have thrown out most of the dimensions (transcripts).

For the *Drosophila*, with all 17,238 transcripts we get an environment signal of about 16.29652 and a sex signal of about 119.5324, while taking only the 3,938 transcripts differentially expressed between each pair of environments, we get an environment signal of about 10.914574 and a sex signal of about 51.5635. This shows a substantial relative increase in the environment signal compared to the sex 'signal', but the actual signal is still small in absolute terms. This shows that even with a large biological sample size, if the initial signal we are interested in is small compared to all the other signals/noise, such as morph vs. lake in the charr, or environment vs. sex in the *Drosophila*, then we will still be unable to perceive it.

5.4 Comparison

In this section we will look at the mathematical theory behind the differential gene expression and gene co-expression methods. We will also consider how they compare to our topological deformation method. As in previous chapters, we will consider our data to come in the form of an $n \times m$ matrix, where n is the number of samples, and m the number of genes, and entries are normalised gene expression values.

5.4.1 Differential Gene Expression

In this section, we will offer a definition of differential gene expression, and then describe methods for assigning importance to genes, based on their difference in expression between conditions of interest. This is to allow us to see which statistical tests can be considered as part of our deformation approach, and which cannot. We will assume that we have accounted for the effect of location.[4]

Definitions

We will start off with some general definitions, which we will illustrate with some more specific examples:

Definition 5.4.1 (Conditions)**.** *Conditions* are the partitions of samples into different sets.

This partition usually arises from a priori knowledge coming from the scientific experiment. For example, we can have as conditions, healthy and diseased, in the case of samples from breast cancer or normal patients (see §2.2.1), or benthic and pelagic, when we are looking at samples from Arctic charr fish (see Tab. 3.1).

Definition 5.4.2 (Differentially Expressed Gene)**.** Given different conditions, a *differentially expressed gene* is a gene with expression values which are statistically significantly different between different conditions.

[4] Alternatively, we can analyse only the samples from a single location.

Which genes are differentially expressed depends on the statistical model used [CSDL17], and the choice of conditions to compare.

Statistical testing first involves the formulation of null and alternative hypotheses, which postulate that the observed data come from random variables with a hypothesised probability distribution.

Definition 5.4.3 (Null Hypothesis)**.** The *null hypothesis*, H_0, for differential gene expression analysis, is that there is no change in the expression of a gene between conditions.

Part of our null hypothesis must include statistical assumptions about, for example, the distribution of the gene expression values. It is very important that the assumptions underlying our null hypothesis is true. Otherwise, we may be rejecting the null hypothesis because, for example, the distribution of gene expression values is different to what we assumed, rather than because there is a difference between the conditions.

Definition 5.4.4 (Alternative Hypothesis)**.** The *alternative hypothesis*, H_1, for differential gene expression analysis, is that there is a change in the expression of a gene between conditions.

Definition 5.4.5 (Statistical Test)**.** A *statistical test*, or, more fully, *statistical hypothesis testing*, involves testing if a given null hypothesis can be rejected in favour of an alternative hypothesis.

The null and alternative hypotheses will depend on what biological question we are asking. For example, if we are only interested in genes with a more than two-fold change in expression between, say, condition A and condition B, then our null hypothesis is that the expression of genes has a less than two-fold change between A and B, and the alternative hypothesis is that there is a greater than two-fold change. In this chapter, we will assume that we are looking for a purely statistically significant change, without regard to the magnitude of the change.

Before we go on to the examples, we will give some desirable properties of the probability distributions typically used in differential gene expression analysis.

Definition 5.4.6 (Desirable Distributions)**.** The probability distributions used for differential gene expression analysis are unimodal,[5] and typically consist of two parameters, the mean, μ, and a parameter related to the variance.

For example, it is typical to assume that log-normalised gene expression data comes from a normal distribution, while count data comes from a negative binomial distribution.

In practice, the testing for differential expression reduces to looking at a test statistic with a known distribution under the null hypothesis. This is usually the difference between the means of the gene expression between the conditions. However, if we're

[5] Have one local maximum.

interested in a more certain difference, we can test if, for example, there is a difference between the upper and lower quantiles of the two conditions.

If the value of the test statistic is sufficiently unlikely then we reject the null hypothesis in favour of the alternative. A typical cut-off for statistical significance in the scientific literature is a probability of < 0.05. This is the typical value for the medical and biological sciences. In situations where it is possible to get large enough sample sizes with small enough errors, smaller probability values can be use. For example, a cut-off of $< 2.87 \times 10^{-7}$ is used in high-energy particle physics [Tan+18, §39.3.2]. See the examples below for some test statistics which are used.

Definition 5.4.7 (Differential Gene Expression)**.** A gene is statistically significantly differentially expressed if the null hypothesis that the gene expression does not differ between conditions is rejected.

We will now outline two different classes of statistical approaches. One is parameter estimation, which is based on estimating the expression of a gene in different conditions, which lets us determine if a gene is differentially expressed in pairwise comparisons, and also gives an estimate of the difference in expression. The other is the likelihood-ratio test, which tests if a gene accounts for statistically significant differences in expression between multiple (two or more) conditions.

Parameter Estimation

Let D be our $n \times m$ data matrix, for n samples and m genes. We will assume that the data have been appropriately normalised.[6] We are interested in the difference between two conditions, and we want to know which genes are statistically significantly different between the conditions. We will determine this with a gene-by-gene basis so it will be enough for us to consider the case of a single gene, say gene k. So for the rest of this example, we will assume that the gene expression values are coming from some column k of the data matrix D. To work out if other genes are differentially expressed, we would do the same statistical test on gene $1, 2, 3, \ldots, m$.

Suppose we have two conditions A and B, with number of samples n_A and n_B, respectively, where $n_A + n_B \leq n$. Let the expression values of the gene under conditions A and B be given by $\{a_1, a_2, \ldots, a_{n_A}\}$ and $\{b_1, b_2, \ldots, b_{n_B}\}$, respectively, where $a_i, b_i \in \mathbb{R}$ are the gene expression values of a particular gene.

Example 5.4.8 (The t-test)**.** In this example, we will go over a commonly used hypothesis test. This is the t-test, more specifically, Welch's t-test [Wel47], for normal distributions with unequal variances.

We will be testing null hypothesis, H_0, that the expression values under conditions A and B have the same mean, $\mu_A = \mu_B$, versus the alternative hypothesis, H_1, that

[6] That is, we assume that the gene expression values have been transformed so they are approximately normally distributed and comparable between samples.

the means are different, $\mu_A \neq \mu_B$. Let A_e and B_e be the random variables for the expression values associated under each respective condition.

Now, if A_e and B_e have finite mean and variance, so $|\mu_A|, |\mu_B|, \sigma_A^2, \sigma_B^2 < \infty$, and we have enough samples, n_A and n_B, we can apply the central limit theorem (see, for instance, [Dur19][Theorem 3.4.1]), which tells us that the estimates of the means will be approximately normally distributed. In particular, $\hat{\mu}_A \sim N\left(\mu_A, \frac{\sigma_A^2}{n}\right)$ and $\hat{\mu}_B \sim N\left(\mu_B, \frac{\sigma_B^2}{n}\right)$, where $\hat{\mu}_A$ and $\hat{\mu}_B$ are estimators of the mean of the gene expression values for condition A and condition B, respectively, given by:

$$\hat{\mu}_A = \frac{\sum_{i=1}^{n_A} a_i}{n_A}$$
$$\hat{\mu}_B = \frac{\sum_{i=1}^{n_B} b_i}{n_B}$$

Furthermore, since we do not know the variance, we will use the estimators, $\hat{\sigma}_A^2$ and $\hat{\sigma}_B^2$ given by:

$$\hat{\sigma}_A^2 = \frac{\sum_{i=1}^{n_A}(a_i - \hat{\mu}_A)^2}{n_A - 1}$$
$$\hat{\sigma}_B^2 = \frac{\sum_{i=1}^{n_B}(b_i - \hat{\mu}_B)^2}{n_B - 1}$$

This allows us to apply the t-test. We can calculate the t-statistic from the estimators as follows:

$$t = \frac{\hat{\mu}_A - \hat{\mu}_B}{\sqrt{\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B}}}$$

In the case where the null hypothesis is true, so $\mu_A = \mu_B$, we know that t has a t-distribution with degrees of freedom approximately given by

$$\nu = \frac{\left(\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B}\right)^2}{\frac{(\hat{\sigma}_A^2/n_A)^2}{n_A-1} + \frac{(\hat{\sigma}_B^2/n_B)^2}{n_B-1}}$$

according to the Welch–Satterthwaite equation [Wel47]. Note that in that paper, Welch uses a slightly more accurate equation: $\nu = \frac{\left(\frac{\hat{\sigma}_A^2}{n_A}+\frac{\hat{\sigma}_B^2}{n_B}\right)^2 - 2\left(\frac{(\hat{\sigma}_A^2/n_A)^2}{n_A+1} + \frac{(\hat{\sigma}_B^2/n_B)^2}{n_B+1}\right)}{\frac{(\hat{\sigma}_A^2/n_A)^2}{n_A+1} + \frac{(\hat{\sigma}_B^2/n_B)^2}{n_B+1}}$, but the test which is coded as the Welch t-test in programs such as Python (in scipy) and Julia (in HypothesisTests.jl) uses the simpler equation given above.

This allows us to calculate a p-value for a given t-statistic. Let T_ν be a random variable with a t-distribution with ν degrees of freedom as above. Then the p-value is given by $\Pr(T_\nu < t) + \Pr(T_\nu > t)$, the probability of getting a more extreme t-statistic under the assumption that the null hypothesis is true.

We reject the null hypothesis if the p-value is below a certain cut-off threshold. In most sciences other than physics, this is 0.05.

When testing multiple genes, we apply this t-test gene-by-gene, and then apply a correction for multiple testing to the resulting p-values. The correction usually used is the Benjamini–Hochberg [BH95] to control the false discovery rate (FDR).

We can relate the t-test to our deformation method by using the fact that the genes which are picked out by the t-test as statistically significantly differentially expressed are exactly the ones which have a high distance between the sample means, taking into account the sample variances.

We note that our deformation method has two steps. The first is coming up with a weight on the genes. The second is coming up with a way to evaluate the resulting weight by seeing if we get a separation between the morphs, whether visually or numerically. We introduced the use of Mapper in this second step, but we can also use the t-test as the inspiration for a way to evaluate the weight on genes numerically.

Recall that in our proof of concept (§5.2) we used the set of genes which were significantly differentially expressed to show that we could visualise the difference between the benthic and pelagic morphs in the Mapper output. This is equivalent to giving weight 1 to the genes (dimensions) which are statistically significant, and weight 0 to the genes which are not. That is, we pick the dimensions with sufficiently low p-value. But this is equivalent to picking dimensions with a high magnitude t-statistic, which means a high value of $\frac{\hat{\mu}_A - \hat{\mu}_B}{\sqrt{\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B}}}$. So we have simply picked dimensions with a high value of difference between the (estimated) means of the expression of conditions A and B, taking into account their estimated standard deviations. In particular, the split into two conditions can be seen in the Figure 5.1, since the dimensions which pass the t-test have a large distance between the centroids of the conditions, relative to their variance.

We can use this as a condition for setting weights in our topological method. To do this, we first modify the metric in $\mathbb{R}^m$ so that we get a distance which more closely corresponds to the t-statistic/p-value. We can do this by weighting each dimension by $\frac{1}{\sqrt{\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B}}}$, where we calculate the estimated variance for conditions A and B for the expression values of the respective gene. This means that if we measure the difference between the means of condition A and condition B for each gene, we will get exactly the magnitude of the t-statistic under this new weighted distance.

Now we can use the joint t-distribution and simulations to figure out what the maximum possible distance between the centroids with a given gene weight is, or alternatively we can use the permutation test. Potential ways of finding gene weightings with the maximum distance include algorithms like simulated annealing, and the genetic algorithm. However, this involves a lot of searching, and we have not found an efficient method to do so.

Likelihood-Ratio Test

In cases where we have more than two morphs, we will not be interested in the difference between two particular morphs, but rather which genes have expressions which are better modelled by taking the morphs into account, rather than ignoring them. The method we will use to measure gene importance shall be the likelihood ratio tests. We will recast this method from a statistical into a geometric definition, to show that we can also describe it under our topological deformation method.

First, let us define likelihood:

Definition 5.4.9 (Likelihood)**.** Given some model (probability distribution) and a sample of data (outcome of an experiment) the *likelihood* is the value of the probability distribution at the outcome.

Now we must pick a model of gene expression. For our purposes, we will pick the normal distribution, but typically more advanced distributions are used [LHA14; Pim+17]. Now, we are testing two hypotheses, the null hypothesis, H_0, is that there is no difference between the morphs, and the alternative hypothesis, H_1, is that there is. Under the H_0, we model expression taking into account only the location of the samples. Under H_1, we model expression taking into account both the location and the morph of the samples.

We then calculate the likelihood of our dataset under the models, and take their ratio, giving us a test statistic.

Example 5.4.10. Let us demonstrate testing statistical significance with an example. In this example, we shall assume that our genes expression comes from a normal distribution.

