Streamline Classification in Human Brain Structural Connectivity

A Thesis submitted in partial satisfaction of the requirements for the degree Bachelors of Sciences in Physics

by

Daniela Koch

Committee in charge:

Dr. Jean Carlson
Dr. Wendy Meiring

June 2017
The thesis of Daniela Koch is approved.

______________________________________________
Jean Carlson

______________________________________________
Wendy Meiring

June 2017
Streamline Classification in Human Brain Structural Connectivity

Copyright © 2017

by

Daniela Koch
# Contents

## List of Figures

## 1 Introduction

1.1 Basic Brain Anatomy ............................................. 2
1.2 Structural Connectivity ........................................ 3
1.3 Motivation ..................................................... 5

## 2 Technical Background Information

2.1 Constructing a Brain Network ................................. 6
2.2 Data Acquisition .............................................. 11

## 3 Methods

3.1 Big Picture .................................................. 20
3.2 Geometric Properties of False Positive Data ................ 21
3.3 Radius Required ............................................. 23
3.4 Functional Data Analysis .................................. 28

## 4 Results and Discussion

4.1 Results ..................................................... 34
4.2 Discussion and Conclusion ................................ 38
4.3 Outlook ................................................... 39

## References

41
List of Figures

1.2 Distribution of grey at white matter regions (a) shows this distribution in a cross-section of the brain. (b) shows how grey matter regions are connected by white matter fibers (axons). ................................................................. 3

1.3 Structural connectivity matrix is an adjacency matrix $A$ with the color of elements $A_{ij}$ determined by the strength of connection between nodes $i$ and $j$. (a) shows a structural network with 250 nodes while (b) shows network with only 50 nodes, highlighting the effect of node choice in the construction of a network. Figure sourced from [1]. ................................................................. 4

2.3 Visualization of directed and weighted networks. (a) shows a directed network where the arrows represent the direction of the edge. (b) shows a weighted directed network, in which the weight of each edge is represented by the thickness of the arrow. ................................................................. 10

2.4 (a) Highlighted regions of interest (ROI’s) (b) White matter fibers connecting those pairwise ROI’s. ................................................................. 11

2.6 Diffusion ellipsoid visualizing the anisotropy and orientation of the diffusion tensor. (Figure sourced from ................................................................. 14

2.7 Voxel diffusion tensor based on diffusion patterns on the sub-voxel scale. (a) shows a coherent diffusion pattern within the voxel, resulting in a highly anisotropic voxel tensor. (b) shows the voxel tensor for a bending diffusion pattern. Both (c) and (d) show the sub-voxel diffusion patterns that can result in an isotropic voxel tensor. Image sourced from [2]. ................................................................. 15

2.8 A diffusion tensor is calculated for each voxel in the brain. The diffusion tensor describes the overall anisotropy and orientation of water flow within that voxel. All the diffusion tensors together create a three dimensional vector field describing the water diffusion pattern throughout the brain. Figures sourced from [3]. ................................................................. 16

2.9 Streamline following path as directed by diffusion tensor. ................................................................. 17

2.10 Streamlines of entire brain generated using DSI studio software with 100,000 seeds. Color indicates direction of streamline. Red indicates left/right, blue indicates superior/inferior (up/down), and green indicates anterior/posterior (front/back). 18
3.1 Example Streamline exhibiting excessive winding. (a) and (b) show same streamline from different angles. ............................................. 21

3.2 Illustration of removing false positive streamlines. (a) shows original streamline data, (b) shows connection with spurious streamlines removed. ................. 22

3.3 (a) Illustration of the radius required for sphere to capture given percent for a spurious streamline (i = 86, shown in orange). The required sphere is small where the streamline is still closely bundled with the others, but increases as it deviates from the others. (b) Shows the radius of the sphere in (a) as a function of streamline index j. Note that the curve $R_i(j)$ is still a curve of discrete data points

3.4 Radius Required for streamline which successfully bundles with other streamlines in connection. (a) is an illustration of required sphere size for a tightly bundled streamline (i = 2) along its trajectory. (b) shows the required radius of sphere as a function of streamline index j. .......................................................... 26

3.5 Radius algorithm that generates curves $R_i(j)$ is iterated over all streamlines in the connection. Each curve corresponds to one streamline. ......................... 27

3.6 Comparison between unregistered and registered curves. (a) Shows the original radius curves $R_i(j)$ (same as Figure 3.5). (b) Shows the registered radius curves $R_i^*(j)$ that are a result of the registration and cutting process. All data has been cut to have 250 indices, the same number of indices as the shortest streamlines in this connection. The expected bundling tendencies are more apparent after registration as curves are now aligned spatially rather than by index. .................... 29

