GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

Fakultät für Physik

Master's Thesis

Network Analysis and hidden Phenotypes in large biological Datasets

Netzwerkanalyse und versteckte Phenotypen in großen biologischen Datensätzen

prepared by

Jana Lasser

from Graz

at the Max Planck Institute for Dynamics and Self-Organization, Göttingen

Thesis period:	28th November 2014 until 28th May 2015
Supervisor:	PhD Eleni Katifori
First referee:	PhD Eleni Katifori
Second referee:	Prof. Dr. Marc Timme

Abstract

In dieser Arbeit wird ein automatisiertes Framework zur Extraktion von Netzwerk-Information aus Bildern von End-Zellen der Drosophila Tracheen sowie die nachfolgende quantitative Analyse des Wachstumsverhaltens selbiger vorgestellt. Der analysierte Datenbestand enthält über 500 Bilder von Tracheen. Die zugehörigen Larven wurden mit unterschiedlichen Mutationen (*Rab8*, *Myospheroid*, *Crumbs*, *Rhea*), genetischen Grundlagen und bei verschiedenen Temperaturen aufgezogen. Ein Datenbestand von solcher Größe war noch nie für quantitative Analyse verfügbar und erlaubt statistisch signifikante Aussagen über die Beschaffenheit der Tracheen. An den als Netzwerken dargestellten Zellen wird mit Hilfe eines Ansatzes aus dem überwachten Lernen festgestellt, wie viel Anteil zur Unterscheidbarkeit der Netzwerke Mutation, genetische Grundlage und Temperatur beitragen. Clustering - ein Ansatz aus dem nicht überwachten Lernen - ermöglicht außerdem das Auffinden von bisher unbekannten Netzwerk- Phänotypen welche unabhängig von den vom Genotyp induzierten Phänotypen auftreten. Der Großteil der in der Netzwerk-Realisation enthaltenen Information bezieht sich auf die Größe der Netzwerke. Durch Analyse der Abweichungen von der Größenabhängigkeit der Netzwerk-Realisation können vier Wachstumscharakteristiken identifiziert werden - die *phänotypischen Trends.* Weiterhin werden zwei Modelle, welche das Verästelungsverhalten und die Verteilung der Zell-Dicken und Zell-Längen beschreiben, vorgestellt.

Stichwörter: Netzwerke, Drosophila, Tracheen

Abstract

We develop a methodology for automated extraction of network information from a large dataset containing images of Drosophila terminal cells. The dataset contains images of larvae grown with different mutations prohibiting the expression of one of four genes: *Rab8*, *Myospheroid*, *Crumbs* and *Rhea.* Larvae are also distinguished based on their genetic background and growing temperature. The dataset is composed of over 500 images which is a novelty for this field of research. This enables us to find statistically highly significant results. We apply a supervised learning approach to quantify the effect on discernability of each of the three growing conditions. Using an unsupervised learning approach we find hidden phenotypes spanning several of the already known phenotypes induced by the larva's genotype. We find that most of the information contained in network growth patterns is strongly tied to network size. By analyzing deviations from the size dependence of network realization we establish four main growth characteristics we call *phenotypic trends.* We are also able to find very simple models describing cell branching behaviour and distributions of tube lengths and tube radii.

Keywords: Networks, Drosophila, Trachea

Contents

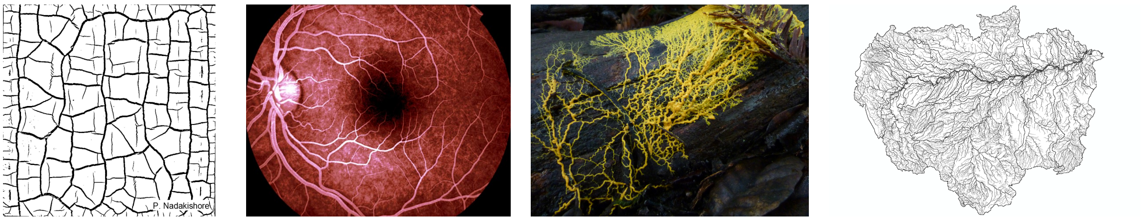
List of Figures

List of Tables

1. Introduction

1.1. Networks as the Lifelines of living Beings

Networks are a widely reoccurring pattern in nature. We can find them from the micro to the macro scale, in the inorganic regime (figure 1.1a), single living beings (figure 1.1b) or even single cells (figure 1.1c) and as organizational structures of large meta-systems (figure 1.1d). Networks are especially important for multiple-cell organisms when it comes to the transport of vital substances. Very prominent examples for these transport networks are the blood vessels and respiratory network of mammals. Understanding how these networks develop and function is central to the understanding of why and how they can fail.



(a) Cracks in dried clay. Image courtesy of Pawan Nandakishore.

(b) Blood vessels in the human retina. Image courtesy of Paul van de Meer.

(c) Physarum slime mold on a tree.

(d) Amazon river network [1].

Figure 1.1. Different examples of networks found in nature: **(a)** Cracks in dried (inorganic) clay. **(b)** Blood vessels in the human retina (multiple cells) **(c)** Network of the Amazon river. **(d)** Slime mold (single cell).

Networks in animals are composed of branched tubular epithelial tissue. This tissue covers all internal and external organs and wraps the lumen - the interior of any hollow organ. There exist various known mechanisms [2], [3], [4] for the formation of those tubes including cell fusion, cell rearrangements, cell migration or cell shape changes. What most of these mechanisms have in common is that they create a tube by enclosing an already existing external or internal space. In this work we are especially interested in the development of *terminal cells* (figure 1.2c) in the *Drosophila Melanogaster* (figure 1.2a) respiratory network, the trachea (figure 1.2b). These cells form tubes *de novo* [5] - they create lumen without enwrapping a preexisting structure and spread in an elongated shape with multiple branches throughout the surrounding tissue. The creation of tubular structures is common in many organisms, be it the small fruit fly or humans. So far, the complexity of mammals has obstructed a clear understanding of how networks form on a molecular level. However, as the underlying mechanisms are believed to be qualitatively the same for every organism, the study of a much simpler system might yield answers of universal validity. For the research of the formation of tubular structures as well as for many other processes [6], [7], [8] the *Drosophila* has proven to be an excellent paradigm.

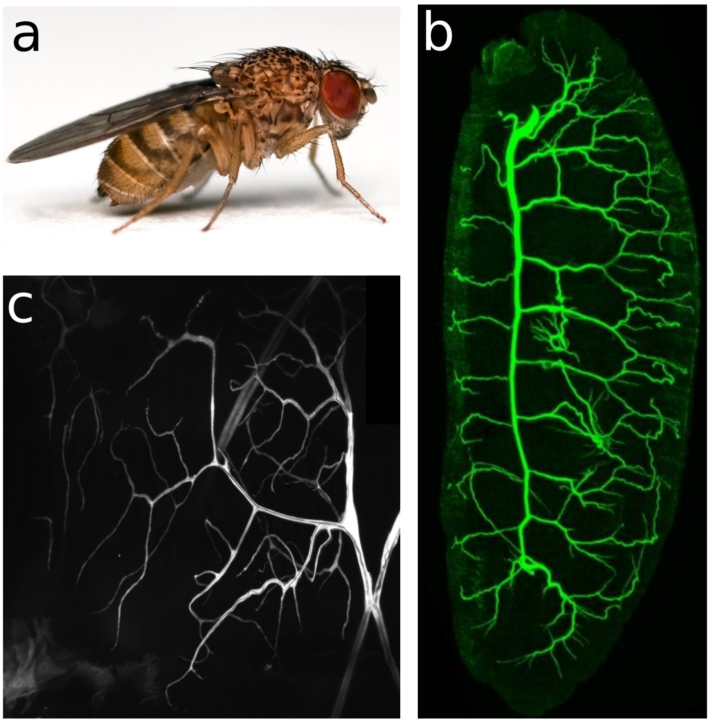


Figure 1.2. **(a)** Image of a *Drosophila Melanogaster*, the common fruit fly. The organism is widely used for biological research.[1] **(b)** Tracheal tube system of a *Drosophila* larva visualized by immunological staining.[2] **(c)** Trachea terminal cell, result of the last of three sequential branching processes in the trachea development.[3]

[1] Image courtesy of Bbski, Wikimedia Commons. `https://en.wikipedia.org/wiki/Drosophila` (20.11.2014) Distributed under the creative commons license.

[2] Image courtesy of Xlab Goettingen. `http://www.xlab-goettingen.de/organogenese.html` (20.11.2014)

[3] Image courtesy of Sara Sigmundbjörnsdottir, EMBL Heidelberg.

1.1.1. Genotype and Phenotype

> **Genotype**: the genetic constitution of an organism. The genotype determines the hereditary potentials and limitations of an individual from embryonic formation through adulthood. [9]
>
> **Phenotype**: all the observable characteristics of an organism, such as shape, size, color, and behaviour, that result from the interaction of its genotype (total genetic inheritance) with the environment. The common type of a group of physically similar organisms is sometimes also known as the phenotype. [10]

The aim of this work is to provide a quantification of the effect of certain growth conditions on the development of tubular structures: mutation of genes involved in tube and lumen creation, genetic background and temperature. We want to define, find and quantify phenotypes in networks in the sense of their common type. This is not an easy task as networks are complex structures which show various geometric and topological characteristics, all contributing to an observable phenotype.

We already know that there are changes in network growth patterns dependent on the genotype of the fly. The genotype is composed of the knockdown gene as well as the genetic background of the animals. In this work we call effects which are caused by the genotype in interaction with the growing temperature *induced phenotypes*.

We want to clarify and quantify the impact of growing conditions on the expression of induced phenotypes. Moreover we want to investigate whether there are *hidden phenotypes*, i.e. formerly unknown similarities, expressed by the networks independent of their genotype or at least spanning several genotypes.

To reach these goals, we consider a single terminal cell with its branching and wide-stretching structure to be a network and apply approaches borrowed from graph theory to understand its complex structure. To quantify different realizations and find distinguishable phenotypes in a large set of networks we apply tools from Bayesian reasoning and machine learning.

1.2. Approach

A broadly accepted problem in research is the possibility of bias in the analysis of data if researchers “know what they are looking for”. This might lead to research being directed only in the most promising or desirable direction and may go as far as the dismissal of inconvenient findings or the over-emphasis of data supporting the desired outcome. Often it is difficult or expensive to generate data suitable for analysis which leads to relatively small sample sizes which are highly susceptible to noise. Moreover, analysis often involves manual processing of data which, even when documented properly, makes it difficult to reproduce the results. During the analysis involved in creating this thesis we were aware of all the above-mentioned problems and were dedicated to implement tools and protocols to generate as unbiased and reproducible results as possible.

The research for this work was conducted in collaboration with Sara Sigmundbjornsdottir and Maria Leptin at EMBL Heidelberg, who were responsible for the preparation of *Drosophila* organisms and the imaging of terminal cells. To circumvent the problem of small sample sizes and high noise, our collaborators went to great lengths to create an extensive dataset containing images of around 500 terminal cells. A dataset of this size has not been available for research so far. Each of the images contains one or both sides of the symmetrically growing terminal cell. The images were recorded with a resolution of 0.3459442 μm × 0.3459442 μm per pixel to capture even very small branches of the terminal cells.

We deliberately chose to learn as little as possible about the particulars of the molecular mechanisms involved in tube formation and expected growth behaviour of the networks. We were provided with the dataset of *Drosophila* terminal cell networks as well as basic labelling information about the mutation, genetic background and growing temperature of each of the samples to enable us to group the samples into families.

Our first goal was to create an automated protocol for the retrieval of network data from the images. We profited by a large degree from previous work we did involving the processing of large amounts of leaf venation patterns [11]. After the digitization of all available network data we applied a very broad range of techniques borrowed from graph theory and machine learning to analyze all available information contained in a single network and in the dataset as a whole.

After finishing our analysis we exchanged results with our collaborators to put them in a biological context.

1.3. Content of the Thesis

The remainder of the thesis is structured as follows: In chapter 2, section 2.2, we give a short description of the digitization technique applied to segment the images of the terminal cells. Adaptations and expansions to the existing framework are described in more detail, as is a newly developed tool for graphical manipulation of networks (section 2.4).

Chapter 3 contains a description of the analysis-methods we applied, as well as a presentation of the results. After giving a short introduction into the methods from graph theory we utilized (section 3.2), we quantify the impact of each of the growing conditions (mutation, genetic background and temperature) on network growth in section 3.3.
In section 3.4 we then localize the phenotypes induced by the network's growing conditions using a supervised learning approach and confirm their discriminative power.
To investigate whether there are hidden phenotypes spanning several network families, in section 3.5 we apply an unsupervised learning approach to find possible clusters in the dataset. We establish that most if not all information contained in different network realizations is information about network size. Using the clusters we found, we are able to isolate new phenotypes distinguished by network size which span several families.
In section 3.6 we then investigate the question whether there is second order information independent of network size. This information has to be contained in the geometry of a network such as how wiggly or clustered a network grows. This second order information we call "phenotypic trends". It can be separated into four main characteristics of network geometry: diameter of branches, length of branches, straightness of branches and growing direction.

After an in-depth analysis of network phenotypes, in section 3.7 we develop a growth model for trachea terminal cells. We split the general model into three sub-models for branching behaviour, radius distribution and growth direction.
Chapter 4 is dedicated to putting our findings into a biological context. In section 4.2 we give a short review of the morphogenesis of insect trachea. We also describe how the genes affected by the mutations in the flies as well as genetic background and

temperature are expected to affect network growth. This information enables us to assess whether our findings support the theoretically expected behaviour. This is described in section 4.3.

We finish with a summary of all our findings and their implications in section 5.1 of chapter 5. During our research we encountered many interesting questions and research directions deviating from our main focus which would be worth further investigation. In section 5.2 we give an outlook on possible future research topics and summarize questions that remained open after our analysis.

2. Data Acquisition

2.1. Introduction

The data used in this work consists of 500 images of *Drosophila* trachea networks, three of these images are shown in figure 2.1. The images were created using confocal light microscopy. For the imaging, larvae were selected based on their developmental stadium, not based on time after egg-lay. This ensured that there was no variation in network realization based on overall larvae development.

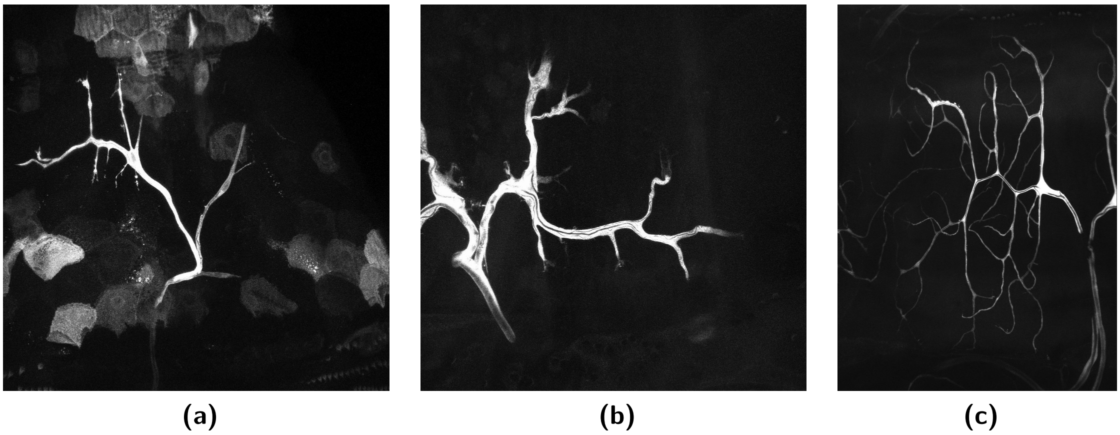


Figure 2.1. Confocal microscopy images of *Drosophila* trachea. **(a)** shows bright parts not belonging to the network. **(c)** shows parts of a neighboring network coming into the picture from the left. **(a)** - **(c)** contain some or all of the other half of the terminal cell to the left or right. All of these artifacts had to be removed manually before automated processing.

To make the networks accessible for analysis, the networks first have to be extracted and digitized. We utilize an algorithm developed in an earlier work, [11], where it was used to digitize the large networks of leaf venation patterns, and modified it to fit our needs. The following chapter consists of a description of the steps we undertook to adapt and expand the digitization framework as well as an illustration of its application to the data. As the networks we analyze in this work are only approximately flat, we use a two-dimensional projection of the data. This in some cases causes the appearance of fake junctions by superposition of two branches growing at different z-coordinates. To get rid of these junctions we developed a graphical user interface for graph manipulation. We also give a description of this GUI as well as some screenshots of the graph-representation using the GUI.

2.2. Image Vectorization

When trying to extract ribbon-like shapes from digital images (e.g. networks), there are two main approaches: Thinning [12] and Vectorization [13]. As described in [14], for the purpose of network digitization, vectorization is the approach of choice. It might be harder to implement but in the end, the network structures created by vectorization are much easier to analyze and handle.

2.2.1. Binarization

The basis of a successful vectorization is a suitable binary representation of the image. This means that foreground features (the features we want to depict in the network structure) and background features are clearly separated and represented as zeros and ones in a binary image. In a perfect binary image, there are no more artifacts from the imaging process or unwanted features present.

Image Processing

To create such an image, several techniques from image processing can be applied to reduce noise, increase contrast and sharpen edges. A detailed description of the techniques applied in network vectorization can be found in [14]. The cornerstones of pre-processing are:

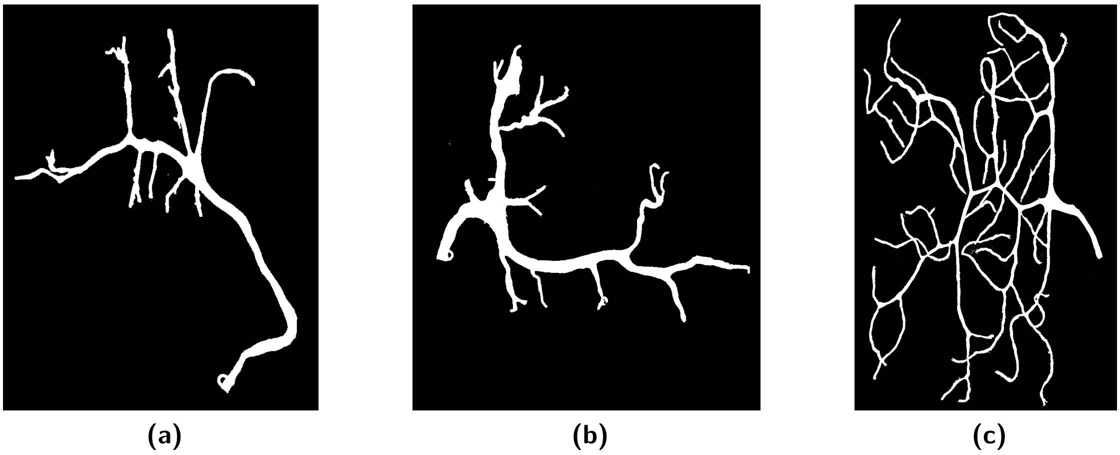
- Gaussian filtering to reduce noise.
- Local histogram equalization to sharpen and equalize global contrast and sharpen edges.
- Otsu-thresholding [15] to create a binary image.

Artifact Removal and manual Processing

The images used in this work commonly contain many features or artifacts we did not want to include in our processing. Processing steps that were performed manually on many of the images include:

- Selection of the region of interest (ROI) excluding parts of neighboring networks from the considered region.
- Division of images which contained both symmetrically growing terminal cells into two separate images for the left and the right cell respectively.
- Manual selection of the threshold used to create the binary image in cases where Otsu-thresholding did not yield satisfactory results.
- Manual removal of stains, flares and parts of neighboring networks intermingled with the focal network.

In general we try to restrict the manual processing to *removing* features. We only add white (foreground) pixels in cases were the binarization algorithm did not recognize crucial connections in the network. Results of the binarization and manual processing can be seen in figure 2.2.



(a) (b) (c)

Figure 2.2. The images from figure 2.1 after the binarization process and artifact removal.

2.2.2. Vectorization

As the *Drosophila* trachea are three-dimensional structures which grow on the surfaces of the insect's organs, the first approach was to slice the structure into several horizontal parts and create z-stacks. The vectorization itself can easily be generalized for n dimensions as it relies on a constrained Delaunay triangulation [16] which is defined for all dimensions $n \geq 1$. It soon became clear that in application, using the z-stacks was not possible for several reasons, including:

- In z-direction the network parts could not reliably be linked.
- Contrast was not high enough.
- A large set of artifacts ranging from imaging artifacts to parts of the organism which did not belong to the network would have forced us to manually process every single z-slice.

Finally, we decided to treat the images as pseudo-two-dimensional and processed the two-dimensional projections of the z-stacks. As the trachea grow on internal surfaces of the flies which can be considered flat, the length distortions introduced by using a projection instead of a three-dimensional image were reasonably small. Only rarely a superposition of two tubes would result in a fake junction which had to be corrected manually after the digitization process.

2.3. Graph Creation

To extract a network from a binary image, we first have to determine what information is needed to satisfactorily describe the network. As we want to analyze the topology of the network as well as its geometry, we have to make sure that the vectorization process provides us with

- two-dimensional coordinates of all nodes (x_i, y_i),
- lengths of all edges l_i,
- radii of all edges r_i,
- neighbors of every node i.e. the network's adjacency matrix,
- a distinguished root node.

Given this information we can reconstruct all information contained in the network growth pattern, as illustrated in figure 2.3.

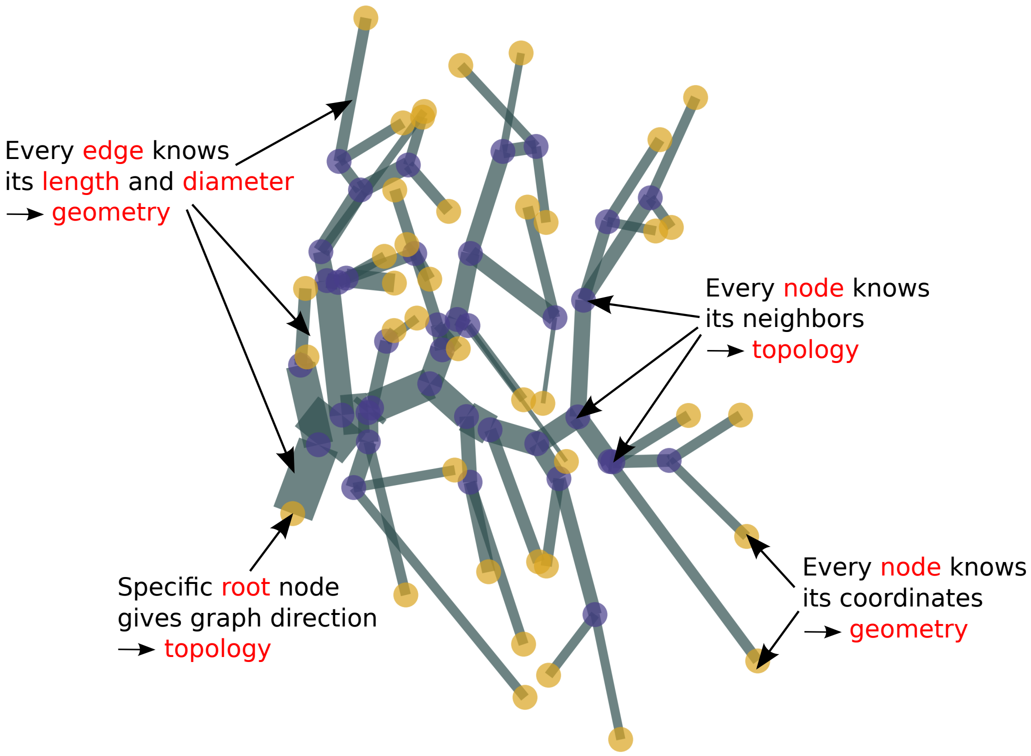


Figure 2.3. Digitized network representation of a *Drosophila* terminal cell illustrating geometry and topology information contained in the network structure.

2.3.1. Approximation of the Skeleton

Digitization of ribbon-like shapes requires finding the skeleton of the shape. The skeleton is defined as a one pixel thin line which has the same distance to all borders of the feature. The vectorization approach accomplishes the extraction of the skeleton by representing the features with vectors and points rather than pixels. The first step is to find the contours of the shapes and approximate them by their dominant points [17], [18]. Using these contour points, we can perform a constrained delaunay triangulation on the shape with the contours as constraints. An example of a triangulation performed on a network contour can be seen in figure 2.5. This leaves us with a series of neighboring triangles represented by the coordinates of their vertices.

Triangle "Centers"

For every triangle we now want to find a point C (the "center" of the triangle) which will then be used to approximates the segments of the skeleton contained

in the triangle. This center-point therefore should have the same distance to both borders of the feature.
To find C for every triangle, we first have to differentiate between four different kinds of triangles. Triangles are distinguished by their number of *external* (shared with the contour) and *internal* (not shared with the contour) edges:

- *Junction* triangles: These are triangles which have three internal edges, they lie in the middle of a junction.
- *Normal* triangles: These are triangles which have two internal edges, they lie on the elongated parts of the shape.
- *End* triangles: These are triangles which have only one internal edge, they lie at the tip of an elongated part of the shape.
- *Isolated* triangles: These are triangles which have no internal edge. They do not have any neighboring triangles and are ignored.

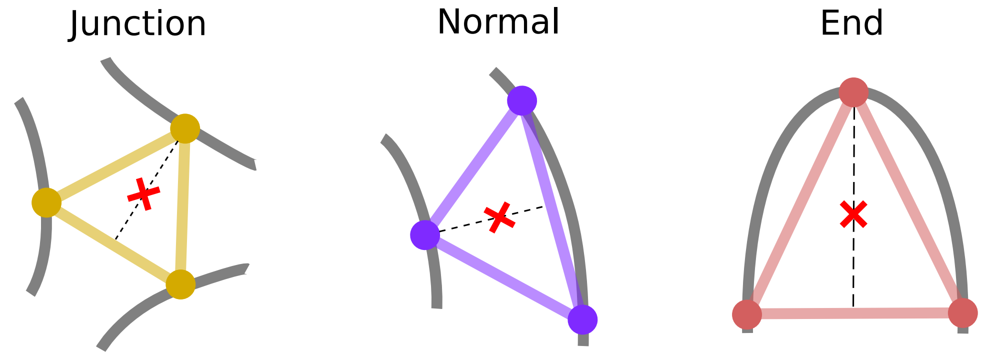


Figure 2.4. Illustration of *Junction*, *Normal* and *End* triangles. The “center point” C is indicated with a red cross.

For *Junction* triangles we expect C to lie on the longest angle bisection line. For *Normal* triangles we expect C to lie on the line bisecting the angle between the two internal edges. For *End* triangles we expect C to lie on the line bisecting the angle between the two external edges. An illustration of the position of C can be seen in figure 2.4.
To find the exact coordinates of C, we superimpose the triangles with the euclidean distance map (EDM)[19] of the binary image. The EDM of an image is a matrix which for each pixel in the image contains the distance to the nearest background pixel. We then look along the angle bisection line for the point with the highest

EDM-value, which signifies that it has the highest distance to the next background pixel. Therefore C maximizes the distance to both borders of the feature, which is required for a point on the skeleton.

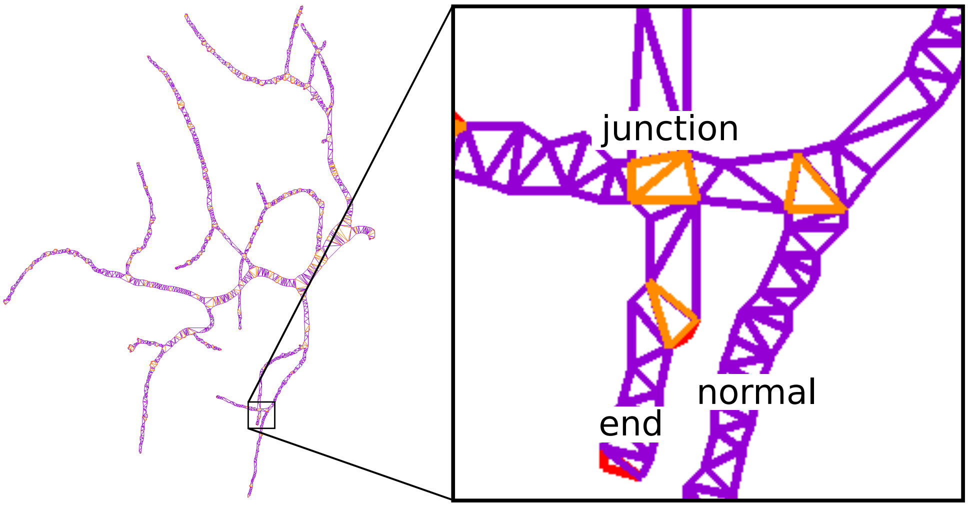


Figure 2.5. Example of a triangulation of a contour belonging to a terminal cell network shape. *Junction*, *Normal* and *End* triangles are indicated in orange, purple and red.

Creating the Graph

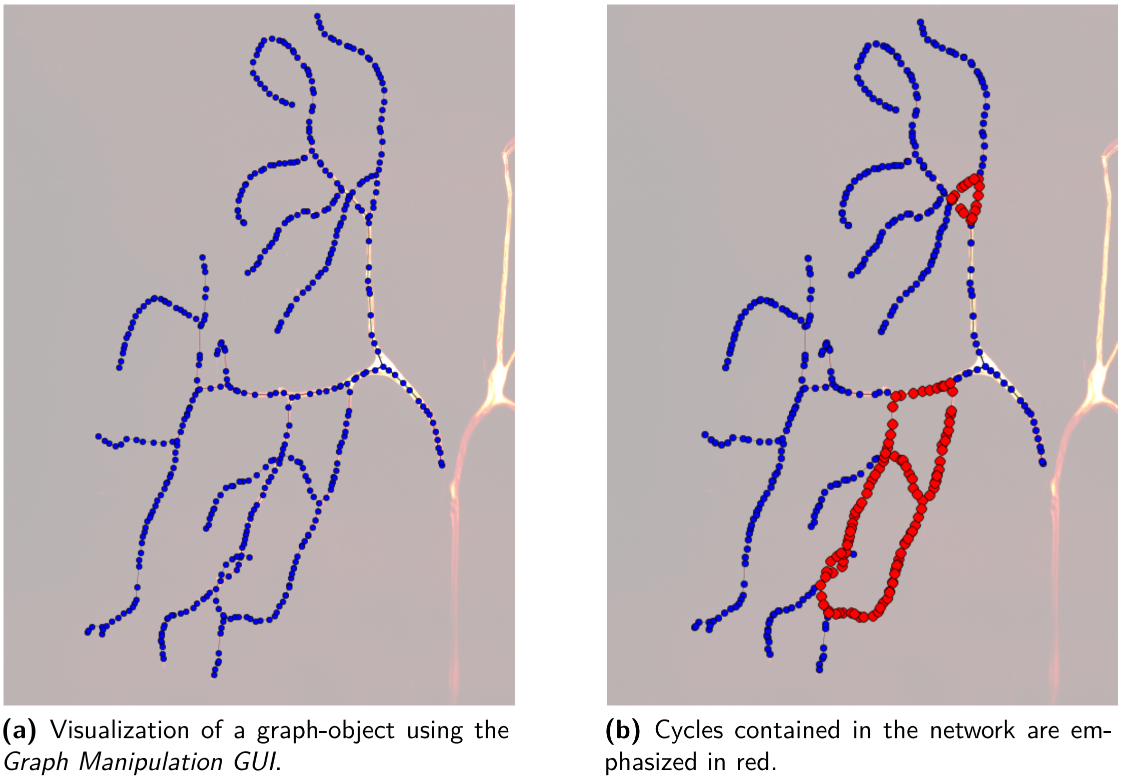
Now we have extracted all the information we need to describe the network. We create a graph by compiling all the information into a `networkx` [20] graph-object. The coordinates for the triangle-centers C_i are the coordinates of the nodes of the graph. The distance to the nearest background pixel at each C_i determined from the EDM is the *thickness* or *radius* of the tube at point C_i. By comparing triangle vertices we can determine that each two triangles which share two of their vertices have to be neighbors. This information we can use to create the graph's adjacency matrix and establish neighborhood relations. The length of the edges between neighboring nodes we calculate by using the euclidean distance between the two nodes. The radius of an edge is defined as the mean of the two radii of its nodes.

This leaves results in a graph that is a close approximation of the network present in the original microscopy image.