Suppose we have two conditions, A and B, with four samples each, and we are looking at a single gene, with expressions $\{1, 2, 2, 3\}$ in condition A, and $\{7, 8, 8, 9\}$ in condition B.

The expression in condition A has sample mean 2 and sample variance $\frac{2}{3}$, while condition B has sample mean 8 and sample variance $\frac{2}{3}$. If we look at the expression from both conditions, we get a sample mean of 5 and a sample variance of $\frac{76}{7}$.

We can now perform our statistical tests. In the first case, we can see if there is a statistically significant difference in mean between the two fitted distributions $N(2, \frac{2}{3})$ and $N(8, \frac{2}{3})$. We can use the Student's t-test which gives us a test statistic of:

$$t = \frac{8-2}{\sqrt{\frac{2}{3}}\sqrt{\frac{2}{4}}} = 6\sqrt{3}$$

which gives us a two-tailed p-value of approximately 4.649×10^{-5}.

In the second case, we will also use the distribution we fit to the entire dataset, and calculate its likelihood, against the likelihood of the distributions fit to each condition. The distribution fit to the entire dataset has a log-likelihood of about -26.75 and when fit to each condition we get a log-likelihood of about -8.608. If we now use the likelihood-ratio test [Wil38], we get a test statistic of about 36.29 giving a p-value of about 1.084×10^{-286}.

Note that the values of the gene expression were chosen to give us extremely significant p-values. Also, in the case where we test tens of thousands of genes, we run into the multiple testing problem. This reduces the ability for us to call genes statistically significantly differentially expressed. However, the greater problem lies in the realm of theory, since when dealing with so many genes, we simply do not know their function, nor how to infer their function.

As a trivial example, we can use the likelihood statistic as input into our topological deformation model to get the same weighting of the genes. That is, we just input a set of significantly differentially expressed genes, as we did for our proof of concept. This does not yield any more insight.

On the other hand, we can try to define a testing regime which will reproduce the result. That is, we use a method for testing how good a set of gene weights is at separating out specified groups. And the method will be such that the gene weights which are good will have high weight for genes with low p-values in the likelihood ratio test. It will take some work to come up with this method, and we try this using measures of clustering quality in §5.4.3.

5.4.2 Gene Co-Expression Networks

In this section we will provide a brief overview of gene co-expression networks and their analysis. In particular, we will look at an implementation in the Weighted Gene Co-expression Network Analysis (WGCNA) R package [LH08].

Definitions

In this section, we will give some definitions of what gene co-expression is. As usual, let D be an $n \times m$ matrix of normalised expression values, where we have n samples (rows) and m genes (columns).

To quantify gene similarity, we require a similarity measure (or, its dual, a dissimilarity/distance measure). A typically used distance is the Pearson correlation distance:

$$d(i, j) = 1 - \operatorname{cor}(i, j)$$

see Definition 2.2.1 for the definition of Pearson correlation. Note that we are taking the Pearson correlation of genes rather than samples, so we must switch the indices and sum over samples.

Note that the Pearson correlation is equivalent to taking the dot product of the two expression vectors, $\mathbf{y}_{i,k}$ and $\mathbf{y}_{j,k}$, after they have been centred by having their means subtracted from them. Spearman correlation distance, mutual information, and Euclidean distance are other commonly used measures.

Definition 5.4.11 (Gene Co-expression)**.** We say that two genes are *co-expressed* if they have similar expression over the samples in our experiment.

Once we have a distance between genes, our aim is to put the genes into clusters of co-expressed genes. This is done by applying a clustering algorithm. Commonly used clustering algorithms are hierarchical clustering, and k-means [AJK18]. Once we have clusters of co-expressed genes, we can redefine gene co-expression in terms of these. For clustering, see Definition 2.1.15.

Definition 5.4.12 (Gene Co-expression)**.** We say that two genes are *co-expressed* if they belong to the same co-expression cluster.

Hypothetically, gene co-expression clusters will also be biologically meaningful. For example, they might correspond to the genes involved in a certain metabolic pathway, or cellular function. The theory is that genes which have similar expression will be involved in similar functions. The way to extract the function from a gene co-expression cluster is to see if there is an unusually high number of genes which are known to have a certain function in that cluster. However, we do not always have functional biological information for genes, especially those of non-model species. So this makes it difficult for us to determine the biological function of gene co-expression clusters.

Note, there are algorithms in use apart from clustering, and some of the more sophisticated ones, such as `clust` [AJK18], extract subsets of genes with similar expression, rather than trying to cluster/partition the entire set of genes. These algorithms are ultimately based on the idea that the clusters/gene modules we find should satisfy constraints which come from the biology, some of which we give below.

Definition 5.4.13 (Desirable Clusters)**.** *Desirable clusters* are those which, at the very least, have low dispersion and a certain minimum size.

As part of finding desirable clusters, we must pick a measure of dispersion (e.g. mean squared error, the average distance of points to the centroid of their cluster) and a cut-off for our measure, and a minimum size. The idea is to have clusters which are large enough that they are unlikely to be spurious, while at the same time still having most of the points occurring in a small vicinity. We may also consider other properties, depending on what we're looking for, for example, we might desire that the expression of genes in clusters do not have much overlap with those in other clusters.

WGCNA

We will outline Weighted Gene Co-expression Network Analysis (WGCNA) [LH08] as an example of a gene network analysis technique. As input, it receives gene expression data suitably quantified and normalised. Usually, for computational reasons, only a subset of genes is analysed (e.g. the 4,000 most variable genes).

Distance Instead of directly defining a distance, WGCNA first defines a similarity measure between the genes. It uses the absolute value of the Pearson correlation raised to a power: $|\mathrm{cor}(i,j)|^{\beta}$, where $\beta \in \mathbb{N}$. The power β is chosen to be the smallest value such that the resulting network is scale-free. The scale-free property is used since it is seen to occur often in biological networks [Alb05], but whether or not there is a biological basis for scale-free networks is disputed [GP12].

WGCNA further modifies the similarity measure by using the *topological overlap*, which considers the relative interconnectedness between nodes. It is defined as follows:

$$\omega_{ij} = \frac{l_{ij} + a_{ij}}{\min\{k_i, k_j\} + 1 - a_{ij}}$$

where $l_{ij} = \sum_{u \neq i,j} a_{iu} a_{uj}$ and $k_i = \sum_{u \neq i} a_{iu}$. The a_{ij} are some measure of similarity between genes (nodes) i and j, and for WGCNA we take $a_{ij} = |\text{cor}(i,j)|^{\beta}$.

Since $\omega_{ij} \in [0,1]$, we can turn the topological overlap into a dissimilarity (pseudometric) measure by subtracting it from 1, so $d(i,j) = 1 - \omega_{ij}$.

Clustering Now that we have distances between the genes, WGCNA uses a Dynamic Tree Cut method [LZH08] involving average linkage hierarchical clustering to form modules. We will give a brief outline of their method here. More detail can be found in the supplementary material of the above-cited paper.

The input is a dendrogram of the topological overlap dissimilarities clustered using average linkage clustering, and four parameters, which define the kind of clusters we want to end up with. The parameters are N_0, $h_{\max}$, $g_{\min}$, and $d_{\max}$, with the choice of a few defaults for use in WGCNA. N_0 is the minimum cluster size, and influences n_c, the number of objects in the core of a cluster, so that

$$n_c = \min\left\{\text{int}\left(N_0/2 + \sqrt{N - N_0/2}\right), N\right\}$$

where N is the number of objects in the cluster under consideration. $h_{\max}$ is the height in the dendrogram above which no joining occurs. $d_{\max}$ is the maximum value allowed of the core scatter $\bar{d}$, which is defined as the average pairwise dissimilarity between all elements of the core. $g_{\min}$ is the minimum gap size, where the gap is defined as the difference between the core scatter, $\bar{d}$, and the height where the cluster joins the rest of the dendrogram.

The algorithm works through the dendrogram from the bottom to the top, and creates clusters whenever it tries to join two branches which both satisfy all the criteria to be considered clusters of themselves. For each cluster, we measure the average of the distance from a point to every other point in the cluster, then take its maximum. Then, looking at all unclustered points and points in clusters which failed to attain the minimum cluster size, we assign them to clusters if the average distance between the unclustered points and the cluster is less than the maximum distance found. That is, suppose we have a cluster indexed by $i \in \mathcal{I}$, where $|\mathcal{I}| = n$. Then to this cluster, we assign the distance:

$$d_{\mathcal{I}} = \max_{i \in \mathcal{I}} \frac{\sum_{j \neq i} d(i,j)}{n-1}$$

and suppose we have an unclustered point p. Then if

$$\frac{\sum_{i \in \mathcal{I}} d(p,i)}{n} < d_{\mathcal{I}}$$

we assign the unclustered point to the cluster $\mathcal{I}$.

As an optional additional step when dealing with gene expression data, we can find the eigengene (first principal component of the expression) of a cluster and merge together clusters with eigengenes that have greater than 0.75 Pearson correlation. In the end, the clusters we get are considered gene co-expression modules, and two genes are co-expressed if they both belong to the same module.

Simplicial Complexes

In this subsubsection, we will give an alternative view of gene co-expression as defined on a simplicial complex. Examples of gene co-expression from clustering and WGCNA viewed from this perspective can be found in §5.4.3. As usual, let D be an $n \times m$ matrix of normalised expression values, where we have n samples (rows) and m genes (columns). We will now consider how this data can be used to build a simplex of gene co-expression, with given desirable properties determined by biological considerations.

Let X be the simplicial complex under consideration. In fact, X will be a $(m-1)$-simplex, with one vertex for each gene. We will consider successively higher dimensions of the simplicial complex, and give a description of them.

Definition 5.4.14 (0-Skeleton)**.** The 0-skeleton of X, X_0, consists of m vertices, one for each gene. Attached to each vertex is a vector of gene expression in $\mathbb{R}^n$. Let this extra data be denoted by s_i, for $i \in \{0, 1, 2, \ldots, n-1\}$. This will be used to determine co-expression in the higher dimensions of the simplicial complex.

In order to define the 1-skeleton of the simplicial complex, X, we will need a notion of dissimilarity between elements in $\mathbb{R}^n$.

Definition 5.4.15 (Dissimilarity Function)**.** A *dissimilarity function* $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_{\geq 0}$, takes two elements of $\mathbb{R}^n$ and gives a non-negative real number. The dissimilarity function should satisfy the following two properties:

1. $d(x, x) = 0$
2. $d(x, y) = d(y, x)$

An example of a dissimilarity function is the Pearson correlation distance $d(i, j) = 1 - \text{cor}(i, j)$ with $\text{cor}(i, j)$ as in Definition 2.2.1.

Definition 5.4.16 (1-Skeleton)**.** The 1-skeleton of X, X_1, is a weighted graph, where the weight of the edge $e(i, j)$ is given by the dissimilarity function $d(s_i, s_j)$.

Standard methods in gene co-expression analysis, such as WGCNA (which we describe in the next subsubsection) take a weighted graph, or, rather, a dissimilarity matrix as input. This is also the kind of data we could input into a clustering algorithm, to end up with clusters which we then interpret as co-expressed gene modules.

We now define higher dimensional structures in the simplicial complex. This requires us to choose some measure of cluster dispersion, which will allow us to associate a number with n-simplices, for $n \geq 2$. We give a definition of cluster dispersion, and some examples of cluster dispersion measures, here.

Definition 5.4.17 (Cluster Dispersion)**.** *Cluster dispersion* is a measure of how loose a cluster is. It is based on the elements in a cluster, and their distance either from each other, or from some common point (say, the barycentre). It is desirable for us to have clusters with low dispersion, meaning each clusters has genes which do not much differ from a given expression profile.

Example 5.4.18 (Mean Squared Error)**.** The *mean squared error*, MSE, of a cluster is given by:

$$\mathrm{MSE} = \sum_{i=1}^{k} \frac{|s_{c_i} - \bar{s}_c|^2}{k}$$

where the c_i are indices of the genes in the cluster under consideration, k is the total number of genes in the cluster, $\bar{s}$ is the barycentre (centroid) of the gene expression of genes in the cluster, and s_{c_i} is the expression of gene c_i.

Note, we can only use MSE in spaces where we can calculate the barycentre.

Example 5.4.19 (Scatter)**.** The *scatter* of a cluster is given by:

$$d_s = \frac{\sum_{i=1}^{k-1} \sum_{j>i}^{k} d(s_{c_i}, s_{c_j})}{\frac{n(n-1)}{2}}$$

where the c_i are indices of the genes in the cluster under consideration, k is the total number of genes in the cluster, and s_{c_i} is the expression of gene c_i.

Definition 5.4.20 (2-Skeleton)**.** The 2-skeleton of X, X_2, consists of triangles between triples of vertices, and all lower dimensional simplices. Now suppose we are given a measure of cluster dispersion, $f : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_{\geq 0}$, for example, the MSE, or scatter. We can use this to associate a weight to each 2-simplex.