3.7 Functional box plot indicating statistical outliers. (a) shows registered and cut radius curves $R_i^*(j)$. (b) shows functional boxplot of registered curves as generated by Matlab code. The pink region indicates the interquartile range of depth measures (deepest 50% of curves). The black line indicates the median curve while the outliers are indicated in red. ......................................................... 31

3.8 Statistical outliers as identified by FDA. (a) shows the original connection. (b) shows the streamlines identified by FDA to be outliers ................................. 32

4.1 Example where algorithm extracted those streamlines which had been visually identified as biologically inviable. (a) shows the original streamline connection, (b) shows the functional boxplot of the registered radius curves for the streamlines in (a). ............................................................... 34

4.2 A biologically viable connection. (a) shows the streamline data, all of which appear to be biologically viable. (b) shows the functional boxplot of the registered radius curves. Although all the streamlines in (a) appear to be viable, the algorithm still identifies outliers. .......................................................... 35

4.3 Effects of registration method. (a) shows the unregistered radius curves. Two streamlines which exhibit excessive winding behavior can be visually identified. However, in (b), the registration process has cut the connection too short for the functional boxplot to identify the winding streamlines as outliers. .......................... 36
4.4 Sub-connections at scale 125 comprising the connection from Figure 3.2 in scale 60. (a)-(d) are each sub-connections of the connection in Chapter 3 used to develop the algorithm.
Acknowledgments

I want to thank Dr. Jean Carlson for being an engaged and inspiring mentor and the rest of the Complex Systems group for guiding me through my first experience with real scientific research. I would also like to thank the Worster family for funding my research over the summer as well as Joshua Mueller for being an exceptional graduate student mentor throughout this entire process. Finally, thank you to my parents and all my friends who have all been so incredibly supportive of my work this year, I really couldn’t have done it without them.
Chapter 1

Introduction

The human brain is an immensely complex organ which has captured the attention of researchers across a wide range of disciplines. From psychology to electrical engineering, each field has contributed new insights and methods for understanding the many intricacies of the brain. These methods include studying the brain on both micro and macroscopic scales ranging from individual neurons to the entire brain. One way to study the brain is to model it as a network. In the same way that the Internet can be thought of as a network of computers connected through the sharing of information, the brain can be modeled as a network of specified regions connected anatomically to one another. The network of these anatomical connections is used to establish the structural connectivity of the human brain [4]. The study of structural connectivity relies on the accuracy of noninvasive techniques to observe the physical structures within the human brain. This is done using magnetic resonance imaging to establish the paths taken by bundles of brain cells to transmit signals between
regions. However, present methods have been shown to introduce inherent uncertainties in the construction of a brain network. In this thesis, we will investigate the reliability of the data used to establish structural connectivity.

1.1 Basic Brain Anatomy

The human brain is a highly complex, well-organized system. At the most fundamental level, it is comprised of brain cells known as neurons. A typical neuron consists of a cell body, dendrites, and an axon (see Figure 1.1).

![A Typical Neuron](https://online.science.psu.edu)

Figure 1.1: Basic brain cell anatomy. The cell nucleus resides in the cell body. Electrochemical signals are passed along the axon to the terminal bulb, which connects to the dendrites of another neuron to pass a signal to the next cell.\(^1\)

The cell body is where the nucleus of the cell and various organelles responsible for metabolic functions reside. One of these is the axon, which transmits electrochemical signals between neurons. This axon is often insulated by myelin sheath to reduce signal decay.

\(^1\)Figure 1.1 source: https://online.science.psu.edu
Additionally, axons tend to travel in bundles between cells, thus further reducing signal decay [5]. On a macroscopic scale, we can comprehensively describe the brain using grey and white matter. *Grey matter* refers to anatomical regions where cell bodies cluster together, located predominantly on the surface of the brain. *White matter* refers to the bundles of axons which connect grey matter regions. A rough idea of the distribution of grey and white matter is shown in Fig. 1.2a. [6]

Figure 1.2: Distribution of grey at white matter regions (a) shows this distribution in a cross-section of the brain. (b) shows how grey matter regions are connected by white matter fibers (axons).

The grey matter regions are predominantly located on the surface of the brain and are connected by white matter fibers which reside in the interior [4] (see Figure 1.2). Next we will discuss the organization of these connections.

### 1.2 Structural Connectivity

*Structural connectivity* refers the anatomical connections between pairwise grey matter regions in the brain. Measures of structural connectivity such as fiber count and fiber density
describe the presence and strength of these connections. A *connection* refers to a group of white matter fibers connecting two specified regions.

Mathematically speaking, the structural connectivity can be represented by a symmetric matrix whose nonzero entries represent the strength of the connection between a pair of regions. See Figure 1.3 for an illustration.