2.4. Graph Manipulation

From the biology of the *Drosophila* trachea we know that the network does not contain any loops. As the image we used to create the graph representation was a two-dimensional projection of a three-dimensional image, the digitization process might have created artificial loops out of superimposed tubes. These need to be removed manually.

Removing false loops and "rewiring" the network is a difficult task to automate, therefore we decided to do it manually. To make this unavoidable manual manipulation step as transparent and fast as possible, we created a graphical user interface, the *Graph Manipulation GUI*, to manipulate the graph using `networkx` and `matplotlib` [21].



(a) Visualization of a graph-object using the *Graph Manipulation GUI*.

(b) Cycles contained in the network are emphasized in red.

Figure 2.6. (a) Example of a digitized network loaded in the *Graph Manipulation GUI*. The nodes (blue) and edges (black) are superimposed onto the original scan. **(b)** Cycles (highlighted in red) can form because of fake junctions.

The tool superimposes the extracted graph object and the original image, as can be seen in figure 2.6a, to make it as easy as possible for the human using the tool to decide whether a junction is a real branching point or just an artifact from the projection. The GUI also provides a functionality to highlight all remaining cycles in the graph which helps to make sure that no cycles are left. Figure 2.6b illustrates a fake junction and its correction.

2.4.1. Graph Manipulation GUI

Options to remove and create nodes and edges to correct fake junctions and some rudimentary tools to measure lengths and radii are also included in the GUI. Attributes for newly created elements of the graph are taken directly from the EDM. An illustration of a fake junction and how it is corrected with the GUI is provided in figure 2.7.

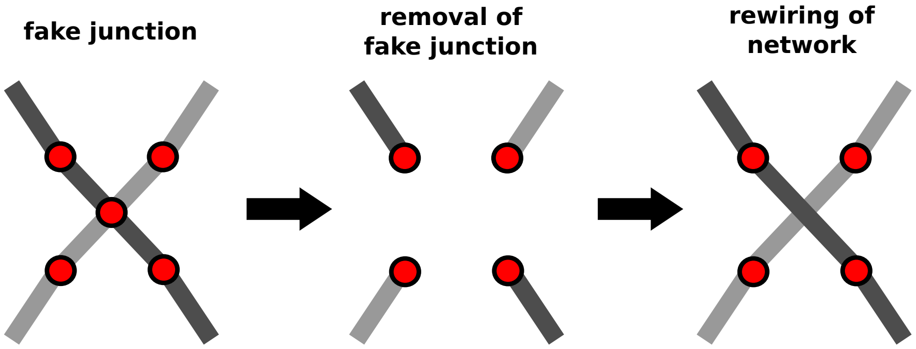


Figure 2.7. Illustration of the rewiring of a fake junction. The central node is falsely created at the point where two superimposed tubes cross. The central node is removed and the remaining nodes are rewired correctly.

2.4.2. Trees and directed Graphs

We know that the terminal cell which forms the network is a single cell that starts its branching process at a spatially fixed point in the trachea. We call this point the *root* of the network. By specifying a distinguished node in the graph as root we can make use of this directional information about the network growth and give the graph a *direction*. We can either select the node at which the veins have the largest radius to be the root node, this is true for most cases - or specify it manually using the *Graph Manipulation GUI* as we did for this work.

2.5. Summary

At the beginning of the digitization process, we used techniques from image processing to create a suitable binary representation of the image. The binary image had to be manually modified in many cases to remove artifacts including parts of neighboring networks and surrounding tissue. We tried to keep manual processing to a minimum and refrained from *adding* foreground pixels wherever possible.

Using the binary image we followed a vectorization approach to extract the skeleton of the ribbon-like network features and create a graph representation of the network. The resulting graph contains geometric information such as node coordinates, edge lengths and edge radii as well as topological information such as neighborhood relations and a distinguished root node. The graphs are known to not contain any loops. As we treated 2D projections of 3D images, we had to correct "fake" loops in the network after which we were able to transform the graph into its final representation as a directed graph.

Examples of the extracted networks can be seen in figure 2.8.

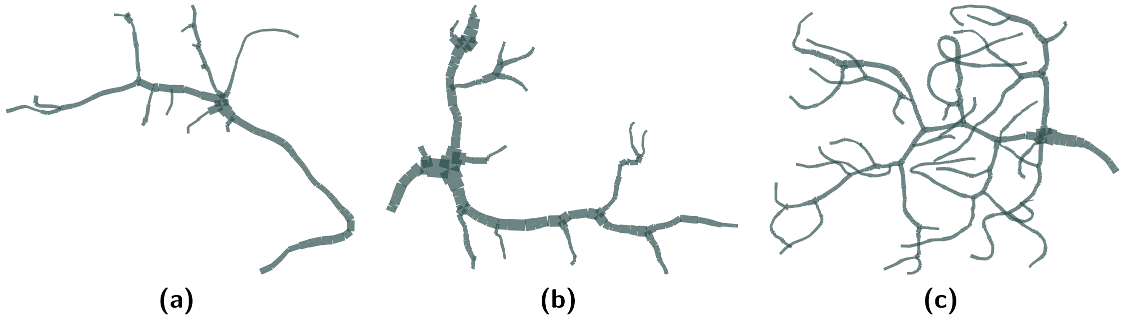


Figure 2.8. Networks extracted from the binary images displayed in figure 2.2.

3. Network Phenotyping

3.1. Introduction

The main focus of this work is to analyze a large set of digitized trachea networks from different genetic backgrounds, grown at different temperatures and expressing different mutations. The data was given to us with the minimum possible information about the biology of network growth. We did not know how the networks were *expected* to behave for certain mutations. Furthermore, the expected impact of different temperatures and genetic backgrounds on the functionality of gene knockdown and network growth in general was not known to keep possible bias in the analysis to a minimum.

Because of this deliberate non-information we had to look in every direction and analyze every possible aspect of the networks. In the following chapter in section 3.2 we will define and explain the metrics we measured for every network including purely topological as well as geometric observables.

We will then go on to describe the analysis methods we used to investigate how different growing conditions affect the growth of the networks and which growing conditions have sufficiently high impact to induce a distinguishable phenotype. In the following we will call the set of growing conditions a cell was grown with its *family* and each of the growing conditions a *characteristic* of the family. We specifically want to investigate whether the networks from the same family clustered in high dimensional metric space and therefore coincided with the phenotype defined by their growing conditions. If this is not the case this would indicate the existence of “hidden phenotypes” and the networks would form clusters independent of their families. For this purpose we first do a very qualitative assessment of the metrics we measured to see which of the family characteristics (mutation, genetic background or temperature) have the largest impact on the growth of the networks and if some of the characteristics can be ignored in later analysis. The results of this assessment are described in section 3.3.

In section 3.4 we determine whether we can reliably differentiate between the networks from different growing conditions by applying *linear discriminant analysis (LDA)*, a supervised learning approach, to the data. This gives us a measure of how reliable we can decide whether a network was grown with a certain characteristic using the measured growth behaviour.

To find possible hidden phenotypes, in section 3.5 we choose an unsupervised learning approach and apply clustering methods (*KMeans* and *GMM*) to the dataset. We investigate whether the clusters found by the algorithm are separated clearly and what the affiliation to a cluster signifies for the growth pattern expressed by the network.

As the clustering reveals that the strongest signal stems from pure network size and the networks fall on a continuum rather than forming separate clusters in metric space, we want to filter out the impact of network size on the metrics. In section 3.6 we do this by first choosing the four most representative geometric metrics for network growth: the length and radius of edges, the area of the convex hull and the length of the path along the tree from the root to the leaves. We then try to find the underlying dependence of each of these metrics on the best representative for network size: the number of branching points. After the dependence is known we can look at significant deviations from the predicted curve. With this approach we are able to find second order effects in network growth which contribute a small but visible amount to the information contained in each network realization. We call this effects *phenotypic trends*.

To ultimately understand how trachea terminal cells grow, in the last section 3.7 we try to set up growth models for the three main aspects of network growth: branching behaviour, radius distribution and direction of growth. We also try to assess whether mutations are able to qualitatively change the underlying growth models.

3.2. Network Metrics

Given the vectorized representations of the networks, we can start measuring certain metrics for each network. In the following, we want to establish a nomenclature for the description of graphs and derive the metrics we were able to measure given the information contained in the graphs.

3.2.1. Definitions

All definitions used in this work can be found in [22].

Graph: Let $G = (V, E)$ be the graph-representation of the network. We call V the *vertices* (nodes) of the graph and the set of vertex pairs $E \subseteq [V]^2$ the *edges* of G. In the following, we always assume $|V| > 1$.

Subgraph: Let $G = (V, E)$ and $G' = (V', E')$ be two graphs. We call G' a *subgraph* of G if $V' \subseteq V$ and $E' \subseteq E$.

Degree: We call the number of edges incident to a vertex the *degree* of the vertex $d_G(v)$.

Path: A *path* is a graph $P = (V, E)$ of the form

$$V = x_0, x_1, ..., x_k \qquad E = x_0x_1, x_1x_2, ..., x_{k-1}x_k,$$

with x_i pairwise different. The vertices x_0 and x_k are called the initial and terminal vertices of P_{x_0,x_k}, they are connected via P_{x_0,x_k}.

Connected: We call a graph G *connected* if for every two nodes $x, y \in V$ there is a path $P \in G$ which connects x and y. In the following, we always assume connected graphs.

Cycle: Let P be a path. We call P a *cycle* if $x_0 = x_k$, therefore P is a cyclic sequence of edges.

Tree: If a graph G does not contain any cycles, we call it a *tree* $T = (V, E)$. If moreover no vertex $v \in T$ has a degree $d_T(v) > 3$ we call the graph a *binary tree.*

Directed Graph: A *directed* graph D is a pair of vertices and edges (V, E) together with two functions init:$E \to V$ and ter:$E \to V$ which for each edge $e \in E$ specify an initial vertex init(e) and a terminal vertex ter(e). The edge is then directed *from* init(e) to ter(e). In a tree, by specifying a distinguished vertex as *root* $r \in V$ of the tree, we can define init and ter such that every edge is pointing away from the root and therefore create a directed tree T_D.

Subtree: A subgraph of a tree T which is also a tree is called *subtree.* In a directed binary tree, for every node with degree $d(v) > 2$ we can define a *left* subtree $T'_L = (V'_L, E'_L)$ and a *right* subtree $T'_R = (V'_R, E'_R)$.

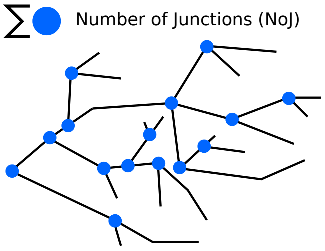
Leaf: Every vertex that has a degree $d_T(v) = 1$ and is not the root we call a *leaf* $l \in L$ of the tree.

The graph-representations of the networks used in this work all are connected, directed binary trees with $|V| > 1$.
We divide the metrics measured on the networks into two categories: Topological metrics are observables which only depend on the topological characteristics of the network i.e. the number of its elements (vertices and edges) and their relations. Geometric metrics are observables which only depend on the geometric characteristics of the network i.e. the coordinates of the vertices and the radii of the edges.

3.2.2. Topological Metrics

We define the topological metrics *Number of Junctions*, *Asymmetry*, *Tree Depth* and *Completeness* based on the information contained in the networks regarding the number and neighborhood-relationship of the network's vertices as well as the identity of the root node r.

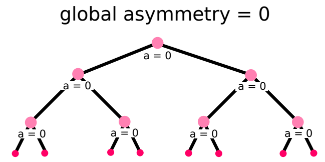
Number of Junctions

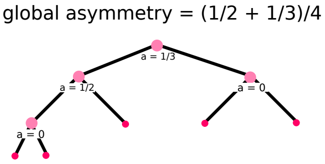


We define the *Number of Junctions* (NoJ) of a graph $G = (V, E)$ as the sum of all the nodes which have a degree larger than two, plus the root node:

$$\mathit{NoJ} = |\{v \in V \mid d_G(v) > 2\}| + 1 \tag{3.1}$$

Asymmetry





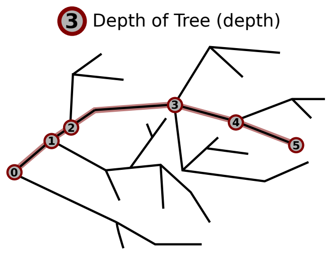
We define the *Asymmetry* asym of a vertex $v \in V$ as the absolute value of the difference between the number of leaves in the left and right subtree $|L_{L,R}| = \left|\{v \in V'_{L,R} \mid d(v) = 1\}\right|$ normalized by $\max(|L_L|, |L_R|)$. The asymmetry of the whole tree is the sum of all vertex-asymmetries normalized by the number of junctions:

$$\mathit{asym} = \frac{1}{\mathit{NoJ}} \cdot \sum_{i=1}^{\mathit{NoJ}} \frac{|L_{L,i} - L_{R,i}|}{\max(L_{L,i}, L_{R,i})} \tag{3.2}$$

Tree Depth

Given a graph G, the *Tree Depth* (or *Depth*) $td(v)$ of a vertex $v \in V$ is the length of the shortest path connecting v and the root r:

$$td(v) = \min\left(|\{P_{r,v} \in G\}|\right) \tag{3.3}$$

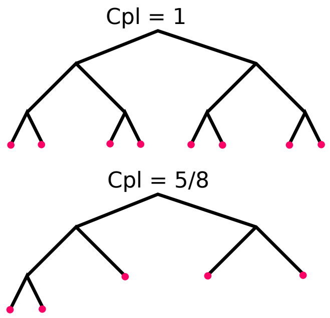


The depth of the tree is given by the maximum vertex depth $td(T) = \max\left[td(v \in V)\right]$.

Completeness

Given a binary tree T width depth $td(T)$, the maximum number of leaves it could have is $|L|_{\max} = 2^{td(T)}$. Its completeness *Cpl* is the ratio between the actual and the maximal number of leaves:

$$Cpl = \frac{|L|}{|L|_{\max}} \tag{3.4}$$



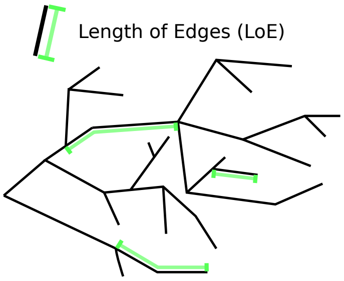
3.2.3. Geometric Metrics

We define the geometric metrics *Length of Edges and Network*, *Radius of Edges*, *Spacefillingness* and *Topological, Geometric and Normalized Distance to Root* based on the information contained in the networks regarding the (x, y)-coordinates of the nodes in the graph as well as the radii of its edges.

Length of Edges and Length of the Network

Given a graph $G = (V, E)$, for every edge $e \in E$ we can define the *Length of the Edge LoE* as the geometric distance between the two vertices $v_{i,j}$ it connects:

$$LoE(e) = \sqrt{(x_{v_i} - x_{v_j})^2 + (y_{v_i} - y_{v_j})^2} \tag{3.5}$$

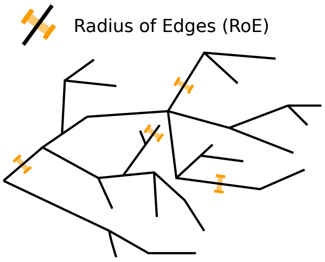


It has to be noted that the edge lengths used for the calculations are not simply the geometric distances between the junctions. The final graph-representation of the network only contains the junctions and leaves - all the

nodes with degree two (corresponding to *Normal* triangles) which originally were produced by the triangulation have been removed. To create a better approximation of the geometry and take curvy growth into account, the actual length of an edge is the sum of the lengths of all the originally created skeleton segments approximating the original network.
The length of the whole network is the sum of the lengths of all edges:

$$LoN = \sum_{e \in E} LoE(e_i) \tag{3.6}$$

Radius of Edges



The *Radius of an Edge RoE* is the mean of the radii of the tube at the two vertices it connects. This radius is not to be confused with the topological definition of the radius of a graph, it is a purely geometric measure of how thick the tube at the position of the vertex is.
To increase the accuracy of the approximation, the radius of an edge contained in the final graph representation also is calculated as a mean of all the radii of the skeleton segments which originally were situated in between the junctions it connects weighted by their lengths.

Area of Edges and Area of the Network

Using the *LoE* and *RoE* we can calculate the *Area of an Edge AoE* and the *Area of the Network AoN*:

$$AoE(e) = 2 \cdot RoE(e) \cdot LoE(e) \; ; \qquad AoN = \sum_{e \in E} AoE(e) \tag{3.7}$$

Normalized Area of Edges and Normalized Area of the Network

The *Normalized Area of Edges NA(e)* is the area of an edge normalized by the minimal edge length present in the graph:

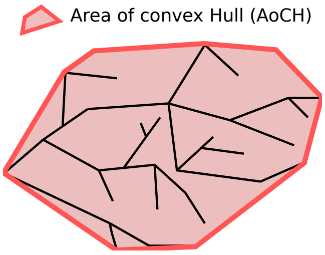
$$NA(e) = \frac{2 \cdot RoE(e) \cdot LoE(e)}{\min(\{LoE(e) \mid e \in E\})} \tag{3.8}$$

The normalized area of the whole graph is the average over all its normalized edge areas:

$$NA = \frac{1}{|E|} \sum_{e \in E} NA(e) \tag{3.9}$$

Area of the Convex Hull

A set of points is defined to be convex if it contains all the line segments connecting each of its points. We define the convex hull CH of a graph $G = (V, E)$ as the minimal convex set containing V. If we order the points of the convex hull counterclockwise we can guarantee that the points represent a non-intersecting polygon for which we can calculate the area



$$AoCH = \frac{1}{2} \cdot \left| \sum_{i=0}^{n-1} x_i y_{i+1} - x_{i+1} y_i \right| , \tag{3.10}$$

where $n = |CH|$ and x_i, y_i are the two-dimensional coordinates of the vertices contained in the convex hull.

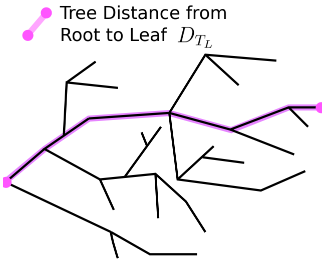
Spacefillingness

The *spacefillingness* Sfn is a measure of how dense a network fills the area it occupies:

$$Sfn = \frac{AoN}{AoCH} \tag{3.11}$$

Tree-Distance to Root

For every vertex $v \in V$ we can define the *Tree-Distance to the Root* $D_T(v)$ as the length of the shortest path $\min(|P_{r,v}|)$ along the network from the root to the vertex:



$$D_T(v) = \sum_{i=0}^{\min(|P_{r,v}|)-1} \sqrt{(x_i - x_{i+1})^2 + (y_i - y_{i+1})^2} \tag{3.12}$$

The *Tree-Distance to the Root* for the whole *network* $D_T(V)$ is the mean of the tree-distances of all its leaves:

$$D_T(V) = \frac{1}{|V|} \sum_{v \in V} D_T(v) \tag{3.13}$$

We can split D_T into three different metrics by looking at different sets of vertices:

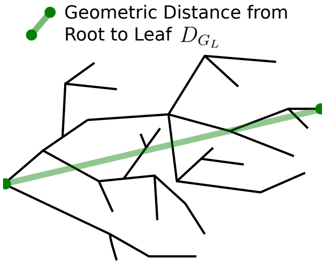
- All vertices: $D_T(V) = D_{T_V}$
- Only junctions: $D_T(J) = D_{T_J}$
- Only leaves: $D_T(L) = D_{T_L}$

As $V = J + L$ it suffices to look at two of the three metrics. Therefore we only use D_{T_V} and D_{T_L} for the following analysis.

Geometric Distance to Root

For every vertex $v \in V$ we can define the *Geometric Distance to the Root* $D_G(v)$ as the euclidean distance between the vertex and the root:

$$D_G(v) = \sqrt{(x_v - x_r)^2 + (y_v - y_r)^2} \tag{3.14}$$



The *Geometric Distance to the Root* for the whole *Network* $D_G(V)$ is the mean of the geometric distances of all its vertices:

$$D_G(V) = \frac{1}{|V|} \sum_{v \in V} D_G(v) \tag{3.15}$$

Again we can split D_G into the three different metrics D_{G_J}, D_{G_V} and D_{G_L}, and use only the last two.

Normalized Distance to Root

For every vertex $v \in V$ we can define the *Normalized Distance to the Root* $D_N(v)$ as the ratio between the tree distance and the geometric distance to the root:

$$D_N(v) = \frac{D_T(v)}{D_G(v)} \tag{3.16}$$

Again we can define the normalized distance to the root for the whole network $D_N(V)$ and split it into three different metrics D_{N_J}, D_{N_V} and D_{N_L} and use only the last two.

3.2.4. Redundancy

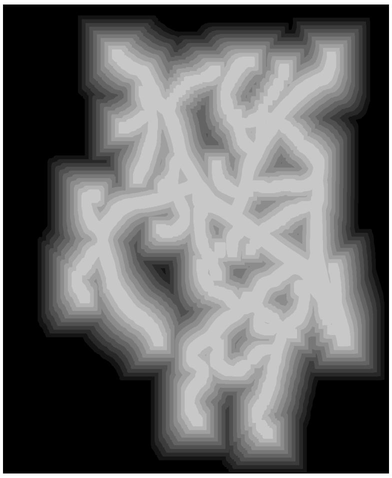


Figure 3.1. Illustration of dilation on a binary image with varying radii r_{diff} of the structuring element.

For every graph we can define the *edge-connectivity*. A graph is *k-edge-connected* if we have to remove a minimum of k edges until the first component of the graph becomes disconnected. In loopy graphs, this can be used as a measure for the graphs *redundancy*. For example, the transport network in plant leaves still works if defects in the form of severed veins are introduced - there is redundancy in the network. For trees-like graphs however, the *k-edge-connectivity* always is one as there are no loops. To define a measure of the redundancy of the trachea, we have to think of something else.
The function of the trachea network is to transport oxygen to the surrounding tissue via diffusion. We therefore define redundancy for this kind of transport network as the area of the surrounding tissue that can be reached via diffusion by more than one tube, given a diffusion range r_{diff}.
We could calculate that area from the geometry and coordinates of the network we already have in the graph object. However it is easier and less approximate to follow a purely graphical approach: As we already have the binary image from previous processing steps, we can utilize the morphological operation of *dilation* [23] to simulate diffusion from the network into the surrounding tissue, as illustrated in figure 3.1. For each dilation range r_{diff} (i.e. kernel size/structuring element of the dilation operator) we can calculate the number of redundant pixels by

- First calculating the maximum number of pixels N_{max} which are reached if we dilute a straight line with the length of the boundary of our foreground structure

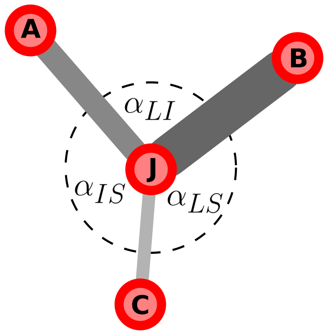
- Then calculating the actual number of pixels N_{act} reached by dilating our foreground structure with a structuring element with radius r_{diff}

- Subtracting N_{act} from $N_{\max}$. We now have calculated the number of pixels N_{Red} which can be reached from more than one tube at diffusion range r_{diff}.

Angles

Every junction J except the root has degree three, therefore at every junction we can measure three angles between edges which meet at the junction. The angle between the edges connecting J to node A and the edge connecting J to node B is

$$\alpha_{AB} = \arccos\left(\frac{\vec{AJ}\cdot\vec{JB}}{\|\vec{AJ}\|\|\vec{JB}\|}\right) . \tag{3.17}$$



We differentiate between three different angles specified by the radius of the edges enclosing the angle:

- The angle between the edge with the largest and the edge with intermediate radius α_{LI}.
- The angle between the edge with the largest and the edge with the smallest radius α_{LS}.
- The angle between the edge with intermediate and the edge with the smallest radius α_{IS}.

3.2.5. Summary

For the purpose of further analysis, in this section a large quantity of different metrics which are measured on the networks from the dataset are defined. The metrics can be divided into purely topological observables only dependent on the number of network elements and their relation such as the number of junctions and the network asymmetry. The other type of metric is purely geometric, only dependent on the two-dimensional coordinates of the nodes and the radii of the edges. We also defined a measure for the redundancy of a tree-like transportation network using its functionality as a diffusive transport system for oxygen.

3.3. Information Content of Network Characteristics

3.3.1. The Dataset

For each network in the dataset, the growing conditions of the larvae were controlled for *temperature* t, *genetic background* B_{Gen} and *mutation* M. Furthermore, for each network, the horizontal (left or right hemisphere) and vertical position in the insect at which the network grew were recorded. For further analysis, horizontal and vertical position were ignored as this information was already being analyzed in another work created in parallel to this thesis [24]. The position in the larva also was assumed to have less impact on the growing pattern of the terminal cell than t, B_{Gen} and M.

As already mentioned, a combination of $\{t, B_{\text{Gen}}, M\}$ is called a *family*, the elements t, B_{Gen} and M are called *characteristics* of the family. In total our dataset consists of 30 different families where each family contains between 10 and 20 networks for a total of 500 different networks. The exact number of networks each family is composed of is listed in table C.1 in the appendix. The number of networks in each family is already quite high compared with other analysis done for trachea terminal cells [25], [26]. This allows for statistically more significant statements, however it has to be kept in mind that the amount of noise still can be quite high and other effects such as possible mechanical deformations of the larvae during preparation for the imaging could distort the results.

Temperature

Each characteristic combination (M, B_{Gen}) was grown at two different temperatures $t \in \{18°C, 29°C\}$. In general, larvae grow faster at higher temperatures but as the time span after which the imaging was performed was not constant but chosen with regard to the developmental stage of each larva, a systematic impact of temperature on the size of the network based on the development of the whole larvae can be ruled out.

Genetic Background

A genetic background is a genetic construct used to tag the larvae for imaging (described in more detail in section 4.2.3). This tagging also influences the RNA-machinery responsible for the mutation. For this work the larvae came from three different genetic backgrounds expressing the fluorescing protein *GFP* and the proteins *Dicer2* and *PIP2*. In the following, the genetic backgrounds will often be abbreviated as G (only GFP), DG (Dicer2 and GFP) and PDG (PIP2, Dicer2 and GFP).

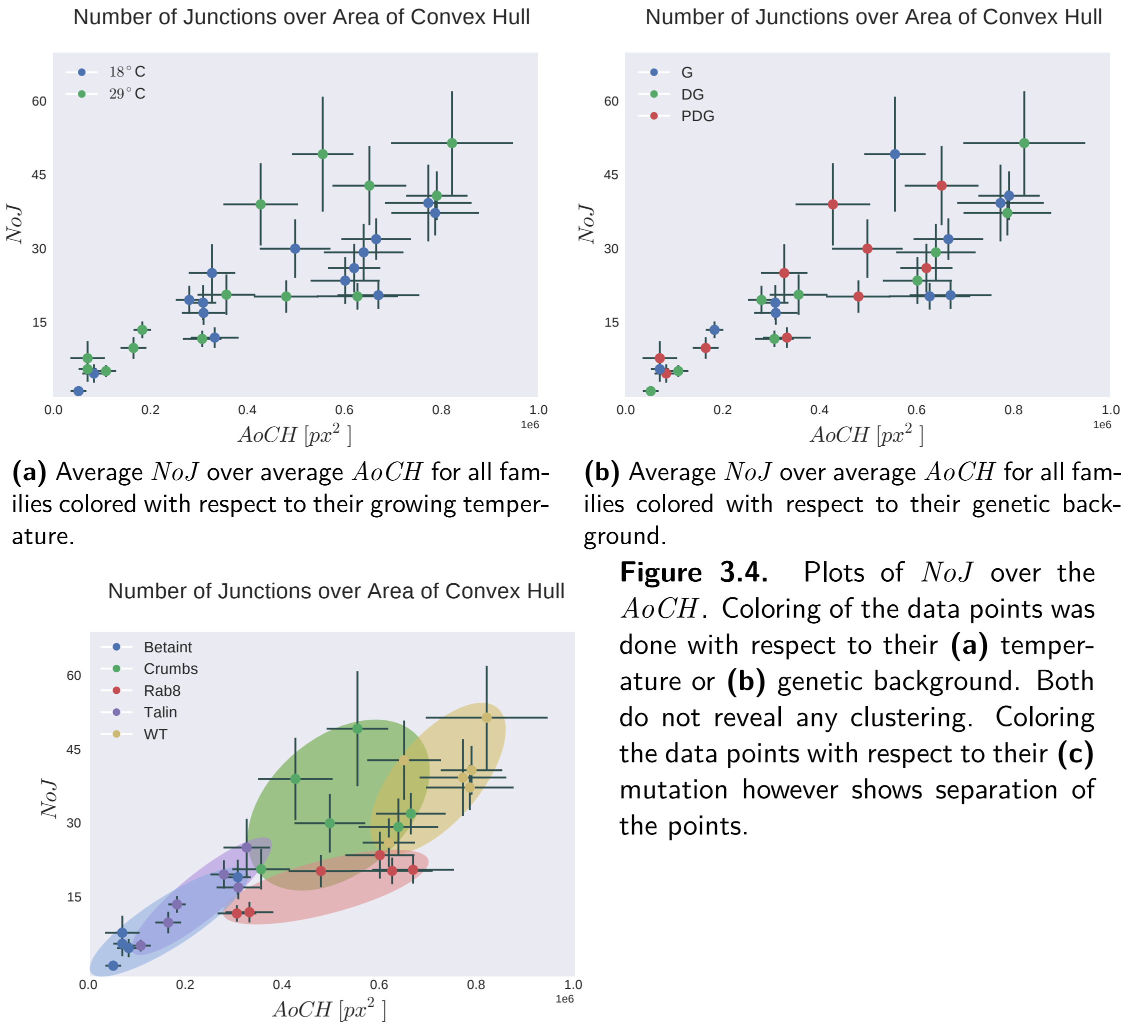
Mutation

For each combination of (t, B_{Gen}) a population with a different mutation was grown. The mutation was realized by interfering with the expression of one of the four genes *Betaint*, *Crumbs*, *Rab8*, or *Talin*, the function of these genes is described in section 4.2.2. As control, for each temperature and background there is also a set of networks from *Wild Type* (WT) larvae.

The means and standard deviations of all topological and geometric metrics for every network family except the angles and redundancy are listed in the appendix C.

3.3.2. Is a single Characteristic enough to discriminate between Networks?

As the number of data points in every family still is very low and noise therefore pretty high, we first wanted to investigate whether all of the characteristics have an impact on network growth. If this is not the case, we can safely collapse some of the families and analyze only within the characteristics that do matter. We can get a first impression of whether we can reliably discriminate between different families based on the discriminative power of the information contained in each of the characteristics.
The first approach was to plot each two metrics we measured against each other and see whether data points expressing a certain characteristic would visibly cluster. An example of that approach can be seen in figure 3.4. We can easily assess that for t and B_{Gen} (sub-figures 3.4a and 3.4b) there is absolutely no separation between data points with different characteristics - points with different characteristics are spread over the whole range of possible values. For M however we see a clear tendency of different mutants to occupy only part of the available range (subfigure 3.4c). This indicates that information contained in the characteristics t and B_{Gen} is not sufficient to discriminate between different families but information contained in the mutation might be. The same behaviour can be seen for all pairings of metrics.



(a) Average NoJ over average $AoCH$ for all families colored with respect to their growing temperature.

(b) Average NoJ over average $AoCH$ for all families colored with respect to their genetic background.

(c) Average NoJ over average $AoCH$ for all families colored with respect to their mutation.

Figure 3.4. Plots of NoJ over the $AoCH$. Coloring of the data points was done with respect to their **(a)** temperature or **(b)** genetic background. Both do not reveal any clustering. Coloring the data points with respect to their **(c)** mutation however shows separation of the points.

3.3.3. Conclusion

The variations in network growth caused by different t or B_{Gen} seem to be small compared to the variations caused by M. A first assessment of the metrics measured for all the networks showed that the networks with the same temperature of genetic background do not cluster whereas networks with the same mutation do. Therefore, using the growth pattern of a network, we might be able to make a statement of whether the larva had a certain mutation or not but we cannot predict its genetic background or growth temperature. Nevertheless, within all the networks with the same M we still might be able to find an impact of t and B_{Gen} in the form of smaller variations in network realization.