Definition 5.4.21 (n-Skeleton)**.** The n-skeleton of X, X_n, consists of n-simplices, and all lower dimensional ones. n-simplices consist of $n+1$ vertices, and we can associate a weight to these using some measure of cluster dispersion, as above for the 2-skeleton, with some function $f : \left(\mathbb{R}^l\right)^{n+1} \to \mathbb{R}_{\geq 0}$, where l here is the number of samples.

Given our simplicial complex X with a weight on each of its component simplices, we can define a subcomplex $C \subset X$ which contains only simplices with weights under a certain cut-off, chosen with biological considerations in mind. In this case, we can give the following definition of gene co-expression:

Definition 5.4.22 (Gene Co-expression)**.** We call a set of genes *co-expressed* if the simplex they form is part of the subcomplex $C \subset X$, where C is a subcomplex of the $(m-1)$-simplex X

Note that we can also use more advanced procedures for choosing the subcomplex $C \subset X$, and in fact we can also have more than one weight on every simplex on X, but we will restrict ourselves to this simpler model for now.

5.4.3 Correspondence to the Topological Method

In this subsection, we will show how the results of the differential gene expression analysis and gene co-expression analysis methods can be approximately reproduced in our method of deforming metrics.

Differential Gene Expresssion

We will begin by showing how we can take into account some of the results of differential gene expression. There are three ways this could occur. Firstly, we can come up with a method which assigns a weight of one to all genes which a given method indicates are differentially expressed. Secondly, we deform the metric (weight the genes) in such a way that, when we look at the genes individually, any gene which has the average difference in expression between conditions differing over a given cut-off is considered differentially expressed. Finally, we can run an optimisation algorithm (e.g. simulated annealing) on the gene weights, where the objective function to optimise is a measure of cluster quality, where the clusters are given by the morphs.

Trivial Case As usual, we begin with D, an $n \times m$ data matrix, with n samples and m genes, where the replicate effect has been subtracted. Now, suppose we have k morphs, so $n_1 + n_2 + \ldots + n_k = n$, where n_i is the number of samples of morph i. Then, if we look gene-by-gene and apply the same statistical test of differential gene expression as in §5.4.1, and define our weighting as 1 if a gene passes this test, and 0 otherwise, then we trivially get back the same set of genes with weight 1 as our given differential gene expression method.

Weight On the other hand, suppose we want to define a metric which is related to certain statistical tests of differential gene expression. A reason to do this is that we desire some sort of generalisability for our gene expression results, and having a metric which has a statistical meaning is one way to achieve this.[7]

We will use relations between distances and probability distributions to approximate the gene-by-gene statistical tests which are done in differential gene expression

[7] In order to be biologically accurate, we would have to have the metric vary depending on what point in the space we are, but I doubt we would ever collect enough experimental, or even observational, data for that.

analysis. We will give an example of a metric deformation, based on the idea of the Mahalanobis distance [Mah36]. Note, a limitation of our approach involving the deformation of metrics is that we can only assign one weight/scaling factor to each independent dimension. This makes it impossible for us to precisely reproduce the results of statistical tests like Welch's t-test [Wel47], where both the test statistic and its distribution depend on the variance of the samples under consideration.

First, we assume that each gene (dimension) has the same variance across different conditions (even if the mean might vary). Then, we can define a weight for each dimension by calculating the pooled variance, which is given by:

$$s_{p,g}^2 = \frac{(n_1-1)s_{1,g}^2 + (n_2-1)s_{2,g}^2 + \ldots + (n_k-1)s_{k,g}^2}{n-k}$$

where $s_{i,g}^2$ is the same standard variance for condition i and gene g, and we calculate this for every gene, g, separately. We can now weight each gene by $1/s_{p,g}^2$. This will give us a distance where, if we project down on each dimension, then a distance of 1 is approximately equal to a standard deviation. And if we do not project down, then this gives us a distance in our original space $\mathbb{R}^m$, where a distance of 1 is about one standard deviation, assuming that all genes are independent.[8]

Note: This does not actually reproduce the t-test on each gene (dimension), since the t-test also considers the sample size, rather than just the estimated variance. If we wished to take this into account, we would have to pick two morphs, say k_1 and k_2, which we are interested in conducting the test between. Then, instead of weighting each dimension by just $1/s_{p,g}^2$, we weight by:

$$\frac{1}{s_{p,g}^2\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

where n_1 and n_2 are the number of samples of morph k_1 and k_2, respectively. While this will produce a weight so that, on each dimension, the weighted distance between the mean expressions of k_1 and k_2 is exactly the t-statistic which is used to test them, I am unaware of any more meaning the weighting might have.

Counterexample We will now go over a counterexample which shows we cannot implement the Welch t-test by a deformation of the metric. An example of the Welch t-test, along with its definition, can be found in Example 5.4.8. We will consider two morphs and two genes, weighted so that the distance between the mean expression of the morphs is precisely the t-statistic. However, the variances will vary between the two genes and morphs, so that, even though the second gene will have a greater t-statistic (in absolute value), it will not be considered statistically significant (p-value

[8] This assumption is usually false. Nevertheless, it is the assumption made when testing for differential gene expression. Especially in non-model organisms, and even in model organisms, the dependence in expression amongst genes is unknown. In fact, it's likely different even in different tissues, or the same tissue under different states, e.g. fed vs. starving.

≥ 0.05), while the first gene will.

So, let A and B be morphs, with 4 samples each, and g_1 and g_2 be genes with expression $g_1 = (1, 2, 3, 4, 3.5, 4.5, 5.5, 6.5)$ and $g_2 = (2, 2, 2, 3, 2, 4, 6, 8)$, where the first four samples belong to morph A, and the last four to morph B. If we now weight each dimension by $1/\hat{\sigma}_{\bar{\Delta}}$, where

$$\hat{\sigma}_{\bar{\Delta}} = \sqrt{\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B}}$$

and $n_A = n_B = 4$ in this case, the distance between the means of g_1 will be approximately 2.74, and between the means of g_2 will be approximately 2.85, which are the t-statistics for the Welch t-tests in each case. However, g_1 has a p-value of 0.03, while g_2 has one of 0.06, where we have rounded both values to two decimal digits. So, despite having the smaller t-statistic, g_1 is statistically significant ($p < 0.05$), but g_2 is not. This is due to the Welch t-test also depending on the estimated variance of the samples in each morph, and the way we constructed this example, the variances are much more unequal in g_2 compared to g_1, which reduces the degrees of freedom used in the t-test, and in turn requires a higher value of the t-statistic in order to reject the null hypothesis. The point is that, despite using a weighting which gives the resulting t-statistic in each dimension, this is not directly related to the statistical significance, so similar ‘distances’ (t-statistics) can give different statistical results.

Validity We will now talk about the statistical tests which can be generalised by weighting the dimensions.

Definition 5.4.23 (Weighting-Compatible)**.** A statistical test is *weighting-compatible* if its test statistic has the same probability distribution for each dimension(/gene).

A weighting-compatible statistical test allows us to weight each dimension so that the value of the test statistic appears as the weighted distance between, for example, the means of different groups, or otherwise, depending on how the test statistic is calculated. Furthermore, having the same probability distribution for each dimension means that the statistical meaning of these distances is comparable between dimensions.

Clustering Quality The second kind of differential expression method we looked at above used likelihood-ratio tests, rather than parameter estimation, to determine if genes are significantly differentially expresssed. We will now consider an implementation of the topological method which takes into account the differential gene expression method using likelihood-ratio tests. We will do this by considering the likelihood-ratio test as analogous to tests of clustering quality, where the partition of our samples into morphs provide the clusters to be tested.

Firstly, we assume we have a data matrix D where the replicate effect has been subtracted. Now, let us assume we have k morphs. The goal is to find a perturbation under which the k morphs are apparent as clusters.

One possible way to assess clustering is to look at the sum of the squared distances from each sample point to the global sample mean (centroid), and compare that to the sum of the squared distances from the morph centroids to the overall centroid (weighted by the number of samples in each morph). Let us now give a definition for these sums of squared distances:

Definition 5.4.24 (Sum of Squared Distances)**.** Let SS_T be the *total sum of squares*, and SS_B be the *between sum of squares.* Then we have:

$$SS_T = \sum_{j=1}^{n} ||\mathbf{x}_j - \mu||^2$$

where n is the number of samples and μ is the mean over all the samples, and:

$$SS_B = \sum_{i=1}^{k} n_i ||\mu_i - \mu||^2$$

where k is the number of morphs, n_i is the number of samples of morph i, μ_i is the mean of the samples of morph i, and μ the mean over all samples.

SS_B will always be a fraction of SS_T, and if the fraction is great, then we have a good clustering, since this means our clusters take into account most of the variation in the data. In the context of finding suitable perturbations, we could pick one under which $\frac{SS_B}{SS_T}$ is sufficiently high.

Another way is to use an ANOVA, which tests if the means of each group (morph in our case) is the same. In this case, the sum of squares is also used, but scaled in the following manner.

Definition 5.4.25 (ANOVA F-test Statistic)**.** The *ANOVA F-test statistic* for a partition of our samples is given by:

$$F = \frac{\frac{SS_B}{k-1}}{\frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} ||\mathbf{x}_{ij} - \mu_i||^2}{n-k}}$$

where SS_B is the between sum of squares (see Definition 5.4.24), k is the number of morphs (sets in the partition), n is the total number of samples, and n_i is the number of samples of morph i.

Note that this is equivalent to the t-test when we have only two groups. (Refer to §5.2 where we had a proof of concept for our topological method using t-tests on the individual genes.)

A way to pick an appropriate perturbation would be one which has a significant enough F-statistic. For example, we might want to find a perturbation which gives us an F-statistic in the top 8×10^{-5} fraction.

We will first show that, if an F-test applied to each dimension separately produces a significant enough F-statistic, then the F-test applied to all dimensions will also produce a significant F-statistic.

For convenience, we will write the sum of squares occurring in the denominator of Definition 5.4.25 as:

$$SS_R = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \|\mathbf{x}_{ij} - \mu_i\|^2$$

which we call the *residual sum of squares.*

Theorem 5.4.26 (F-test)**.** *Assume we have a data matrix A, with the replicate effect subtracted, and a partition of the n samples (rows) into k groups. Suppose that we want an F-statistic of a certain significance value, p, so that the F-statistic is greater than a certain value F_p. Then, if we have genes $\mathbf{g}_1, \ldots, \mathbf{g}_l$ where the F-test applied to each gene individually gives an F-statistic greater than F_p, then the F-test applied to all the genes at the same time will also give an F-statistic greater than F_p.*

F-test. Let F_o be the value of the F-statistic for the gene o. Then, by assumption, each $F_o \geq F_p$. In particular, we have $F_o = \frac{\frac{SS_{B_o}}{k-1}}{\frac{SS_{R_o}}{n-k}} \geq F_p$. Furthermore, since we can obtain the sum of squares over multiple dimensions by simply adding up the individual dimension, we have $SS_B = \sum_{o=1}^{l} SS_{B_o}$ and $SS_R = \sum_{o=1}^{l} SS_{R_o}$,where SS_B and SS_R are the between and residual sum of squares of all the l genes under consideration, respectively. Now, since we have $\frac{\frac{SS_{B_o}}{k-1}}{\frac{SS_{R_o}}{n-k}} \geq F_p$ for all $o \in \{1, 2, \ldots l\}$, we know that $\frac{SS_B}{k-1} = \frac{\sum_{o=1}^{l} SS_{B_o}}{k-1} \geq F_p \frac{SS_B}{n-k} = F_p \frac{\sum_{o=1}^{l} SS_{B_o}}{n-k}$. So, if F is the F-statistic for the l genes under consideration, then we have $F = \frac{\frac{SS_B}{k-1}}{\frac{SS_R}{n-k}} \geq F_p$, as desired. □

A way that this method extends the standard differential gene expression analysis method is that, instead of looking gene-by-gene, we can look at (weighted) gene sets. This flexibility also allows us to consider gene sets with genes which would not make the cut-off we have chosen. We can combine this with looking at correlated gene sets, so we can tell if these are statistically different or not between the morphs.

In fact, this method allows us to test arbitrary gene sets, rather than just an over-representation analysis on a subset of differentially expressed genes, or on a list of genes ordered by significance.

Gene Co-Expression

It is not possible to take into account gene co-expression by deforming metrics, assuming that we are dealing with D a $n \times m$ data matrix, with n samples and m genes and the replicate effect subtracted, which we consider as n points in $\mathbb{R}^m$. Furthermore, it is not possible to interpret the output, which will be a set of weights on the dimensions, as having to do with gene modules. The best we could do is to attempt to find all the genes in a single module, and weight these genes one, and the others zero. Then we could remove the genes which have been assigned to the module, and repeat the procedure. Alternatively, it may be possible to define some number of abstract modules, and weight genes between zero and one by their membership in these modules.

We will instead discuss in the section other methods for implementing gene coexpression analyses with a topological bent.

Firstly, we require a measure of dissimilarity/distance between genes. It is typical to consider the genes as m points in $\mathbb{R}^n$, where the values of the coordinates are given by their expression in the samples. From here, usually correlation distances are used, since we are only interested if genes have the same shape of expression as each other.