![Figure 1.3](image-url)

Figure 1.3: Structural connectivity matrix is an adjacency matrix $A$ with the color of elements $A_{ij}$ determined by the strength of connection between nodes $i$ and $j$. (a) shows a structural network with 250 nodes while (b) shows network with only 50 nodes, highlighting the effect of node choice in the construction of a network. Figure sourced from [1].

In practice, dividing the grey matter of the brain into anatomically distinct regions is a nontrivial task. There are several methods for dividing the brain into such regions [7], and thus the aforementioned matrix will vary depending on the choice of method.

Regardless of region choice, structural connectivity can be viewed as a network of white matter fibers connecting those regions. The ultimate goal of understanding the structural connectivity of the human brain is to accurately and quantitatively establish this network.
1.3 Motivation

The human brain is a highly complex entity which makes it difficult to determine structural connectivity with high precision [8, 1]. Researchers are finding that it is difficult to find consistency across different structural network models and that large amounts of the data used in developing such models are unreliable [9, 10]. Thus we propose to examine more closely some of the sources of uncertainty in the presently available methods for establishing a structural network. Some challenges include defining grey matter regions, observing white matter in a noninvasive fashion, and processing large amounts of data, all of which contribute to the overall uncertainty. In this thesis, we will attempt to quantify some of the uncertainty that arises from the noninvasive data generation process and propose an algorithmic criterion to eliminate them. Specifically, we will focus on identifying and extracting certain inaccurate connections between grey matter regions.
Chapter 2

Technical Background Information

In the following section we discuss the technical background information necessary for the subsequent discussion of our work. Specifically, we will formalize the model of the brain as a structural network and then present the noninvasive data acquisition and measurement process used to estimate the connections of such a network from diffusion tensor imaging measurements of a human brain. We will then outline some of the specific sources of uncertainty inherent in the data acquisition process and how those uncertainties can lead to false connections in the structural network of the brain.

2.1 Constructing a Brain Network

To model structural connectivity of the brain as a network, the grey matter must be divided into well-defined, non-overlapping regions. Once these regions have been established, the
network will describe the connections between these regions. In the following we will first discuss the parcellation of the grey matter. Next we will introduce graph theory to provide a formal footing for the subsequent discussion of the brain as a network.

Brain Parcellation

Structural connectivity, as defined in Section 1.2, describes the brain as a network of connections between regions of grey matter, as shown in Figure 2.1. However, there are several biologically reasonable methods for defining such regions [7, 11, 9]. Naturally, the choice of parcellation method will directly affect the structure of the network. Additionally, subjects exhibit differences in grey matter region size, making consistent parcellation across subjects more difficult [12].

Figure 2.1: Visualization of a brain parcellated into distinct grey matter regions regions¹. Note that actual parcellation methods generally use much smaller region sizes [7].
Methods for brain parcellation and implications of parcellation choice can be found in references [7, 9, 8]. Since there are several methods of brain parcellation, brain regions can differ in number and size. In this paper, we will refer to parcellation scales 60 and 125, which refer to lower and higher parcellation resolutions respectively.

In order to study the brain as a network more formally, we will now introduce graph theory.

**Graph Theory**

Graph theory is a mathematical way of examining the pairwise relationships between nodes. A graph is comprised of a collection of points called *nodes* which are connected to each other by *edges* as shown below in Figure 2.2

---

1. Figure 2.1 source: http://brainybehavior.com/neuroimaging
2. Figures 2.2 and 2.3 are sourced from http://mathinsight.org
Graph theory can be applied to model entire systems as networks [13]. For example, we can apply graph theory to the Internet as a network (as mentioned in the introduction) by defining each computer as a node, and the sharing of information between computers as edges.

Edges can be directed or undirected. A directed graph specifies the direction in which an edge connects two nodes (see Figure 2.3a), whereas an undirected graph gives no such specification indicating a bidirectional connection.

Additionally, we can take into account the edges weight, which indicates the strength of the pairwise connection (see Figure 2.3b). For example, the weight of edges between two computers could be defined as the number of messages sent between them. Thus, the more messages that are sent between a pair of computers, the greater the weight of the edge connecting them. A weighted network is therefore one which accounts for the weight of these connections. Conversely, in an unweighted network, all edges are given binary weights. (DK: consider wording)
CHAPTER 2. TECHNICAL BACKGROUND INFORMATION

Figure 2.3: Visualization of directed and weighted networks. (a) shows a directed network where the arrows represent the direction of the edge. (b) shows a weighted directed network, in which the weight of each edge is represented by the thickness of the arrow.

Other relevant network characteristics include size and average node degree. Where network size corresponds to the number of nodes $N$ in the network, and node degree is the number of edges connected to a given node. Therefore the greater the degree, the more connected that node is within the network.