3.4. Quantification of Class Discernability using Linear Discriminant Analysis

After the first very qualitative assessment of the data generated by measuring all the above-mentioned metrics on the networks, in the next section we want to follow a more quantitative approach. Our goal is to quantify the discriminative power of the characteristics using *Linear Discriminant Analysis*.

3.4.1. Linear Discriminant Analysis

Methodology

Linear discriminant analysis (LDA) is a supervised learning approach to discriminate between different (known) classes in a dataset. It was first described by Fischer [27] and later generalized for more than two classes by Rao [28]. LDA tries to find a linear combination of class features (in our case the metrics we measured) which characterizes a certain class best. From this we can also derive linear decision-boundaries or hyperplanes in the multidimensional metric space which optimally separate the classes from each other. LDA can be visualized by projecting the data onto the most discriminative directions, in the following called the first and second LDA component.

Assumptions

For the application of LDA it is assumed that the data-population of each class follows a Gaussian distribution and the covariance matrices of every population are equal. If the data is multimodal, modeling it with a single Gaussian is a poor approximation and might lead to high class overlap. In general using a more complicated distribution is possible but might lead to a very complicated optimization problem. Therefore in this work we model distributions as Gaussian.

Even if the covariance matrices are not equal, LDA works quite well in most of the cases. Problems caused by non-equal covariance matrices can be circumvented by using the *Mahalanobis* distance [29] between a point and a distribution instead of the symmetric squared distance.

A more detailed description of the approach and the algorithm used for this work as well as an explanation of the scenarios in which the LDA-algorithm might fail is given in the appendix A.2.

3.4.2. LDA Class Analysis

To further assess how well we can discriminate between different families, we performed LDA on the dataset. The metrics we used for the LDA were

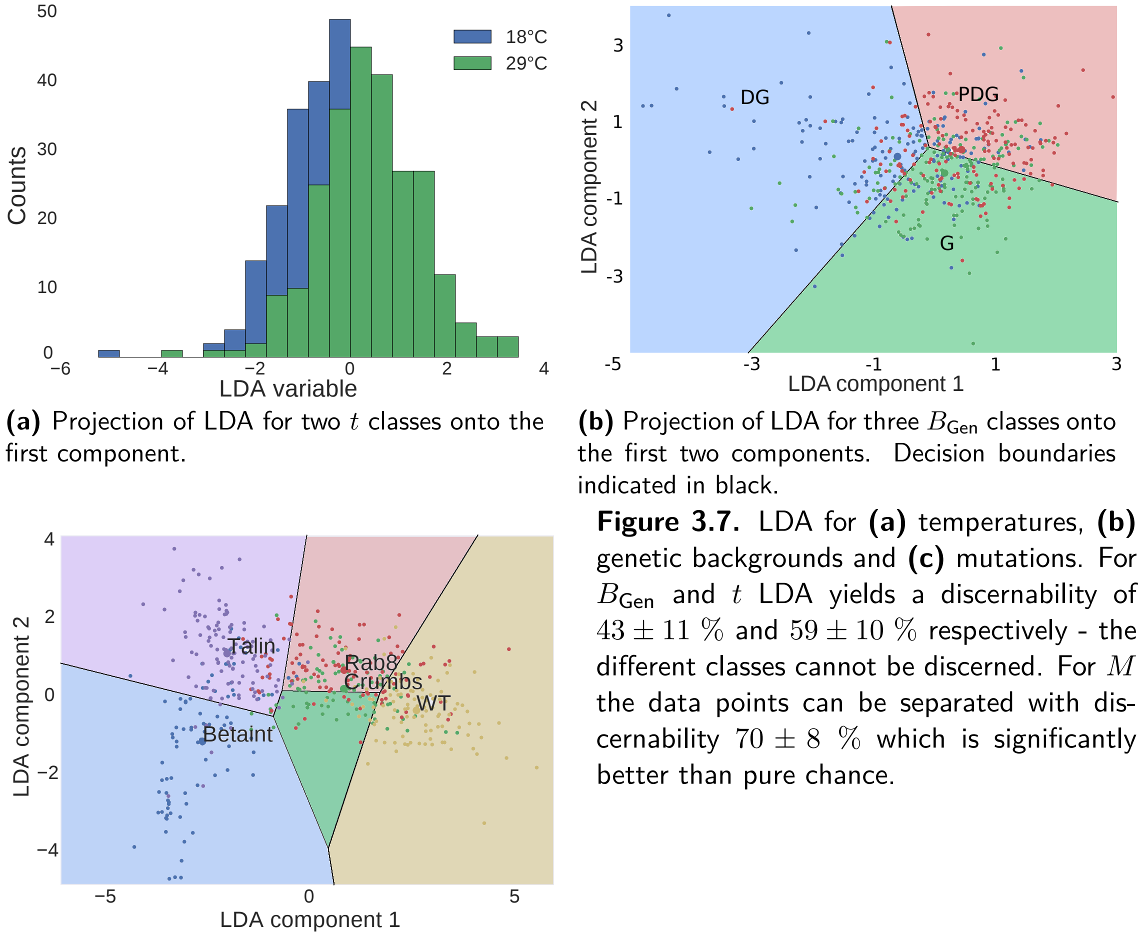
- NoJ, $asym$, Cpl
- LoE, LoN, AoN, RoE, NA, $AoCH$, Sfn
- D_{T_L}, D_{T_V}, D_{G_L}, D_{T_V}, ND

for a total of 15 different metrics.
As demonstrated in figure 3.7c, LDA is able to discriminate between networks from different mutants (five different classes) with a success rate of 70 ± 8 % true positive and true negative predictions. Taking into account the different temperatures too and therefore performing LDA for 10 different classes worsens the discernability to 52 ± 7 %. Additionally splitting the classes for genetic background (30 classes) even brings discernability down to 37 ± 8 %. Using only temperature (two classes, figure 3.7a) or genetic background (three classes, figure 3.7b) as class label yields a discernability of 59 ± 10 % and 43 ± 11 % respectively which is not better than guessing randomly. Therefore the mutation is the only label which enables us to discriminate between networks with a success rate significantly better than chance.

3.4.3. LDA Metric Analysis

In figures 3.8a and 3.8b the weights used for the linear combination of metrics forming the first and second LDA component are visualized. It can be seen that the radius of edges RoE and the spacefillingness Sfn have by far the largest influence. This is an interesting finding as both metrics are purely geometric. Topological metrics like $asym$ or Cpl do not seem to have much impact on the network discernability.

We now want to investigate whether some of the metrics are obsolete or contribute more to discernability than others. To find out whether this is the case, we perform LDA using varying numbers of metrics, starting with only two and then

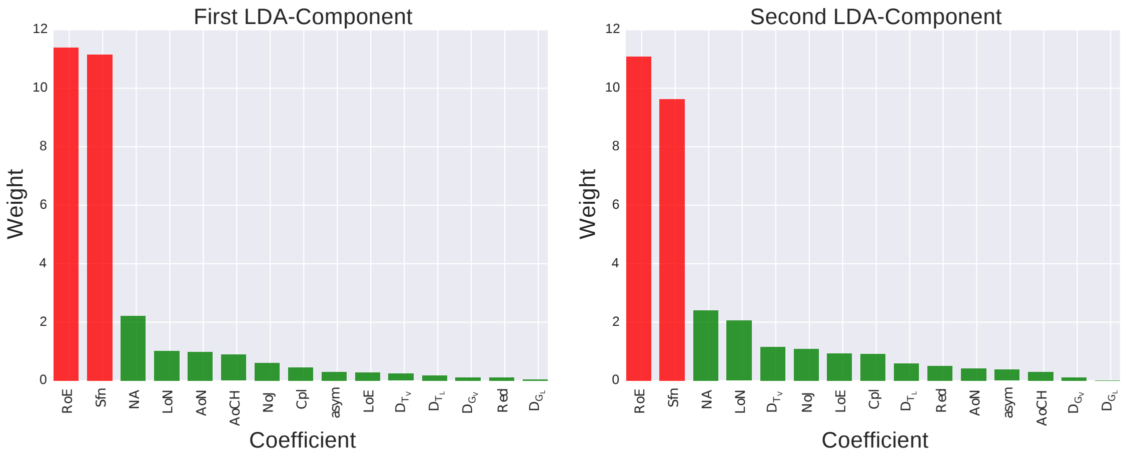


(a) Projection of LDA for two t classes onto the first component.

(b) Projection of LDA for three B_{Gen} classes onto the first two components. Decision boundaries indicated in black.

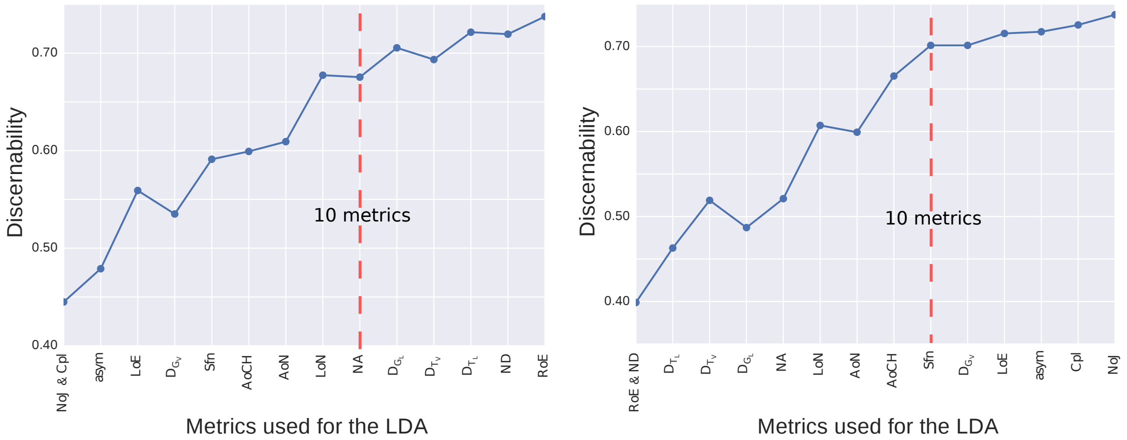
(c) Projection of LDA for five M classes onto the first two components. Decision boundaries indicated in black.

Figure 3.7. LDA for **(a)** temperatures, **(b)** genetic backgrounds and **(c)** mutations. For B_{Gen} and t LDA yields a discernability of 43 ± 11 % and 59 ± 10 % respectively - the different classes cannot be discerned. For M the data points can be separated with discernability 70 ± 8 % which is significantly better than pure chance.



(a) Coefficients of the first LDA component. **(b)** Coefficients of the second LDA component.

Figure 3.8. Analysis of the coefficients of the metrics present in the first **(a)** and second **(b)** LDA component. It can be seen that RoE and Sfn have the largest influence.



(a) Discernability for LDA with increasing metric number.

(b) Discernability for LDA with increasing metric number and reversed order to **(a)**.

Figure 3.9. Discernability for LDA performed on different sets of metrics. The LDA was started with two metrics and then successively more metrics were added. The x-axis shows the metrics added in each step. **(a)** Shows a certain order of metrics starting with the topological ones. **(b)** reverses this order. For both trajectories, a qualitatively similar trend is visible.

successively adding more metrics, one at a time. Figure 3.9a shows the development of the discernability with increasing number of metrics for a certain order of metrics. Figure 3.9b shows the same for the reversed order. It is noticeable that the trajectory both times exhibits qualitatively similar behaviour: It reaches around 70 % discernability after around 10 metrics and then seems to saturate.

We checked other metric orders and combinations and found the same behaviour:

It is more or less not important *which* metrics are used for the LDA, only their raw number matters. From this we can conclude, that the metrics share a lot of information on the network. This is quite natural as some of them are derived from other metrics and most of them depend on the number of junctions present in the network. Nevertheless, the picture only becomes complete after we have added approximately 10 metrics to the LDA. This indicates that the metrics also each contribute an individual part of information.

3.4.4. Conclusion

Using an approach from supervised learning, namingly linear discriminant analysis shows that based on the already known class labels for the different mutants, we can discriminate networks with a 70 ± 8 % accuracy which means that the probability $P[\text{correct classification}] = 0.70$. Using only temperature yields a much worse discernability of 59 ± 10 % as does discrimination based on genetic background with a discernability of 43 ± 11 %. Increasing the number of classes by using a combination of mutation with temperature or mutation with genetic background also does not yield a discernability better than chance.

The discernability saturates after utilization of about 10 out of the 15 metrics, the order and identity of the metrics used does not matter. This points to a high correlation between the different metrics.

3.5. Hidden Phenotypes

After applying a supervised learning method to the dataset to discriminate between networks from different mutants, we want to investigate whether we can find grouping in the networks leaving aside the known labels. For this we use an unsupervised learning approach where we try to assign class labels to networks based on the growth characteristics they exhibit. We use clustering on the dataset with the same metrics as in the LDA approach to see whether the networks cluster within the phenotype induced by their mutation. We also want to investigate, if the clusters are separated clearly and what the optimum number of clusters is.

3.5.1. Clustering

Clustering Algorithm

To rule out an influence of the clustering method on the results we tried two different popular clustering algorithms: a centroid based clustering method, the *KMeans* [30] algorithm, and a distribution based clustering method, the *Gaussian Mixture Model* [31] (GMM). The approaches for both KMeans and GMM as well as the actual implementation used in this work are described in the appendix in sections A.3.1 and A.3.2. Both clustering methods showed very similar results and only very small if any quantitative deviations. We therefore continued to only use *KMeans* in the following clustering analysis.

Silhouette Score

To answer the question of which number of clusters is optimal, we look at the *Silhouette Score* S [32] of each clustering for different numbers of clusters:

$$S = \frac{a - b}{\max(a, b)} \ ,$$

where a is the mean intra-cluster distance and b is the mean nearest-cluster distance. Values of S lie in $[1, -1]$, were 1 signifies good clustering, zero signifies overlapping clusters and -1 signifies false clustering.

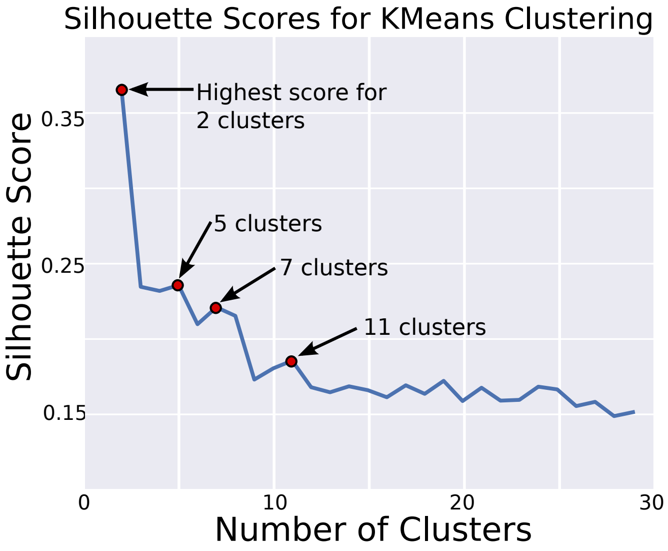
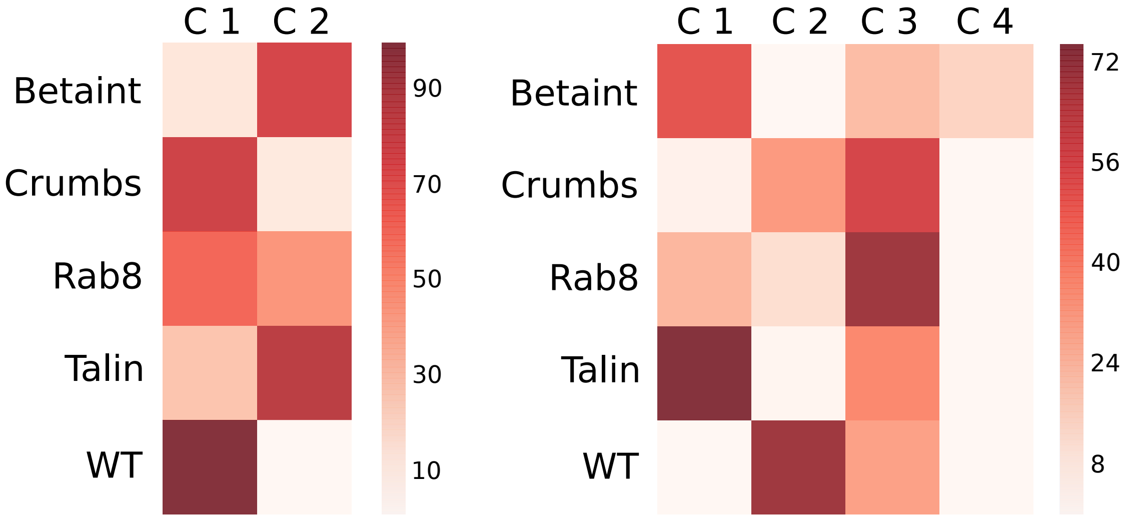


Figure 3.10. Behaviour of the *Silhouette Score* for the *KMeans* clustering algorithm with increasing cluster sizes. The first four significant local maxima are indicated by arrows and red dots. The Silhouette Score never lies above 0.5 which indicates overlapping clusters.

Figure 3.10 shows the behaviour of the silhouette score for different cluster sizes. It can be seen that the silhouette score always lies below 0.5 which indicates that the clusters are not separated clearly. Local maximal in the silhouette score signify a more feasible number of clusters than the neighboring numbers of clusters. The first significant local maxima are 2, 5, 7 and 11 clusters.

Clustering into two and four Clusters

For further analysis, we looked at the first two maxima of S. Five clusters turned out to be only four *proper* clusters and one separate cluster for a single outlier. Figures 3.11a and 3.11b show clustering with two clusters and four clusters. At first glance the clustering already reveals that networks with the same mutation are *not* all affiliated with the same cluster but spread between two or more different clusters.



(a) Clustering for two possible clusters.

(b) Clustering for four possible clusters.

Figure 3.11. Heatmaps of the clustering into two **(a)** and four **(b)** clusters respectively. The color code signifies number of networks which have been assigned to each of the clusters. Cluster numbers are indicated by labels C1 - C4.

Four Clusters

We continued with clustering into four clusters. S might be lower but forcing the networks to divide into four different clusters gives a more differentiated picture of similarities and dissimilarities between networks than using only two clusters.

Using the LDA approach described in section 3.4 we found that classifying networks based only on their temperature and genetic background was not possible. Nevertheless we could not rule out an impact of temperature and genetic background on network realization within a mutation. The unsupervised learning approach allows us to differentiate between different temperatures again and see whether networks from the same mutant but grown at different temperatures and with different genetic backgrounds will be affiliated with different clusters.

Figure 3.12 illustrates that temperature can have a high influence on the clustering of networks with the same mutation. For *Talin* and *Betaint* the majority of networks grown at different temperatures are affiliated with completely different clusters. For the other mutants, the impact is not that strong and could be attributed to noise. Genetic background has some influence on the clustering of the networks from all mutants and even the *Wild Type* but the influence is not as strong as that of temperature and might be caused by noise. One notable exception is *Betaint* 18° were each different genetic backgrounds is affiliated with a different cluster.

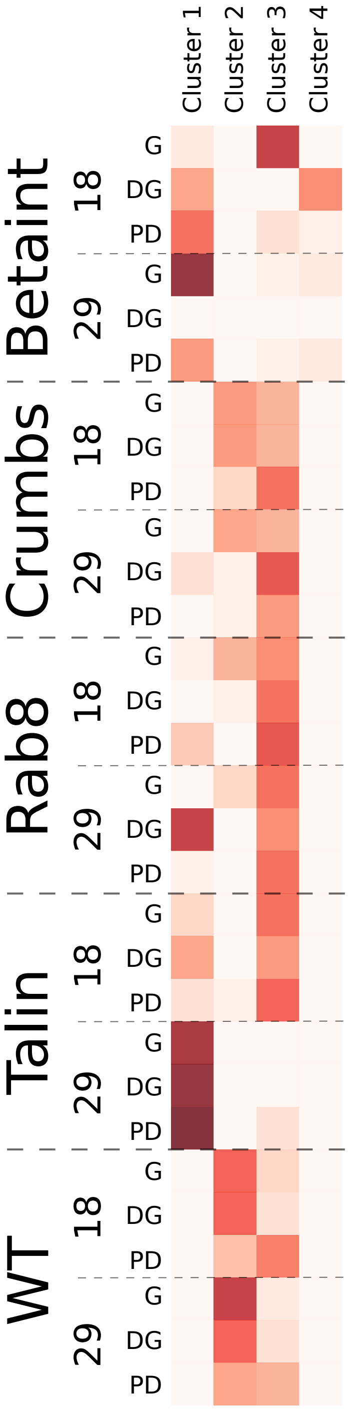


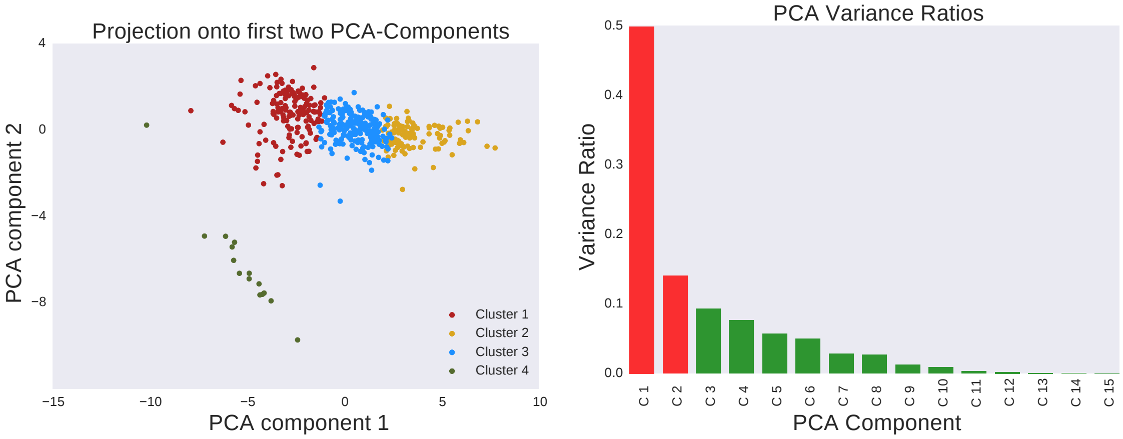
Figure 3.12. Clustering into four clusters with network families broken apart into mutation, temperature and genetic background.

3.5.2. PCA

We have answered the questions of whether the clusters fall onto the original phenotypes induced by the mutation (they do not) and if temperature and genetic background have an influence on the clustering. The low silhouette score also already indicated overlapping clusters therefore in the next step, we want to visualize the clustering to see how the clusters are located in high-dimensional metric space.

Clustering Visualization

Principal Component Analysis (PCA) [33] is a feasible way to make the clustering visible. In appendix A.1 we give a more detailed description of the functionality of PCA and its implementation in this work.
We perform PCA on the same dataset as the clustering before, then project the dataset onto the first two principal components and color the data points with regard to their affiliation to a specific cluster.



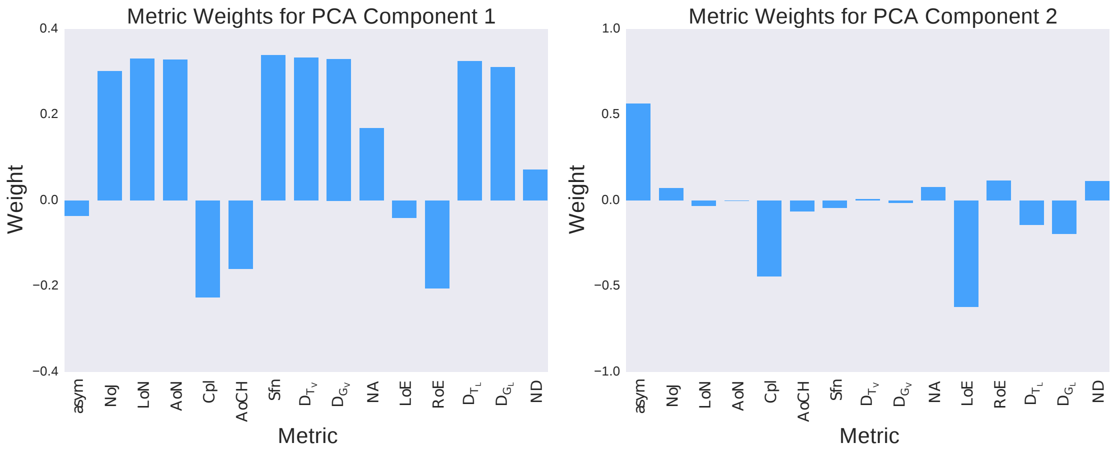
(a) Projection onto the first two PCA components. Coloring according to cluster affiliation: Cluster 1=red, Cluster 2=yellow,Cluster 3=blue, Cluster 4=green.

(b) Contribution to variance of each of the PCA-components $C1 - C15$.

Figure 3.13. **(a)** Projection of the dataset onto the first two PCA components. The colors were chosen according to the affiliation of each data point with the clusters created in the previous sections. **(b)** Visualization of the contribution to variance in the dataset by the different PCA components. The first two components are emphasized in red, they already contribute 50% and 15 % of the variance respectively.

Figure 3.13a shows the projection of the PCA onto the first two components. It is clearly visible that the first three clusters form a continuum along the first PCA

component rather than clearly separated clusters. Only cluster four is separated from the others along the second PCA component.
Looking at the metric weights for the first two PCA components in figures 3.14a and 3.14b reveals that all metrics except *asym*, *LoE* and *ND* contribute to the first component with more or less the same absolute weight whereas the second component is heavily influenced by *asym*, *LoE* and *Cpl*.



(a) Metric weights for the 1st PCA-component. **(b)** Metric weights for the 2nd PCA component.

Figure 3.14. **(a)** and **(b)** Analysis of the influence of the different metrics on the first and second PCA component.

To explain the separation of the fourth cluster from the other clusters along the second PCA component and the strong influence of *asym*, *Cpl* and *LoE* on this component, we anticipate the analysis performed in section 3.5.3 and look at the growth patterns of the specific networks forming the fourth cluster.
We can see that only *Betaint* 18° *Dicer GFP* occupies the fourth cluster. These networks show a degenerate growing pattern: they do not branch and therefore consist of only the root node and a leaf. This edge case causes especially the topological metrics *asym* and *Cpl* to yield extreme results: for a network with no branches $asym = 1$ whereas if branches are present it always lies around 0.5. $Cpl = 0.5$ for this degenerate case whereas for larger networks it tends to be smaller by at least one order of magnitude.
The influence of the edge length on the second component also fits well into the picture of the degenerate networks: these networks will exhibit an extremely high value for *LoE* as there is only one long branch which is not averaged with other

smaller edges.

Therefore, as soon as a network starts to have at least some branches it immediately falls into a continuum of network realizations along one dimension in metric space which is influenced by all metrics except *asym*, *LoE* and *ND*. The underrepresentation of *ND* could be explained by its dependance on D_{T_L} and D_{G_L}.

As the origin of the spread in the second PCA component stems from the degeneracy of the networks and metrics which diverge for networks with no branches, we can safely ignore the second component and only look at the first. This approach is also supported by the fact that the first PCA component already contributes half of the variance present in the dataset as is illustrated in figure 3.13b.

A closer inspection of the projection onto the first PCA component in figure 3.15 still shows absolutely no separation between adjacent clusters. The overlap between adjacent clusters is not large however neither is the distance. Clusters 4 and 2 as well as clusters 1 and 2 can be separated from each other quite clearly but cluster 3 completely spans the space separating the other clusters.

Had we chosen fewer or maybe different mutation for our analysis, the clusters might have separated more clearly but for the selection of *Betaint*, *Crumbs*, *Rab8*, and *Talin* this is not the case.

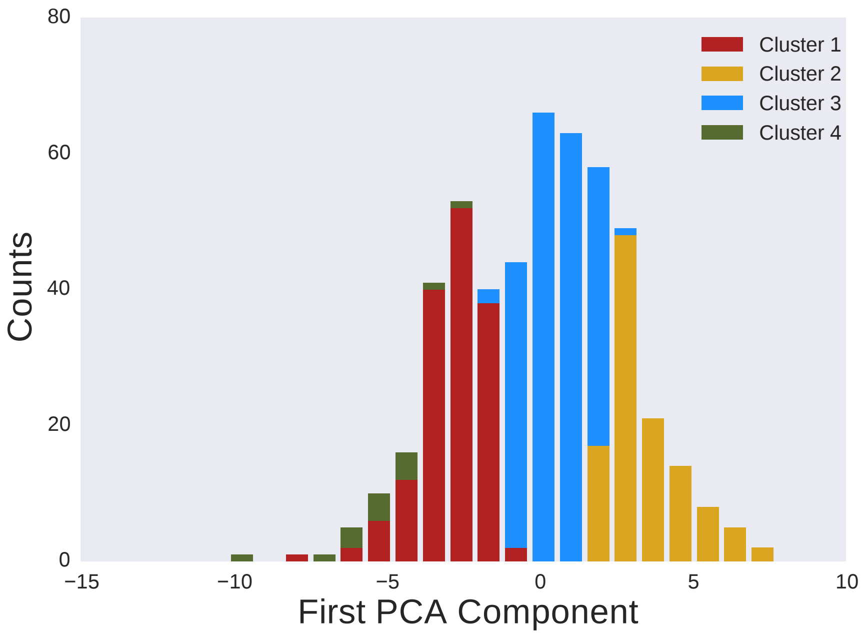


Figure 3.15. Histogram of the data points projected onto the PCA component with the highest variance. The coloring was chosen according to the affiliation of the data points to the different clusters.

3.5.3. New Phenotypes

After establishing that the networks form a continuum rather than separate clusters, we can proceed to investigate what growth characteristics the networks belonging to a certain cluster express. For a qualitative indication of network traits we look at scatterplots of different metrics. How are the networks belonging to a cluster arranged in these plots? Figure 3.16 is the most telling of these scatterplots: it reveals that the clustering is based to a very large degree on network size. This makes sense as most of the metrics are also dependent on network size.

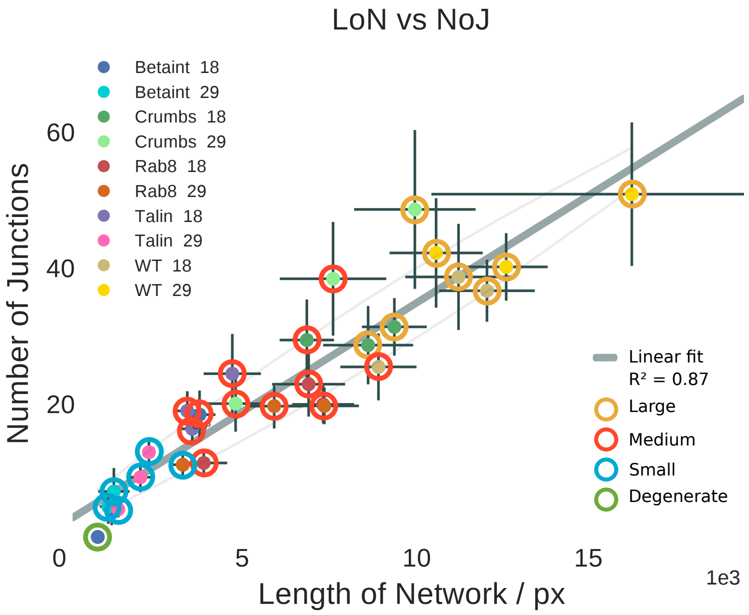


Figure 3.16. Scatterplot of the NoJ over NoJ. The data points all fall on a line, the clustering divides this linear distribution into four parts based on network size. Cluster affiliation is indicated by colored circles around the data points: Large (yellow) = cluster 2, Medium (red) = cluster 1, small (blue) = cluster 3 and degenerate (green) = cluster 4.