Spearman's Correlation We have a choice of distances to use between genes, when trying to determine gene sets. A good choice appears to be Spearman correlation [BVG15], which is define as follows:

Definition 5.4.27 (Spearman Correlation)**.** The Spearman correlation between two genes i and j is given by

$$\operatorname{cor}_s(i, j) = \operatorname{cor}(\mathrm{rg}_{\mathbf{y}_i}, \mathrm{rg}_{\mathbf{y}_j})$$

where cor is the Pearson correlation (see Definition 2.2.1), $\mathbf{y}_i$ is the expression vector of gene i, and $\mathrm{rg}_{\mathbf{y}_i}$ is a vector of the ranks of $\mathbf{y}_i$.

In order to determine modules of genes, we can pick a cut-off for the Spearman correlation, and then produce a network and take its connected components. This is the same as single-linkage clustering on the Spearman correlation distance.

Mapper If we recall §2.2.3, we know that it is possible for the topological data visualisation Mapper to find tightly linked subgroups of the data which cannot be found with standard clustering methods. Could it be the case that applying Mapper to genes, rather than samples, also allows us to find subgroups of genes which are not found by clustering?

To investigate, we apply Mapper and WGCNA to genes from the Arctic charr. We picked the top 10,000 most significantly differentially expressed genes between the benthic and pelagic ecotypes, due to computational constraints. We process these genes first with the pipeline outlined by the package authors on their webpage: `https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/`. The distance used for Mapper is the toplogical overlap calculated by WGCNA based on $|\operatorname{cor}(i, j)|^6$, the absolute Pearson correlation raised to the sixth power. For WGCNA, we continue with average-linkage hierarchical clustering, and then find modules with a minimum module size of 30, combining modules with eigengenes which are correlated with greater than 0.75 Pearson correlation.

Figure 5.7 shows the results of Mapper applied to the top 10,000 most significantly differentially expressed genes between the benthic and pelagic ecotypes. We use the cluster cut-off of 0.998, which corresponds to the $h_{\max}$ used by WGCNA (see §5.4.2 for

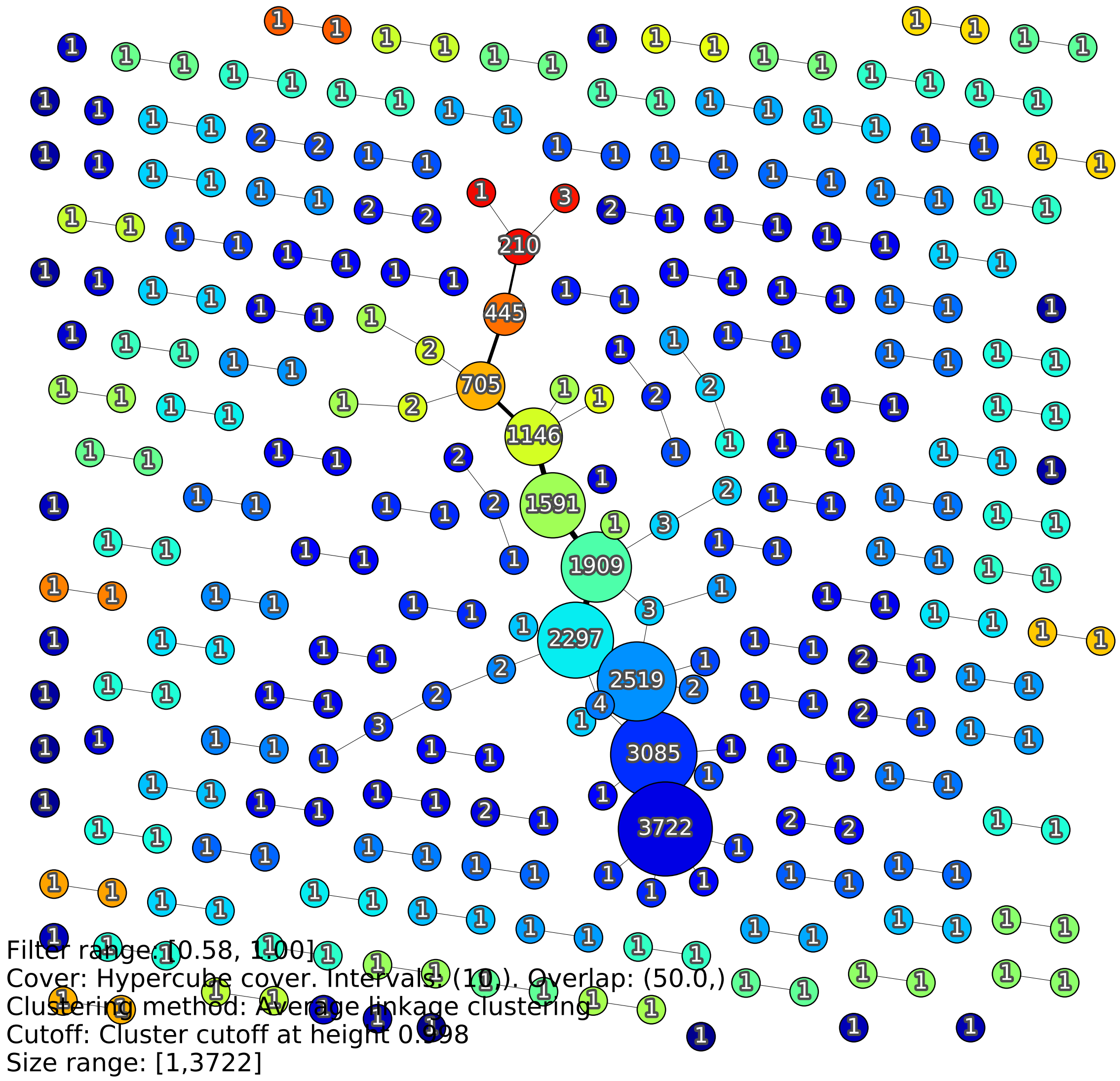

210
445
705
1146
1591
1909
2297
2519
3085
3722
Filter range: [0.58, 1.00]
Cover: Hypercube cover. Intervals: (10,). Overlap: (50.0,)
Clustering method: Average linkage clustering
Cutoff: Cluster cutoff at height 0.998
Size range: [1,3722]

Figure 5.7: Mapper output. The filter function is $1-p$-value, with 10 intervals overlapping by 50%. Clustering function is average linkage clustering, with a cut-off of 0.998, which was the $h_{\max}$ parameter WGCNA used when run on the same data. In the case of Mapper, we end up with basically one large cluster, which is visualised as a line, since the gene have been binned by $1 - p$-value.

more details). There is basically only one cluster, split by the filter function into nodes of different p-value. So, we see now interesting subgroups of genes in this case. When we apply WGCNA to the same data, as we see in Figure 5.8, we get eleven modules, showing that WGCNA is doing something different than Mapper.

Simplicial Complexes We now move on to examples of gene co-expression expressed in the language of simplicial complexes, as define in §5.4.2.

Example 5.4.28 (Clustering)**.** First, we describe how to view clustering from this perspective. We will only consider the example of hierarchical clustering here, since the implementation of other clustering algorithms will depend on the algorithm. Suppose the hierarchical clustering algorithm on the m genes is based on the dissimilarity measure d. We begin with the 1-skeleton with m vertices and edge-weights coming

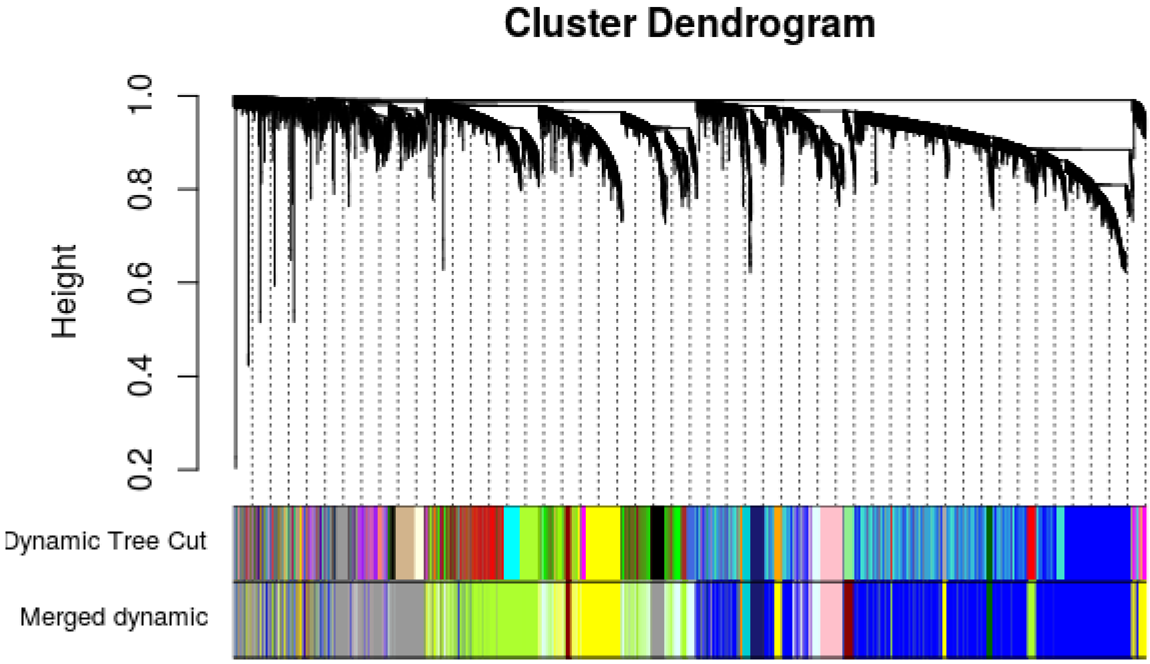


Figure 5.8: WGCNA applied to the top 10,000 most significantly differentially expressed genes between the benthic and pelagic ecotypes. We followed the procedure in the WGCNA tutorial on the authors' webpage: `https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/`.

from some distance/dissimilarity function between gene expression profiles. A hierarchical clustering algorithm builds up a dendrogram by successively joining clusters at a certain height. We can build up a simplicial complex X by following the cluster dendrogram, and including simplices on those clusters, weighted by the weight at which the cluster is formed in the dendrogram. We can then form a subcomplex $C \subset X$ simply by picking a cut-off height, h, and considering only simplices with weight $w < h$. The resulting subcomplex will have simplices corresponding exactly to the cluster we would get by cutting the cluster dendrogram at the same height h.

Example 5.4.29 (WGCNA)**.** Now we consider WGCNA from this perspective. We start by using the topological overlap dissimilarity to weight the edges in the 1-skeleton. Then we can apply average-linkage hierarchical clustering, with parameters N_0, $h_{\max}$, $g_{\min}$, and $d_{\max}$ as in §5.4.2 where we describe WGCNA. The extra parameters can be considered as weights on the simplices in the simplicial complex X. In this case, we'd have a weight for the core scatter, $\bar{d}$, and the gap height, as well as the height as we would when implementing normal hierarchical clustering.

5.5 Discussion

Following the set-back of Mapper being unable to visualise our Arctic charr dataset, we were inspired to find a better way of visualising the data. We decided to try a method inspired by topology, which involves weighting the genes, and then visualising the resultant perturbed gene expression space. Since we noticed in the Arctic charr that the effect of lake dominates the results, we generalise this to a *replicate effect* [Def. 5.1.1] which we define as a vector of gene expressions associated to each replication. Once we have this replication effect, we subtract it to remove its influence on the results.

When the lake effect was removed from our Arctic charr, we still found that, upon applying Mapper, we did not get the expected split into two morphs. To help us visualise the difference between the morphs, we perturb the gene expression space, until we find a perturbation where the two morphs are apparent in the visualisation. This idea can be generalised to where we are looking for any interesting groups in a dataset, from morphs, to different tissue types, to *Drosophila* bred in different seasons, and so on. We provide a proof-of-concept of the perturbation method, by showing that we can pick out a perturbation which makes the benthic and pelagic morphs of the Arctic charr distinct in visualisations.

Following this, we conduct some experiments, testing the assumptions we made when we were coming up with our perturbation method. One of these was that our approximation to the replicate effect (the centroid of the gene expressions of the samples in a given replication) actually removed the replicate effect. To check, we look for a residual replicate effect, which amounts to checking if we can tell the difference between replicates in our groups of interest. We find that there are genes for which we can, but overall, there is no significant residual replicate effect, so our naïve first approximation to the replicate effect does seem to get rid of most of it.

The last experiment we conduct is looking at the effect of sample size, and if increasing sample size would make it easier for us to distinguish any biologically interesting signals from biological noise. We also test the ability for perturbations to increase the signal-to-noise ratio. We find that, in the case of the *Drosophila*'s 726 samples compared to our Arctic charr's 32, we still see no significant increase in the signal-to-noise ratio when perturbing due to there being so much biological variation between the samples.