Having discussed a few fundamental properties of a network within graph theory, we are now prepared to apply these concepts to the brain.

**Brain Networks**

Applying brain parcellation and graph theory allows us to model the brain as a network. The parcellated regions act as the nodes of the network while the white matter fibers connecting these regions represent the edges [12]. Specifically, one edge between two regions is comprised of all the white matter fibers between them as illustrated in Figure 2.4. We can then examine
pairwise connections between regions by considering the white matter fibers whose endpoints reside in the same regions.

Figure 2.4: (a) Highlighted regions of interest (ROI’s) (b) White matter fibers connecting those pairwise ROI’s.

The number of white matter fibers, and thus the strength of the connection between two nodes will vary depending on the two nodes of interest. Furthermore, as we will discuss the following section, the data acquisition methodology does not provide directional information about the edges. Therefore, we are considering the structural network of the brain to be weighted and undirected.

2.2 Data Acquisition

The study of structural connectivity requires accurate knowledge of white matter fiber paths within the brain. However, we cannot directly observe white matter fibers in the brain without physically dissecting it. Instead, researchers use state-of-the-art magnetic resonance imaging technology to obtain high-resolution data about brain tissue (including white matter
fibers). From this process, inferred paths called *streamlines* can be generated which represent the white matter fibers connecting grey matter regions.

The data used throughout this thesis is obtained from the Human Connectome Project\(^3\) (HCP).

**Magnetic Resonance Imaging**

Magnetic resonance imaging (MRI) scans provide researchers with high-resolution data about brain tissue. A subject is placed in the MRI scanner (see Figure 2.1) which uses nuclear magnetic resonance \(^1\) to measure water diffusion patterns in the subject’s brain.

Figure 2.5: MRI scan setup\(^4\). Subject is placed in MRI scanner until scan is completed. This process can last anywhere between a few minutes \(^5\) or a few hours \(^6\).

---

\(^3\)Human Connectome Project website: http://www.humanconnectomeproject.org/
\(^4\)Figure 2.5 source: https://www.medventura.com/healthaffairs/mri-scanner
Studies show that the nuclear magnetic resonance signal has the greatest attenuation (i.e. it diffuses faster) when the magnetic field used in an MRI is aligned with the direction of the fiber. Conversely, when the field is perpendicular to the fiber, the signal is attenuated least [14, 15]. It is also known that water does not diffuse across the lipid layers of the Myelin sheath [5], and thus diffuses instead in parallel with the axon. Thus we can use signal attenuation to infer white matter fiber paths by tracing out the direction of water flow as measured by the MRI [3]. It is important to note here that these paths are only coarse estimates of prominent white matter paths and are not of high enough resolution to represent individual axons. This will be further explained in the following section.

Diffusion Tensor Imaging

In order to generate streamlines from the aforementioned water flow, these diffusion patterns must be examined more closely. To do so, consider partitioning the brain into small cubes, called voxels. Typically, the sidelength of a voxel is on the scale of 1-3 mm [3]. Within each voxel, the overall water diffusion can be quantified by a diffusion tensor, $D$. This tensor can be represented as a $3 \times 3$ matrix where each $D_{ij}$ indicates the diffusion coefficient in that direction [17].

$$D = \begin{bmatrix}
D_{xx} & D_{xy} & D_{xz} \\
D_{yx} & D_{yy} & D_{yz} \\
D_{zx} & D_{zy} & D_{zz}
\end{bmatrix} = \begin{bmatrix}
e_1 & e_2 & e_3
\end{bmatrix} \begin{bmatrix}
\lambda_1 & 0 & 0 \\
0 & \lambda_2 & 0 \\
0 & 0 & \lambda_3
\end{bmatrix} \begin{bmatrix}
e_1 \\
e_2 \\
e_3
\end{bmatrix}$$
The three diagonal elements \((D_{xx}, D_{yy}, D_{zz})\) are the diffusion coefficients in along the principal axes \(x, y,\) and \(z\) in the lab frame. On the right hand side is the eigendecomposition of the tensor \(D\) where \(e_1, e_2, e_3\) are the eigenvectors corresponding to eigenvalues \(\lambda_1 \leq \lambda_2 \leq \lambda_3\) [18, 17]. These eigenvalues and eigenvectors determine the tensors anisotropy and its orientation, as illustrated in 2.6. For more information on computing the eigenvalues and eigenvectors of diffusion tensors, see references [19, 20, 21].

![Figure 2.6: Diffusion ellipsoid visualizing the anisotropy and orientation of the diffusion tensor.](source)

The orientation of the diffusion tensor is determined by the first principle axis \(e_1\) while its anisotropy is determined its eigenvalues. That is, if \(\lambda_1 = \lambda_2 = \lambda_3\) the tensor will be isotropic. However, if \(\lambda_1 > \lambda_2 \geq \lambda_3\) the tensor will be anisotropic. Examples of both isotropic and
anisotropic diffusion tensors can be found in Figure 2.7 labeled as "voxel tensors" in parts (a) and (d) respectively.