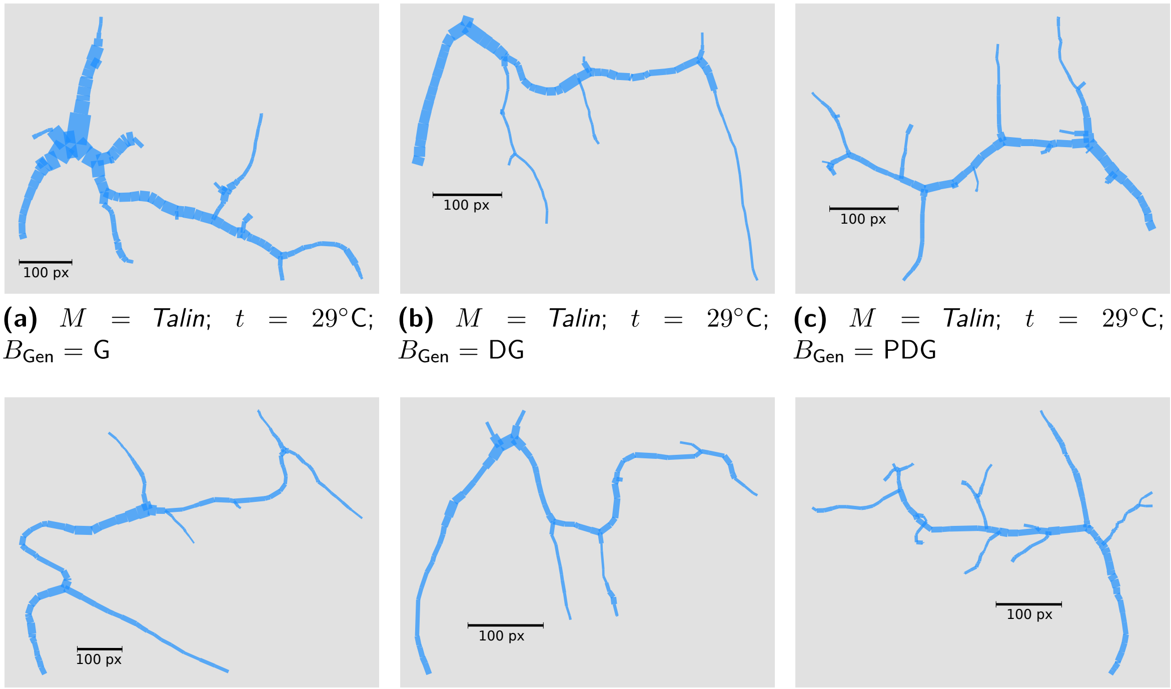
With this in mind, we conclude that the metric most important to clustering has to be the number of junctions present in a network. Its influence on the other metrics will be described in more detail in section 3.6. Based on this, we are able to define a main growth characteristic for every cluster: the main trait of networks affiliated with clusters one to three is to be of *small*, *medium* or *large* size whereas the networks in cluster four are *degenerate* and contain only a root and a leaf.

Cluster 1 or the "*small*" Phenotype

The networks belonging to cluster one all exhibit a relatively small number of junctions. Cluster one is mostly occupied by

- All of
  - *Talin* 29°C
  - *Betaint* 29°C
- Most of
  - *Rab8* 29°C DG
  - *Betaint* 18°C PDG
- Half of
  - *Talin* 18°C DG
  - *Betaint* 18°C DG

Figure 3.17 shows some examples of networks belonging to the *small* phenotype.



(a) M = *Talin*; t = 29°C; B_{Gen} = G

(b) M = *Talin*; t = 29°C; B_{Gen} = DG

(c) M = *Talin*; t = 29°C; B_{Gen} = PDG

(d) M = *Betaint*; t = 29°C; B_{Gen} = G

(e) M = *Betaint*; t = 29°C; B_{Gen} = G

(f) M = *Betaint*; t = 29°C; B_{Gen} = PDG

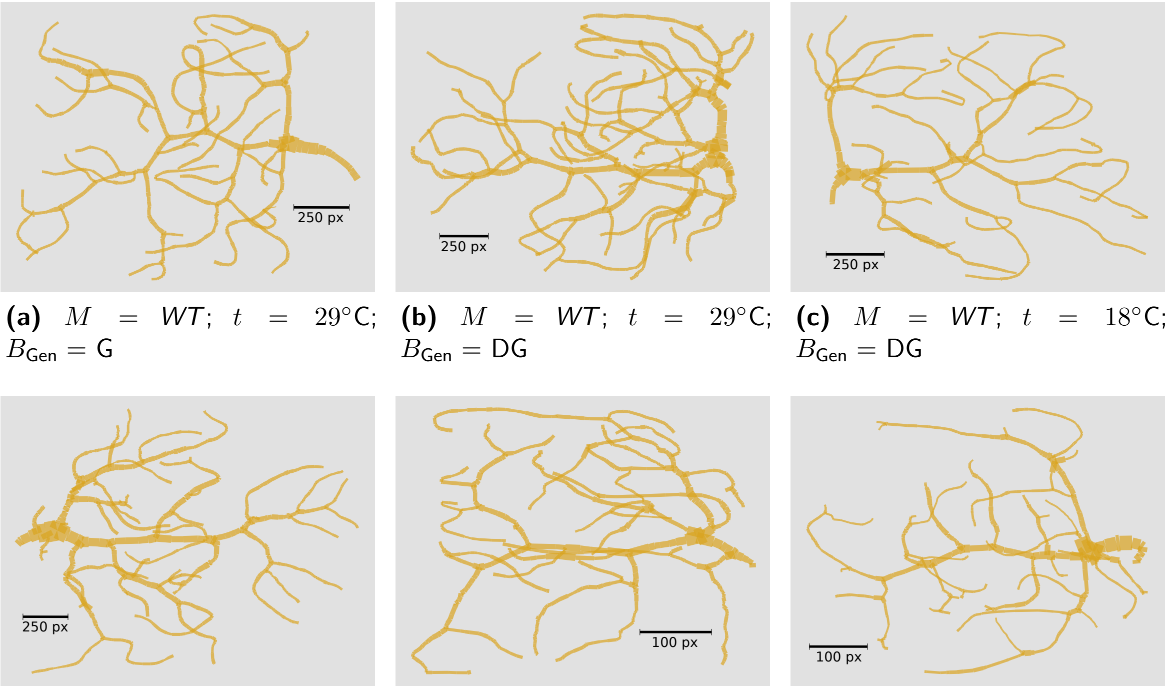
Figure 3.17. Networks from the *small* phenotype belonging to *Talin* 29° and *Betaint* 29°.

Cluster 2 or the “*large*” Phenotype

Networks from the *large* phenotype are the largest networks present in the dataset. The second cluster is composed of networks from

- All of
  - *WT* 29°C G, DG
  - *WT* 18°C G, DG
- Half of
  - *Crumbs* 18°C G, DG
  - *Crumbs* 29°C G
  - *Rab8* 18°C G
  - *WT* 29°C PDG
- Some of
  - *WT* 18°C PDG
  - *Rab8* 29°C G

Figure 3.18 shows some examples of networks belonging to the *large* phenotype.



(a) $M = WT$; $t = 29°C$; $B_{\text{Gen}} = \text{G}$ **(b)** $M = WT$; $t = 29°C$; $B_{\text{Gen}} = \text{DG}$ **(c)** $M = WT$; $t = 18°C$; $B_{\text{Gen}} = \text{DG}$

(d) $M = WT$; $t = 18°C$; $B_{\text{Gen}} = \text{G}$ **(e)** $M = Crumbs$; $t = 18°C$; $B_{\text{Gen}} = \text{G}$ **(f)** $M = Crumbs$; $t = 29°C$; $B_{\text{Gen}} = \text{G}$

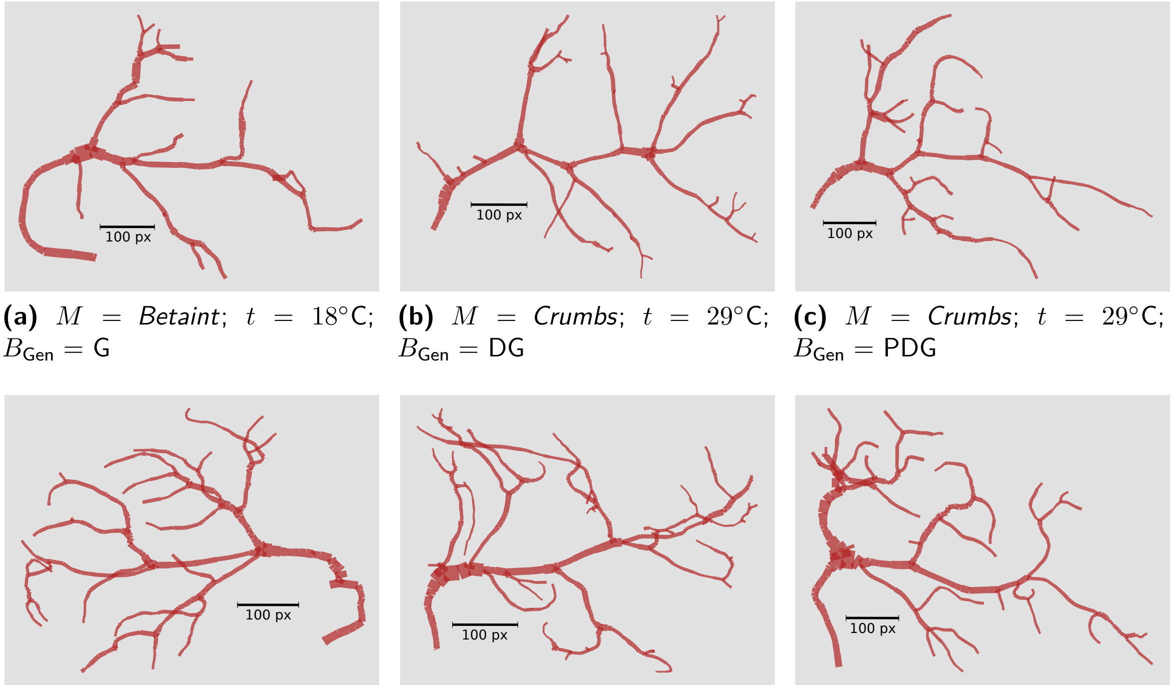
Figure 3.18. Networks from the *large* phenotype mostly belonging to *WT* and *Crumbs*.

Cluster 3 or the "*medium*" Phenotype

The third cluster or *medium* phenotype is the cluster containing the most networks. It spans the widest range on the first PCA component and contains networks with intermediate sizes from

- All of
  - *Betaint* 18°C G
  - *Crumbs* 29°C DG, PDG
  - *Rab8* 18°C DG
  - *Rab8* 29°C PDG
  - *Talin* 18°C PDG
- Most of
  - *Crumbs* 18°C PDG
  - *Rab8* 29°C G
  - *Talin* 18°C G
  - *WT* 18°C PDG
- Half of
  - *Crumbs* 18°C G, DG
  - *Crumbs* 29°C G
  - *Rab8* 18°C G
  - *WT* 29°C PDG

Figure 3.19 shows some examples of networks belonging to the *medium* phenotype.



(a) M = *Betaint*; t = 18°C; B_{Gen} = G

(b) M = *Crumbs*; t = 29°C; B_{Gen} = DG

(c) M = *Crumbs*; t = 29°C; B_{Gen} = PDG

(d) M = *Rab8*; t = 18°C; B_{Gen} = DG

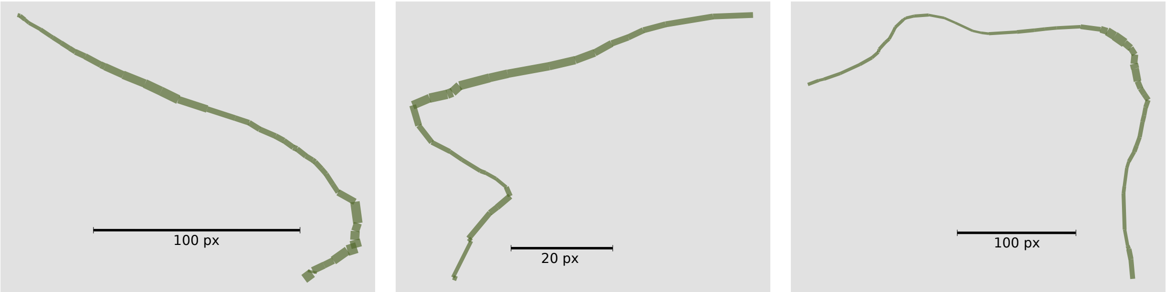
(e) M = *Rab8*; t = 29°C; B_{Gen} = PDG

(f) M = *Talin*; t = 18°C; B_{Gen} = PDG

Figure 3.19. Networks from the *medium* phenotype mostly belonging to *Betaint*, *Rab8*, *Crumbs* and *Talin*.

Cluster 4 or the "*degenerate*" Phenotype

The fourth cluster represents the *degenerate* phenotype. It consists of networks that do not branch. The only family that expresses such networks is *Betaint* 18°C DG. Figure 3.20 shows some examples of networks belonging to the *degenerate* phenotype.



(a) M = *Betaint*; t = 18°C; $B_{\mathsf{Gen}} = \mathsf{DG}$ **(b)** M = *Betaint*; t = 18°C; $B_{\mathsf{Gen}} = \mathsf{DG}$ **(c)** M = *Betaint*; t = 18°C; $B_{\mathsf{Gen}} = \mathsf{DG}$

Figure 3.20. Networks from the *degenerate* phenotype all belonging to *Betaint* 18°C *Dicer GFP*.

3.5.4. Fake Mutants

To assess how large the influence of network size on the clustering is, we perform an experiment: We want to know if newly created networks controlled only for their size, so called "fake mutants", will cluster in the same way as the real mutants do. For this purpose, we measure the *NoJ* for *Betaint*, *Crumbs*, *Rab8* and *Talin* and average over each mutant. Then we create fake mutants by removing junctions farthest away from the root from the *Wild Type* networks until these networks reach the target *NoJ* of the respective mutant. We then perform the same clustering as described in section 3.5.1 on all the real and the fake mutants. The result can be seen in figure 3.21: as expected, if the clustering is only dependent on network size, the fake mutants cluster in the same way as the real mutants do. The only qualitative differences can be seen in the clustering of the fake *Betaint* 18°C, the fake *Rab8* 29°C and the fake *Talin* 29°C. In all three cases, the difference can be explained by the averaging process involved in calculating *NoJ*: as the target *NoJ* was calculated by averaging over all families with a certain mutation, there is no possibility of creating for example degenerate networks, as the average *NoJ* for all networks from *Betaint* 18°C will always be larger than one.

3.5.5. Conclusion

The networks do not cluster in the phenotypes induced by their mutations, instead they form new - formerly hidden - phenotypes. The growing characteristics of these new phenotypes seem to be heavily influenced by the network size. The networks cannot be separated into clearly discernible clusters but form a continuum on the dimension with the highest variance of the high-dimensional space spanned by the PCA components. Temperature and genetic background have an influence on the position of the networks in this continuum: Being grown at a different temperature can cause a network to be affiliated with networks with a completely different mutation. The effect of temperature on network growth is larger than that of the genetic background with the notable exception of *Betaint*.

We assumed that the clustering is solely dependent on network size which can best be represented by a network's *NoJ*. This assumption was verified by an experiment in which we created fake mutants by removing elements from *Wild Type* networks. The fake mutants which were created by controlling solely for their *NoJ* clustered qualitatively in the same way as their real counterparts. The only differences were explainable by the averaging involved in creating the fake mutants.

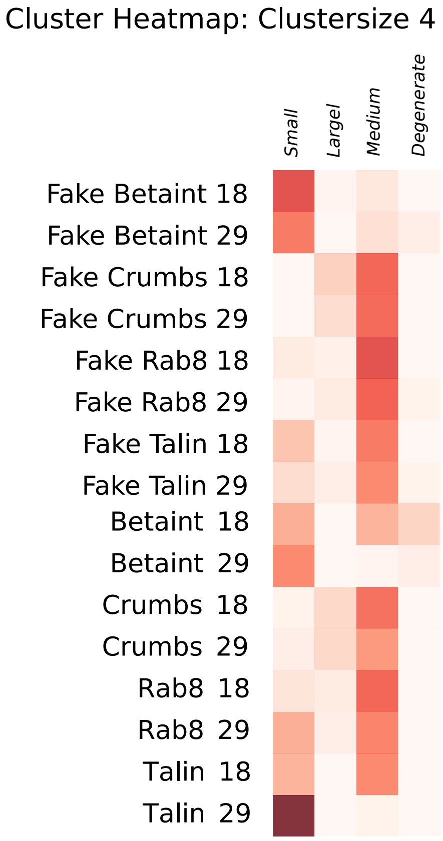


Figure 3.21. Clustering performed on the "fake mutants" created by removing elements from the *WT* networks and real mutants.

3.6. Phenotypic Trends

In the previous sections we established that most of the information contained in the networks of trachea terminal cells is related to network size. In the following section we want to investigate whether there are significant second order signals in the network's growth patterns independent of network size.

3.6.1. Size Dependence

To get rid of the size dependence of the different metrics, we try to find the dependence of the metrics on network size. After this dependence is known, we can look for significant deviations from the predicted trend and define “phenotypic trends”. If a certain family or cluster follows a phenotypic trend, its networks tend to deviate from the predicted growth pattern in a way indicated by the trend.
We have already established that genetic background and temperature do have an impact on how large a network will grow. While looking for deviations from the general trend we can also gather information about trends in the influence of t and B_{Gen} on NoJ.

3.6.2. Significant Deviations

Analysis Approach

Figure 3.22 shows a characteristic example of the described analysis process: first the dependence of LoN on NoJ has to be found. Looking at the dependence of single edge lengths on the tree-depth of the edge (shown in figure B.1) shows that LoE does not significantly change with depth. Therefore the presence of more junctions should linearly increase the size of the network. We can now perform linear regression on the data points and get an indication of the 95% confidence interval around this fit. As we are only interested in the identification of *trends*, identification of deviations

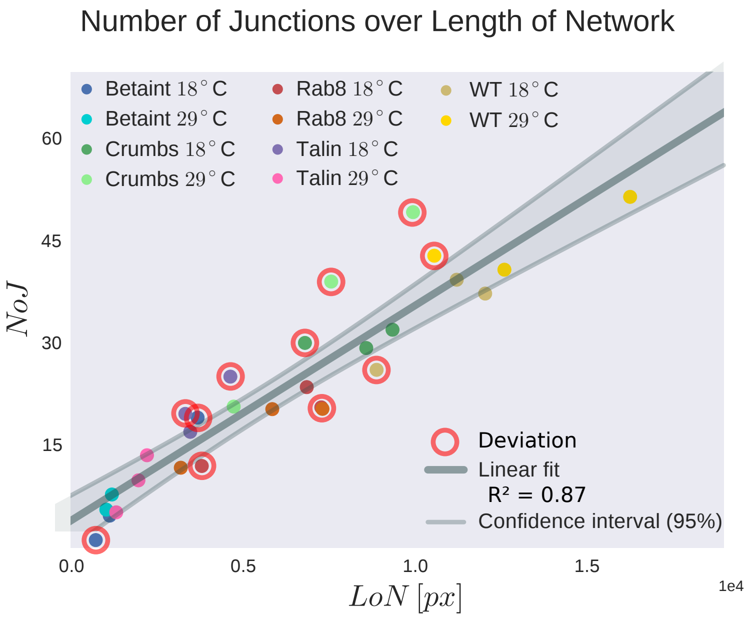


Figure 3.22. Dependence of LoN on NoJ. The dependence is clearly linear. The gray line and area represent a linear fit to the data points and a 95% confidence interval. Data points which lie outside the confidence interval have been marked with red circles. Networks grown at different temperatures have been indicated with lighter and darker shades of the respective color. The three data points per color correspond to the three genetic backgrounds.

as data points lying outside the confidence interval by eye was sufficient. By looking at significant deviations from the size-dependence we can identify the data points which lie significantly *above* the linear fit as families which tend to grow *shorter* than average. Data points which lie *below* the fit correspond to networks that grow *longer* than average.

Dependence of NoJ on t and B_{Gen}

Additionally, the data points in 3.22 have been colored to depict the temperature at which the networks were grown. With this additional information we can draw conclusions about the dependence of network growth on the temperature: The *Wild Type* shows a quite clear dependence on temperature - all networks grown at 18°C have less junctions than the ones grown at 29°C. *Talin* shows an inverse dependence of NoJ on temperature whereas *Betaint*, *Rab8* and *Crumbs* do not show a clear trend for different temperatures.

We also performed this analysis for the genetic backgrounds but there were no significant trends visible.

3.6.3. Four Phenotypic Trends as Deviations from basic growth Characteristics

We can define four basic geometric properties of network growth: edge length and radius, angle between edges and how straight edges grow. We can also find metrics which we think best represent these geometric properties: LoE, RoE, $AoCH$ and D_{T_L}.

Based on the deviations of these representative metrics from their dependence on NoJ, we are able to define four qualitative phenotypic trends. The geometric realizations of these trends are illustrated in figure 3.23.

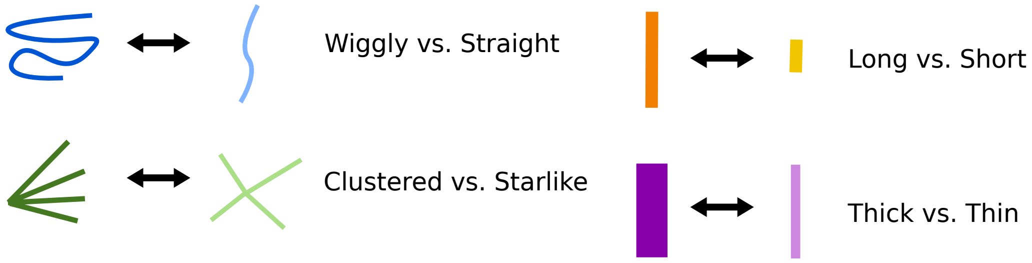


Figure 3.23. Phenotypic trends as a result of deviations of the four basic geometric characteristics from their dependence on the number of junctions.

The observed dependencies and their phenotypic trends are

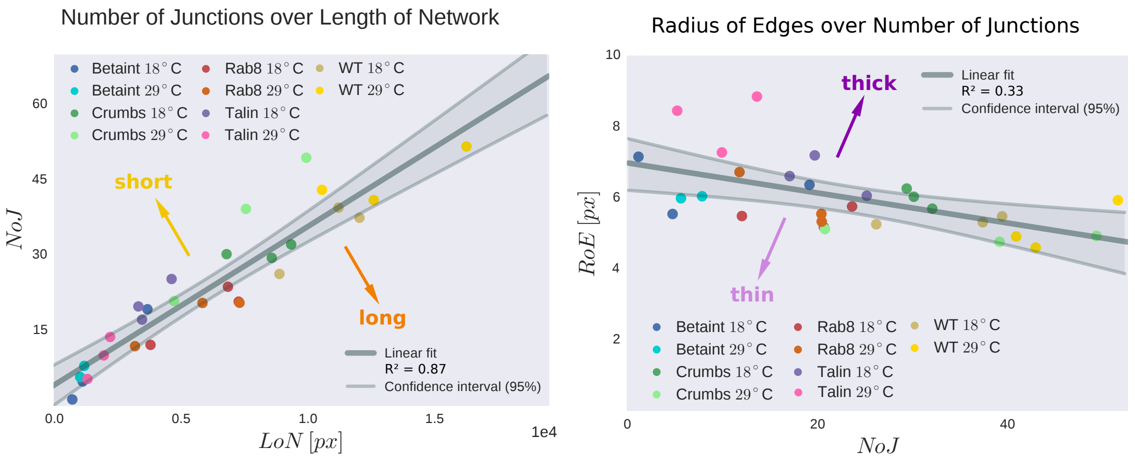
- Edge length: LoE seems to be independent of tree-depth. Deviations from the dependence of LoE or LoN on NoJ indicate a trend on the *short* ↔ *long* axis (figure 3.24a).

- Edge radius: RoE seems to be linearly dependent on NoJ. Deviations from the dependence of RoE indicate a trend on the *thin* ↔ *thick* axis (figure 3.24b).

- Straightness: D_{T_L} linearly depends on NoJ. A deviation from this dependence indicates a trend on the *straight* ↔ *wiggly* axis (figure 3.25a).

- Angle at junctions: $AoCH$ also depends linearly on NoJ. Deviations from this dependence indicate a trend on the *starlike* ↔ *clustered* axis (figure 3.25b).

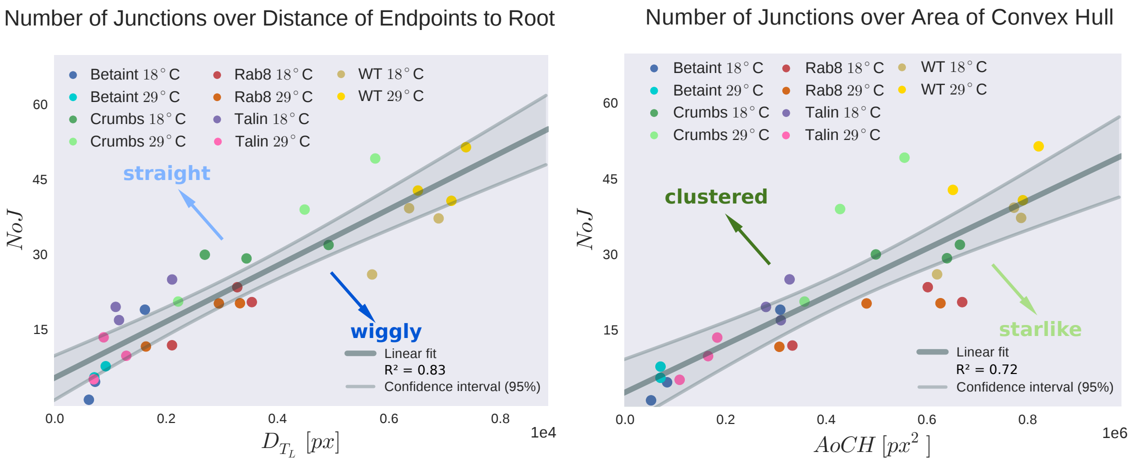
An overview of the dependencies of each of the representatives on NoJ as well as significant deviations from this dependency is given in figures 3.24 and 3.25.



(a) Dependence of LoN on NoJ is linear. Deviations indicate *longer* or *shorter* growth.

(b) Dependence of RoE on NoJ is linear. Deviations indicate *thicker* or *thinner* growth.

Figure 3.24. Dependence of representatives of the two basic geometric properties on NoJ: **(a)** LoN represents *edge length*. **(b)** RoE represents *edge radius*.



(a) Dependence of D_{T_L} on NoJ is linear. Deviations indicate *wigglier* or *straighter* growth.

(b) Dependence of $AoCH$ on NoJ is linear. Deviations indicate more *starlike* or *clustered* growth.

Figure 3.25. Dependence of representatives of the two basic geometric properties on NoJ: **(a)** D_{T_L} represents *straightness*. **(b)** $AoCH$ represents *compactness*

3.6.4. Trends for the different Mutants

For each of the four mutants and the *Wild Type* we gathered all significant deviations and compiled them in table 3.1.

Mutant	t	NoJ dep. on t	LoN	RoE	$AoCH$	D_{G_L}
Betaint	18°C	-	-	-	-	-
Betaint	29°C	-	-	thin	-	-
Crumbs	18°C	-	-	-	-	-
Crumbs	29°C	-	short	-	clustered	straight
Rab8	18°C	-	long	thin	starlike	wiggly
Rab8	29°C	-	long	-	starlike	wiggly
Talin	18°C	negative	short	thick	-	straight
Talin	29°C	negative	-	thick	-	-
Wild Type	18°C	positive	-	-	-	wiggly
Wild Type	29°C	positive	-	-	clustered	-

Table 3.1. Compilation of temperature dependence and phenotypic trends for all four mutants and the *Wild Type*.

A deviation was recorded if more than one family from a (M, B_{Gen}) combination would significantly deviate from the trend in one direction and if no family from the combination deviated in the other direction. We split the networks into groups of

(M, B_{Gen}) as different mutation and growth temperature had the highest impact on the expression of different trends.
The table aims to form a picture of growth trends but it is hard to interpret in a biological context as the underlying mechanisms influencing characteristics like wiggly growth are unknown.

3.6.5. Conclusion

From the two-dimensional geometry of the networks we were able to infer four basic growth characteristics: edge length and radius, angle between edges and how straight an edge grows. These basic characteristics were used to define four "phenotypic trends" in network growth: networks can grow longer/shorter, thinner/thicker, more spread out/clustered and straighter/wigglier. We found a metric representing each of these characteristics and the dependence of these metrics on network size. We identified significant deviations from the general dependency and gathered these deviations to form growth-tendencies for each of the families. A discrimination between networks grown at different temperatures was necessary because these networks often exhibited different or even contrary trends. Genetic background, however did not seem to have a major impact on these trends.

3.7. Growth Models

To better understand how certain mutations affect network growth, it is desirable to have a model that describes the growth process. This model should cover all the aspects that characterize a network realization:

- Branching behaviour.
- The distribution of edge radii over the network.
- The growth direction of an edge.

We assume that knocking down certain genes in the larvae does not change the underlying growth model but only changes the parameters which scale its behaviour. In the following, we try to model each of the above-mentioned aspects and test the assumption of the unchanged growth model against the data. This works quite well for the branching behaviour and radius distributions but for the growth direction the modeling is more complicated.

In the following analysis we always asses the significance of our findings using the Lilliefors test for normality [34]. This test generalizes the Kolmogorow-Smirnov test for distributions with theoretical mean and variance unknown. To get an indication of the significance of our findings, we first have to compute the maximum difference D between the sample cumulative distribution function (cdf) and the theoretical cdf. We calculate the theoretical cdf using the mean and variance of the data-sample. We can then translate D into p-values using tables provided for example by [35] which link critical values of D, sample size N and a significance level p. We assume that we can ignore the impact of t and B_{Gen} and look at samples composed of all networks with a certain mutation. Therefore we have around 100 networks for each mutant. If we were to split the dataset into 30 families, the number of networks in each family would be too small and the results too noisy to fit a distribution. For all our findings we encounter values of $D \propto 10^{-2}$. For $N = 100$ this corresponds to a significance level of $p < 0.01$.

3.7.1. Branching Behaviour

Assumptions

To describe the branching behaviour of terminal cells we assumed that:

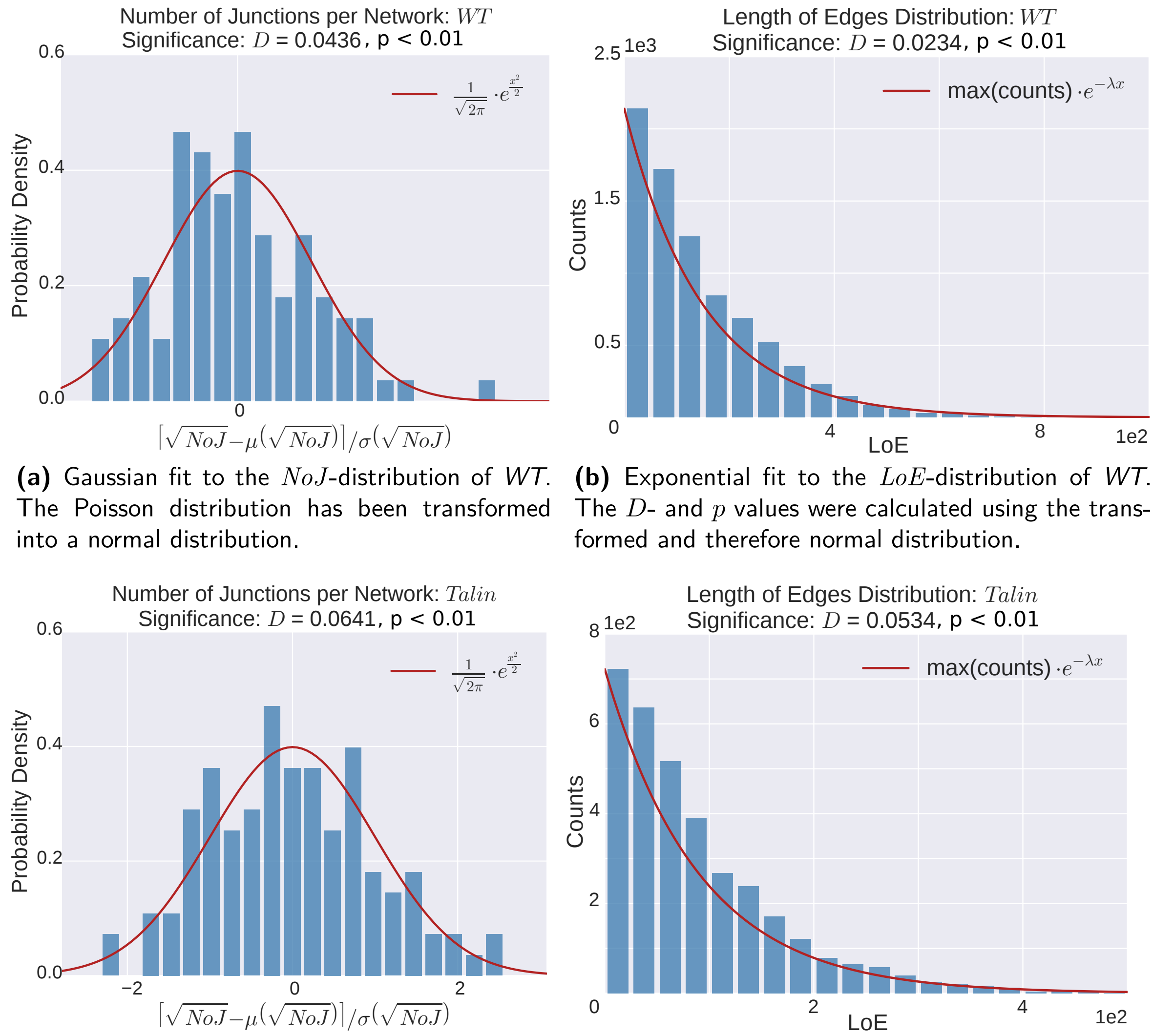
- The growth speed of edges is constant.
- Branching probability is independent of edge radius and tree depth.
- The length of a branch is independent of tree depth (this was already indicated by the data presented in figure B.1).
- Branch length is also independent of the length of preceding branches.
- Branching events are rare events and therefore can be modeled as *Bernoulli Process.*

Predictions

If the above assumptions are true and branching indeed is a Bernoulli process, the number of junctions should follow a Poisson distribution. The length of the edges in-between the junctions should therefore follow an exponential distribution.