Finally, we turn to an overview of theory, looking at the two major current methods of analysing gene expression data, by differential gene expression (DGE) or by gene co-expression networks, and giving a mathematical description of them, which we use to compare these methods to our perturbation method. We find that while our method can cover some simpler methods of DGE, like Student's t-test, more complicated methods like Welch's t-test will need something more than perturbing gene expression space by weighting the genes. We also find that we can relate gene co-expression networks to clustering, and through them to Mapper and the idea of simplicial complexes, which one can consider as higher-dimensional clusters.

In our last chapter, we will discuss more about the difficulties we had, working with non-model species and a small sample size, and propose a follow-up experiment which could be done in the future.

Chapter 6

The Shape of Things to Come

In this final chapter, we discuss our results, the application of Mapper to the Arctic charr, our success in visualising the difference between the benthic and pelagic morphs with a pre-selected gene set. The project aimed to overcome challenges of working with a non-model species and small sample sizes. We tested a topological application on a new data type and found it did not resolve biological patterns. This is due to limitations of that approach when applied to a non-model species in an ecological/evolutionary context. These include limitations with small sample size, which mean we are unable to find any novel subgroups in the Arctic charr data we do have. We combined information across lakes to maximise biological and ecological information. However, this added a layer of evolutionary complexity that the Mapper approach was not able to deal with. In fact, we find (apparently) new "noise". This is expected due to evolutionary distance, and inherent in these kinds of datasets.

To address these challenges and propose a direction for the future, we present an potential experiment, assuming there are fewer limitations on data collection, such as a lower cost of sequencing. This will involve sequencing the RNA of the Arctic charr as they develop on either a benthic or pelagic diet, and using the changes in gene expression between these two experiments to work out which pathways are, finally, responsible for the division into two ecomorphs, at least from a developmental plasticity point of view. These will give us candidate genes which we can look at to work out what the underlying genetic causes of these two morphs might be.

6.1 An Evolving Field

An early motivation for the project was Nicolau *et al.*'s application of Mapper, a topological data analysis-based visualisation algorithm, that provided an important advance by discovering of a new subtype of breast cancer that other bioinformatics approaches had missed. Aspects of their approach which we wanted to utilise are the overcoming of the noise of gene expression data, and the distilling of subtle but biologically important signals. These are particularly relevant to data from an ecological and evolutionary context, where we expect to have a subtle signal and a lot of local variation. In seeking this new application of an already quite untested topological approach, we found it to be lacking in several aspects. Firstly, there was a healthy vs. diseased distinction in the breast cancer data, and this asymmetry does not translate over to the benthic vs. pelagic morphs. That is, for the breast cancer data set, we could consider healthy expression as baseline, and diseased expression as deviation from baseline expression, while for the Arctic charr, there is no obvious choice for picking either benthic or pelagic to be baseline and the other to be deviation from baseline. To solve this issue, we introduce the notion of lake effect (§3.4.1), which can be viewed as the baseline expression of each lake, and remove it from the gene expression data, so we are able to see the difference between the morphs more clearly. This also helps to solve the problem of the lake signal being greater than the morph signal, as we can see in Fig. 3.4. This gets us to the data of interest, which is the morph signal.

Secondly, in the breast cancer data, they found a distinct subgroup of tissue samples which had an unusual trait (100% survival) and were able to show statistically that this constituted a subtype by isolating these tissues and conducting statistical tests comparing their expression with the rest of the breast cancer tissues. We did not find any consistent new subgroups in the Arctic charr data, either because our sample size (36 fish) is too low, or there are no such consistent subgroups, or because there are different pathways to the same benthic/pelagic effect. That is, we were not able to resolve if the lack of repeated signal from the topological approach is biological reality from possible limited power. However, traditional approaches have suggested that there are differences, but these were not picked up, see, for example, the work of Jacobs *et al.* [Jac+19].

The fact that subgroups indicating the presence of two morphs were not inferred by the Mapper approach, inspired us to look at the gene expression data by deforming the space. In particular, we had expected Mapper to at least be able to pick out the two morphs, which are apparent when we use a more traditional approach. Deforming the gene expression space came up as a topological elaboration of the more traditional approach of selecting a set of genes by differential gene expression analysis. Practically speaking, our deforming the space is done by weighting the dimensions corresponding to genes, and visualising the resulting points. We can see in Figure 5.1 that we do get the morphs into two different groups when we weight just the differentially expressed genes by one, and the rest by zero. However, we find in §5.3.2 that this occurs only by reducing much of the strength of the signals in the data. This implies that, to get the desired visualisation, we must ignore a lot of potential signal in the data, which may be of biological relevance.

An additional complication when working with non-model species is that there will not be many details about what the genes do. There are no gene ontology (GO) annotations or KEGG pathway information about the Arctic charr. If we look instead at a more well-studied close relative, the Atlantic salmon (*Salmo salar*) we see that of the $\sim 80,000$ gene products, $\sim 60,000$ have GO annotations, but of these only 454 are manual, and the rest have been inferred electronically using algorithms. Regarding KEGG pathways, only 13,514 of 55,214 genes are in a pathway. Nevertheless, we can still use these genes to help inform on the functions of any genes of importance that we find.

There has also been work by various groups to associated regions in Salmonid (including Arctic charr) genomes to physiological traits, with the work summarised in a paper of Jacobs *et al.* [Jac+17]. This shows that the amount of information on Arctic charr genomes is building up over time.

While the project has been ongoing, there have been developments which put our assumption of removing the lake effect into question. For example, work by Jacobs *et al.* [Jac+19] have indicated that the difference in gene expression is not expected to be concordant across different lakes. Further discussion with biologists have also led to the difference between lakes being more of a concern. In particular, there is the thought that there are different mechanisms between lakes, or at least the differences might be in different genes, but involved in the same pathway.

Considering all these developments, our aim is to propose an experiment that would overcome the current limitations of data size and address the outstanding questions about the association between the development of the benthic and pelagic morphs, and the gene expression of the fish as they develop.

6.2 Experiment

Considering the challenges described in the previous sections, we will now design an experiment to overcome the limitations of existing approaches of working with a non-model species with a small sample size, in order to elucidate the genetic mechanisms behind a potential incipient speciation event. This will require the currently unrealistic scenario of being able to sequence thousands of samples,[1] and there is also an information gap between model and non-model which will require many experiments and much data to bridge.

We will be taking inspiration from developmental biology [Sch+17], drawing on an analogy between developing organisms and cells. Figure 6.1, [Sch+17, Fig. 1], shows the idea behind Schiebinger *et al.*'s using optimal transport to work out what genes are involved in the regulation of cell differentiation. The general idea is that there is a developmental landscape by which cells develop (shown by railway tracks (**A**) and valleys in a landscape (**B**) in 6.1). The primary difference between their experiment, and ours previous ones, is the use time-series data. This additional dimension lets them shed light on the genes and regulatory pathways involved in the differentiation of

[1] It should be noted here that sequencing an Arctic charr transcriptome with library preparation and to 30M reads per individual takes about £500, so doing one run of this experiment will take about £500,000. This is despite the sharp decrease in sequencing cost over recent decades. Partly, this is due to the Arctic charr's large transcriptome (the salmonids all underwent a recent gene duplication). Another factor is also the use of external agencies for sequencing, which have their own costs.

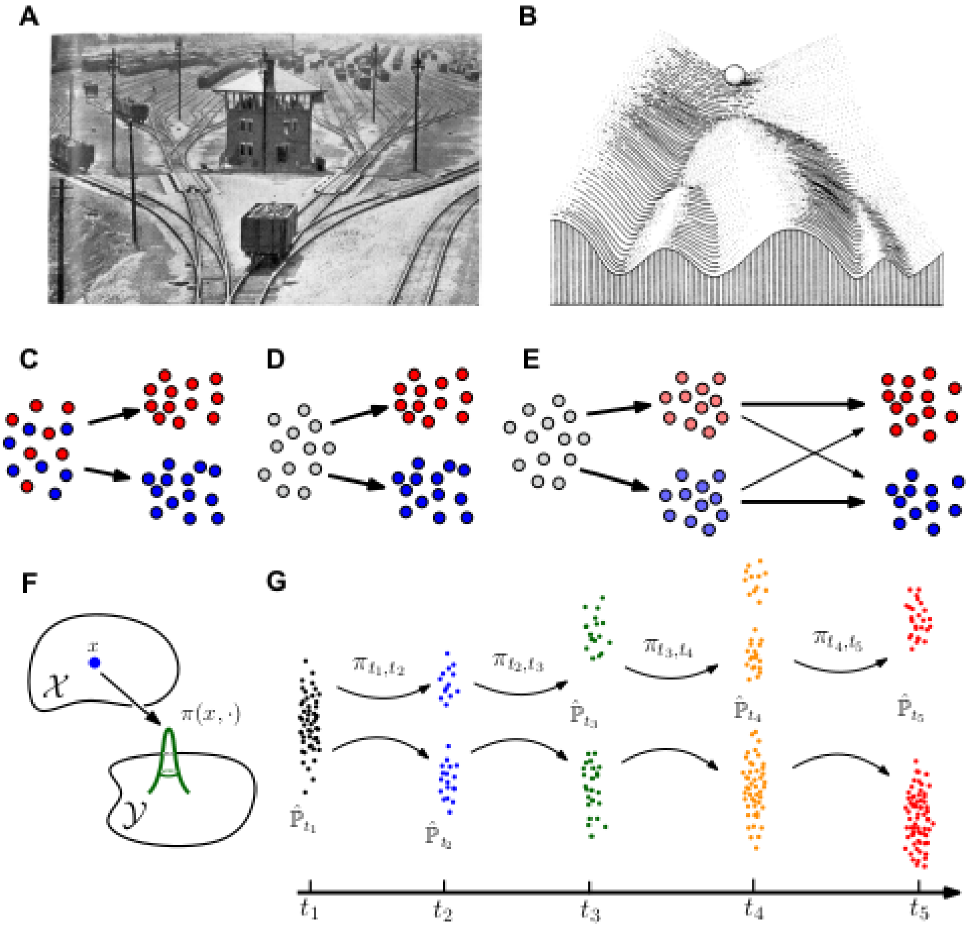


Figure 6.1: [Sch+17, Fig. 1], with (**A–B**) showing Waddington's analogies of cells undergoing differentiation, as like railroad cars switching tracks (**A**) or marbles rolling down a landscape (**B**). (**C–E**) show processes of cell differentiation which are either predetermined (**C**), not determined (**D**) or progressively determined (**E**). (**F**) shows a transport map, π, from a single point x at stage $\mathcal{X}$ to a probability distribution at a subsequent stage $\mathcal{Y}$. (**G**) shows transport maps computed from samples taken at different timepoints.

stem cells by using optimal transport to find which genes best predict the expression of the next time-point from the previous one. By analogy, we wish to shed light on the genes and regulatory pathways involved in the development of the benthic and pelagic morphs. Note that this divergence along the depth axis in fishes is the most pervasive and abundant way fish diversify and therefore represents a pattern of biodiversity origins that is hugely representative, at least for fish.

We will describe an experiment which focuses on the plasticity of Arctic charr, by feeding hatchlings from a single morph in a single lake on either pelagic or benthic diets. The results will still have a bearing on the evolution of these morphs, since we expect the regulation of the same pathways to be involved both in the plasticity and the evolution of these morphs, through the mechanism of genetic canalisation [WAD42]. An alternative experiment we could perform is to start with two populations of hatchlings, one pelagic and one benthic, and feed them all on the same diet. Figure 6.2 gives a

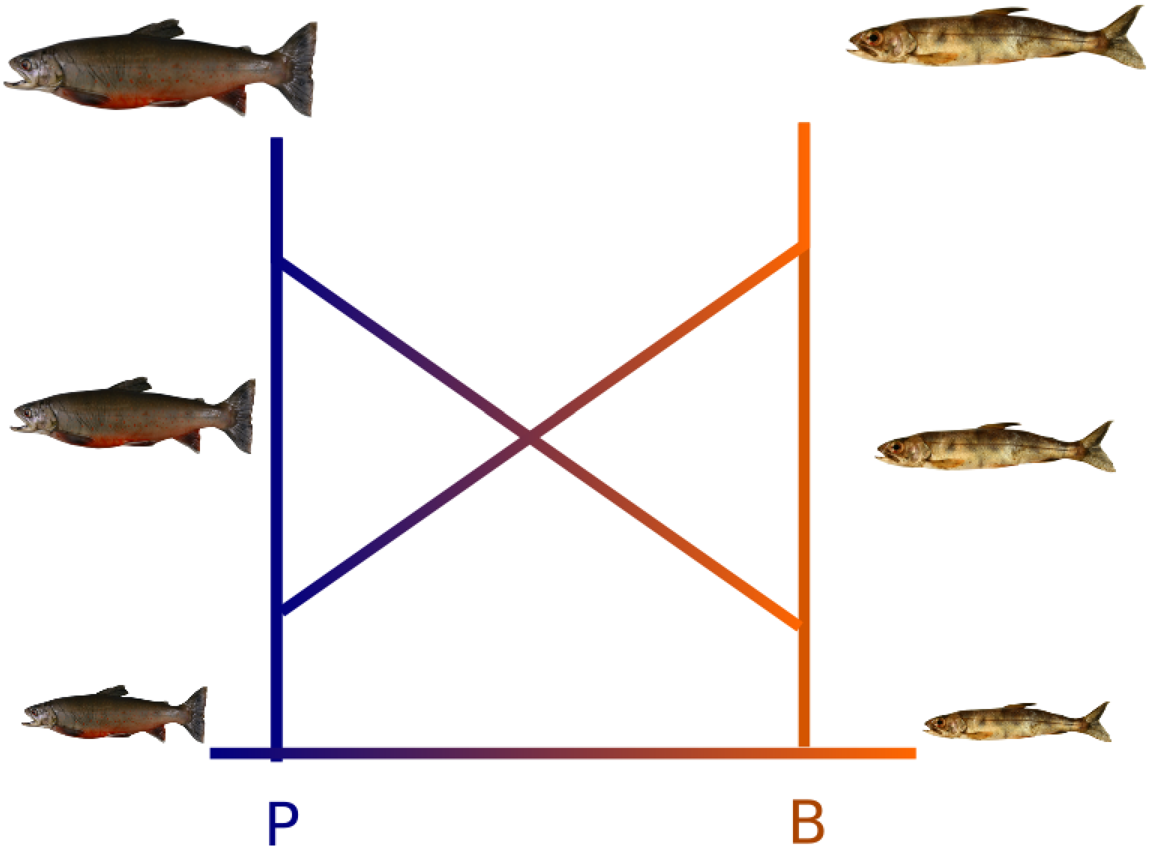


Figure 6.2: Schematic showing experiments for working out the genes involved in development of the Arctic charr morphs. They involve taking the morphs, feeding them on either the benthic or pelagic diet, and seeing how the gene expression changes over time.

schematic for these experiments.