Because orientation of the diffusion ellipsoid is equal to that of the orientation of the first principal axis ($e_1$), the ellipsoid cannot provide directionality, as $-e_1$ is an equally valid eigenvector. There are several limiting factors associated with DTI. For example, the diffusion tensor will not contain information about water diffusion on a sub-voxel scale [3, 22]. Thus, the anisotropy of water diffusion on such a scale may be overlooked by the diffusion tensor for a given voxel. Additionally, crossing fibers can result in what appears to be an
There are several different methods for constructing the diffusion tensor within a voxel. The Human Connectome Project uses High Angular Resolution Diffusion Imaging (HARDI) [24].

Figure 2.8: A diffusion tensor is calculated for each voxel in the brain. The diffusion tensor describes the overall anisotropy and orientation of water flow within that voxel. All the diffusion tensors together create a three dimensional vector field describing the water diffusion pattern throughout the brain. Figures sourced from [3].

The principal axis of the diffusion tensor in voxel of the brain defines a three dimensional vector field quantitatively describing water diffusion as shown in Figure 2.8. It is important to note that this vector field only provides information about the orientation of the water flow, but not its directionality [18]. A tracking algorithm can be used in conjunction with this vector field to generate the white matter fiber bundle estimations called *streamlines*.

\footnote{There are several approaches to handling fiber crossings. One approach to solving this problem is the use of diffusion spectrum magnetic resonance imaging (DSI) which is discussed in further detail in [23].}

\footnote{For more information on HARDI, see [25].}
Streamline Generation

To generate a streamline, a 'seed' is planted in a grey matter region and its path is according to the vector field as it flows from voxel to voxel [3]. Because the diffusion tensor analysis does not provide directionality, any tracked streamlines are undirected. That is, we do not know if a streamline is traveling from Region A to Region B, from Region B to Region A, or both.

Figure 2.9: Streamline following path as directed by diffusion tensor.

The tracking algorithm used to impose parameters on the generated streamline within each voxel is designed to ensure biological viability [3]. Such parameters include anisotropy value and streamline trajectory angle. A streamline will terminate if it reaches a voxel with
an anisotropy value typical of grey matter (0.1-0.2) as this indicates that the generated streamline has successfully connected a pair of grey matter regions [15]. Within a voxel, the angle between the incoming and outgoing streamline trajectories must be greater than 60 degrees to be considered viable. By definition of our tracking algorithm, an angle less than this is anatomically unreasonable for a streamline’s path [3]. Thus a smaller angle will result in streamline termination in that voxel.

The seeding process is repeated randomly throughout the grey matter regions of the brain until the desired number of streamlines is reached. The raw DTI data used in this study was obtained from the Human Connectome Project. The seeding process is typically repeated 100 thousand to one million times for a given subject. Each seed generates one streamline. The streamline data is what we will use throughout this thesis. An example of generated streamlines based on DTI data from the HCP is shown in Figure 2.10.

Figure 2.10: Streamlines of entire brain generated using DSI studio software with 100,000 seeds. Color indicates direction of streamline. Red indicates left/right, blue indicates superior/inferior (up/down), and green indicates anterior/posterior (front/back).
CHAPTER 2. TECHNICAL BACKGROUND INFORMATION

These collection of streamlines serve as edges between the parcellated grey matter regions in the construction of a structural network. The weight of this connection corresponds to the number of streamlines connecting the two nodes.

Problem Statement

The ultimate goal is to construct an accurate network describing the structural connectivity of the human brain. It is therefore essential that the generated streamlines used in the construction of such a network be accurate. It is the goal of this thesis to eliminate false positive streamlines. Due to the size of the data sets provided by the Human Connectome Project, we must develop an algorithm to systematically remove spurious streamlines automatically. This algorithm will be described in detail throughout the next chapter.

\footnote{Software can be found at \url{http://dsi-studio.labsolver.org/}}
Chapter 3

Methods

3.1 Big Picture

Streamlines are used to estimate the paths taken by white matter fibers in the development of a structural network between grey matter regions. Their validity is essential for building an accurate network. Any false positive streamlines will alter the weight of the edges within the structural connectivity network and must therefore be identified and removed. As discussed previously, the streamline generation process introduces inherent uncertainties. The approach we will take to address these uncertainties can be summarized as follows:

- Understand geometric properties of potentially false positive streamlines

- Develop a quantitative metric for path similarity

- Automatize extraction process of spurious streamlines based on path similarity metric
These three steps are outlined in detail in the subsequent sections.