Comparison with Data

Figure 3.26 shows two examples of distributions of NoJ and LoE. As examples we took the networks from *Wild Type* and *Talin* which illustrates that the distributions do not qualitatively change regardless of mutation or network size. The NoJ and LoE in both cases follow a normal distribution after a variance stabilizing transformation from a Poisson and exponential distribution respectively.



(a) Gaussian fit to the NoJ-distribution of *WT*. The Poisson distribution has been transformed into a normal distribution.

(b) Exponential fit to the LoE-distribution of *WT*. The D- and p values were calculated using the transformed and therefore normal distribution.

(c) Gaussian fit to the NoJ-distribution of *Talin*. The Poisson distribution has been transformed into a normal distribution.

(d) Gaussian fit to the LoE-distribution of *Talin*. The D- and p values were calculated using the transformed and therefore normal distribution.

Figure 3.26. **(a)** and **(b)** distribution of the NoJ and LoE of all networks from *Wild Type*. **(c)** and **(d)** distribution of NoJ and LoE of all networks from *Talin*. Significance levels were calculated using the Lilliefors test for normality on the distributions after a variance stabilizing transformation.

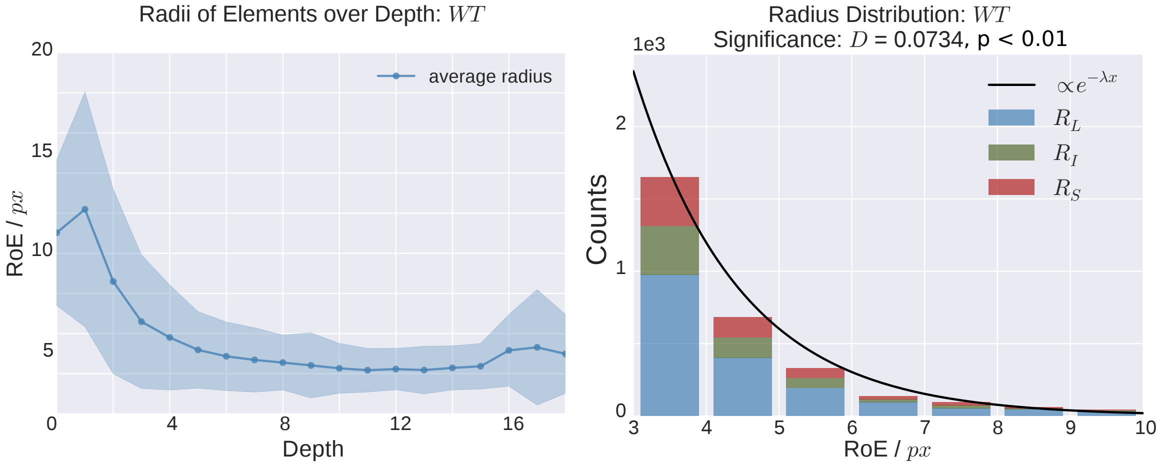
3.7.2. Edge Radii

Assumptions

Finding an adequate growth model for the radii at first seems to be more difficult than for the branching process. Radii are not independent of tree depth and radius of preceding branches. Nevertheless, we can assume the branch radius along an edge to be constant. Modeling the distribution of edge radii can be reduced to finding a model of how edge radii behave at junctions. Given the radius of the mother branch we have to find out how the radii of the two daughter branches behave.

Findings

At each junction we can discriminate between the radius of the largest, intermediate and smallest edge connecting at the junction (R_L, R_I, R_S). Figure 3.27a shows that the radii decay exponentially with depth. Figure 3.27b shows the distribution of all radii in a stacked histogram colored by the proportions of R_L, R_I and R_S for each bin. The figure shows an exponential decay of the number of edges with increasing radius. Bearing in mind that radius exponentially decays with depth this distribution can be explained by the exponentially increasing number of edges with depth. Therefore, edges are thickest close to the root and the radius thereafter decays exponentially with tree-depth.



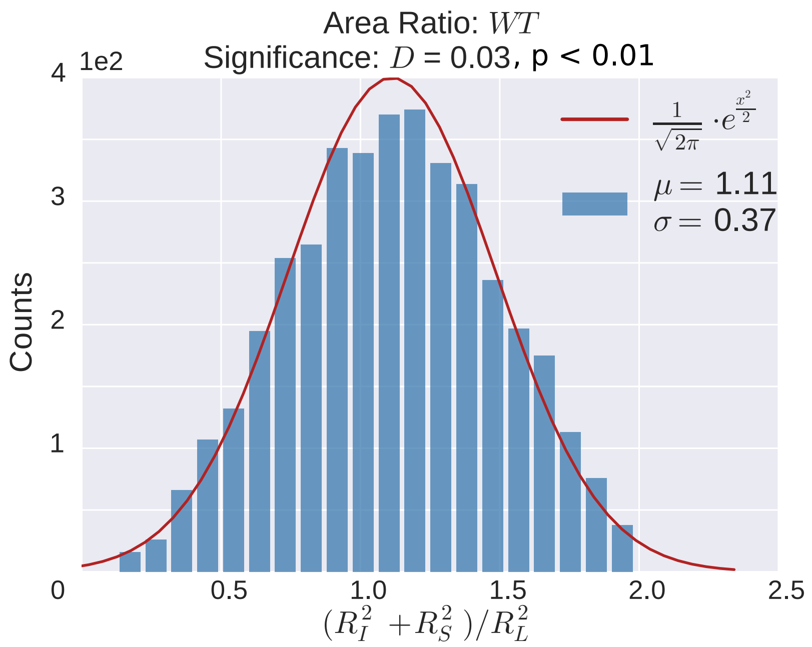
(a) Development of average radius over tree depth. **(b)** Exponential fit to *WT* radius distribution.

Figure 3.27. **(a)** Radii are smaller the deeper in the tree the edge is located - the decay seems to be exponential. **(b)** Histogram showing the radius distribution split into R_L (blue), R_I (green) and R_I (red). Radii are distributed exponentially: more edges have a smaller radius.

A very simple model for the behaviour of the radius at junctions is the assumption that the total intersection area of the edges is conserved, therefore $R_L^2 = R_I^2 + R_S^2$. To verify this assumption, in figure 3.30 we have plotted the distribution of the area ratios $AR = (R_I^2 + R_S^2)/R_L^2$.

We can see that for the *Wild Type* the area ratios follow a normal distribution around $\mu = 1$. The behaviour is the same for the mutants. Averaged over all families, the area ratio $AR = 1.02 \pm 0.07$ is very close to one which means that *da Vinci's rule* [36] holds.

This is surprising as we would intuitively expect the radius of the branches to further grow for some time after branching has happened. This expectation comes from the intuition that the network as a whole is still growing and all tube radii are still getting larger. This would shift area ratios to values smaller than one. The finding indicates that the radius of a branch is dependent on the radius of the mother branch at branching time and then conserved. Nevertheless there is a slight trend visible indicating that larger networks have higher mean area ratios than smaller networks (as can be seen in table B.1 which lists all AR values independently).



(c) Area ratio distribution for *WT*.

Figure 3.30. **(c)** Area ratios at junctions follow a normal distribution with mean $\mu \approx 1$. The depicted distribution shows the AR for the *Wild Type* but other mutants behave the same with a slight trend towards higher average AR for larger networks.

3.7.3. Angles

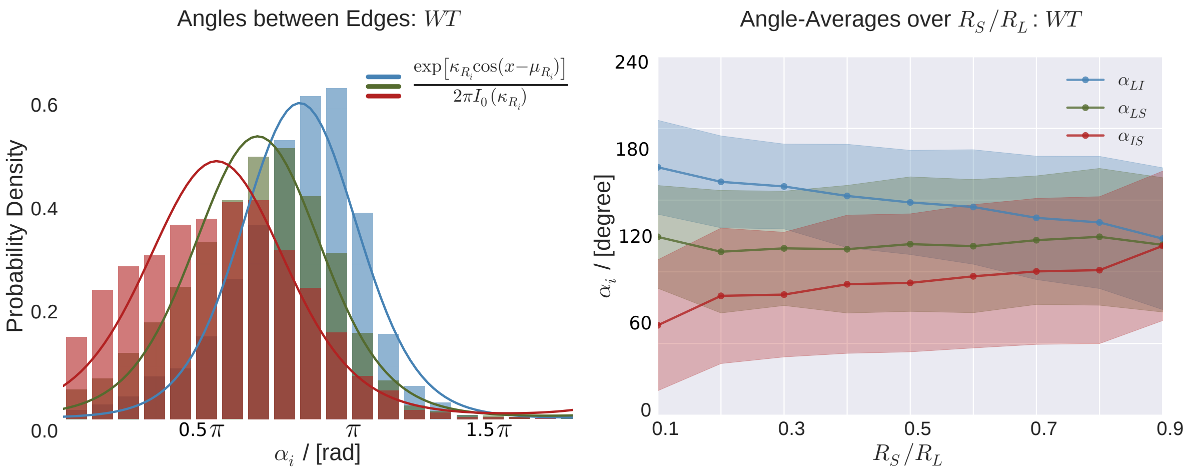
Assumptions

Now that we can model when an edge will sprout a new branch and what radii the edges will have after branching, the only thing missing to completely model the whole network growth is the growing direction of branches. The model can be divided into two sub-models describing the growing direction directly at the branching point and the growth farther away from junctions.

In the following, we will present some findings related to the growing direction directly at the branching point. We were not able to completely model how branches grow farther away from junctions. We suspect that growth farther away from junctions is heavily influenced by the surrounding tissue. The cells grow on muscle tissue which is inhomogeneous and structured. To model the growth behaviour within this tissue, we do not have enough information about the character of this tissue.

Angles at Junctions

To analyze the angles between branches, we first distinguish between three different angles (as described in section 3.2.4): The angle between the thickest and intermediate branch α_{LI}, the angle between the largest and the smallest branch α_{LS} and the angle between the intermediate and the smallest branch α_{IS}.



(a) Angle distribution split for α_{LI} (blue), α_{LS} (green) and α_{IS} (red).

(b) Average angles α_{LI} (blue), α_{LS} (green) and α_{IS} (red) over the radius ratio R_S/R_L.

Figure 3.31. **(a)** The angles α_{LI}, α_{LS} and α_{IS} follow a von Mises Distribution. **(b)** Angles converge to 120° the more equal the radii of the three edges at a junction are.

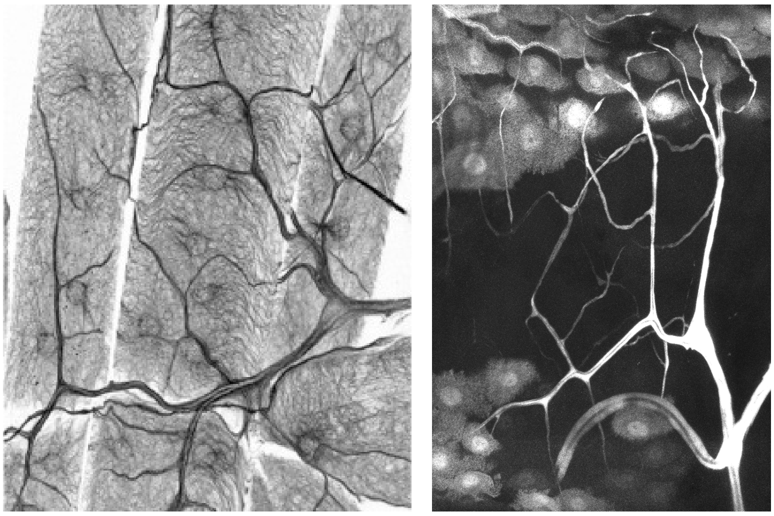
Figure 3.31a shows that each of the three angles follows a von Mises Distribution [37]. In figure 3.31b we can see that the more equal the radii of the branches at a junction are, the closer the angles between the branches get to 120°. This indicates a repelling force between branches at a junction.

The question remains whether this force is purely due to the mechanics of branch growth or if it has a chemical aspect such as branches growing in the direction of lowest oxygen saturation. Both could create a repelling force dependent on the radius ratios between branches.

Redundancy

Understanding the behaviour of network growth with regard to an edge's growth direction is quite difficult given our limited knowledge of the tissue the networks grow in. Figure 3.34a shows the muscle tissue surrounding the terminal cell. It is structured and shows regions of increased density and preferred directions of growth. We have no means of assessing the influence of this tissue on network growth.

Figure 3.34b gives an indication of the cells which are supplied with oxygen by the trachea. These images illustrate the *function* of the networks: every cell in the organism that needs oxygen has to be in diffusion range of at least one network branch. Using the *redundancy* defined in section 3.2 we can at least estimate whether the networks express something like a preferred distance between branches - a tube frequency in space. Indeed, the *additional number of redundant pixels* for each di-



(a) Image courtesy of S. Sigmundbjörnsdottir, EMBL Heidelberg

(b) Image taken from the dataset.

Figure 3.34. Microscopy images of the tissue surrounding the trachea terminal cells: **(a)** shows that the networks grow in structured muscle tissue. Several cell nuclei can be identified as darker circular shapes. **(b)** also shows other cells in the neighborhood of the terminal cell which are supplied with oxygen by the network.

lation step always shows a global minimum around 200 pixels which corresponds to $\approx 70\ \mu$m.

We can now solve the diffusion equation for a spherical tube and a target concentration $c(r) = 0.1$ mol/m^2 which is the oxygen saturation found in the guts of herbivorous insects [38]. This yields a diffusion range of

$$r = \frac{c_0 r_0}{c(r)} = 100\ \mu\text{m} \mathrel{\hat{=}} 300\ \text{px}\ ,$$

with a starting concentration of $c_0 = 44$ mol/m^3 which corresponds to the concentration of oxygen in normal air and a tube radius $r_0 = 3\ \mu$m. We neglected any effects related to membrane permeability and assumed the diffusion coefficient of the tissue to be sufficiently close to that of water. We find it quite noticeable that our predicted diffusion range so closely matches the observed redundancy minimum even with this very rough calculation. This leads us to speculate that the trachea networks try to optimize for lowest possible redundancy: At a diffusion range close to a range we assumed to be the maximum range a spherical tube can supply surrounding tissue with oxygen the networks express a redundancy minimum.

3.7.4. Conclusion

We were able to find growth models which quite successfully describe the branching behaviour and radius distribution described by trachea terminal cells. We also found that with regard to branch growth direction directly at the junction there is a repelling force between branches dependent on their radius ratios. To accurately describe branch growth farther away from a junction we would need more information about the surrounding tissue as we cannot assume it to be homogeneous.

We can get an indication of the large-scale growth pattern by looking at network redundancy: the networks seem to optimize their growth with regard to lowest possible redundancy, therefore expressing a preferred distance between branches. This distance is close to the maximum range a branch can supply with oxygen. We also found that all of the networks follow qualitatively similar distributions, therefore mutation does not seem to change the underlying growth model.

3.8. Summary

This chapter was dedicated to describing the approaches we followed to analyze the dataset and the results of this analysis. As during the analysis we did not know “what we were looking for”, we choose a very broad approach by defining and measuring a large set of metrics.

Metrics and Dataset

In section 3.2 we introduced the definitions of these metrics and found that the networks contained in our dataset are directed, binary trees. The metrics were divided into topological and geometric metrics and a measure for the network’s redundancy. Our dataset is composed of 500 networks which are divided into 30 families with regard to their mutation, genetic background and growing temperature.

Assessment of known Phenotypes

The first and foremost question we wanted to answer is whether there are “hidden phenotypes” besides the ones induced by each network’s genotype. To answer this question, in section 3.3 we first looked at the already known phenotypes and assessed the discriminative power of mutation, genetic background and growing temperature. The goal was to investigate which of the already known growing parameters had the capacity to define a distinguishable phenotype for network growth. We found that a network’s mutation influences its growth so drastically that it is possible to discriminate between networks with different mutations using a qualitative assessment of the metrics measured on them. This is not the case for temperature and genetic background.

We were further able to quantify this result by performing linear discriminant analysis, a form of supervised learning, on the dataset. The process is described in section 3.4. The results also show that we can discern between networks based on their mutation with an accuracy of 70 ± 8 %. If we only use temperature or genetic background, discernability is not better than chance.

Hidden Phenotypes

We have established that a network’s mutation induces a well-distinguished phenotype for the growth pattern. Using an unsupervised learning approach, in section 3.5 we applied clustering to see whether the networks express clustering behaviour different from the already known phenotypes. We found that the networks indeed form clusters beyond their mutations. We also saw that these clusters are not separated clearly but rather form a continuum in metric space. Furthermore we were able to identify the meaning of the direction with the highest variance in metric space as network size. This means that networks with different mutations can still have the same size and therefore be affiliated with the same cluster. We also found that all of the metrics measured on the networks are more or less dependent on the size of the network.

Phenotypic Trends

We suspected that there is weaker second order information describing *how* a network grows. To be able to assess this information, we first had to remove the size dependence of all metrics. In section 3.6 we established that there are three main independent geometric characteristics of network growth: edge length, edge radius, growth direction and a measure of how "wiggly" the network grows. The best representative for network size is its number of junctions. To each characteristic we assigned one of the metrics as best representative. We then found the dependence of each of these metrics on the number of junctions and looked for significant deviations from this dependence. These significant deviations are the "phenotypic trends" a family of networks exhibits.

Growth Models

In the last part of this chapter, in section 3.7, we tried to find growth models explaining the dependencies and distributions of the metrics we measured. We divided the global growth model into three sub-models describing branching behaviour and edge length, radius distribution and growth direction. The branching process was assumed to be a Bernoulli process and the data supports this assumption.

To model the radius distribution within the network, we looked at the behaviour of radii at branching points. We found that the intersection area on average is conserved at junctions and that edge radius decays exponentially with tree-depth.

The model for growth direction can be divided into two regimes: behaviour of branch growth close to a junction, and farther away from it. We found that close to the junction we see a repelling force between branches depending on their radius ratios. To describe the growth behaviour away from branching points we would need more information about the tissue the branch grows in. Using the previously defined measure for redundancy we got an idea of the general large-scale growth pattern: the terminal cell tries to minimize overlap between diffusion ranges of its branches.

4. Biological Context

4.1. Introduction

The respiratory system of *Drosophila* flies has long been a suitable paradigm for the study of network-like structures involved in the transport of oxygen. In contrast to the much more complex structures found in mammals, the trachea of *Drosophila* are a much simpler system. For the flies, the cellular mechanisms and genetic programs which guide the development of the trachea are understood, at least in part. Many of these mechanisms also can be found in mammals. Therefore, the formation of the *Drosophila* trachea over the last decades has turned out to be a rewarding research target for the understanding of branching morphogenesis. We hope the answers provided by this research will give deeper insight in the development of network structures in organisms of higher complexity.

In the following we will first give a short introduction into trachea morphogenesis. Before, we have used mutation, genetic background and growing temperature as variables without any expectations or meaning attached to them. Now we will explain the meaning of these growing conditions in a biological context.

In the second part we will put the results we obtained in the previous chapter into a biological perspective: how, from a purely biological point of view, would we expect the terminal cells to behave for different growing conditions?

4.2. Biology of *Drosophila* Trachea

The *Drosophila* trachea are a network of interconnected epithelial tubes which can be divided into a primary, secondary and terminal part [39].

In this work, we are only interested in the growth patterns of trachea terminal cells and the impact certain mutations have on their growth. In the following, we give a short overview over the trachea morphogenesis. We also describe the mechanism causing mutations in the genetic expression of the larvae. Finally we look at the genetic backgrounds and growing temperatures and their impact on network growth.

4.2.1. Trachea Morphogenesis

Tracheal tubes consist of a simple epithelial monolayer wrapped around a central lumen which contains the gas transported throughout the system. At the spiracular openings on the insect's surface, oxygen enters the network and then diffuses through the network until it reaches its target tissue. The trachea network originates from ≈10 cell clusters in early developmental stages of the larva. Traces of these clusters can later be found in a segmentation of the resulting network along the length of the larva. The trachea display a bilateral symmetry and every terminal cell itself also has a left and a right part.

Tracheal branches are formed by a sequence of branching events forming successively thinner branches. The primary and secondary branching stages express branches with rather well defined growth structures and positions. However, the growth patterns of the terminal branches are believed to be completely random except for the fact that they always grow two well-distinguishable parts. The primary and secondary branches of the trachea consist of around 80 cells [40] whereas the terminal branches are formed by the outgrowths of only one single cell.

As already mentioned, there exist several ways for lumen formation within primary and secondary trachea branches [2], [3], [4] which generate lumen by wrapping around or changing an already existing external or internal space. The lumen formation in terminal cells however is believed to be *de novo* [5] which is a process were a tube is created *without* an external space. During growth of the larva, the terminal cells undergo a hundredfold increase in size. This and the *de novo* lumen formation indicate that there has to be a very efficient mechanism in place to provide the cell with the material to create new tubes. During its growth, the terminal cell is believed to respond to hypoxic signals from target tissue [41] by sprouting new branches to deploy oxygen. It is also known that the terminal cells do not only grow at the tips of the branches but along the whole length of the tube. Most of the mutations we used in this work affect the vesicular transport mechanisms of molecules to the growing terminal cells. This results in different lumen morphology - tangled or faulty lumen (which was not analyzed in this work) and different branching patterns. The latter was subject to extensive analysis in the previous chapter.

4.2.2. Mutants

To investigate the effect of several genes known to play a functional role in trachea terminal branching, mutants were created where these genes were knocked down. This inactivation of genes was obtained via RNA mediated interference (RNAi). The selected genes were

- *Rab8* which regulates diverse aspects of vesicular transport in the cells including: the transport of cargoes from the trans-Golgi network to the apical or basolateral plasma membrane, regulation of recycling endosomes and regulation of apical protein localization [42].
- *Myospheroid* which encodes βPS-integrin (in the analysis called *Betaint*) which plays a role in macrophage migration [43].
- *Crumbs* which plays a role as an apical determinant required for tube formation [44].
- *Rhea* which encodes *Talin* which plays a role in the ability of the network to attach itself to surrounding tissue [45].

It is important to mention that the larvae expressing a lack of the above-mentioned genes were not *real* mutants missing the gene completely. They rather were organisms in which the expression of these genes was diminished. Therefore, the larvae could still express some reduced level of the knocked down genes. The effectiveness of the knockdown is believed to depend on other growing conditions such as genetic background or temperature. Therefore the actual expression levels of the knocked down gene are unknown.
As a control group, for each growing condition there also was a population of *Wild Type* flies for which images of the trachea terminal cells were recorded.

4.2.3. Genetic Backgrounds and Temperatures

To facilitate imaging of the terminal cells, they were tagged with a fluorescent protein. The protein chosen for this task was the *Green Fluorescent Protein GFP* [46] which fluoresces green when exposed to light in the blue to ultraviolet spectrum. Expression of GFP interacts with the RNAi machinery responsible for the knockdown and can change its efficiency. To test the influence of GFP and two other proteins on network growth and knockdown efficiency, larvae expressing different combinations of them were grown. The selection of these proteins the larva expresses is called its *Genetic Background*. The proteins used are *Dicer2* [47], *PIP2* [48] and *GFP*, of which *Dicer2* is believed to increase knockdown efficiency. The genetic backgrounds are

- Dicer2 and GFP also called DG,
- Dicer2, PIP2 and GFP also called PDG,
- Only GFP also called G.

Another condition that can influence the efficiency of the gene knockdown is the temperature at which the larvae are grown. To create larvae expressing a range of knockdown-efficiencies, for each mutant, larvae from all three genetic backgrounds were grown at temperatures of 18° C and 29° C. This yields a total of 30 different growing conditions composed of combinations of mutation, genetic background and temperature.

4.3. Analysis Results in a biological Context

4.3.1. Mutation Effect on Network Growth

The analysis we performed in section 3.5 clearly shows that the by far strongest effect of mutations on network growth is the influence on network size by the reduction of branches. This can be explained by the functions the knocked down genes normally fulfill: *Betaint*, *Rab8* and *Crumbs* all play a crucial role in the transport of material for the formation of new tubes. If less material is provided, the tubes cannot form the extensive networks we see in the *Wild Type*. Without *Talin* the cell is no longer able to attach to the surrounding tissue. It might create several branches but during growth they tend to collapse into fewer but thicker branches.

If our dataset consisted of networks with different or fewer mutations, we might have been able to find network phenotypes clearly separated by their number of junctions. However, with the mutations our larvae expressed, the networks happened to all fall on a continuum with regard to their size.

In this continuum we see a clear trend of *Betaint* most drastically impeding network growth. This goes as far as the expression of networks with only a few or even only one branch. The larvae grown with *Betaint* DG at 29° C were not able to

survive, therefore we have no data from this family. This lethality can be seen as a continuation of the trend of drastic size reduction.
Talin also interferes with network growth quite strongly although not as severe as *Betaint*. *Rab8* has a clearly measurable impact on network growth. However, it does not seem to impede functions necessary for network growth in the same drastic way as *Betaint* and *Talin*.
Crumbs is the mutation which has the least impact on network branching. The networks with knocked down *Crumbs* are nearly as large as the *Wild Type*. Some of these networks are even indistinguishable from *Wild Type* networks.

4.3.2. Effect of M, B_{Gen} and t on the Mutation

We have learned that genetic background and temperature are expected to influence the efficiency of the RNA interference and therefore create small-scale variations in the expressed phenotypes. In sections 3.3 and 3.4 we were able to confirm this expectation. We were not able to discern networks only based on their genetic background and growing temperature.

Genetic Background

Out of B_{Gen} and t the genetic background seems to have the least impact on the efficiency of the gene knockdown. There is no clear general trend indicating which of the genetic backgrounds is most favorable to the gene knockdown.
However the variations between genetic backgrounds for the same mutation and temperature seem to be clearly visible for some cases. Nevertheless, it has to be kept in mind that single families of networks only contain 10-20 data points, therefore noise is very high. Also, it is not clear whether these variations are caused by the impact of the genetic background on the knockdown mechanism. The genetic background could also influence network growth on its own in other, yet unknown ways. The second statement is supported by the finding that we can also see varying network growth throughout *Wild Type* families.
To see whether significant variations in branching numbers were more common for mutants, we performed a two sided t-test for every two genetic backgrounds within a (M,t) combination. Table 4.1 shows the results of these tests. p-values larger than 0.1 are highlighted in red to show where distributions are similar and therefore

variation is not significant. It is noticeable that for the *Wild Type* and *Rab8*, two out of three tests indicated similar distributions. The other three mutants express much larger variations in their branching numbers with no or only one test indicating a similar distribution.

This leads us to the conclusion that variation in branching numbers due to genetic background is caused by two separate effects: impact of the genetic background itself on growth behaviour cannot be ruled out as there still are significant variations in the *Wild Type*. Nevertheless variations are much higher in the mutants, therefore there is an influence of the genetic background on the gene knockdown. We can also infer that the knockdown process of *Rab8* is not affected as much by genetic background as the knockdown process of the other mutants.

Variations in NoJ for different genetic backgrounds						
t	B_{Gen}	*WT*	*Betaint*	*Crumbs*	*Rab8*	*Talin*
18° C	G vs. DG	$t = -0.34$ $p > 0.5$	$t = 6.61$ $p < 0.01$	$t = -3.06$ $p < 0.01$	$t = 0.12$ $p > 0.75$	$t = 2.52$ $p < 0.1$
18° C	G vs. PDG	$t = 1.08$ $p > 0.50$	$t = 10.05$ $p < 0.01$	$t = 0.80$ $p > 0.25$	$t = -1.22$ $p > 0.1$	$t = -1.48$ $p > 0.1$
18° C	DG vs. PDG	$t = 0.45$ $p > 0.25$	$t = 3.10$ $p < 0.01$	$t = 3.35$ $p < 0.01$	$t = -1.32$ $p > 0.1$	$t = -3.92$ $p < 0.01$
29° C	G vs. DG	$t = 4.63$ $p < 0.01$	$t = -4.33$ $p < 0.01$	$t = -2.79$ $p < 0.01$	$t = -0.20$ $p > 0.75$	$t = -7.63$ $p < 0.01$
29° C	G vs. PDG	$t = 1.36$ $p > 0.1$	$t = -4.59$ $p < 0.01$	$t = -4.85$ $p < 0.01$	$t = -5.47$ $p < 0.01$	$t = -3.71$ $p < 0.01$
29° C	DG vs. PDG	$t = -3.89$ $p < 0.01$	$t = -1.84$ $p < 0.1$	$t = -1.96$ $p < 0.1$	$t = -4.82$ $p < 0.01$	$t = 5.79$ $p < 0.01$

Table 4.1. t-values and p-values of two-sided t-tests for the null hypothesis that the two samples are *not* drawn from the same distribution. Tests were performed for every two NoJ distributions for different B_{Gen} within the same t and M. p-values larger than 0.1 corresponding to similar distributions are emphasized in red.

Temperature

We expected higher temperatures to be more favorable to gene knockdown. This is the case as the mechanism used for the RNAi originally stems from yeast. It is most efficient at the temperature optimal for yeast growth which is 30°C. According to our analysis we can confirm this assumption of increased efficiency. Indeed, we found a large impact of temperature on network growth which goes so far as to cause networks grown at different temperatures but same mutation to be clustered with networks of a different mutation.

The variation in network growth caused by different growing temperatures seems to be caused only by the temperature's impact on the gene knockdown: Table 4.2 shows p-values for a two sided t-test assuming the branching numbers were drawn from different distributions. For the tests we averaged over all families for the same (M, t) combination. We can see that only the *Wild Type* expresses no significant variation for different growth temperatures.

The analysis of branching numbers also showed us how large the variation in network size *within* one family can be: for example *WT 29° C DG* contains both a network with 17 and a network with 106 branches which is a six-fold difference in junctions.

Variations in NoJ for different temperatures					
t	*WT*	*Betaint*	*Crumbs*	*Rab8*	*Talin*
18° C vs. 29° C	$t = -0.26$ $p > 0.75$	$t = 2.71$ $p < 0.01$	$t = 6.58$ $p < 0.01$	$t = 2.49$ $p < 0.1$	$t = 2.52$ $p < 0.1$

Table 4.2. t-values and p-values of two-sided t-tests for the null hypothesis that the two samples are *not* drawn from the same distribution. Tests were performed for NoJ distributions averaged over all families from one mutation-temperature-combination. p-values larger than 0.1 corresponding to similar distributions are emphasized in red.

4.3.3. Second Order Information

After exchanging results with our collaborators, who created the data set and analyzing the experimental process, we concluded that we are not able to reliably identify second order trends. This is the case because the error introduced by mechanical deformation of larvae during the imaging process is so large that given our still relatively small sample sizes the noise should be far stronger than the signal. The phenotypic trends we defined are especially prone to error by deformation as characteristics like straight or wiggly growth, how spread out the network is and even edge radius can easily be changed by slightly squeezing the larva. Variations in the length of the edges can also be affected by these deformations as the networks are only pseudo two-dimensional and can be squeezed shorter without the imaging picking up on it.