We will begin with a population showing a large (phenotypic) divergence between the benthic and pelagic ecomorphs, such as the Arctic charr from Loch Tay. We will initially use hatchlings from a single morph, say the benthic morph. We then divide up hatchlings into two sections of a common tank. One section will be fed a benthic diet (bloodworms) and the other will be fed a pelagic diet (plankton). This is inspired by an experiment on Madeleine Carruthers' on developmental plasticity in Salmonid species (unpublished data). We take about 640 fish in each section, and will take samples of 40 fish each at 16 different timepoints.[2] We will be particularly interested in the difference in gene expression between the two sections of fish over the different timepoints. We expect to find in the set of differentially expressed genes ones which could plausibly drive the development of the different ecomorphs.

As a note, we see in §A.1 that, even if there is a continuous range of possibilities, if we want to measure with a degree of statistical certainty, then we are forced to consider a discrete number of finite forms. This concludes the data generating part of our thought experiment

[2] These are numbers used since they will leave us with 320 fish in each diet experiment, which we could profitably visualise using Mapper.

Issues

Some issues with the experiment include the fact that there are already differences in gene expression even before the Arctic charr hatch [Gud+18]. Nevertheless, we will still proceed, since there has been work showing that switching the diets of Arctic charr in a juvenile state is enough to produce ecomorphs which are between benthic and pelagic.

A secondary experiment we could conduct would be to switch the diets after 8 timepoints and see the difference in gene expression this causes. This would give more evidence for genes involved in the development of the two ecomorphs if we see the same genes switch expression to ones matching the other morph in the subsequent timepoints.

Another extension we could perform is to sequence the RNA from specific tissues, whose development is important for the distinct ecomorphs, such as bones and muscles. We would expect genes associated with growth and remodelling to be differentially expressed between the two ecomorphs. Genes related to these functions have been previously found to be significant in the difference between these ecomorphs [Jac+19].

Another follow-up experiment, could be to start with a pelagic population of fish, rather than a benthic one, and switch from a pelagic to a benthic diet at different timepoints, and see what effect this has on their path in gene expression space. We expect the gene expression to shift towards the benthic morphs, but not completely, and with almost no effect once the fish reach adulthood.

6.2.1 Mathematical Framework

Having given the outline of some experiments we could do to gather gene expression data, we will now give a mathematical framework for analysing the resultant data. We first give mathematical definitions for the gene expression space, and the shape-space (morphometrics) of the developing fry.

Definition 6.2.1 (Genetic Developmental Trajectory)**.** We consider a fish a sample $x(d) \in \mathbb{R}^m$, where m is the number of genes under consideration, and d is the time from hatching measured in days, where we allow fractions. A *genetic developmental trajectory* for this fish is a continuous function:

$$x : [0, D) \to \mathbb{R}^m$$

where we note that a single fish can have different developmental trajectories, depending on the tissue the RNA-sample is extracted from.

Definition 6.2.2 (Morphological Developmental Trajectory)**.** We use 7 linear measurements and fork length (length from tip of the snout to the end of the middle caudal fin rays), corrected for length as in [Jac+19], to assign each fish sample a point in an eight-dimensional morphology space. A *morphological developmental trajectory* for a fish is then a continuous function:

$$y : [0, D) \to \mathbb{R}^8$$

Optimal Transport

We will now give some definitions, inspired by ideas from [Sch+17]. We will only give an overview of optimal transport here, those interested in more details can consult the supplementary sections of the cited paper. Now, we give the definition of a transport map.

Definition 6.2.3 (Transport Map)**.** Given a pair of probability distributions $\mathbb{P}$ and $\mathbb{Q}$ on $\mathbb{R}^G$, a *transport map* (also called a *coupling*) is a probability distribution on $\mathbb{R}^G \times \mathbb{R}^G$ with $\mathbb{P}$ and $\mathbb{Q}$ as its two marginals.

We will give a discrete example to illustrate the notion of a transport map.

Example 6.2.4 (Discrete Transport Map)**.** Suppose we have a pair of sets of points $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ and $Y = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m\}$ in $\mathbb{R}^l$, where $n, m, l \in \mathbb{N}$. Then a transport map is a probability distribution $p : \{1, 2, \dots, n\} \times \{1, 2, \dots, m\} \to \mathbb{R}$ where $p(i, \cdot) = \sum_{j=1}^{m} p(i, j) = \frac{1}{n}$ and $p(\cdot, j) = \frac{1}{m}$. Note, here we have implicitly assumed that the marginal distribution is uniform on the points of X and Y.

Now, to define optimal transport.

Definition 6.2.5 (Optimal Transport)**.** Suppose we have a pair of probability distributions $\mathbb{P}$ and $\mathbb{Q}$ on $\mathbb{R}^G$, and a cost function $c : \mathbb{R}^G \times \mathbb{R}^G \to \mathbb{R}_{\geq 0}$ where $c(x, y)$ is the cost of transporting a unit mass from x to y, then the cost of a transport map π is given by:

$$\iint c(x, y)\pi(x, y)dxdy$$

An *optimal* transport map π is one which minimises this cost.

Regulatory Networks

Schiebinger *et al.* [Sch+17] use the optimal transport map to work out what genes are involved in the regulation of cell differentiation. We will give a short outline of the method here.

Definition 6.2.6 (Regulatory Network Regression)**.** Consider a pair of timepoints t_i, t_{i+1}. Let $\pi_{[t_i, t_{i+1}]}$ be the transport map between the times t_i and t_{i+1}, and let $(X_{t_i}, X_{t_{i+1}})$ be a pair of random variables distributed according to $\pi_{[t_i, t_{i+1}]}$. Finally, let $\mathcal{F}$ be a class of functions. Then *regulatory network regression* is finding:

$$\min_{f \in \mathcal{F}} E_{\pi_{[t_i, t_{i+1}]}} \left\| \frac{X_{t_{i+1}} - X_{t_i}}{t_{i+1} - t_i} - f(X_i) \right\|^2$$

so that f is the function which best predicts the change in gene expression in the coming timestep, where we assume that best means minimising the expected squared Euclidean norm of the difference between f and the actual change in gene expression, normalised for the change in time.

Practically speaking, Schiebinger *et al.* use linear combinations of transcription factor expression values, passed through a general logistic function, and then take linear combinations again. That is, the function family they use is given by:

$$f(x) = U\ell(WT\mathbf{x})$$

where $\mathbf{x}$ is the vector of gene expression levels, T is a matrix which picks out the transcription factors, W is some matrix, ℓ is a generalised logistic function applied element-wise, and U is another matrix mapping into gene expression space.

6.2.2 Morphology Experiment

Inspired by the optimal transport technique used by Schiebinger *et al.* in [Sch+17], we can analagously use it here to find the morphological developmental trajectory. Furthermore, we can find genes regulating this by using the gene expression at a certain timepoint t_i to predict the change in morphology between t_i and the next timepoint t_{i+1}.

Firstly, we run the experiment with 640 Tay benthic fry on each diets, and at timepoints of 30, 40, 50,. . . , 190 days we sample 40 fish each. When we sample, we measure the morphology of the fish. The goal is to work out when the fish stop changing in morphology as they grow. This gives us a total of 1,280 fish samples over 16 timepoints, so we have 1,280 points in $\mathbb{R}^8$.

We can now look at the change in morphology, and work out what is the crucial time period in the 30–190 days in which the changes between the two morphs occur. A way we could do this is to calculate optimal transport maps between the timepoints, so from days 30 to 40, 40 to 50, etc. and then see the cost between the timepoints. We can use the squared Euclidean distance as a cost function, or try a some others, such as Euclidean distance based on the first two principal components [Mit+04], based on which metric captures the change in shape best. Then we can work out when there is a high cost time period, indicating a large change in morphology between those timepoints. Once we have found this time period using the data from this initial experiment, we can use it to refine the time period we study in the next experiment, which will involve also collecting gene expression data.

6.2.3 Gene Expression Experiment

Once we have found the time-frame in which the morphology changes in developing Arctic charr, we can repeat the experiment, but with the 16 timepoints occurring only in this time-frame of morphology change. In addition to collecting morphology, in the experiment, we also collect gene expression by sequencing, say, white-muscle tissue (or some other tissue).[3]

Again, we have 1,280 points in $\mathbb{R}^8$ over 16 timepoints in morphology space, but we also have the same 1,280 points in gene expression space $\mathbb{R}^m$, where m is the number of genes under consideration.

We again apply the optimal transport idea of [Sch+17] to the fish samples in morphology space. The result of applying optimal transport to the fish in morphology space will be to produce a coupling between samples in adjacent timepoints. This will be in the form of probabilities of pairs $(a_{t_i}, b_{t_{i+1}})$ occurring, where a and b are integers between 1 and 40 (inclusive), indexing the samples of either the fish reared on a benthic diet or a pelagic diet.

One we have this coupling between timepoints, we can, again following [Sch+17], predict which genes may be involved, by finding a function from gene expression space to (change in) morphology space which best fits the observed coupled differences between timepoints. We note Definition 6.2.6 here, except we are considering the function f from gene expression space to morphology space which best predicts the change in morphology between timepoints. This will tell us which genes are important to consider for the change in morphology from one timepoint to the next.

Now that we can focus on a list of genes, we can go out into the field. With developments in technology, we will be able to do qPCR on a limited number of genes on Arctic charr in the field.

6.2.4 Field Experiment

Assuming developments in technology, we will outline an experiment which could be done in the field. We will here assume that we have the technology to do mRNA sequencing in the field, so we can get readings of gene expression levels. We will furthermore assume that we do not need to kill the fish to obtain a sample for sequencing, so we can resample fishes. This technology does not yet exist, but there are developments now for DNA sequencing in the field, and we believe this thought experiment will be useful to consider for the future. We also assume that we can age the fish rather accurately, and that we can catch fish which are still developing.

For the field experiment, suppose we pick a lake and tag each fish we catch, so we can keep track of whether we catch the same fish or not. Furthermore, suppose we sample the lake to the point where we have resampled, say, about twenty fish over four times each. This will allow us to estimate how much of the variation in gene expression is due to day-to-day changes (e.g. time of day sampled, temperature, etc.) and how much is due to genetic differences. We can do this by focusing on the fish which we have resampled, and seeing how their gene expression has changed when they've been

[3] The issue with sequencing multiple types of tissue, is that the sequencing cost multiples times as much.

caught at different times, since the differences in their gene expression will be all due to the environment.

Once we have some idea of how gene expression varies due to day-to-day environmental changes, we can catch and age fish over a longer period of time, over their entire growth period, and, focusing on the genes which we have discovered to be important from the previous section, we can measure how their expression changes as the fish age. Along with morphology measurements, this will allow us to validate the importance and effect of the genes on Arctic charr morphology as they develop.

Finally, we can compare to other lakes. We repeat the above experiments, and see how the expression of our genes of interest change as benthic and pelagic morphs of Arctic charr in other lakes develop. Being able to track the expression of genes, along with morphology, will allow us to develop a model of the gene expression's association with Arctic charr morphological development. Once we have this model for different lakes, we can see which genes behave in the same way. These genes should show, or indicate a way to show, the pattern of development leading to the benthic and pelagic morphs across different lakes. We will not combine information across the lakes. Instead, we will use what we learn about the development ot these ecomorphs in different lakes to work out what are the regulatory pathways and mechanisms in common, leading to this divergence in Arctic charr over diverse locations.