### 3.2 Geometric Properties of False Positive Data

The tracking algorithm used to generate the streamlines (as outlined in Section 2.2) imposes constraints only on each step of the streamline in question. That is, each voxel provides a set of parameters which dictates the streamline’s trajectory at that point but not its geometry as a whole. Biologically speaking, however, there are geometric limitations on the overall path white matter fibers take. For example, the transmission of electrochemical signals requires energy and thus fibers will tend to take the most efficient path between regions. Due to the complexity of the brain’s anatomy, this does not necessarily mean a straight path, but it excludes paths that loop or wind excessively as shown in Figure 3.1.

![Figure 3.1: Example Streamline exhibiting excessive winding. (a) and (b) show same streamline from different angles.](image)
Additionally, streamlines traveling between the same regions (i.e. those in the same connection) tend to form bundles, thereby reducing signal decay (as discussed in Section 1.1). This tendency can be observed in the connection of streamlines shown in Figure 3.2. The left panel shows a single streamline generated from DTI data obtained from the Human Connectome Project. Notice there are two streamlines which deviate from the bundle. Streamlines of this nature as well as those that exhibit excessive winding are not biologically viable and can therefore be labeled as false positives.

Figure 3.2: Illustration of removing false positive streamlines. (a) shows original streamline data, (b) shows connection with spurious streamlines removed.

Due to the large size of these data sets (ranging from 100 thousand to one million streamlines), our objective is to create an algorithm to identify and extract spurious streamlines. This algorithm will involve two steps. First we will develop a quantitative metric of path similarity between streamlines in the same connection. Then, we will apply functional data analysis to identify those which deviate most from the path as statistical outliers. Let us now discuss the first step in more detail.
3.3 Radius Required

In identifying a consensus along which the bundle of streamlines travels, we must account for the variability in streamline geometry between different connections. Even the same connection might look different across various subjects. Thus, we cannot hard code a consensus path from which to measure the deviation of each streamline. Instead, we must develop an algorithm which allows us to identify and remove spurious streamlines based on their similarity to the other streamlines in the same connection. If a streamline follows the consensus path, it should be spatially close to the other streamlines in that connection. Therefore, the algorithm must determine the spatial proximity of any given streamline to all the others within its connection at each point along its path.

Here we propose to quantify this spatial proximity using the following method: At a given point along the streamline we define a sphere. Note that streamline data are provided as sets of spatial coordinates \((x_{ij}, y_{ij}, z_{ij})\) where the index \(i\) denotes the streamline in question and the index \(j\) indicates the point along that streamline. The radius of the sphere will be determined such that a given percentage \(X\) of all the other streamlines in the connection will pass through it. Given a choice for the percentage\(^1\) \(X\), this algorithm will compute a specific value of the radius \(R_{ij}\) for each point \(j\) along each streamline \(i\). The algorithm is set to iterate over all indices of a streamline. At this time it is convenient to define

\(^1\)Obviously the percentage \(X\) is a parameter of this algorithm which needs to be chosen with some care. The extreme values \(X = 0\%\) and \(X = 100\%\) are both meaningless. The former will not capture any streamlines, while the latter will capture all and thus render our algorithm useless. In this study we will use \(X = 70\%\).
$R_i(j) = \{ R_{ij} \mid j \in [1, N_i] \}$, where $N_i$ is the total number of indices $j$ of a given streamline. In this regard, $R_i(j)$ represents the entire radius curve generated by our algorithm. Figure 3.3b shows the result of iterating over the spurious streamline from Figure 3.3a. In Figure 3.4 this process is illustrated for a streamline which remains tightly bundled throughout its trajectory.
Figure 3.3: (a) Illustration of the radius required for sphere to capture given percent for a spurious streamline ($i = 86$, shown in orange). The required sphere is small where the streamline is still closely bundled with the others, but increases as it deviates from the others. (b) Shows the radius of the sphere in (a) as a function of streamline index $j$. Note that the curve $R_i(j)$ is still a curve of discrete data points.
To avoid penalizing streamlines for fanning at the endpoints, we should extract only those
streamlines whose required radii deviate significantly from the other streamlines. Figure 3.5 shows the resulting radius curves $R_i(j)$ for each streamline in the connection (each curve corresponds to a single streamline). The two curves which deviate drastically from the rest of the set (shown in yellow and orange) correspond to the two streamlines identified visually to be spurious in Figure 3.2.

![Figure 3.5: Radius algorithm that generates curves $R_i(j)$ is iterated over all streamlines in the connection. Each curve corresponds to one streamline.](image)

This process again requires visual inspection to remove outliers. However, as discussed
previously, our aim is to automate this process in order to cope with the vast size of the data set. Now that we have reduced the dimensionality of our data, we can apply functional data analysis to automatically extract outlying curves.