With this information we can only conclude that we cannot make any significant statements about second order growth effects. Nevertheless, the definition of main geometric growing characteristics seems to be quite concise and might be worth further investigation if larger sample sizes are available.

4.3.4. Growth Models

The models for branching behaviour and radius distribution in a network we created were approved by the biologists amongst our collaborators. It would be interesting to see if we can grow artificial networks using these growth models and compare them to their real counterparts. For metrics only dependent on edge length, number of junctions and radius, this is already possible.
However, we still miss a model which describes growth direction to be able to create a complete artificial network. Regarding the networks' tendency to minimize redundancy it was pointed out to us that the observed distance of 70 μm at which the additional redundancy is minimal also matches the rough size of a cell (cell sizes in eukaryotes range from 10-100 μm) in the surrounding tissue. Combined with the fact that the cells in the surrounding tissue can be assumed to be close-packed in space, this indicates that the network might designate roughly one branch to support one cell with oxygen which would also result in the average spread of branches we observed.

4.4. Summary

Trachea Morphogenesis

In the first part of this chapter we have given a brief introduction into the morphogenesis and function of *Drosophila* trachea. Trachea grow in several sequential branching steps of which we are only interested in the terminal one. The terminal branches consist of a single cell which is believed to express random growth patterns. The growth of the terminal cell is influenced by several genes. The expression of these genes is mediated to create *mutants*. The genes knocked down via RNA-interference are *Rab8*, *Betaint*, *Crumbs* and *Talin*. In addition to their mutation, larvae also have a distinct genetic background composed of the expression of one or a combination of *Dicer2*, *PIP2* and *GFP*. GFP is used to tag the cell's lumen for imaging. In theory, the genetic background influences the RNA-interference ma-

chinery and therefore can lead to a variation in efficiency of the gene knockdown. This also is true for the temperature at which the larvae are grown.
By varying genetic background and temperature we hoped to create different mutation severities and therefore a range of small variations within the *phenotypes*.

Analysis Results in biological Context

In the second part we tried to put the results from the previous extensive quantitative analysis into a biological context. We were able to confirm that the mutation had a much larger impact on network growth than temperature or genetic background. This is consistent with the expected behaviour.
We also assessed that the effect of the genetic background on network growth is twofold: firstly it affects network growth on its own, as has been shown by quantifying variations in *Wild Type* networks. Secondly it also affects the efficiency of the gene knockdown which causes the mutants *Betaint*, *Crumbs* and *Talin* to express more variations in network growth. The mechanism of *Rab8* knockdown however does not seem to be affected by genetic background.
We assumed that increased temperature would increase efficiency of gene knockdown. We can confirm this assumption, as all the mutants expressed a significant trend in networks growing smaller at higher temperatures.
Given the possibility of mechanical deformations during the imaging process and resulting increased noise levels, we had to reject our findings concerning phenotypic trends. We still think that the definition of principal growth characteristics is promising but to find phenotypic trends in these networks we would need larger data sets.
Our growth models were approved by our collaborators. With regard to the network's preferred spatial branch frequency it was pointed out to us that this also corresponds to the size of cells in the surrounding tissue. If every cell is on average supplied with oxygen by one branch, this also yields a length close to the redundancy minimum we found.

5. Conclusion

5.1. Summary

During the course of this work our aim was to get a better understanding of the networks forming vital transport systems in living organisms. We concentrated on the formation of tubes and the impact on network growth caused by mutations of genes crucial for vesicular transport. Humans or even small mammals are far too complex to understand the interplay of genes and molecular processes in network formation. Therefore, we used the processes observed in Drosophila terminal cells, the final extensions of the respiratory network of the fly, as paradigm. The genes *βPS-integrin*, *Crumbs*, *Rab8* and *Talin* are known to play a role in network growth in many organisms including humans. Therefore it is hoped that answers obtained by observing their impact on network realization in fruit flies bear general validity. We employed a highly automated and reproducible approach for the segmentation and analysis of network data. A high degree of automation was necessary as the dataset we were provided with was large: Our collaborators in Heidelberg created a collection of 500 images of trachea terminal cells divided into 30 different families labeled by mutation, genetic background and growing temperature. The extensive size of the dataset is new to the field and provides us with the unique possibility to make highly significant statistical statements about the networks.

Network Digitization

To extract the network information contained in the images we adapted an already existing framework for the digitization of leaf venation patterns. As the terminal cell networks are only approximately flat, we used a two-dimensional projection for the processing, and needed to do a manual rewiring of false junctions. For this task we developed a graphical user interface for the manipulation of graphs, the *Graph Manipulation GUI*. We described the functionality of the digitization framework as well as the adaptations we undertook in section 2.2. A description of the GUI's functionality and of the necessary manual processing can be found in section 2.4.

Quantification of known Phenotypes

Our main focus was to quantify the impact of mutation as well as the genetic background of the organism and its growing temperature on the formation of trachea terminal cells. We call the visibly different network realizations *network phenotypes* induced by the larva's growing conditions. We continued to ask the question whether we could find hidden phenotypes spanning several of the induced phenotypes.
To quantify a network's properties we defined a large set of metrics we were able to measure on the networks. These metrics include topological measures like a network's number of junctions and asymmetry as well as geometric measures like its length and area of convex hull. We defined and measured 15 different metrics which yielded a high-dimensional set of data points to which in the course of the analysis we applied techniques from machine learning and Bayesian reasoning for dimensionality reduction.
In section 3.4 we assessed the discriminative power of the impact of mutation, genetic background and temperature on network growth. Using the supervised learning approach of linear discriminate analysis we were able to distinguish between networks with different mutations. For temperature and genetic background the discernability was not better than chance. This confirms the assumption that the phenotype is primarily induced by the mutation whereas genetic background and temperature serve to create subtle variations of this phenotype.

Detection of hidden Phenotypes

To investigate the existence of possible hidden phenotypes in section 3.5 we used clustering, an unsupervised learning approach, to see which networks were close in metric space. We identified each cluster with a phenotype expressed by the networks and found that some clusters span two or more mutations. This confirms the existence of hidden phenotypes in the dataset. To find the directions of highest variation in the dataset we applied PCA. We visualized the clustering by projecting it onto the first two PCA components which already contributed around 50 % and 15 % to variance.
The visualization made clear that the clusters are not well-separated but rather lie on a continuum along the first PCA-component. We suspected the information contained in variation along this component to be the network's size. By performing

the same clustering on fake mutants created from *Wild Type* networks we were able to confirm this suspicion: The fake mutants created by controlling *Wild Type* networks only for their size clustered in exactly the same way as their real counterparts did.

Phenotypic Trends

After the existence of hidden phenotypes was confirmed we were interested in the question whether there is information contained in the growth pattern of a network independent of its size. We identified four main dimensions in the set of metrics we defined earlier: information about edge radius, edge length and edge straightness as well as growing direction.

In section 3.6 we determined metrics as representatives for these four dimensions and assessed the metrics' dependence on network size. By identifying significant deviations from this dependance - "phenotypic trends" - we tried to uncover whether certain network families would exhibit distinctive growing patterns.

Our collaborators at EMBL in Heidelberg pointed out that these growing patterns are especially prone to deformations of the larvae during the imaging process. We therefore can not reliably state that the phenotypic trends we found are significant or merely a product of mechanical forces acting on the larvae from the outside. Moreover as the phenotypic trends are only second order information and their signal is much smaller than the signal coming from network size we would need a much larger dataset to make robust statements.

Growth Models

After investigating the realizations of terminal cell networks in such great detail we asked ourselves if we could create a model that describes the growth of these structures. In section 3.7 we split the general model into three sub-categories dealing with branching behaviour, radius distribution and growth direction separately.

We assumed the event of branching to be a rare event and predicted that the number of branches follows a Poisson distribution. Therefore the distribution of lengths of the tubes in between branching points follows an exponential distribution. Analysis of the data shows that the predicted distributions are in very good agreement to the measured ones.

To model the radius distributions we reduced the problem to the question of how radii behave at a branching point, given the radius of the mother branch R_L. We

assumed that at each branching point, the intersection area is conserved and therefore $R_L^2 = R_I^2 + R_S^2$. Analyzing the area ratio $R_L^2/(R_I^2 + R_S^2)$ showed a Gaussian distribution of area ratios with mean very close to one. This supports the assumption that the intersection area at least on average is conserved.
For an accurate modeling of growth direction we lack information about the structure of the tissue the networks grow in. Using an approach evolving around the networks redundancy as defined in 3.2.4 we were at least able to find that the networks seem to have a preferred spacing between branches. This can be explained by the spacing of cells in the surrounding tissue which need to be supplied with oxygen.

5.2. Outlook

For the analysis of the terminal cell networks we followed a very broad approach. Nevertheless there are some analysis methods and research directions which we were not able to investigate due to lack of information or time constraints. In the following we would like to give an outlook on topics that could be worth further investigation.

Dimensionality Reduction

We found out that all of the metrics are somehow dependent on each other but the dimensionality of information contained in the network structure is still larger than one. During our analysis of the phenotypic trends we established four main dimensions for geometric characteristics. It would be interesting to investigate whether four dimensions already contain all the information and if we chose the right interpretations of these directions in metric space. We would suggest an approach involving factor analysis [49] to see how many of the dimensions collapse and what the dimensions that are left signify for network growth.

Growth Direction

The part that is missing from a complete model of network growth is a robust understanding of the direction in which a branch will grow. We established that close to junctions a branch is subject to a repelling force from the other branches dependent on their relative radius. We still do not know if this force is of mechanical or chemical origin and therefore is caused by the growth mechanics or oxygen gradients in the surrounding tissue. To investigate this further we would suggest measurements

of oxygen concentrations close to junctions to see whether a gradient could be responsible for determination of the growth direction.
To model branch growth farther away from junctions we would first need a working model of growth in inhomogeneous tissue. With this model and information about the structure of the surrounding tissue we could predict in which direction a branch is most likely to grow. On the other hand a model based on the functionality of the network could be possible, predicting not the growth direction of every single branch but a spatial distribution and spacing between branches.

Phenotypic Trends

We still believe that the networks we analyzed contain more information than only their size. We have seen differences in straightness of branches or how clustered a network will grow but we have not been able to quantify these differences. To be able to make robust statements about this kind of second-order information we would need a much larger dataset (maybe around 100 samples for each of the families) to rule out systematic error due to mechanical deformation of the larva and minimize the effect of noise.

Additional Mutants

For this work we used four mutants known to affect the molecular dynamics of tube creation. Nevertheless there are around 70 different genes known so far which have a confirmed impact on the machinery involved in network assembly in organisms. It would be interesting to do the same analysis for additional mutants to create a more complete picture of the possible network phenotypes. It would also be interesting to see if any of these mutants are able to modify the underlying growth models in a fundamental way, for example preventing branching entirely without hampering edge length.

3D-Expansion and further Application

In theory it is possible to expand the digitization process to three dimensions. Given images with high enough quality it would be great to develop a tool which can extract three dimensional networks from image z-stacks. This could have many applications in the research of three dimensional networks, including

- blood vessels in the organs of mammals,

- neurons in the brain,
- mammal respiratory networks,
- roots of trees,
- animal dens.

With our extensive quantitative analysis of this very large dataset we were able to make statistically significant statements about the phenotypes induced by mutations, the impact of mutations on network size and the functionality of network growth. The results of this work give a quantitative confirmation of the effects of mutation on network growth patterns predicted from a biological point of view. We hope that our results help to shed some light on the underlying processes governing the formation of transport networks in organisms.

A. Methods for Data Analysis

The following chapter contains a description of the methods used to analyze the dataset of trachea terminal cell networks in more detail, including principal component analysis, linear discriminant analysis and two clustering methods: KMeans and the Gaussian mixture model.

In the following, each network is called a *data point* $\mathbf{x}^j$. We assume the networks to be independent of each other because they either stem from entirely different larvae or are sufficiently separated in space to not influence each others growth. Therefore we are dealing with independent data points. Each data point is composed of a number of *observations*, the metrics we measured for each network such as its number of junctions or its length. Therefore $\mathbf{x}^j = (NoJ, LoN, \ldots)$. The single observations are *not* independent of each other as they are all measured within the same network and might be correlated to a large degree.

A.1. Principal Component Analysis

The following description of principal component analysis (PCA) can be found in more detail in [33].

Algorithm

Principal Component Analysis (PCA) is an unsupervised dimensionality reduction method. PCA assumes that the data points we are looking at all lie close to a linear subspace of the high dimensional space spanned by all metrics. Our aim, therefore, is to discover a low-dimensional coordinate system which can approximately describe the data points. We are looking for a coordinate system which captures most of the *variance* present in the data as illustrated in figure A.1. The approximate representation of the data points in the reduced coordinate system assuming that

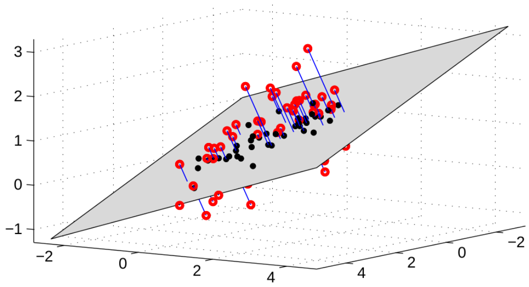


Figure A.1. Illustration of a linear subspace fitted to the data points (red) in a way such that the squared distances of the projections onto the subspace (black) are minimal. Figure borrowed from [33] P.316.

the data is centered is given by

$$\mathbf{x}^n \approx \sum_{j=1}^{M} y_j^n \mathbf{b}^j \equiv \tilde{\mathbf{x}}^n \ , \tag{A.1}$$

Where the $\mathbf{b}^j$ are base vectors that span the linear subspace which are also called *principal component coefficients* and the y_j^n are the low dimensional coordinates of the data which form a lower dimension $\mathbf{y}^n$ for each data point n. We can collectively write those lower dimensional vectors as $\mathbf{Y} = [\mathbf{y}^1, \ldots, \mathbf{y}^N]$ Given the dimensionality of the data-space $D = \dim\ (x)$ we hope to describe the data with only a small number $M \ll D$ of coordinates $\mathbf{y}$.

We now want to find the optimal base $B = [\mathbf{b}^1, \ldots, \mathbf{b}^M]$ by minimizing the squared distance error $E(\mathbf{B}, \mathbf{Y})$ between $\mathbf{x}$ and its reconstruction $\tilde{\mathbf{x}}$

$$\begin{aligned} E(\mathbf{B}, \mathbf{Y}) &= \sum_{n=1}^{N} \sum_{i=1}^{D} (x_i^n - \tilde{x}_i^n)^2 = \sum_{n=1}^{N} \sum_{i=1}^{D} \left(x_i^n - \sum_{j=1}^{M} y_j^n b_i^j \right) \\ &= \operatorname{trace}\left[(\mathbf{X} - \mathbf{BY})^T (\mathbf{X} - \mathbf{BY}) \right] \ , \end{aligned} \tag{A.2}$$

where $\mathbf{X} = [\mathbf{x}^1, \ldots, \mathbf{x}^n]$. Without loss of generality we can constrain $\mathbf{B}$ to be an orthonormal matrix. We can now find the minimal squared distance error by differentiating equation A.2 with respect to y_k^n and using the orthonormality constraint on $\mathbf{B}$, we then obtain

$$\begin{aligned} -\frac{1}{2} \frac{\partial}{\partial y_k^n} E(\mathbf{B}, \mathbf{Y}) &= \sum_i x_i^n b_i^k - y_k^n = 0 \\ \Rightarrow y_k^n &= \sum_i x_i^n b_i^k \quad \text{which can be written as} \quad \mathbf{Y} = \mathbf{B}^T \mathbf{X} \ . \end{aligned}$$

We can now substitute solution A.3 into equation A.2 and receive the squared error as a function of only $\mathbf{B}$

$$(\mathbf{X}-\mathbf{B}\mathbf{Y})^T(\mathbf{X}-\mathbf{B}\mathbf{Y}) = \mathbf{X}^T\mathbf{X} - \mathbf{X}^T\mathbf{B}\mathbf{B}^T\mathbf{X} - \cancel{\mathbf{X}^T\mathbf{B}\mathbf{B}^T\mathbf{X}} + \cancel{\mathbf{X}^T\mathbf{B}\mathbf{B}^T\underbrace{\mathbf{B}^T\mathbf{B}}_{\mathbf{I}}\mathbf{X}} \; . \tag{A.3}$$

Using trace($\mathbf{ABC}$) = trace($\mathbf{CAB}$) lets us write the above equation as

$$E(\mathbf{B}) = \text{trace}\left[\mathbf{X}\mathbf{X}^T(\mathbf{I}-\mathbf{B}\mathbf{B}^T)\right] = (N-1)\left[\text{trace}(\mathbf{S}) - \text{trace}(\mathbf{S}\mathbf{B}\mathbf{B}^T)\right] \; , \tag{A.4}$$

with $\mathbf{S}$ the sample covariance matrix of the data. We use a set of Lagrange multipliers $\mathbf{L}$ to minimize equation A.4 under the orthonormality constraint for $\mathbf{B}$ while neglecting the constant prefactor $(N-1)$ and the trace(S) term. We therefore want to minimize

$$-\text{trace}\left[\mathbf{S}\mathbf{B}\mathbf{B}^T\right] + \text{trace}\left[\mathbf{L}(\mathbf{B}^T\mathbf{B}-\mathbf{I})\right] \; , \tag{A.5}$$

which yields, assuming that $\mathbf{L}$ is symmetric,

$$\mathbf{S}\mathbf{B} = \mathbf{B}\mathbf{L} \; . \tag{A.6}$$

One case where the matrices $\mathbf{B}$ and $\mathbf{L}$ satisfy above matrix equation is when $\mathbf{L}$ is diagonal in which case the matrix equation is a form of Eigen-equation and the columns of $\mathbf{B}$ are the corresponding Eigen-vectors of $\mathbf{S}$. In this case trace($\mathbf{S}\mathbf{B}\mathbf{B}^T$) = trace($\mathbf{L}$) and equation A.4 can be written as

$$\frac{1}{N-1}E(\mathbf{B}) = \text{trace}(\mathbf{S}) - \text{trace}(\mathbf{L}) = -\sum_{i=1}^{M}\lambda_i + \text{const} \tag{A.7}$$

with Eigenvalues λ_i. As our goal is to minimize $E(\mathbf{B})$ we choose the base with the largest corresponding Eigenvalues. Ordering the Eigenvalues according to their value $\lambda_1 \geq \lambda_2, \ldots$ gives us

$$\frac{1}{N-1}E(\mathbf{B}) = \sum_{i=1}^{D}\lambda_i - \sum_{i=1}^{M}\lambda_i = \sum_{i=M+1}^{D}\lambda_i \; , \tag{A.8}$$

but this solution to the Eigen-problem only serves to define the solution subspace - we still may rotate and rescale $\mathbf{B}$ and $\mathbf{Y}$ such that $E(\mathbf{B})$ remains the same. We can justify selecting the non-rotated solution by introducing another requirement: that of maximal variance along the base vectors, this means that the base vectors we choose represent *interesting* directions in the dataset.

Implementation

In this work we use the PCA algorithm provided by `scikit learn` which can be found at `http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html` which implements the method of Thomas P. Minka [50].

A.2. Linear Discriminant Analysis

The following description of linear discriminant analysis (LDA) can be found in more detail in [51].

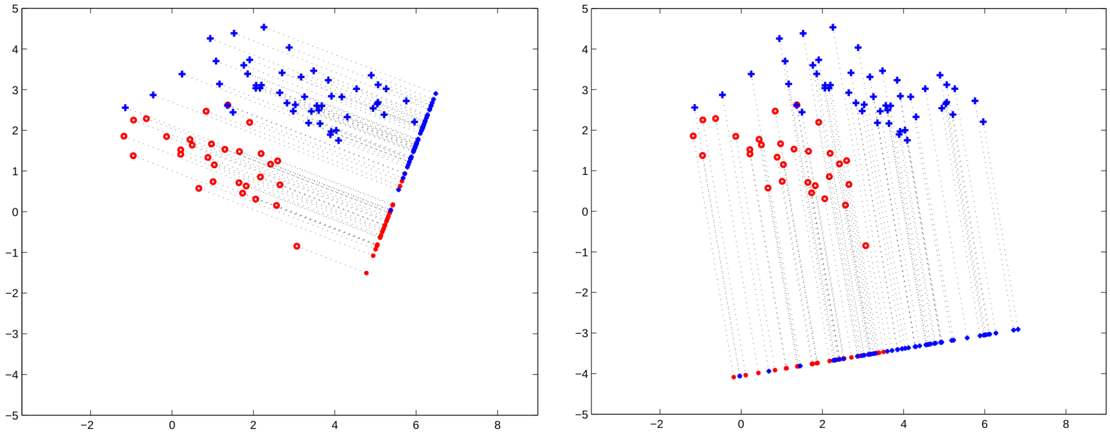
LDA or Fischer's linear discriminant is a method from supervised learning. Its ultimate goal is the reduction of dimensionality for improved classification of data. We can use LDA if we have data points with a high number of observations (high dimensionality of the problem) and also have some class information about the data available. If we used PCA for this problem, when subsequently using the projected data in a classification problem we would be unable to make use of the class labels and therefore lose information. This loss of information leads to a lower dimensional representation and therefore might lead to a suboptimal separation of the different classes. This is illustrated in figure A.2 where dimension reduction using LDA and PCA is shown.

A.2.1. Supervised Linear Projection

To understand how LDA works, let us at first look at the simplest case of two classes $\chi_i = \{\mathbf{x}_i^1, \dots, \mathbf{x}_i^{N_j}\}, i \in (1,2)$ with binary class data $\mathbf{x}_i^j = (a, b)$. We now want to find a linear projection

$$\mathbf{y} = \mathbf{w}^T \mathbf{x}$$

where $\dim \mathbf{w} = D \times L, L < D$. Additionally we want for two data points $\mathbf{x}_i^m, \mathbf{x}_i^n$ from the same class that the distance between their projections $\mathbf{y}_i^m, \mathbf{y}_i^n$ should be small. If we take two data points from different classes $\mathbf{x}_i^m, \mathbf{x}_j^m$ the distance between their



(a)

(b)

Figure A.2. Large crosses and circles in blue and red represent data from two classes, a projection on one dimension is given by their smaller counterparts. **(a)** projection using Fischer's linear discriminant, the projection shows little overlap between classes. **(b)** unsupervised dimension reduction using PCA, the projection shows considerable overlap. Picture borrowed from [51] P. 344

projections $\mathbf{y}_i^m, \mathbf{y}_j^m$ should be large. This we can already use for classification: if we add a new data point $\mathbf{x}^*$ and its projection $\mathbf{y}^* = W^T\mathbf{x}^*$ is close to the projection of data points from a certain class, we can assume that $\mathbf{x}^*$ belongs to that class.
By applying this projection approach, we form the supervised projection where only the class discriminative parts of the data are retained.

A.2.2. Fischer's Linear Discriminant

Algorithm

We again look at only two classes χ_1 and χ_2 and model the data from each class with a Gaussian, therefore

$$p(\mathbf{x}_1) = N(\mathbf{x}_1|\mathbf{m}_1, \mathbf{S}_1) \ , \qquad p(\mathbf{x}_2) = N(\mathbf{x}_2|\mathbf{m}_2, \mathbf{S}_2) \ , \tag{A.9}$$

where $\mathbf{m}_i$ are the sample means of class data and $\mathbf{S}_i$ is the sample covariance and $N(\mathbf{x}_i|\mathbf{m}_i, \mathbf{S}_i)$ is the Gaussian distribution of a continuous variable $\mathbf{x_i}$

$$N(\mathbf{x}_i|\mathbf{m}_i, \mathbf{S}_i) = \frac{1}{\sqrt{\det(2\pi\mathbf{S}_i)}}(\mathbf{S}_i)^n \cdot \exp\left[-(\mathbf{x}_i - \mathbf{m}_i)^T\mathbf{S}_i^{-1}(\mathbf{x}_i - \mathbf{m}_i)\right] \ . \tag{A.10}$$

Projections of the points are given by

$$\mathbf{y}_i^n = \mathbf{w}^T \mathbf{x}_i^n \ . \tag{A.11}$$

The projections $\mathbf{w}$ are linear operations, therefore the projected distributions are still Gaussian:

$$p(\mathbf{y}_i) = N(\mathbf{y}_i|\mu_i, \sigma_i^2) \ , \qquad \mu_i = \mathbf{w}^T \mathbf{m}_i \ , \qquad \sigma_i^2 = \mathbf{w}^T \mathbf{S}_i \mathbf{w} \ . \tag{A.12}$$

We now want to search for a projection $\mathbf{w}$ where the projected distributions of both classes have minimal overlap with each other. We can achieve this when the projected means are maximally separated and therefore $(\mu_1 - \mu_2)^2$ is maximal. However if the variances σ_i are also large, there is still the possibility of large overlap. We therefore define an objective function using the fraction π_i of data points which are in the class χ_i

$$\frac{(\mu_1 - \mu_2)^2}{\pi_1 \sigma_1 + \pi_2 \sigma_2} \ . \tag{A.13}$$

In terms of the projection W the objective equation reads

$$\begin{gathered} F(W) = \frac{\mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{w}}{\mathbf{w}^T(\pi_1 \mathbf{S}_1 + \pi_2 \mathbf{S}_2)\mathbf{w}} = \frac{\mathbf{w}^T \mathbf{A} \mathbf{w}}{\mathbf{w}^T \mathbf{B} \mathbf{w}} \\ \mathbf{A} = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \ , \qquad \mathbf{B} = \pi_1 \mathbf{S}_1 + \pi_2 \mathbf{S}_2 \ . \end{gathered} \tag{A.14}$$

We can now find the optimal projection by differentiating with regards to v

$$\frac{\partial}{\partial \mathbf{w}} \frac{\mathbf{w}^T \mathbf{A} \mathbf{w}}{\mathbf{w}^T \mathbf{B} \mathbf{w}} = \frac{2}{(\mathbf{w}^T \mathbf{B} \mathbf{w})^2} \Big[(\mathbf{w}^T \mathbf{B} \mathbf{w}) \mathbf{A} \mathbf{w} - (\mathbf{w}^T \mathbf{A} \mathbf{w}) \mathbf{B} \mathbf{w} \Big] = 0 \tag{A.15}$$

$$\Rightarrow (\mathbf{w}^T \mathbf{B} \mathbf{w}) \mathbf{A} \mathbf{w} = (\mathbf{w}^T \mathbf{A} \mathbf{w}) \mathbf{B} \mathbf{w} \ . \tag{A.16}$$

We now multiply by $\mathbf{B}^{-1}$ and get

$$\mathbf{B}^{-1}(\mathbf{m}_1 - \mathbf{m}_2) \underbrace{(\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{w}}_{a} = \underbrace{\frac{\mathbf{w}^T \mathbf{A} \mathbf{w}}{\mathbf{w}^T \mathbf{B} \mathbf{w}}}_{b} \mathbf{w} \ , \tag{A.17}$$

where a, b are scalars, therefore

$$\mathbf{w} \propto \mathbf{B}^{-1}(\mathbf{m}_1 - \mathbf{m}_2) \ . \tag{A.18}$$

The proportionality factor k still depends on $\mathbf{w}$ but since the objective function $F(\mathbf{w})$ in equation A.14 is invariant to rescaling of $\mathbf{w}$ we still can take k to be constant. It is common to rescale $\mathbf{w}$ such that $\mathbf{w}^T\mathbf{w} = 1$, therefore

$$\mathbf{w} = k\mathbf{B}^{-1}(\mathbf{m}_1 - \mathbf{m}_2) \tag{A.19}$$

$$k = \frac{1}{\sqrt{(\mathbf{m}_1 - \mathbf{m}_2)^T\mathbf{B}^{-2}(\mathbf{m}_1 - \mathbf{m}_2)}} \ . \tag{A.20}$$

Implementation

Using LDA for non-binary class data is no problem given the above derivation of the method. However using LDA for more than two classes and projections onto more than one dimension requires a generalization. Fischer's method is generalized by the Canonical Variates to projections of more than one dimension and more than two classes [28]. Also, there can arise problems with the actual derivation of the objective function: the matrix $\mathbf{B}$ might not be invertible which is the case when there are fewer data points than observations. Also $\mathbf{B}$ can have zero entries when there are elements in the input vectors that never vary. This causes the denominator of the objective function A.14 to become zero and therefore the problem is ill defined. The generalization as well as problem handling is implemented in `scikit learn`'s [52] implementation of LDA which is used in this work and can be found at `http://scikit-learn.org/stable/modules/generated/sklearn.lda.LDA.html#sklearn.lda.LDA`.

A.3. Cluster Analysis

Cluster analysis is a form of unsupervised learning as we treat a dataset where we explicitly do not know (or in our case do not utilize) class affiliations of data points within the dataset. The task of cluster analysis is to group the data in such a way that data points belonging to the same cluster are more similar to each other than data points belonging to other clusters. In this work we utilize two different clustering algorithms: the centroid-based approach *KMeans* and the distribution based approach of a *Gaussian Mixture Model* (GMM).

A.3.1. k Means

Algorithm

The aim of the KMeans algorithm is to divide N data points into k clusters where each cluster is represented by its mean. KMeans tries to minimize the within-cluster sum of squares of data points $(\mathbf{x_i}, \ldots, \mathbf{x}_N)$ belonging to k sets $\mathbf{S} = \{S_1, \ldots, S_k\}$, therefore the objective function is

$$\sum_{i}^{k} \sum_{x \in S_i} \|\mathbf{x}_i - \mu_i\| \ . \tag{A.21}$$

Solving this problem is NP-hard, therefore the algorithms implemented for finding the best clustering utilize a heuristic approach which converge to a local optimum. The standard algorithm also called *Lloyd's Algorithm* alternates between two steps until it converges. If we have a set of k means $\mu_1, \ldots, \mu_k$ in the first step we assign each data point to the cluster (represented by its mean) which yields the smallest within cluster sum of squares, therefore for each time-step t the assignment step is

$$S_i^{(t)} = \{x_p : \|x_p - \mu_i^{(t)}\|^2 \leq \|x_p - \mu_j^{(t)}\|^2 \ \forall j, 1 \leq j \leq k\} \ .$$

After the assignment an update step is performed which calculates the new means as representatives for the clusters:

$$\mu_i^{(t+1)} = \frac{1}{|S_i^t|} \sum_{x_j \in S_i^{(t)}} x_j \tag{A.22}$$

Implementation

In this work, the KMeans implementation of `scikit learn` is used which implements the method of Arthur, D. and Vassilvitskii, S. [53] for increased convergence speed and can be found at `http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html`.

A.3.2. Gaussian Mixture Model

Model

The following description of Gaussian Mixture Models (GMM) can be found in more detail in [31].
Given the definition of a Gaussian distribution of a continuous variable $\mathbf{x}$ in equation A.10, a mixture of Gaussians with mixture weight $p(i)$ can be expressed as

$$p(\mathbf{x}) = \sum_{i=1}^{H} p(\mathbf{x}|\mathbf{m}_i, \mathbf{S}_i)p(i) \ . \tag{A.23}$$

For each set of data points $\chi = \mathbf{x}^1, \ldots, \mathbf{x}^N$ the log probability is then given as

$$\log p(\chi, \theta) = \sum_{n=1}^{N} \log \sum_{i=1}^{H} p(i) \frac{1}{\sqrt{\det(2\pi \mathbf{S}_i)}} \exp\left[-\frac{1}{2}(\mathbf{x}^n - \mathbf{m}_i)^T \mathbf{S}_i^{-1} (\mathbf{x}^n - \mathbf{m}_i)\right] \ , \tag{A.24}$$

where the optimal parameters θ can be found using an expectation-maximization algorithm as described in [31] P.415, where the component index i plays the role of a latent variable.

Implementation

For this work, the GMM implementation of `scikit learn` is used. Its implementation corresponds to the frequentist (non-Bayesian) formulation of Gaussian Mixture Models and can be found at `http://scikit-learn.org/stable/modules/generated/sklearn.mixture.GMM.html`.