In more generality, the methods developed in such an experiment could tell us more about any case where we have a wild species with morphs. In the first instance, we could see if the same regulatory pathways are also involved in the benthic/pelagic differentiation in other species of fishes. Furthermore, this work can help to understand any situation where two morphs arise in nature, and we do not have a case-control morph to study. For example, there could be morphs of mosquitoes which either carry or do not carry a certain disease, and the analysis we conduct in Chapter 5 and here provides a way to, theoretically, work out what gene sets are important for these two morphs.

6.3 Conclusion

In conclusion, we have seen how a data visualisation technique, Mapper, inspired by topology, has found a new subtype of breast cancer by being applied to a breast cancer gene expression dataset. Mapper was used, essentially, as a combination of breaking the breast cancer samples up into degree of 'diseased' expression, and then clustering within those intervals. The validation of the new subtype was purely based on biology and statistics, with no input from topology.

We ported this method over to a non-model ecological and evolutionary study system Arctic charr, found it to be lacking in several aspects. This is due, at the very least, to lacking a large enough sample size, and the lack of case-control in samples, like the normal samples in the breast cancer. These are typical challenges when dealing with non-model data from wild systems, and the reason for this project. To account for the complexity of ecological and evolutionary systems, we have introduced another topologically-inspired perspective for looking at the gene expression data by weighting genes. The idea is to get a weighting of genes which shows the difference between the morphs that we know is present. We show a ‘proof of concept’, but find that both sample size and noise are still problems, even when we experimented on a large (~ 700) Dropsophila gene expression dataset.

Finally, we have ended with a discussion about a thought experiment, inspired by the work of Schiebinger *et al.* on optimal transport for reconstructing developmental landscapes. This experiment integrates the developments in the field of evolutionary biology, sequencing technologies, and analysis of gene expression data with the increasing amount of genomic resources available for non-model species, such as Arctic charr. This provides the basis for future work to apply what Schiebinger *et al.* have done in analysing developing human stem-cells to developing Arctic charr.

On a wider scale, this opens opportunities and approaches for applying new methods to the challenges of extracting the meaningful core genes from biologically critical but temporally variable pathways. This kind of experiment can work whenever there is a developmental component to morphs arising in the wild.

Appendix A

Miscellanea

A.1 Forced Discretisation

In this section, we will discuss the effect of choosing categories and a statistical measure on a continuous space. As a toy example, we will consider categorising points arising as independent identically distributed (i.i.d.) samples from two different random variables, call them B and P. In particular, this means we are only considering one dimension, $\mathbb{R}$. Furthermore, suppose we have that B and P are both normally distributed, with the same variance σ^2, and means μ_B and μ_P, with $\mu_P < \mu_B$, so $B \sim N(\mu_B, \sigma^2)$ and $P \sim N(\mu_P, \sigma^2)$.

Now, suppose that we want to decide with a certain probability, p, whether a sample, given as a point $x \in \mathbb{R}$ came from the random variable B or P. Then, assuming we use intervals of equal length around the mean, we can give the radius of the intervals we require in terms of standard deviations, σ. That is, the radius will be:

$$\sigma \cdot \Phi^{-1}\left(\frac{1-p}{2}\right), \quad p \in (0,1)$$

where Φ^{-1} is the quantile function of the standard normal distribution, which has no closed form solution. It has the property that, if $Z \sim N(0,1)$, then $\Phi^{-1}(q) = z_q$, where z_q is a number such that $\Pr(Z \leq z_q) = q$.

For example, suppose we want intervals which will capture 95% of the samples coming from B or P. Then we would use intervals with radius $\sim 2\sigma$ centred around μ_B and μ_P, respectively. Then, assuming that $\mu_B - \mu_P = l\sigma$, where $l > 4$, we can fit a total of $\lfloor \frac{l}{4} - 1 \rfloor$ intermediate categories between B and P without overlapping intervals. The top of Figure A.1 shows a schematic of this situation.

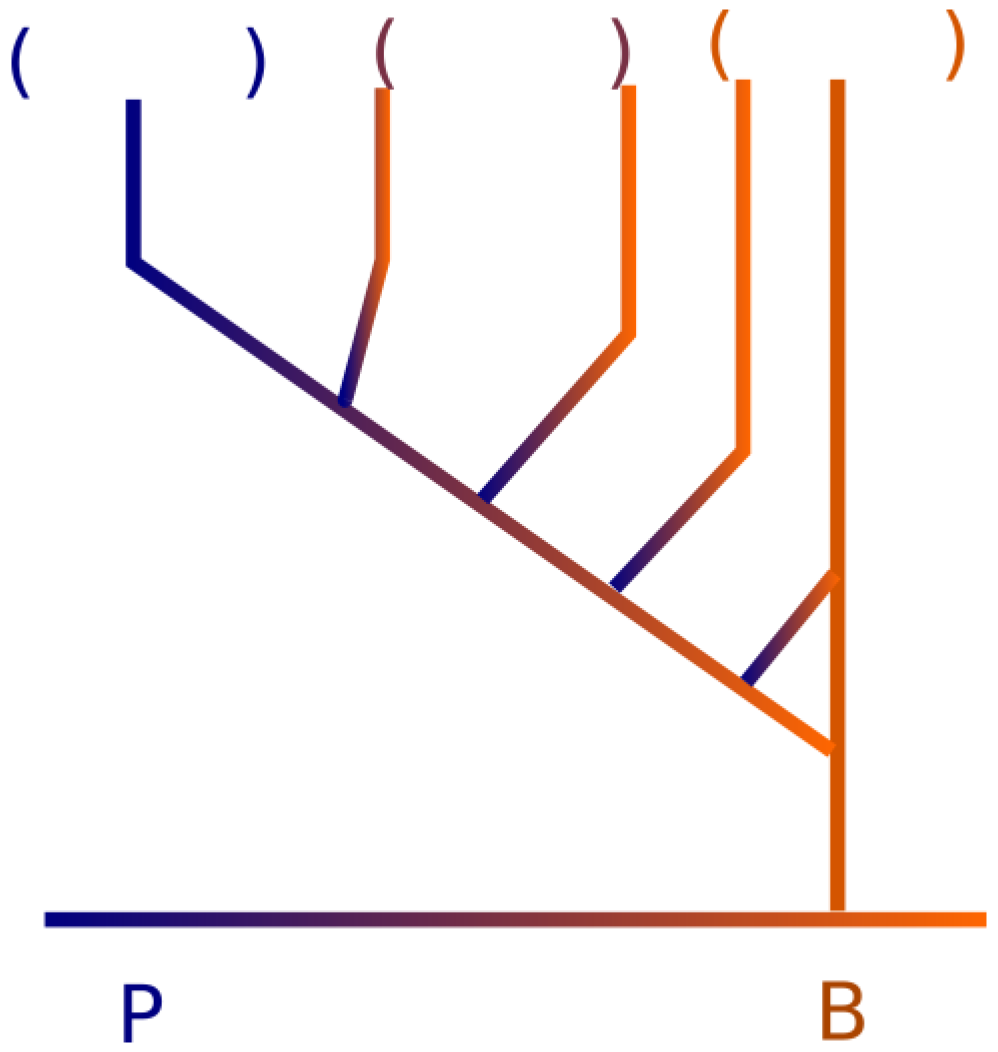


Figure A.1: A schematic showing what happens if we want to classify intermediate samples between B and P with a certain statistical confidence. In this case, with the intervals shown as brackets at the top of the figure have the desired confidence, we see that we can only have three classifications without overlap.

Bibliography

[Afg+16] Enis Afgan et al. "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update". In: *Nucleic Acids Research* 44.Web Server issue (Mar. 2016), W3–W10. ISSN: 1362-4962. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4987906/.

[AJK18] Basel Abu-Jamous and Steven Kelly. "Clust: automatic extraction of optimal co-expressed gene clusters from gene expression data". In: *Genome biology* 19.1 (Oct. 2018), pp. 172–172. ISSN: 1474-7596. URL: https://www.ncbi.nlm.nih.gov/pubmed/30359297.

[Alb05] Réka Albert. "Scale-free networks in cell biology". In: *Journal of Cell Science* 118.21 (2005), pp. 4947–4957. ISSN: 0021-9533. DOI: 10.1242/jcs.02714. eprint: http://jcs.biologists.org/content/118/21/4947.full.pdf. URL: http://jcs.biologists.org/content/118/21/4947.

[BC64] G. E. P. Box and D. R. Cox. "An analysis of transformations. (With discussion)". In: *J. Roy. Statist. Soc. Ser. B* 26 (1964), pp. 211–252. ISSN: 0035-9246. URL: http://links.jstor.org/sici?sici=0035-9246(1964)26:2<211:AAOT>2.0.CO;2-6&origin=MSN.

[BCV13] Y. Bengio, A. Courville, and P. Vincent. "Representation Learning: A Review and New Perspectives". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8 (2013), pp. 1798–1828. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2013.50.

[BH95] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *J. Roy. Statist. Soc. Ser. B* 57.1 (1995), pp. 289–300. ISSN: 0035-9246. URL: http://links.jstor.org/sici?sici=0035-9246(1995)57:1<289:CTFDRA>2.0.CO;2-E&origin=MSN.

[BVG15] S. Ballouz, W. Verleyen, and J. Gillis. "Guidance for RNA-seq co-expression network construction and analysis: safety in numbers". In: *Bioinformatics* 31.13 (2015), pp. 2123–2130. DOI: 10.1093/bioinformatics/btv118. eprint: /oup/backfile/content_public/journal/bioinformatics/31/13/10.1093_bioinformatics_btv118/3/btv118.pdf. URL: http://dx.doi.org/10.1093/bioinformatics/btv118.

[Car+18] Madeleine Carruthers et al. "De novo transcriptome assembly, annotation and comparison of four ecological and evolutionary model salmonid fish species". In: *BMC Genomics* 19 (Jan. 2018), pp. 32–. ISSN: 1471-2164. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5759245/.

[Car09] Gunnar Carlsson. "Topology and data". In: *Bull. Amer. Math. Soc. (N.S.)* 46.2 (2009), pp. 255–308. ISSN: 0273-0979. DOI: 10.1090/S0273-0979-09-01249-X. URL: https://doi.org/10.1090/S0273-0979-09-01249-X.

[CM+04] Kree Cole-McLaughlin et al. "Loops in Reeb graphs of 2-manifolds". In: *Discrete Comput. Geom.* 32.2 (2004), pp. 231–244. ISSN: 0179-5376. DOI: 10.1007/s00454-004-1122-6. URL: http://dx.doi.org/10.1007/s00454-004-1122-6.

[CO18] Mathieu Carrière and Steve Oudot. "Structure and Stability of the One-Dimensional Mapper". In: *Foundations of Computational Mathematics* 18.6 (2018), pp. 1333–1396. ISSN: 1615-3383. DOI: 10.1007/s10208-017-9370-z. URL: https://doi.org/10.1007/s10208-017-9370-z.

[CS13] L. M. Cook and I. J. Saccheri. "The peppered moth and industrial melanism: evolution of a natural selection case study". eng. In: *Heredity* 110.3 (Mar. 2013), pp. 207–212. ISSN: 0018-067X. URL: https://www.ncbi.nlm.nih.gov/pubmed/23211788.

[CSDL17] Juliana Costa-Silva, Douglas Domingues, and Fabricio Martins Lopes. "RNA-Seq differential expression analysis: An extended review and a software tool". In: *PloS one* 12.12 (Dec. 2017), e0190152–e0190152. ISSN: 1932-6203. URL: https://www.ncbi.nlm.nih.gov/pubmed/29267363.

[Cám16] Pablo G. Cámara. "Topological methods for genomics: present and future directions". In: *Current Opinion in Systems Biology* (2016), pp. –. ISSN: 2452-3100. DOI: http://dx.doi.org/10.1016/j.coisb.2016.12.007. URL: //www.sciencedirect.com/science/article/pii/S2452310016300270.

[DO14] Nadia M. Davidson and Alicia Oshlack. "Corset: enabling differential gene expression analysis for de novo assembled transcriptomes". In: *Genome Biology* 15.7 (July 2014), p. 410. ISSN: 1465-6914. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4165373/.

[Dur19] Rick Durrett. *Probability: Theory and Examples*. 5th ed. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[Elm+10] Kathryn R Elmer et al. "Local variation and parallel evolution: morphological and genetic diversity across a species complex of neotropical crater lake cichlid fishes". In: *Philosophical Transactions of the Royal Society B:*

Biological Sciences 365.1547 (June 2010), pp. 1763–1782. ISSN: 1471-2970. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2871887/.

[Elm+14] Kathryn R. Elmer et al. “Parallel evolution of Nicaraguan crater lake cichlid fishes via non-parallel routes”. In: *Nature Communications* 5 (Oct. 2014), p. 5168. URL: https://doi.org/10.1038/ncomms6168.

[GP12] Jesse Gillis and Paul Pavlidis. “"Guilt by association" is the exception rather than the rule in gene networks”. In: *PLoS computational biology* 8.3 (Mar. 2012), e1002444–e1002444. ISSN: 1553-734X. URL: https://www.ncbi.nlm.nih.gov/pubmed/22479173.