3.4 Functional Data Analysis

In the previous section, we introduced a method that reduces our data from three to two dimensions, which allows us to apply functional data analysis (FDA) techniques.\(^2\) FDA is a statistical tool which provides a quantitative comparison between sets of curves [26]. We will use Matlab code\(^3\) provided by Dr. Jim Ramsay of the Department of Psychology at McGill University to identify and extract spurious streamlines through the use of functional box plots.

Data Registration and Curve Fitting

The Matlab code we use to apply functional data analysis requires that each curve in the set have an equal number of data points. In the literature, this is often referred to as having equal time steps [27]. However, in our case the number of data points for a given radius curve \(R_i(j)\) is equal to the number of indices \(j\) of streamline \(i\) and is therefore not equal for all streamlines within the connection. Therefore, we must first develop a method which

\(^2\)Note that the reason we do not apply FDA directly to the streamlines as spatial data is because applying FDA in three dimensions becomes computationally excessive.

\(^3\)Matlab code used for functional boxplots can be found at http://www.psych.mcgill.ca/misc/fda/software.html.
provides us with curves with an equal number of indices.

To achieve this, we do the following: First we find the point in space where the streamlines are most tightly bundled. Second, we identify the shortest streamline of the connection and shift the indices of every other streamline such that every streamline’s index \( j \) is equal at this point. Finally, we cut any data which goes beyond the indices of the shortest streamline. The result of this process, to which we will subsequently refer to as registration, is shown in Figure 3.6. We are now at a necessary point to introduce the new definition of a \textit{registered} radius curve \( R_i^*(j) = \{ R_{ij} \mid j \in [1, N_{\text{min}}] \} \), where \( N_{\text{min}} \) denotes the total number of indices of the \textit{shortest} streamline.

(a) \hspace{2cm} (b)

Figure 3.6: Comparison between unregistered and registered curves. (a) Shows the original radius curves \( R_i(j) \) (same as Figure 3.5). (b) Shows the registered radius curves \( R_i^*(j) \) that are a result of the registration and cutting process. All data has been cut to have 250 indices, the same number of indices as the shortest streamlines in this connection. The expected bundling tendencies are more apparent after registration as curves are now aligned spatially rather than by index.
At this stage, the data is in the correct format for the functional boxplot code. Note that the registered data is still defined as a set of discrete values. Functional data analysis, on the other hand, is based on continuous functions. Therefore, the first step in the functional boxplot code is to use spline interpolation\(^4\) to fit a continuous function to each curve in the data set. Next, functional data analysis is used to generate functional boxplots.

**Functional Boxplots**

After the above preprocessing procedure, the data is now suitable for functional data analysis, which will determine curve similarity based on statistical depth. Statistical depth indicates the “centrality” of a curve with respect to the remainder of the curves in the data set\(^{[28]}\) (for further information on the calculation of depth, see \(^{[27]}\)). A functional boxplot describes the maximum, median, and interquartile range of acceptable depth measures \(^{[29]}\). Any curves with a depth measure greater than the maximum is considered an outlier. An example of a functional boxplot can be found in Figure 3.7b.

\(^{4}\)This is specific to this thesis, other methods of curve fitting can also be used.
Figure 3.7: Functional box plot indicating statistical outliers. (a) shows registered and cut radius curves $R_i^*(j)$. (b) shows functional boxplot of registered curves as generated by Matlab code. The pink region indicates the interquartile range of depth measures (deepest 50% of curves). The black line indicates the median curve while the outliers are indicated in red.

The outliers extracted by the functional boxplot are the curves identified as spurious in the previous section. The corresponding streamlines are shown in Figure 3.8. In this example, the algorithm has correctly identified the streamlines that deviate from the bundle and which should be removed from the original data set.
Figure 3.8: Statistical outliers as identified by FDA. (a) shows the original connection. (b) shows the streamlines identified by FDA to be outliers.

We have now outlined our algorithm for identifying and removing false positive streamlines. For the example connection used in this chapter, it has served its purpose by correctly identifying those which had been qualitatively deemed to be spurious. However, this is only one connection of many. We must still investigate the algorithm’s performance on other connections.
Chapter 4

Results and Discussion

Our ultimate goal is to improve the accuracy of tractography methods to ensure that structural connectivity of the human brain is reliably extracted from the data. Thus far we have developed an additional preprocessing procedure to identify and remove false positive streamlines based on characteristics specific to the connection group in which they reside. Next, we will investigate performance of the proposed algorithm across different connections. As discussed earlier, each connection in the brain looks anatomically different (has distinct geometric properties) and our method aims to capture and accommodate those distinctions. However, the proposed method is effective only for connections which posses certain characteristics, which shall discuss below.
4.1 Results

Thus far, we have examined only a single connection. We will now provide examples of connections which highlight the strengths and shortcomings of our method. First, let us present example of a connection for which this method is very successful, meaning the algorithm extracted only those streamlines deemed to be spurious based on the biological restrictions described in previous sections.