B. Supporting Information

Edge Length over Depth

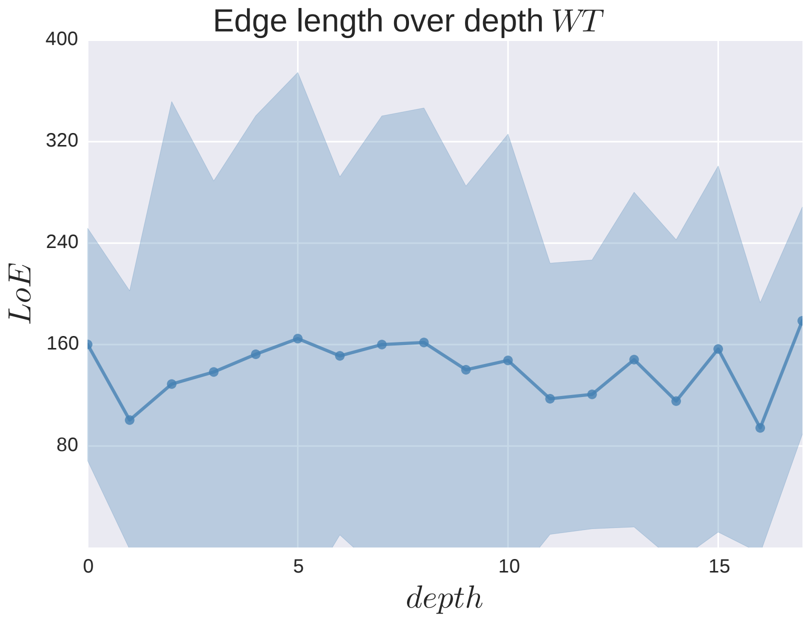


Figure B.1. The length of edges in the graph does not significantly change with the edge's position in the graph, i.e., its depth.

Area Ratios for different Mutants

AR	*WT*	*Betaint*	*Crumbs*	*Rab8*	*Talin*
μ	1.11	0.96	1.07	1.05	0.93
σ	0.37	0.37	0.38	0.38	0.37

Table B.1. Means and variances of the area ratios for different mutants ant the *Wild Type*.

C. Dataset

Number of Networks						
temperature	gen. background	*WT*	*Betaint*	*Crumbs*	*Rab8*	*Talin*
18° C	G	17	19	16	18	16
	DG	19	23	15	16	19
	PDG	16	18	16	13	17
29° C	G	17	10	18	27	20
	DG	17	16	16	19	17
	PDG	15	12	10	13	24

Table C.1. Number of networks contained in the dataset for each family.

AoCH $[px^2]$						
t	B_{Gen}	*WT*	*Betaint*	*Crumbs*	*Rab8*	*Talin*
18° C	G	$\mu = 6.67 \cdot 10^5$ $\sigma = 1.31 \cdot 10^5$	$\mu = 6.71 \cdot 10^5$ $\sigma = 1.65 \cdot 10^5$	$\mu = 3.11 \cdot 10^5$ $\sigma = 8.14 \cdot 10^4$	$\mu = 7.74 \cdot 10^5$ $\sigma = 1.69 \cdot 10^5$	$\mu = 3.10 \cdot 10^5$ $\sigma = 5.46 \cdot 10^4$
18° C	DG	$\mu = 6.41 \cdot 10^5$ $\sigma = 1.49 \cdot 10^5$	$\mu = 6.03 \cdot 10^5$ $\sigma = 1.13 \cdot 10^5$	$\mu = 2.81 \cdot 10^5$ $\sigma = 5.24 \cdot 10^4$	$\mu = 7.88 \cdot 10^5$ $\sigma = 1.64 \cdot 10^5$	$\mu = 5.28 \cdot 10^4$ $\sigma = 3.19 \cdot 10^4$
18° C	PDG	$\mu = 5.00 \cdot 10^5$ $\sigma = 1.32 \cdot 10^5$	$\mu = 3.34 \cdot 10^5$ $\sigma = 9.98 \cdot 10^4$	$\mu = 3.28 \cdot 10^5$ $\sigma = 9.05 \cdot 10^4$	$\mu = 6.21 \cdot 10^5$ $\sigma = 1.02 \cdot 10^5$	$\mu = 8.47 \cdot 10^4$ $\sigma = 4.37 \cdot 10^4$
29° C	G	$\mu = 5.56 \cdot 10^5$ $\sigma = 1.11 \cdot 10^5$	$\mu = 6.28 \cdot 10^5$ $\sigma = 1.52 \cdot 10^5$	$\mu = 1.84 \cdot 10^5$ $\sigma = 3.65 \cdot 10^4$	$\mu = 7.92 \cdot 10^5$ $\sigma = 1.28 \cdot 10^5$	$\mu = 7.16 \cdot 10^4$ $\sigma = 4.18 \cdot 10^4$
29° C	DG	$\mu = 3.58 \cdot 10^5$ $\sigma = 1.16 \cdot 10^5$	$\mu = 3.08 \cdot 10^5$ $\sigma = 9.89 \cdot 10^4$	$\mu = 1.10 \cdot 10^5$ $\sigma = 4.35 \cdot 10^4$	$\mu = 8.23 \cdot 10^5$ $\sigma = 2.37 \cdot 10^5$	$\mu = 0.00 \cdot 10^0$ $\sigma = 0.00 \cdot 10^0$
29° C	PDG	$\mu = 4.28 \cdot 10^5$ $\sigma = 1.02 \cdot 10^5$	$\mu = 4.81 \cdot 10^5$ $\sigma = 1.06 \cdot 10^5$	$\mu = 1.66 \cdot 10^5$ $\sigma = 6.19 \cdot 10^4$	$\mu = 6.53 \cdot 10^5$ $\sigma = 1.33 \cdot 10^5$	$\mu = 7.16 \cdot 10^4$ $\sigma = 5.37 \cdot 10^4$

Table C.2. Area of convex hull for each family: the table contains the mean μ as well as the standard deviation σ.

AoN $[px^2]$						
t	B_{Gen}	*WT*	*Betaint*	*Crumbs*	*Rab8*	*Talin*
18° C	G	$\mu = 4.77 \cdot 10^4$ $\sigma = 8.82 \cdot 10^3$	$\mu = 3.27 \cdot 10^4$ $\sigma = 7.33 \cdot 10^3$	$\mu = 2.20 \cdot 10^4$ $\sigma = 4.71 \cdot 10^3$	$\mu = 5.47 \cdot 10^4$ $\sigma = 1.75 \cdot 10^4$	$\mu = 2.30 \cdot 10^4$ $\sigma = 5.64 \cdot 10^3$
18° C	DG	$\mu = 4.92 \cdot 10^4$ $\sigma = 1.35 \cdot 10^4$	$\mu = 3.41 \cdot 10^4$ $\sigma = 7.91 \cdot 10^3$	$\mu = 2.31 \cdot 10^4$ $\sigma = 4.80 \cdot 10^3$	$\mu = 5.45 \cdot 10^4$ $\sigma = 1.16 \cdot 10^4$	$\mu = 5.36 \cdot 10^3$ $\sigma = 1.61 \cdot 10^3$
18° C	PDG	$\mu = 3.73 \cdot 10^4$ $\sigma = 8.77 \cdot 10^3$	$\mu = 1.88 \cdot 10^4$ $\sigma = 5.91 \cdot 10^3$	$\mu = 2.58 \cdot 10^4$ $\sigma = 8.54 \cdot 10^3$	$\mu = 4.02 \cdot 10^4$ $\sigma = 8.93 \cdot 10^3$	$\mu = 6.56 \cdot 10^3$ $\sigma = 2.55 \cdot 10^3$
29° C	G	$\mu = 4.57 \cdot 10^4$ $\sigma = 1.45 \cdot 10^4$	$\mu = 3.42 \cdot 10^4$ $\sigma = 9.00 \cdot 10^3$	$\mu = 1.94 \cdot 10^4$ $\sigma = 3.33 \cdot 10^3$	$\mu = 5.65 \cdot 10^4$ $\sigma = 1.26 \cdot 10^4$	$\mu = 6.54 \cdot 10^3$ $\sigma = 3.05 \cdot 10^3$
29° C	DG	$\mu = 2.30 \cdot 10^4$ $\sigma = 8.05 \cdot 10^3$	$\mu = 1.94 \cdot 10^4$ $\sigma = 5.20 \cdot 10^3$	$\mu = 1.09 \cdot 10^4$ $\sigma = 2.50 \cdot 10^3$	$\mu = 8.95 \cdot 10^4$ $\sigma = 6.31 \cdot 10^4$	$\mu = 0.00 \cdot 10^0$ $\sigma = 0.00 \cdot 10^0$
29° C	PDG	$\mu = 3.46 \cdot 10^4$ $\sigma = 8.35 \cdot 10^3$	$\mu = 2.99 \cdot 10^4$ $\sigma = 7.42 \cdot 10^3$	$\mu = 1.36 \cdot 10^4$ $\sigma = 5.83 \cdot 10^3$	$\mu = 4.60 \cdot 10^4$ $\sigma = 1.16 \cdot 10^4$	$\mu = 7.64 \cdot 10^3$ $\sigma = 3.59 \cdot 10^3$

Table C.3. Area of the network for each family: the table contains the mean μ as well as the standard deviation σ.

				LoN $[px]$		
t	B_{Gen}	*WT*	*Betaint*	*Crumbs*	*Rab8*	*Talin*
18° C	G	$\mu = 9.37 \cdot 10^3$ $\sigma = 1.71 \cdot 10^3$	$\mu = 7.30 \cdot 10^3$ $\sigma = 1.76 \cdot 10^3$	$\mu = 3.48 \cdot 10^3$ $\sigma = 7.15 \cdot 10^2$	$\mu = 1.12 \cdot 10^4$ $\sigma = 2.93 \cdot 10^3$	$\mu = 3.69 \cdot 10^3$ $\sigma = 9.35 \cdot 10^2$
18° C	DG	$\mu = 8.61 \cdot 10^3$ $\sigma = 2.38 \cdot 10^3$	$\mu = 6.87 \cdot 10^3$ $\sigma = 1.69 \cdot 10^3$	$\mu = 3.34 \cdot 10^3$ $\sigma = 7.24 \cdot 10^2$	$\mu = 1.21 \cdot 10^4$ $\sigma = 2.53 \cdot 10^3$	$\mu = 7.29 \cdot 10^2$ $\sigma = 2.15 \cdot 10^2$
18° C	PDG	$\mu = 6.82 \cdot 10^3$ $\sigma = 1.44 \cdot 10^3$	$\mu = 3.82 \cdot 10^3$ $\sigma = 1.38 \cdot 10^3$	$\mu = 4.65 \cdot 10^3$ $\sigma = 1.58 \cdot 10^3$	$\mu = 8.91 \cdot 10^3$ $\sigma = 2.09 \cdot 10^3$	$\mu = 1.14 \cdot 10^3$ $\sigma = 3.85 \cdot 10^2$
29° C	G	$\mu = 9.97 \cdot 10^3$ $\sigma = 3.09 \cdot 10^3$	$\mu = 7.33 \cdot 10^3$ $\sigma = 1.83 \cdot 10^3$	$\mu = 2.22 \cdot 10^3$ $\sigma = 3.31 \cdot 10^2$	$\mu = 1.26 \cdot 10^4$ $\sigma = 2.44 \cdot 10^3$	$\mu = 1.04 \cdot 10^3$ $\sigma = 5.59 \cdot 10^2$
29° C	DG	$\mu = 4.75 \cdot 10^3$ $\sigma = 1.65 \cdot 10^3$	$\mu = 3.20 \cdot 10^3$ $\sigma = 9.58 \cdot 10^2$	$\mu = 1.33 \cdot 10^3$ $\sigma = 3.86 \cdot 10^2$	$\mu = 1.63 \cdot 10^4$ $\sigma = 1.10 \cdot 10^4$	$\mu = 0.00 \cdot 10^0$ $\sigma = 0.00 \cdot 10^0$
29° C	PDG	$\mu = 7.58 \cdot 10^3$ $\sigma = 2.06 \cdot 10^3$	$\mu = 5.87 \cdot 10^3$ $\sigma = 1.31 \cdot 10^3$	$\mu = 1.97 \cdot 10^3$ $\sigma = 7.12 \cdot 10^2$	$\mu = 1.06 \cdot 10^4$ $\sigma = 2.37 \cdot 10^3$	$\mu = 1.20 \cdot 10^3$ $\sigma = 6.89 \cdot 10^2$

Table C.4. Length of the network for each family: the table contains the mean μ as well as the standard deviation σ.

				NoJ		
t	B_{Gen}	*WT*	*Betaint*	*Crumbs*	*Rab8*	*Talin*
18° C	G	$\mu = 3.21 \cdot 10^1$ $\sigma = 7.69 \cdot 10^0$	$\mu = 2.07 \cdot 10^1$ $\sigma = 5.51 \cdot 10^0$	$\mu = 1.71 \cdot 10^1$ $\sigma = 4.39 \cdot 10^0$	$\mu = 3.95 \cdot 10^1$ $\sigma = 1.47 \cdot 10^1$	$\mu = 1.92 \cdot 10^1$ $\sigma = 7.17 \cdot 10^0$
18° C	DG	$\mu = 2.94 \cdot 10^1$ $\sigma = 1.05 \cdot 10^1$	$\mu = 2.37 \cdot 10^1$ $\sigma = 7.58 \cdot 10^0$	$\mu = 1.98 \cdot 10^1$ $\sigma = 5.46 \cdot 10^0$	$\mu = 3.74 \cdot 10^1$ $\sigma = 8.28 \cdot 10^0$	$\mu = 1.22 \cdot 10^0$ $\sigma = 1.81 \cdot 10^0$
18° C	PDG	$\mu = 3.02 \cdot 10^1$ $\sigma = 1.08 \cdot 10^1$	$\mu = 1.21 \cdot 10^1$ $\sigma = 4.23 \cdot 10^0$	$\mu = 2.52 \cdot 10^1$ $\sigma = 1.10 \cdot 10^1$	$\mu = 2.62 \cdot 10^1$ $\sigma = 9.30 \cdot 10^0$	$\mu = 4.81 \cdot 10^0$ $\sigma = 3.38 \cdot 10^0$
29° C	G	$\mu = 4.94 \cdot 10^1$ $\sigma = 2.04 \cdot 10^1$	$\mu = 2.05 \cdot 10^1$ $\sigma = 4.90 \cdot 10^0$	$\mu = 1.37 \cdot 10^1$ $\sigma = 3.45 \cdot 10^0$	$\mu = 4.09 \cdot 10^1$ $\sigma = 1.00 \cdot 10^1$	$\mu = 5.70 \cdot 10^0$ $\sigma = 5.75 \cdot 10^0$
29° C	DG	$\mu = 2.08 \cdot 10^1$ $\sigma = 8.05 \cdot 10^0$	$\mu = 1.19 \cdot 10^1$ $\sigma = 4.17 \cdot 10^0$	$\mu = 5.30 \cdot 10^0$ $\sigma = 2.47 \cdot 10^0$	$\mu = 5.16 \cdot 10^1$ $\sigma = 1.99 \cdot 10^1$	$\mu = 0.00 \cdot 10^0$ $\sigma = 0.00 \cdot 10^0$
29° C	PDG	$\mu = 3.92 \cdot 10^1$ $\sigma = 1.11 \cdot 10^1$	$\mu = 2.05 \cdot 10^1$ $\sigma = 5.23 \cdot 10^0$	$\mu = 1.00 \cdot 10^1$ $\sigma = 5.06 \cdot 10^0$	$\mu = 4.30 \cdot 10^1$ $\sigma = 1.40 \cdot 10^1$	$\mu = 7.92 \cdot 10^0$ $\sigma = 5.20 \cdot 10^0$

Table C.5. Number of junctions for each family: the table contains the mean μ as well as the standard deviation σ.

		D_{T_L} $[px]$				
t	B_{Gen}	*WT*	*Betaint*	*Crumbs*	*Rab8*	*Talin*
18° C	G	$\mu = 4.93 \cdot 10^3$ $\sigma = 4.14 \cdot 10^3$	$\mu = 3.55 \cdot 10^3$ $\sigma = 2.92 \cdot 10^3$	$\mu = 1.16 \cdot 10^3$ $\sigma = 1.18 \cdot 10^3$	$\mu = 6.37 \cdot 10^3$ $\sigma = 5.13 \cdot 10^3$	$\mu = 1.63 \cdot 10^3$ $\sigma = 1.48 \cdot 10^3$
18° C	DG	$\mu = 3.45 \cdot 10^3$ $\sigma = 2.87 \cdot 10^3$	$\mu = 3.29 \cdot 10^3$ $\sigma = 2.98 \cdot 10^3$	$\mu = 1.10 \cdot 10^3$ $\sigma = 1.03 \cdot 10^3$	$\mu = 6.90 \cdot 10^3$ $\sigma = 4.62 \cdot 10^3$	$\mu = 6.23 \cdot 10^2$ $\sigma = 1.27 \cdot 10^2$
18° C	PDG	$\mu = 2.70 \cdot 10^3$ $\sigma = 2.70 \cdot 10^3$	$\mu = 2.12 \cdot 10^3$ $\sigma = 1.99 \cdot 10^3$	$\mu = 2.12 \cdot 10^3$ $\sigma = 1.89 \cdot 10^3$	$\mu = 5.71 \cdot 10^3$ $\sigma = 4.39 \cdot 10^3$	$\mu = 7.34 \cdot 10^2$ $\sigma = 5.18 \cdot 10^2$
29° C	G	$\mu = 5.76 \cdot 10^3$ $\sigma = 4.93 \cdot 10^3$	$\mu = 3.34 \cdot 10^3$ $\sigma = 2.91 \cdot 10^3$	$\mu = 8.88 \cdot 10^2$ $\sigma = 7.97 \cdot 10^2$	$\mu = 7.13 \cdot 10^3$ $\sigma = 4.60 \cdot 10^3$	$\mu = 7.20 \cdot 10^2$ $\sigma = 6.27 \cdot 10^2$
29° C	DG	$\mu = 2.23 \cdot 10^3$ $\sigma = 2.02 \cdot 10^3$	$\mu = 1.65 \cdot 10^3$ $\sigma = 2.21 \cdot 10^3$	$\mu = 7.09 \cdot 10^2$ $\sigma = 4.75 \cdot 10^2$	$\mu = 7.39 \cdot 10^3$ $\sigma = 5.05 \cdot 10^3$	$\mu = 0.00 \cdot 10^0$ $\sigma = 0.00 \cdot 10^0$
29° C	PDG	$\mu = 4.50 \cdot 10^3$ $\sigma = 3.83 \cdot 10^3$	$\mu = 2.96 \cdot 10^3$ $\sigma = 2.56 \cdot 10^3$	$\mu = 1.29 \cdot 10^3$ $\sigma = 1.56 \cdot 10^3$	$\mu = 6.53 \cdot 10^3$ $\sigma = 4.12 \cdot 10^3$	$\mu = 9.25 \cdot 10^2$ $\sigma = 9.31 \cdot 10^2$

Table C.6. Averaged tree-distance from leaf to root for each family: the table contains the mean μ as well as the standard deviation σ.

		D_{G_L} $[px]$				
t	B_{Gen}	*WT*	*Betaint*	*Crumbs*	*Rab8*	*Talin*
18° C	G	$\mu = 5.76 \cdot 10^2$ $\sigma = 2.17 \cdot 10^2$	$\mu = 5.84 \cdot 10^2$ $\sigma = 2.16 \cdot 10^2$	$\mu = 4.85 \cdot 10^2$ $\sigma = 1.82 \cdot 10^2$	$\mu = 5.71 \cdot 10^2$ $\sigma = 2.29 \cdot 10^2$	$\mu = 4.81 \cdot 10^2$ $\sigma = 1.83 \cdot 10^2$
18° C	DG	$\mu = 5.95 \cdot 10^2$ $\sigma = 2.35 \cdot 10^2$	$\mu = 5.22 \cdot 10^2$ $\sigma = 1.84 \cdot 10^2$	$\mu = 4.34 \cdot 10^2$ $\sigma = 1.66 \cdot 10^2$	$\mu = 5.70 \cdot 10^2$ $\sigma = 2.20 \cdot 10^2$	$\mu = 4.40 \cdot 10^2$ $\sigma = 1.03 \cdot 10^2$
18° C	PDG	$\mu = 5.13 \cdot 10^2$ $\sigma = 2.10 \cdot 10^2$	$\mu = 4.68 \cdot 10^2$ $\sigma = 1.92 \cdot 10^2$	$\mu = 5.02 \cdot 10^2$ $\sigma = 2.02 \cdot 10^2$	$\mu = 5.41 \cdot 10^2$ $\sigma = 2.24 \cdot 10^2$	$\mu = 4.04 \cdot 10^2$ $\sigma = 1.44 \cdot 10^2$
29° C	G	$\mu = 4.87 \cdot 10^2$ $\sigma = 1.96 \cdot 10^2$	$\mu = 5.53 \cdot 10^2$ $\sigma = 2.05 \cdot 10^2$	$\mu = 3.62 \cdot 10^2$ $\sigma = 1.59 \cdot 10^2$	$\mu = 5.81 \cdot 10^2$ $\sigma = 2.24 \cdot 10^2$	$\mu = 3.35 \cdot 10^2$ $\sigma = 1.17 \cdot 10^2$
29° C	DG	$\mu = 4.92 \cdot 10^2$ $\sigma = 1.89 \cdot 10^2$	$\mu = 4.33 \cdot 10^2$ $\sigma = 1.90 \cdot 10^2$	$\mu = 3.27 \cdot 10^2$ $\sigma = 1.37 \cdot 10^2$	$\mu = 5.84 \cdot 10^2$ $\sigma = 2.39 \cdot 10^2$	$\mu = 0.00 \cdot 10^0$ $\sigma = 0.00 \cdot 10^0$
29° C	PDG	$\mu = 4.76 \cdot 10^2$ $\sigma = 2.24 \cdot 10^2$	$\mu = 4.91 \cdot 10^2$ $\sigma = 1.82 \cdot 10^2$	$\mu = 3.73 \cdot 10^2$ $\sigma = 1.68 \cdot 10^2$	$\mu = 5.49 \cdot 10^2$ $\sigma = 2.35 \cdot 10^2$	$\mu = 3.38 \cdot 10^2$ $\sigma = 1.50 \cdot 10^2$

Table C.7. Averaged geometric distance from leaf to root for each family: the table contains the mean μ as well as the standard deviation σ.

				LoE [*px*]		
t	B_{Gen}	*WT*	*Betaint*	*Crumbs*	*Rab8*	*Talin*
18° C	G	$\mu = 1.44 \cdot 10^{2}$ $\sigma = 1.22 \cdot 10^{2}$	$\mu = 1.72 \cdot 10^{2}$ $\sigma = 1.31 \cdot 10^{2}$	$\mu = 9.87 \cdot 10^{1}$ $\sigma = 9.13 \cdot 10^{1}$	$\mu = 1.40 \cdot 10^{2}$ $\sigma = 1.22 \cdot 10^{2}$	$\mu = 9.37 \cdot 10^{1}$ $\sigma = 8.23 \cdot 10^{1}$
18° C	DG	$\mu = 1.44 \cdot 10^{2}$ $\sigma = 1.25 \cdot 10^{2}$	$\mu = 1.42 \cdot 10^{2}$ $\sigma = 1.05 \cdot 10^{2}$	$\mu = 8.23 \cdot 10^{1}$ $\sigma = 7.63 \cdot 10^{1}$	$\mu = 1.59 \cdot 10^{2}$ $\sigma = 1.16 \cdot 10^{2}$	$\mu = 2.12 \cdot 10^{2}$ $\sigma = 2.38 \cdot 10^{2}$
18° C	PDG	$\mu = 1.11 \cdot 10^{2}$ $\sigma = 1.03 \cdot 10^{2}$	$\mu = 1.51 \cdot 10^{2}$ $\sigma = 1.11 \cdot 10^{2}$	$\mu = 9.03 \cdot 10^{1}$ $\sigma = 7.76 \cdot 10^{1}$	$\mu = 1.67 \cdot 10^{2}$ $\sigma = 1.43 \cdot 10^{2}$	$\mu = 1.06 \cdot 10^{2}$ $\sigma = 1.34 \cdot 10^{2}$
29° C	G	$\mu = 9.99 \cdot 10^{1}$ $\sigma = 9.52 \cdot 10^{1}$	$\mu = 1.75 \cdot 10^{2}$ $\sigma = 1.48 \cdot 10^{2}$	$\mu = 7.83 \cdot 10^{1}$ $\sigma = 6.96 \cdot 10^{1}$	$\mu = 1.52 \cdot 10^{2}$ $\sigma = 1.35 \cdot 10^{2}$	$\mu = 8.36 \cdot 10^{1}$ $\sigma = 9.48 \cdot 10^{1}$
29° C	DG	$\mu = 1.11 \cdot 10^{2}$ $\sigma = 9.16 \cdot 10^{1}$	$\mu = 1.30 \cdot 10^{2}$ $\sigma = 1.13 \cdot 10^{2}$	$\mu = 1.14 \cdot 10^{2}$ $\sigma = 1.06 \cdot 10^{2}$	$\mu = 1.56 \cdot 10^{2}$ $\sigma = 2.77 \cdot 10^{2}$	$\mu = 0.00 \cdot 10^{0}$ $\sigma = 0.00 \cdot 10^{0}$
29° C	PDG	$\mu = 9.54 \cdot 10^{1}$ $\sigma = 8.75 \cdot 10^{1}$	$\mu = 1.40 \cdot 10^{2}$ $\sigma = 1.15 \cdot 10^{2}$	$\mu = 9.39 \cdot 10^{1}$ $\sigma = 8.36 \cdot 10^{1}$	$\mu = 1.22 \cdot 10^{2}$ $\sigma = 1.14 \cdot 10^{2}$	$\mu = 7.12 \cdot 10^{1}$ $\sigma = 7.18 \cdot 10^{1}$

Table C.8. Average edge length for each family: the table contains the mean μ as well as the standard deviation σ.

				RoE [*px*]		
t	B_{Gen}	*WT*	*Betaint*	*Crumbs*	*Rab8*	*Talin*
18° C	G	$\mu = 5.70 \cdot 10^{0}$ $\sigma = 2.84 \cdot 10^{0}$	$\mu = 5.18 \cdot 10^{0}$ $\sigma = 3.02 \cdot 10^{0}$	$\mu = 6.61 \cdot 10^{0}$ $\sigma = 4.18 \cdot 10^{0}$	$\mu = 5.48 \cdot 10^{0}$ $\sigma = 3.05 \cdot 10^{0}$	$\mu = 6.37 \cdot 10^{0}$ $\sigma = 3.50 \cdot 10^{0}$
18° C	DG	$\mu = 6.26 \cdot 10^{0}$ $\sigma = 3.05 \cdot 10^{0}$	$\mu = 5.76 \cdot 10^{0}$ $\sigma = 3.71 \cdot 10^{0}$	$\mu = 7.20 \cdot 10^{0}$ $\sigma = 5.11 \cdot 10^{0}$	$\mu = 5.31 \cdot 10^{0}$ $\sigma = 3.13 \cdot 10^{0}$	$\mu = 7.16 \cdot 10^{0}$ $\sigma = 2.77 \cdot 10^{0}$
18° C	PDG	$\mu = 6.03 \cdot 10^{0}$ $\sigma = 3.06 \cdot 10^{0}$	$\mu = 5.49 \cdot 10^{0}$ $\sigma = 2.59 \cdot 10^{0}$	$\mu = 6.06 \cdot 10^{0}$ $\sigma = 3.98 \cdot 10^{0}$	$\mu = 5.26 \cdot 10^{0}$ $\sigma = 2.69 \cdot 10^{0}$	$\mu = 5.55 \cdot 10^{0}$ $\sigma = 3.01 \cdot 10^{0}$
29° C	G	$\mu = 4.93 \cdot 10^{0}$ $\sigma = 3.04 \cdot 10^{0}$	$\mu = 5.33 \cdot 10^{0}$ $\sigma = 3.27 \cdot 10^{0}$	$\mu = 8.85 \cdot 10^{0}$ $\sigma = 5.42 \cdot 10^{0}$	$\mu = 4.91 \cdot 10^{0}$ $\sigma = 2.30 \cdot 10^{0}$	$\mu = 5.99 \cdot 10^{0}$ $\sigma = 2.99 \cdot 10^{0}$
29° C	DG	$\mu = 5.12 \cdot 10^{0}$ $\sigma = 3.34 \cdot 10^{0}$	$\mu = 6.73 \cdot 10^{0}$ $\sigma = 4.00 \cdot 10^{0}$	$\mu = 8.45 \cdot 10^{0}$ $\sigma = 5.16 \cdot 10^{0}$	$\mu = 5.94 \cdot 10^{0}$ $\sigma = 3.29 \cdot 10^{0}$	$\mu = 0.00 \cdot 10^{0}$ $\sigma = 0.00 \cdot 10^{0}$
29° C	PDG	$\mu = 4.76 \cdot 10^{0}$ $\sigma = 2.75 \cdot 10^{0}$	$\mu = 5.55 \cdot 10^{0}$ $\sigma = 3.04 \cdot 10^{0}$	$\mu = 7.28 \cdot 10^{0}$ $\sigma = 4.69 \cdot 10^{0}$	$\mu = 4.60 \cdot 10^{0}$ $\sigma = 2.28 \cdot 10^{0}$	$\mu = 6.05 \cdot 10^{0}$ $\sigma = 3.05 \cdot 10^{0}$

Table C.9. Average edge radius for each family: the table contains the mean μ as well as the standard deviation σ.

		$NA\ [px]$				
t	B_{Gen}	*WT*	*Betaint*	*Crumbs*	*Rab8*	*Talin*
18° C	G	$\mu = 1.39 \cdot 10^3$ $\sigma = 2.73 \cdot 10^3$	$\mu = 7.57 \cdot 10^2$ $\sigma = 1.67 \cdot 10^3$	$\mu = 2.05 \cdot 10^2$ $\sigma = 4.29 \cdot 10^2$	$\mu = 2.79 \cdot 10^3$ $\sigma = 6.33 \cdot 10^3$	$\mu = 3.47 \cdot 10^2$ $\sigma = 9.50 \cdot 10^2$
18° C	DG	$\mu = 1.07 \cdot 10^3$ $\sigma = 2.40 \cdot 10^3$	$\mu = 5.96 \cdot 10^2$ $\sigma = 1.71 \cdot 10^3$	$\mu = 2.13 \cdot 10^2$ $\sigma = 5.66 \cdot 10^2$	$\mu = 2.01 \cdot 10^3$ $\sigma = 4.26 \cdot 10^3$	$\mu = 1.21 \cdot 10^2$ $\sigma = 2.77 \cdot 10^2$
18° C	PDG	$\mu = 9.65 \cdot 10^2$ $\sigma = 2.43 \cdot 10^3$	$\mu = 4.49 \cdot 10^2$ $\sigma = 1.59 \cdot 10^3$	$\mu = 5.61 \cdot 10^2$ $\sigma = 1.51 \cdot 10^3$	$\mu = 1.20 \cdot 10^3$ $\sigma = 2.39 \cdot 10^3$	$\mu = 1.71 \cdot 10^2$ $\sigma = 4.99 \cdot 10^2$
29° C	G	$\mu = 1.83 \cdot 10^3$ $\sigma = 4.27 \cdot 10^3$	$\mu = 7.11 \cdot 10^2$ $\sigma = 1.51 \cdot 10^3$	$\mu = 1.99 \cdot 10^2$ $\sigma = 5.85 \cdot 10^2$	$\mu = 2.09 \cdot 10^3$ $\sigma = 4.41 \cdot 10^3$	$\mu = 1.28 \cdot 10^2$ $\sigma = 3.23 \cdot 10^2$
29° C	DG	$\mu = 5.40 \cdot 10^2$ $\sigma = 1.44 \cdot 10^3$	$\mu = 2.48 \cdot 10^2$ $\sigma = 6.63 \cdot 10^2$	$\mu = 1.37 \cdot 10^2$ $\sigma = 2.91 \cdot 10^2$	$\mu = 1.24 \cdot 10^3$ $\sigma = 3.82 \cdot 10^3$	$\mu = 0.00 \cdot 10^0$ $\sigma = 0.00 \cdot 10^0$
29° C	PDG	$\mu = 1.15 \cdot 10^3$ $\sigma = 2.63 \cdot 10^3$	$\mu = 7.46 \cdot 10^2$ $\sigma = 1.55 \cdot 10^3$	$\mu = 3.50 \cdot 10^2$ $\sigma = 2.01 \cdot 10^3$	$\mu = 1.88 \cdot 10^3$ $\sigma = 3.38 \cdot 10^3$	$\mu = 1.35 \cdot 10^2$ $\sigma = 2.82 \cdot 10^2$

Table C.10. Average normalized area of edges for each family: the table contains the mean μ as well as the standard deviation σ.