[Gud+18] Jóhannes Gudhbrandsson et al. “Differential gene expression during early development in recently evolved and sympatric Arctic charr morphs”. In: *PeerJ* 6 (Feb. 2018), e4345–e4345. ISSN: 2167-8359. URL: https://www.ncbi.nlm.nih.gov/pubmed/29441236.

[Hat02] Allen Hatcher. *Algebraic topology.* Cambridge University Press, Cambridge, 2002, pp. xii+544. ISBN: 0-521-79160-X; 0-521-79540-0.

[Her+08] Jason I Herschkowitz et al. “The functional loss of the retinoblastoma tumour suppressor is a common event in basal-like and luminal B breast carcinomas”. In: *Breast Cancer Research : BCR* 10.5 (Sept. 2008), R75–R75. ISSN: 1465-542X. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2614508/.

[Hub85] Peter J. Huber. “Projection pursuit”. In: *Ann. Statist.* 13.2 (1985). With discussion, pp. 435–525. ISSN: 0090-5364. DOI: 10.1214/aos/1176349519. URL: http://dx.doi.org/10.1214/aos/1176349519.

[Jac+17] Arne Jacobs et al. “Significant Synteny and Colocalization of Ecologically Relevant Quantitative Trait Loci Within and Across Species of Salmonid Fishes”. In: *Genetics* 207.2 (Oct. 2017), pp. 741–754. ISSN: 0016-6731. URL: https://www.ncbi.nlm.nih.gov/pubmed/28760747.

[Jac+19] Arne Jacobs et al. “Convergence in form and function overcomes non-parallel evolutionary histories in a Holarctic fish”. In: *bioRxiv* (2019). DOI: 10.1101/265272. eprint: https://www.biorxiv.org/content/early/2019/02/22/265272.full.pdf. URL: https://www.biorxiv.org/content/early/2019/02/22/265272.

[Kle10] Anders Klemetsen. “The Charr Problem Revisited: Exceptional Phenotypic Plasticity Promotes Ecological Speciation in Postglacial Lakes”. In: *Freshwater Reviews* 3.1 (June 2010), pp. 49–74. ISSN: 1755-084X. DOI: 10.1608/frj-3.1.3. URL: http://www.bioone.org/doi/abs/10.1608/FRJ-3.1.3.

[KWG11] Karin S. Kassahn, Nic Waddell, and Sean M. Grimmond. "Sequencing transcriptomes in toto". In: *Integrative Biology* 3.5 (Feb. 2011), pp. 522–528. ISSN: 1757-9708. DOI: 10.1039/c0ib00062k. eprint: http://oup.prod.sis.lan/ib/article-pdf/3/5/522/27326076/c0ib00062k.pdf. URL: https://doi.org/10.1039/c0ib00062k.

[Lan+07] Anita Langerød et al. "TP53 mutation status and gene expression profiles are powerful prognostic markers of breast cancer". In: *Breast Cancer Research* 9.3 (May 2007), R30–R30. ISSN: 1465-542X. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1929092/.

[LH08] Peter Langfelder and Steve Horvath. "WGCNA: an R package for weighted correlation network analysis". In: *BMC Bioinformatics* 9 (Dec. 2008), pp. 559–559. ISSN: 1471-2105. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2631488/.

[LHA14] Michael I. Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15.12 (Nov. 2014), p. 550. ISSN: 1465-6914. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4302049/.

[Lin+15] Yanzhu Lin et al. "Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual Drosophila melanogaster". In: *BMC Genomics* 17 (Dec. 2015), p. 28. ISSN: 1471-2164. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702322/.

[LS12] Ben Langmead and Steven L. Salzberg. "Fast gapped-read alignment with Bowtie 2". In: *Nat Meth* 9.4 (Apr. 2012), pp. 357–359. ISSN: 1548-7091. URL: http://dx.doi.org/10.1038/nmeth.1923.

[Lum+13] P. Y. Lum et al. "Extracting insights from the shape of complex data using topology". In: *Scientific Reports* 3 (Feb. 2013), p. 1236. URL: http://dx.doi.org/10.1038/srep01236.

[LZH08] Peter Langfelder, Bin Zhang, and Steve Horvath. "Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R". In: *Bioinformatics* 24.5 (2008), pp. 719–720. DOI: 10.1093/bioinformatics/btm563. eprint: /oup/backfile/content_public/journal/bioinformatics/24/5/10.1093_bioinformatics_btm563/2/btm563.pdf. URL: http://dx.doi.org/10.1093/bioinformatics/btm563.

[Mah36] Prasanta Chandra Mahalanobis. "On the generalised distance in statistics". In: *Proceedings of the National Institute of Sciences of India* 2.1 (1936), pp. 49–55.

[MB13] Daniel Müllner and Aravindakshan Babu. *Python Mapper: An open-source toolchain for data exploration, analysis and visualization.* 2013. URL: http://danifold.net/mapper.

[Mit+04] Philipp Mitteroecker et al. “Comparison of cranial ontogenetic trajectories among great apes and humans”. In: *Journal of Human Evolution* 46.6 (2004), pp. 679 –698. ISSN: 0047-2484. DOI: `https://doi.org/10.1016/j.jhevol.2004.03.006`. URL: `http://www.sciencedirect.com/science/article/pii/S0047248404000521`.

[Mun00] James R. Munkres. *Topology*. Second edition of [MR0464128]. Prentice Hall, Inc., Upper Saddle River, NJ, 2000, pp. xvi+537. ISBN: 0-13-181629-2.

[MW15] E. Munch and B. Wang. “Convergence between Categorical Representations of Reeb Space and Mapper”. In: *ArXiv e-prints* (Dec. 2015). arXiv: `1512.04108 [cs.CG]`.

[Nic+07] Monica Nicolau et al. “Disease-specific genomic analysis: identifying the signature of pathologic biology”. In: *Bioinformatics* 23.8 (2007), pp. 957–965. DOI: `10.1093/bioinformatics/btm033`. eprint: `http://bioinformatics.oxfordjournals.org/content/23/8/957.full.pdf+html`. URL: `http://bioinformatics.oxfordjournals.org/content/23/8/957.abstract`.

[NLC11] Monica Nicolau, Arnold J Levine, and Gunnar Carlsson. “Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival”. In: *Proceedings of the National Academy of Sciences of the United States of America* 108.17 (Apr. 2011), pp. 7265–7270. ISSN: 1091-6490. URL: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3084136/`.

[Pim+17] Harold Pimentel et al. “Differential analysis of RNA-seq incorporating quantification uncertainty”. In: *Nat Meth* 14.7 (July 2017), pp. 687–690. ISSN: 1548-7091. URL: `http://dx.doi.org/10.1038/nmeth.4324`.

[Ree46] Georges Reeb. “Sur les points singuliers d’une forme de Pfaff complètement intégrable ou d’une fonction numérique”. In: *C. R. Acad. Sci. Paris* 222 (1946), pp. 847–849.

[Rob+17] Fiona M Robertson et al. “Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification”. In: *Genome Biology* 18 (May 2017), pp. 111–. ISSN: 1474-760X. URL: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5470254/`.

[Sch+17] Geoffrey Schiebinger et al. “Reconstruction of developmental landscapes by optimal-transport analysis of single-cell gene expression sheds light on cellular reprogramming.” In: *bioRxiv* (2017). DOI: `10.1101/191056`. eprint: `https://www.biorxiv.org/content/early/2017/09/27/191056.full.pdf`. URL: `https://www.biorxiv.org/content/early/2017/09/27/191056`.

[SMC07] Gurjeet Singh, Facundo Memoli, and Gunnar Carlsson. "Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition". In: *Eurographics Symposium on Point-Based Graphics*. Ed. by M. Botsch et al. The Eurographics Association, 2007. ISBN: 978-3-905673-51-7. DOI: `10.2312/SPBG/SPBG07/091-100`.

[Sør+03] Therese Sørlie et al. "Repeated observation of breast tumor subtypes in independent gene expression data sets". In: *Proceedings of the National Academy of Sciences of the United States of America* 100.14 (June 2003), pp. 8418–8423. ISSN: 1091-6490. URL: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC166244/`.

[Tan+18] M. Tanabashi et al. "Review of Particle Physics". In: *Phys. Rev. D* 98 (3 2018), p. 030001. DOI: `10.1103/PhysRevD.98.030001`. URL: `https://link.aps.org/doi/10.1103/PhysRevD.98.030001`.

[Tib+02] Robert Tibshirani et al. "Diagnosis of multiple cancer types by shrunken centroids of gene expression". In: *Proceedings of the National Academy of Sciences of the United States of America* 99.10 (Feb. 2002), pp. 6567–6572. ISSN: 1091-6490. URL: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC124443/`.

[Tro+01] Olga Troyanskaya et al. "Missing value estimation methods for DNA microarrays". In: *Bioinformatics* 17.6 (2001), pp. 520–525. DOI: `10.1093/bioinformatics/17.6.520`. eprint: `http://bioinformatics.oxfordjournals.org/content/17/6/520.full.pdf+html`. URL: `http://bioinformatics.oxfordjournals.org/content/17/6/520.abstract`.

[TTC01] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. "Significance analysis of microarrays applied to the ionizing radiation response". In: *Proceedings of the National Academy of Sciences of the United States of America* 98.9 (Feb. 2001), pp. 5116–5121. ISSN: 1091-6490. URL: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC33173/`.

[Vee+02] Laura J. van 't Veer et al. "Gene expression profiling predicts clinical outcome of breast cancer". In: *Nature* 415.6871 (Jan. 2002), pp. 530–536. ISSN: 0028-0836. URL: `http://dx.doi.org/10.1038/415530a`.

[Vij+02] Marc J. van de Vijver et al. "A Gene-Expression Signature as a Predictor of Survival in Breast Cancer". In: *New England Journal of Medicine* 347.25 (2002). PMID: 12490681, pp. 1999–2009. DOI: `10.1056/NEJMoa021967`. eprint: `http://dx.doi.org/10.1056/NEJMoa021967`. URL: `http://dx.doi.org/10.1056/NEJMoa021967`.

[VS19] Hendrik Jacob van Veen and Nathaniel Saul. *KeplerMapper*. http://doi.org/10.5281/zenodo.10544 2019.

[WAD42] C. H. WADDINGTON. “CANALIZATION OF DEVELOPMENT AND THE INHERITANCE OF ACQUIRED CHARACTERS”. In: *Nature* 150.3811 (Nov. 1942), pp. 563–565. ISSN: 1476-4687. URL: https://doi.org/10.1038/150563a0.

[Wan+05] Yixin Wang et al. “Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer”. In: *The Lancet* 365.9460 (2005), pp. 671 –679. ISSN: 0140-6736. DOI: http://dx.doi.org/10.1016/S0140-6736(05)17947-1. URL: http://www.sciencedirect.com/science/article/pii/S0140673605179471.

[Wan+17] Zengmiao Wang et al. “VCNet: vector-based gene co-expression network construction and its application to RNA-seq data”. In: *Bioinformatics* 33.14 (2017), pp. 2173–2181. DOI: 10.1093/bioinformatics/btx131. eprint: /oup/backfile/content_public/journal/bioinformatics/33/14/10.1093_bioinformatics_btx131/2/btx131.pdf. URL: http://dx.doi.org/10.1093/bioinformatics/btx131.

[War63] Joe H. Ward Jr. “Hierarchical grouping to optimize an objective function”. In: *J. Amer. Statist. Assoc.* 58 (1963), pp. 236–244. ISSN: 0162-1459. URL: http://links.jstor.org/sici?sici=0162-1459(196303)58:301<236:HGTOAO>2.0.CO;2-9&origin=MSN.

[Wel47] B. L. Welch. “The generalization of ‘Student’s’ problem when several different population variances are involved”. In: *Biometrika* 34 (1947), pp. 28–35. ISSN: 0006-3444.

[WGS09] Zhong Wang, Mark Gerstein, and Michael Snyder. “RNA-Seq: a revolutionary tool for transcriptomics”. eng. In: *Nature reviews. Genetics* 10.1 (Jan. 2009), pp. 57–63. ISSN: 1471-0056. URL: https://www.ncbi.nlm.nih.gov/pubmed/19015660.

[Wil38] S. S. Wilks. “The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses”. In: *Ann. Math. Statist.* 9.1 (Mar. 1938), pp. 60–62. DOI: 10.1214/aoms/1177732360. URL: https://doi.org/10.1214/aoms/1177732360.

[Wol78] Svante Wold. “Cross-validatory estimation of the number of components in factor and principal components models”. In: *Technometrics* 20.4 (1978), pp. 397–405.

[Zha+04] Hongjuan Zhao et al. “Different Gene Expression Patterns in Invasive Lobular and Ductal Carcinomas of the Breast”. In: *Molecular Biology of the Cell* 15.6 (2004), pp. 2523–2536. DOI: 10.1091/mbc.E03-11-0786. eprint: http://www.molbiolcell.org/content/15/6/2523.full.pdf+html. URL: http://www.molbiolcell.org/content/15/6/2523.abstract.

[R C17] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2017. URL: `https://www.R-project.org/`.