		D_N				
t	B_{Gen}	*WT*	*Betaint*	*Crumbs*	*Rab8*	*Talin*
18° C	G	$\mu = 1.84 \cdot 10^3$ $\sigma = 3.25 \cdot 10^3$	$\mu = 1.38 \cdot 10^3$ $\sigma = 2.30 \cdot 10^3$	$\mu = 5.51 \cdot 10^2$ $\sigma = 8.39 \cdot 10^2$	$\mu = 2.32 \cdot 10^3$ $\sigma = 4.13 \cdot 10^3$	$\mu = 7.04 \cdot 10^2$ $\sigma = 1.10 \cdot 10^3$
18° C	DG	$\mu = 1.35 \cdot 10^3$ $\sigma = 2.24 \cdot 10^3$	$\mu = 1.27 \cdot 10^3$ $\sigma = 2.25 \cdot 10^3$	$\mu = 5.13 \cdot 10^2$ $\sigma = 7.54 \cdot 10^2$	$\mu = 2.50 \cdot 10^3$ $\sigma = 4.11 \cdot 10^3$	$\mu = 3.55 \cdot 10^2$ $\sigma = 2.77 \cdot 10^2$
18° C	PDG	$\mu = 1.07 \cdot 10^3$ $\sigma = 1.95 \cdot 10^3$	$\mu = 8.63 \cdot 10^2$ $\sigma = 1.47 \cdot 10^3$	$\mu = 8.74 \cdot 10^2$ $\sigma = 1.42 \cdot 10^3$	$\mu = 2.09 \cdot 10^3$ $\sigma = 3.61 \cdot 10^3$	$\mu = 3.80 \cdot 10^2$ $\sigma = 4.31 \cdot 10^2$
29° C	G	$\mu = 2.09 \cdot 10^3$ $\sigma = 3.86 \cdot 10^3$	$\mu = 1.30 \cdot 10^3$ $\sigma = 2.23 \cdot 10^3$	$\mu = 4.17 \cdot 10^2$ $\sigma = 5.94 \cdot 10^2$	$\mu = 2.58 \cdot 10^3$ $\sigma = 4.18 \cdot 10^3$	$\mu = 3.52 \cdot 10^2$ $\sigma = 4.71 \cdot 10^2$
29° C	DG	$\mu = 9.08 \cdot 10^2$ $\sigma = 1.51 \cdot 10^3$	$\mu = 6.95 \cdot 10^2$ $\sigma = 1.46 \cdot 10^3$	$\mu = 3.46 \cdot 10^2$ $\sigma = 4.06 \cdot 10^2$	$\mu = 2.67 \cdot 10^3$ $\sigma = 4.44 \cdot 10^3$	$\mu = 0.00 \cdot 10^0$ $\sigma = 0.00 \cdot 10^0$
29° C	PDG	$\mu = 1.66 \cdot 10^3$ $\sigma = 2.99 \cdot 10^3$	$\mu = 1.15 \cdot 10^3$ $\sigma = 1.97 \cdot 10^3$	$\mu = 5.57 \cdot 10^2$ $\sigma = 1.05 \cdot 10^3$	$\mu = 2.36 \cdot 10^3$ $\sigma = 3.80 \cdot 10^3$	$\mu = 4.22 \cdot 10^2$ $\sigma = 6.65 \cdot 10^2$

Table C.11. Averaged normalized distance from leaf to root for each family: the table contains the mean μ as well as the standard deviation σ.

				D_{G_V} $[px]$		
t	B_{Gen}	*WT*	*Betaint*	*Crumbs*	*Rab8*	*Talin*
18° C	G	$\mu = 6.22 \cdot 10^0$ $\sigma = 3.44 \cdot 10^0$	$\mu = 5.84 \cdot 10^0$ $\sigma = 3.85 \cdot 10^0$	$\mu = 6.74 \cdot 10^0$ $\sigma = 4.70 \cdot 10^0$	$\mu = 5.94 \cdot 10^0$ $\sigma = 3.72 \cdot 10^0$	$\mu = 6.51 \cdot 10^0$ $\sigma = 4.09 \cdot 10^0$
18° C	DG	$\mu = 6.73 \cdot 10^0$ $\sigma = 3.86 \cdot 10^0$	$\mu = 6.34 \cdot 10^0$ $\sigma = 4.26 \cdot 10^0$	$\mu = 7.26 \cdot 10^0$ $\sigma = 5.55 \cdot 10^0$	$\mu = 5.75 \cdot 10^0$ $\sigma = 3.77 \cdot 10^0$	$\mu = 6.62 \cdot 10^0$ $\sigma = 3.56 \cdot 10^0$
18° C	PDG	$\mu = 6.36 \cdot 10^0$ $\sigma = 3.59 \cdot 10^0$	$\mu = 5.94 \cdot 10^0$ $\sigma = 3.49 \cdot 10^0$	$\mu = 6.17 \cdot 10^0$ $\sigma = 4.38 \cdot 10^0$	$\mu = 5.67 \cdot 10^0$ $\sigma = 3.54 \cdot 10^0$	$\mu = 5.79 \cdot 10^0$ $\sigma = 3.49 \cdot 10^0$
29° C	G	$\mu = 5.41 \cdot 10^0$ $\sigma = 3.66 \cdot 10^0$	$\mu = 6.15 \cdot 10^0$ $\sigma = 4.07 \cdot 10^0$	$\mu = 8.62 \cdot 10^0$ $\sigma = 6.22 \cdot 10^0$	$\mu = 5.51 \cdot 10^0$ $\sigma = 3.03 \cdot 10^0$	$\mu = 6.16 \cdot 10^0$ $\sigma = 3.78 \cdot 10^0$
29° C	DG	$\mu = 5.69 \cdot 10^0$ $\sigma = 4.28 \cdot 10^0$	$\mu = 7.25 \cdot 10^0$ $\sigma = 4.98 \cdot 10^0$	$\mu = 8.17 \cdot 10^0$ $\sigma = 5.99 \cdot 10^0$	$\mu = 5.71 \cdot 10^0$ $\sigma = 3.68 \cdot 10^0$	$\mu = 0.00 \cdot 10^0$ $\sigma = 0.00 \cdot 10^0$
29° C	PDG	$\mu = 5.08 \cdot 10^0$ $\sigma = 3.38 \cdot 10^0$	$\mu = 6.14 \cdot 10^0$ $\sigma = 3.83 \cdot 10^0$	$\mu = 7.39 \cdot 10^0$ $\sigma = 5.30 \cdot 10^0$	$\mu = 5.10 \cdot 10^0$ $\sigma = 2.95 \cdot 10^0$	$\mu = 6.15 \cdot 10^0$ $\sigma = 3.70 \cdot 10^0$

Table C.12. Averaged geometric distance from each node to the root for each family: the table contains the mean μ as well as the standard deviation σ.

				D_{T_V} $[px]$		
t	B_{Gen}	*WT*	*Betaint*	*Crumbs*	*Rab8*	*Talin*
18° C	G	$\mu = 6.22 \cdot 10^0$ $\sigma = 3.44 \cdot 10^0$	$\mu = 5.84 \cdot 10^0$ $\sigma = 3.85 \cdot 10^0$	$\mu = 6.74 \cdot 10^0$ $\sigma = 4.70 \cdot 10^0$	$\mu = 5.94 \cdot 10^0$ $\sigma = 3.72 \cdot 10^0$	$\mu = 6.51 \cdot 10^0$ $\sigma = 4.09 \cdot 10^0$
18° C	DG	$\mu = 6.73 \cdot 10^0$ $\sigma = 3.86 \cdot 10^0$	$\mu = 6.34 \cdot 10^0$ $\sigma = 4.26 \cdot 10^0$	$\mu = 7.26 \cdot 10^0$ $\sigma = 5.55 \cdot 10^0$	$\mu = 5.75 \cdot 10^0$ $\sigma = 3.77 \cdot 10^0$	$\mu = 6.62 \cdot 10^0$ $\sigma = 3.56 \cdot 10^0$
18° C	PDG	$\mu = 6.36 \cdot 10^0$ $\sigma = 3.59 \cdot 10^0$	$\mu = 5.94 \cdot 10^0$ $\sigma = 3.49 \cdot 10^0$	$\mu = 6.17 \cdot 10^0$ $\sigma = 4.38 \cdot 10^0$	$\mu = 5.67 \cdot 10^0$ $\sigma = 3.54 \cdot 10^0$	$\mu = 5.79 \cdot 10^0$ $\sigma = 3.49 \cdot 10^0$
29° C	G	$\mu = 5.41 \cdot 10^0$ $\sigma = 3.66 \cdot 10^0$	$\mu = 6.15 \cdot 10^0$ $\sigma = 4.07 \cdot 10^0$	$\mu = 8.62 \cdot 10^0$ $\sigma = 6.22 \cdot 10^0$	$\mu = 5.51 \cdot 10^0$ $\sigma = 3.03 \cdot 10^0$	$\mu = 6.16 \cdot 10^0$ $\sigma = 3.78 \cdot 10^0$
29° C	DG	$\mu = 5.69 \cdot 10^0$ $\sigma = 4.28 \cdot 10^0$	$\mu = 7.25 \cdot 10^0$ $\sigma = 4.98 \cdot 10^0$	$\mu = 8.17 \cdot 10^0$ $\sigma = 5.99 \cdot 10^0$	$\mu = 5.71 \cdot 10^0$ $\sigma = 3.68 \cdot 10^0$	$\mu = 0.00 \cdot 10^0$ $\sigma = 0.00 \cdot 10^0$
29° C	PDG	$\mu = 5.08 \cdot 10^0$ $\sigma = 3.38 \cdot 10^0$	$\mu = 6.14 \cdot 10^0$ $\sigma = 3.83 \cdot 10^0$	$\mu = 7.39 \cdot 10^0$ $\sigma = 5.30 \cdot 10^0$	$\mu = 5.10 \cdot 10^0$ $\sigma = 2.95 \cdot 10^0$	$\mu = 6.15 \cdot 10^0$ $\sigma = 3.70 \cdot 10^0$

Table C.13. Averaged tree-distance from each node to the root for each family: the table contains the mean μ as well as the standard deviation σ.

Depth						
t	B_{Gen}	*WT*	*Betaint*	*Crumbs*	*Rab8*	*Talin*
18° C	G	$\mu = 1.22 \cdot 10^1$ $\sigma = 2.13 \cdot 10^0$	$\mu = 9.89 \cdot 10^0$ $\sigma = 1.70 \cdot 10^0$	$\mu = 9.62 \cdot 10^0$ $\sigma = 1.69 \cdot 10^0$	$\mu = 1.15 \cdot 10^1$ $\sigma = 1.79 \cdot 10^0$	$\mu = 9.00 \cdot 10^0$ $\sigma = 2.03 \cdot 10^0$
18° C	DG	$\mu = 1.08 \cdot 10^1$ $\sigma = 1.74 \cdot 10^0$	$\mu = 9.92 \cdot 10^0$ $\sigma = 2.34 \cdot 10^0$	$\mu = 1.04 \cdot 10^1$ $\sigma = 1.91 \cdot 10^0$	$\mu = 1.29 \cdot 10^1$ $\sigma = 1.85 \cdot 10^0$	$\mu = 2.17 \cdot 10^0$ $\sigma = 1.71 \cdot 10^0$
18° C	PDG	$\mu = 1.16 \cdot 10^1$ $\sigma = 1.83 \cdot 10^0$	$\mu = 7.47 \cdot 10^0$ $\sigma = 1.57 \cdot 10^0$	$\mu = 1.05 \cdot 10^1$ $\sigma = 2.95 \cdot 10^0$	$\mu = 1.01 \cdot 10^1$ $\sigma = 2.04 \cdot 10^0$	$\mu = 4.81 \cdot 10^0$ $\sigma = 2.21 \cdot 10^0$
29° C	G	$\mu = 1.25 \cdot 10^1$ $\sigma = 1.89 \cdot 10^0$	$\mu = 9.50 \cdot 10^0$ $\sigma = 1.32 \cdot 10^0$	$\mu = 9.26 \cdot 10^0$ $\sigma = 1.55 \cdot 10^0$	$\mu = 1.25 \cdot 10^1$ $\sigma = 1.43 \cdot 10^0$	$\mu = 5.39 \cdot 10^0$ $\sigma = 2.79 \cdot 10^0$
29° C	DG	$\mu = 9.06 \cdot 10^0$ $\sigma = 2.32 \cdot 10^0$	$\mu = 7.48 \cdot 10^0$ $\sigma = 1.69 \cdot 10^0$	$\mu = 5.40 \cdot 10^0$ $\sigma = 1.46 \cdot 10^0$	$\mu = 1.32 \cdot 10^1$ $\sigma = 2.90 \cdot 10^0$	$\mu = 0.00 \cdot 10^0$ $\sigma = 0.00 \cdot 10^0$
29° C	PDG	$\mu = 1.18 \cdot 10^1$ $\sigma = 1.60 \cdot 10^0$	$\mu = 8.62 \cdot 10^0$ $\sigma = 1.33 \cdot 10^0$	$\mu = 7.83 \cdot 10^0$ $\sigma = 2.66 \cdot 10^0$	$\mu = 1.26 \cdot 10^1$ $\sigma = 1.89 \cdot 10^0$	$\mu = 5.83 \cdot 10^0$ $\sigma = 2.70 \cdot 10^0$

Table C.14. Depth of the tree for each family: the table contains the mean μ as well as the standard deviation σ.

Asymmetry						
t	B_{Gen}	*WT*	*Betaint*	*Crumbs*	*Rab8*	*Talin*
18° C	G	$\mu = 4.10 \cdot 10^1$ $\sigma = 4.81 \cdot 10^2$	$\mu = 4.35 \cdot 10^1$ $\sigma = 4.81 \cdot 10^2$	$\mu = 4.43 \cdot 10^1$ $\sigma = 4.99 \cdot 10^2$	$\mu = 3.98 \cdot 10^1$ $\sigma = 3.31 \cdot 10^2$	$\mu = 4.03 \cdot 10^1$ $\sigma = 6.93 \cdot 10^2$
18° C	DG	$\mu = 4.29 \cdot 10^1$ $\sigma = 4.92 \cdot 10^2$	$\mu = 4.25 \cdot 10^1$ $\sigma = 8.22 \cdot 10^2$	$\mu = 4.64 \cdot 10^1$ $\sigma = 7.52 \cdot 10^2$	$\mu = 4.28 \cdot 10^1$ $\sigma = 5.11 \cdot 10^2$	$\mu = 5.23 \cdot 10^1$ $\sigma = 5.85 \cdot 10^2$
18° C	PDG	$\mu = 4.41 \cdot 10^1$ $\sigma = 4.78 \cdot 10^2$	$\mu = 4.28 \cdot 10^1$ $\sigma = 8.43 \cdot 10^2$	$\mu = 4.32 \cdot 10^1$ $\sigma = 5.86 \cdot 10^2$	$\mu = 4.05 \cdot 10^1$ $\sigma = 5.18 \cdot 10^2$	$\mu = 4.89 \cdot 10^1$ $\sigma = 9.80 \cdot 10^2$
29° C	G	$\mu = 3.90 \cdot 10^1$ $\sigma = 5.60 \cdot 10^2$	$\mu = 4.23 \cdot 10^1$ $\sigma = 6.10 \cdot 10^2$	$\mu = 4.66 \cdot 10^1$ $\sigma = 7.30 \cdot 10^2$	$\mu = 4.15 \cdot 10^1$ $\sigma = 5.45 \cdot 10^2$	$\mu = 5.31 \cdot 10^1$ $\sigma = 9.96 \cdot 10^2$
29° C	DG	$\mu = 3.97 \cdot 10^1$ $\sigma = 5.66 \cdot 10^2$	$\mu = 4.45 \cdot 10^1$ $\sigma = 8.16 \cdot 10^2$	$\mu = 5.28 \cdot 10^1$ $\sigma = 9.49 \cdot 10^2$	$\mu = 4.29 \cdot 10^1$ $\sigma = 5.01 \cdot 10^2$	$\mu = 0.00 \cdot 10^0$ $\sigma = 0.00 \cdot 10^0$
29° C	PDG	$\mu = 4.16 \cdot 10^1$ $\sigma = 6.89 \cdot 10^2$	$\mu = 3.82 \cdot 10^1$ $\sigma = 5.69 \cdot 10^2$	$\mu = 5.08 \cdot 10^1$ $\sigma = 8.65 \cdot 10^2$	$\mu = 3.85 \cdot 10^1$ $\sigma = 5.10 \cdot 10^2$	$\mu = 4.80 \cdot 10^1$ $\sigma = 8.68 \cdot 10^2$

Table C.15. Asymmetry of the tree for each family: the table contains the mean μ as well as the standard deviation σ.

Cpl						
t	B_{Gen}	*WT*	*Betaint*	*Crumbs*	*Rab8*	*Talin*
18° C	G	$\mu = 7.15 \cdot 10^3$ $\sigma = 8.79 \cdot 10^3$	$\mu = 1.61 \cdot 10^2$ $\sigma = 1.17 \cdot 10^2$	$\mu = 1.85 \cdot 10^2$ $\sigma = 2.09 \cdot 10^2$	$\mu = 1.01 \cdot 10^2$ $\sigma = 1.05 \cdot 10^2$	$\mu = 2.94 \cdot 10^2$ $\sigma = 2.40 \cdot 10^2$
18° C	DG	$\mu = 1.29 \cdot 10^2$ $\sigma = 1.57 \cdot 10^2$	$\mu = 2.33 \cdot 10^2$ $\sigma = 1.98 \cdot 10^2$	$\mu = 1.41 \cdot 10^2$ $\sigma = 1.96 \cdot 10^2$	$\mu = 4.66 \cdot 10^3$ $\sigma = 6.41 \cdot 10^3$	$\mu = 2.12 \cdot 10^1$ $\sigma = 7.01 \cdot 10^2$
18° C	PDG	$\mu = 6.76 \cdot 10^3$ $\sigma = 4.96 \cdot 10^3$	$\mu = 4.68 \cdot 10^2$ $\sigma = 3.03 \cdot 10^2$	$\mu = 1.83 \cdot 10^2$ $\sigma = 1.69 \cdot 10^2$	$\mu = 2.09 \cdot 10^2$ $\sigma = 2.07 \cdot 10^2$	$\mu = 1.24 \cdot 10^1$ $\sigma = 8.60 \cdot 10^2$
29° C	G	$\mu = 6.29 \cdot 10^3$ $\sigma = 4.57 \cdot 10^3$	$\mu = 2.04 \cdot 10^2$ $\sigma = 1.50 \cdot 10^2$	$\mu = 1.70 \cdot 10^2$ $\sigma = 1.36 \cdot 10^2$	$\mu = 5.47 \cdot 10^3$ $\sigma = 4.35 \cdot 10^3$	$\mu = 1.05 \cdot 10^1$ $\sigma = 9.04 \cdot 10^2$
29° C	DG	$\mu = 3.59 \cdot 10^2$ $\sigma = 3.68 \cdot 10^2$	$\mu = 4.78 \cdot 10^2$ $\sigma = 3.31 \cdot 10^2$	$\mu = 8.47 \cdot 10^2$ $\sigma = 6.27 \cdot 10^2$	$\mu = 7.06 \cdot 10^3$ $\sigma = 8.65 \cdot 10^3$	$\mu = 0.00 \cdot 10^0$ $\sigma = 0.00 \cdot 10^0$
29° C	PDG	$\mu = 8.52 \cdot 10^3$ $\sigma = 6.60 \cdot 10^3$	$\mu = 3.52 \cdot 10^2$ $\sigma = 2.62 \cdot 10^2$	$\mu = 4.24 \cdot 10^2$ $\sigma = 4.06 \cdot 10^2$	$\mu = 6.48 \cdot 10^3$ $\sigma = 6.84 \cdot 10^3$	$\mu = 9.68 \cdot 10^2$ $\sigma = 7.44 \cdot 10^2$

Table C.16. Completeness of the tree for each family: the table contains the mean μ as well as the standard deviation σ.

Sfn						
t	B_{Gen}	*WT*	*Betaint*	*Crumbs*	*Rab8*	*Talin*
18° C	G	$\mu = 7.24 \cdot 10^2$ $\sigma = 9.89 \cdot 10^3$	$\mu = 4.94 \cdot 10^2$ $\sigma = 8.20 \cdot 10^3$	$\mu = 7.34 \cdot 10^2$ $\sigma = 1.28 \cdot 10^2$	$\mu = 7.05 \cdot 10^2$ $\sigma = 1.50 \cdot 10^2$	$\mu = 7.40 \cdot 10^2$ $\sigma = 1.20 \cdot 10^2$
18° C	DG	$\mu = 7.69 \cdot 10^2$ $\sigma = 1.33 \cdot 10^2$	$\mu = 5.66 \cdot 10^2$ $\sigma = 8.54 \cdot 10^3$	$\mu = 8.27 \cdot 10^2$ $\sigma = 1.23 \cdot 10^2$	$\mu = 6.95 \cdot 10^2$ $\sigma = 8.10 \cdot 10^3$	$\mu = 1.46 \cdot 10^1$ $\sigma = 9.19 \cdot 10^2$
18° C	PDG	$\mu = 7.55 \cdot 10^2$ $\sigma = 7.53 \cdot 10^3$	$\mu = 5.63 \cdot 10^2$ $\sigma = 7.21 \cdot 10^3$	$\mu = 7.87 \cdot 10^2$ $\sigma = 1.49 \cdot 10^2$	$\mu = 6.55 \cdot 10^2$ $\sigma = 1.35 \cdot 10^2$	$\mu = 8.38 \cdot 10^2$ $\sigma = 1.80 \cdot 10^2$
29° C	G	$\mu = 8.10 \cdot 10^2$ $\sigma = 1.63 \cdot 10^2$	$\mu = 5.47 \cdot 10^2$ $\sigma = 7.67 \cdot 10^3$	$\mu = 1.07 \cdot 10^1$ $\sigma = 1.58 \cdot 10^2$	$\mu = 7.11 \cdot 10^2$ $\sigma = 8.75 \cdot 10^3$	$\mu = 1.04 \cdot 10^1$ $\sigma = 3.07 \cdot 10^2$
29° C	DG	$\mu = 6.42 \cdot 10^2$ $\sigma = 9.10 \cdot 10^3$	$\mu = 6.44 \cdot 10^2$ $\sigma = 9.25 \cdot 10^3$	$\mu = 1.14 \cdot 10^1$ $\sigma = 4.19 \cdot 10^2$	$\mu = 1.15 \cdot 10^1$ $\sigma = 9.62 \cdot 10^2$	$\mu = 0.00 \cdot 10^0$ $\sigma = 0.00 \cdot 10^0$
29° C	PDG	$\mu = 8.17 \cdot 10^2$ $\sigma = 1.19 \cdot 10^2$	$\mu = 6.20 \cdot 10^2$ $\sigma = 7.70 \cdot 10^3$	$\mu = 8.33 \cdot 10^2$ $\sigma = 1.73 \cdot 10^2$	$\mu = 7.04 \cdot 10^2$ $\sigma = 8.24 \cdot 10^3$	$\mu = 1.96 \cdot 10^1$ $\sigma = 2.09 \cdot 10^1$

Table C.17. Spacefillingness of the network for each family: the table contains the mean μ as well as the standard deviation σ.

Bibliography

[1] E. Mayorga et al. *Amazon River Basin Land and Stream Drainage Direction Maps.* Nov. 2014. URL: `http://daac.ornl.gov`.

[2] A. Jazwinska, C. Ribeiro, and M. Affolter. "Epithelial tube morphogenesis during *Drosophila* tracheal development requires Piopio, a luminal ZP protein." In: *Nature Cell Biol.* 5 (2003), pp. 895–901.

[3] M. L. Keister. "The morphogenesis of the tracheal system of *Sciara*". In: *J. Morphol.* 83 (1948), pp. 373–423.

[4] S. A. Shafiq. "Electron Microscopy of the development of tracheoles in *Drosophila melanogaster*". In: *Q. J. Mocrosc. Sci.* 104 (1963), pp. 135–140.

[5] D. M. Bryant et al. "A molecular network for *de novo* generation of the apical surface and lumen." In: *Nature Cell Biol.* 12 (2010), pp. 1035–1045.

[6] E. Siegfried and N. Perrimon. "Drosophila wingless: A paradigm for the function and mechanism of Wnt signaling." In: *BioEssays* 16 (2005), pp. 395–404.

[7] O. Sayeed and S. Benzer. "Behavioral genetics of thermosensation and hygrosensation in Drosophila." In: *PNAS* 93 (1996), pp. 6079–6084.

[8] E. Segal et al. "Predicting expression patterns from regulatory sequence in Drosophila segmentation." In: *nature* 451 (2008), pp. 535–540.

[9] Genotype. *Encyclopaedia Britannica Online.* Encyclopaedia Britannica Inc., 2014. URL: `http://www.britannica.com/EBchecked/topic/229258/genotype`.

[10] Phenotype. *Encyclopaedia Britannica Online.* Encyclopaedia Britannica Inc., 2014. URL: `http://www.britannica.com/EBchecked/topic/455632/phenotype`.

[11] H. Ronellenfitsch et al. "Topological phenotypes of leaf vascular networks". In: *in preparation* (2014).

[12] L. Lam, S. W. Lee, and C.Y. Suen. “Thinning Methodologies-A Comprehensive Survey.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14 (1992), pp. 869–885.

[13] K. C. Fan, D. F. Chen, and M. G. Wen. “A new vectorization-based approach to the skeletonization of binary images.” In: *Proceedings of 3rd International Conference on Document Analysis and Recognition* (1995), p. 627.

[14] Jana Lasser. “Framework zur computergestützten Analyse von Bildern vaskulärer Netzwerke”. B. S. Thesis. Germany: Georg-August University Göttingen, 2012.

[15] N. Otsu. “A threshold selection method from gray-level histograms”. In: *IEEE Trans. Sys., Man., Cyber* 9 (1979), pp. 62–66.

[16] L. P. Chew. “Constrained delaunay triangulations.” In: *Algorithmica* 4 (1989), pp. 97–108.

[17] S. Suzuki and K. Abe. “Topological Structural Analysis of Digitized Binary Images by Border Following.” In: *CVGIP* 30 (1985), pp. 32–46.

[18] C. H. Teh and R. T. Chin. “On the Detection of Dominant Points on Digital Curve.” In: *PAMI* 11 (1989), pp. 859–872.

[19] P. E. Danielsson. “Euclidean distance mapping.” In: *Computer Graphics and Image Processing* 14 (1980), pp. 227–248.

[20] A. A. Hagberg, D. A. Schult, and P. J. Swart. “Exploring network structure, dynamics, and function using NetworkX.” In: *Proceedings of the 7th Python in Science Conference (SciPy2008)* (2008), pp. 11–15.

[21] J. D. Hunter. “Matplotlib: A 2D graphics environment”. In: *Computing In Science & Engineering* 9 (2007), pp. 90–95.

[22] R. Diestel, ed. *Graphentheorie.* Heidelberg: Springer, 2010.

[23] Jean Serra. *Image Analysis and Mathematical Morphology.* Academic Pr, 1983.

[24] S. Sigurbjoernsdottir. “Complex cell shape: Molecular mechanisms of tracheal terminal cell development in Drosophila melanogaster.” PhD thesis. EMBL Heidelberg, 2014.

[25] T.A. Jones and M.M. Metzstein. “Examination of Drosophila larval tracheal terminal cells by light microscopy.” In: *J Vis Exp* 77 (2013).

[26] A.S. Ghabrial, B.P. Levi, and M.A. Krasnow. “A systematic screen for tube morphogenesis and branching genes in the Drosophila tracheal system.” In: *PLoS genetics* 7 (2011).

[27] R. A. Fischer. “The Use of Multiple Measurements in Taxonomic Problems.” In: *Annals of Eugenics* 7 (1936), pp. 179–188.

[28] R. C. Rao. “The utilization of multiple measurements in problems of biological classification.” In: *Journal of the Royal Statistical Society, Series B* 10 (1948), pp. 159–203.

[29] P. C. Mahalanobis. “On the generalised distance in statistics.” In: *Proceedings of the National Institute of Sciences of India* 2 (1936), pp. 49–55.

[30] D. MacKay, ed. Cambridge: Cambridge University Press, 2003. Chap. 20. An Example Inference Task: Clustering.

[31] D. Barber. *Bayesian Reasoning and Machine Learning.* Cambridge University Press, 2012. Chap. 20.3: The Gaussian Mixture Model.

[32] P. J. Rousseeuw. “Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis.” In: *Computational and Applied Mathematics* 20 (1987), pp. 53–65.

[33] D. Barber. *Bayesian Reasoning and Machine Learning.* Cambridge University Press, 2012. Chap. 15.2: Principal Components Analysis.

[34] H. Lilliefors. “On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown.” In: *Journal of the American Statistical Association.* 62 (1967), pp. 399–402.

[35] R. L. Edgeman and R. C. Scott. “Lilliefors tests for transformed variables.” In: *Brazilian Journal of Probability and Statistics* 1 (1987), pp. 101–112.

[36] C. A. Long. “Leonardo da Vinci’s Rule and Fractal Complexity in Dichotomous Trees.” In: *Journal of Theoretical Biology* 167 (1994), pp. 107–113.

[37] K. V. Mardia and P. E. Jupp. *Directional Statistics.* Wiley, 1999. ISBN: 0471953334.

[38] K. S. Johnson and R. V. Barbehenn. “Oxygen levels in the gut lumens of herbivorous insects.” In: *Journal of Insect Physiology* 46 (2000), pp. 897–903.

[39] M. Galvez-Santisteban et al. “Synaptotagmin-like proteins control the formation of a single apical membrane domain in epithelial cells.” In: *Nature Cell Biol.* 14 (2012), pp. 838–849.

[40] C. Cabernard, M. Neumann, and M. Affolter. “Cellular and molecular mechanisms involved in branching morphogenesis of the Drosophila tracheal system.” In: *Journal of applied physiology* 97 (2004), pp. 2347–2353.

[41] T. D. Pollard, W. C. Earnshaw, and J. Lippineott-Schwartz. *Cell Biology.* W. B. Saunders Co.., 2008.

[42] L. A. Huber et al. “Rab8, a small GTPase involved in vesicular traffic between the TGN and the basolateral plasma membrane.” In: *JBC* 123 (1993), pp. 35–35.

[43] A. J. MacKrell et al. “The lethal myospheroid gene of Drosophila encodes a membrane protein homologous to vertebrate integrin beta subunits.” In: *PNAS* 85 (1988), pp. 2633–2637.

[44] A. Wodarz et al. “Expression of crumbs confers apical character on plasma membrane domains of ectodermal epithelia of drosophila.” In: *Cell* 82 (1995), pp. 67–76.

[45] N. H. Brown et al. “Talin Is Essential for Integrin Function in Drosophila.” In: *Developmental Cell* 3 (2002), pp. 569–579.

[46] M. Chalfie. “Green Fluorescent Protein.” In: *Photochemistry and Photobiology* 62 (1995), pp. 651–656.

[47] Y. S. Lee et al. “Distinct Roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA Silencing Pathways.” In: *Cell* 117 (2004), pp. 69–81.

[48] R. Wong et al. “PIP2 Hydrolysis and Calcium Release Are Required for Cytokinesis in Drosophila Spermatocytes.” In: *Current Biology* 15 (2005), pp. 1401–1406.

[49] H. H. Harman. “Modern Factor Analysis.” In: *University of Chicago Press.* (1976), pp. 175–176.

[50] T. P. Minka. “Automatic Choice of Dimensionality for PCA.” In: *NIPS* (2000), pp. 598–604.

[51] D. Barber. *Bayesian Reasoning and Machine Learning.* Cambridge University Press, 2012. Chap. 16.2: Fischer’s Linear Discriminang.

[52] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[53] D. Arthur and S. Vassilvitskii. “k-means++: the advantages of careful seeding.” In: *ACM-SIAM symposium on Discrete algorithms* (2007).

Erklärung nach §18(8) der Prüfungsordnung für den Bachelor-Studiengang Physik und den Master-Studiengang Physik an der Universität Göttingen:

Hiermit erkläre ich, dass ich diese Abschlussarbeit selbständig verfasst habe, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe und alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen wurden, als solche kenntlich gemacht habe.
Darüberhinaus erkläre ich, dass diese Abschlussarbeit nicht, auch nicht auszugsweise, im Rahmen einer nichtbestandenen Prüfung an dieser oder einer anderen Hochschule eingereicht wurde.

Göttingen, den February 10, 2015

(Jana Lasser)