![Figure 4.1: Example where algorithm extracted those streamlines which had been visually identified as biologically inviable.](image)

Figure 4.1: Example where algorithm extracted those streamlines which had been visually identified as biologically inviable. (a) shows the original streamline connection, (b) shows the functional boxplot of the registered radius curves for the streamlines in (a).

We can qualitatively observe at least one outlier in this connection (see Figure 4.1a) which our algorithm successfully identifies as spurious (see Figure 4.1b). However, our algorithm is not always successful. For example, Figure 4.2a shows a connection in which all streamlines are biologically viable, yet the functional box plot still identifies two streamlines as spurious.
Figure 4.2: A biologically viable connection. (a) shows the streamline data, all of which appear to be biologically viable. (b) shows the functional boxplot of the registered radius curves. Although all the streamlines in (a) appear to be viable, the algorithm still identifies outliers.

Another shortcoming of this algorithm is in its inability to extract the winding behavior explained in Section 3.1. In the connection shown in Figure 4.3a, there are two curves (in yellow and red) which have a large number of time steps while remaining at a relatively small radius. This indicates such winding behavior as they never deviate far from the others. From the panel on the left, it is clear that the corresponding streamlines are spurious. However, once the registration process is applied, they no longer deviate from the rest and are thus not extracted as outliers (the outliers indicated in Figure 4.3b are different streamlines).
(a) (b)

Figure 4.3: Effects of registration method. (a) shows the unregistered radius curves. Two streamlines which exhibit excessive winding behavior can be visually identified. However, in (b), the registration process has cut the connection too short for the functional boxplot to identify the winding streamlines as outliers.

Finally, let us discuss the effect of changing the parcellation resolution. To this end, we will consider once more the connection used in Chapter 3. This connection was taken at parcellation scale 60. By increasing the resolution to scale 125, the connection can be broken down into four distinct sub-connections which are shown in Figure 4.1.
Figure 4.4: Sub-connections at scale 125 comprising the connection from Figure 3.2 in scale 60. (a)-(d) are each sub-connections of the connection in Chapter 3 used to develop the algorithm.

The two streamlines which comprise sub-connection 4 in Figure 4.4d are those which the algorithm had previously identified as spurious during our analysis at scale 60. This indicates a sensitivity of our algorithm to the parcellation scale.
4.2 Discussion and Conclusion

The method for reducing dimensionality of the problem—using the radius plots correctly—illustrates the deviation of spurious streamlines from the consensus path. Additionally, prior to the registration process, it is able to extract streamlines that wind excessively which we may not be able to identify by examining the original connection. After applying functional data analysis, the method works successfully in some cases. However, as demonstrated in our results, there are instances in which our algorithm fails to identify the outliers we had visually identified in either the original connection or in the unregistered radius curves.

The registration method we applied works better for some connections than others. In particular, as demonstrated in Figure 4.3, the registration process cuts tightly wound streamlines to such a short length that their spurious qualities are no longer detectable to FDA.

Concerning the sensitivity of our algorithm to the parcellation scale, it seems advisable to work at a sufficiently high resolution to avoid accidental elimination of entire connections. However, increasing parcellation resolution decreases the streamline count per connection. Therefore, the resolution must remain low enough to maintain a sufficient sample size for functional data analysis.

In summary, we have developed an algorithm to extract spurious streamline data. The algorithm reduced the three dimensional streamlines to one dimensional radii. This allowed for a manageable application of functional data analysis. We found that this algorithm successfully extracts streamlines qualitatively deemed biologically inviable in some cases.
We also identified shortcomings such as registration and parcellation sensitivity which require future development. On a positive note, while our algorithm at present has limited capability to analyze the entire brain, it allows us to extract spurious streamlines between strongly connected regions.

4.3 Outlook

While our work has made considerable progress towards a quantitative analysis of brain imaging data, there is clearly a need for further research. Studies predict anywhere from ten to sixty percent of all streamlines in a given subject to be spurious [10], yet the method we have developed above typically extracts far fewer. This indicates that our algorithm is not sufficient to extract all of the false positive streamlines. Further study could entail the development of a different metric entirely, especially for the connections which are too sparse for this algorithm. An interesting next step would be to examine the geometric properties of sparse connections that are also rare among the subject population. A rare, sparse connection is likely to be false and could therefore provide insight into the characteristics of the individual spurious streamlines itself beyond is behavior as compared to a group of streamlines.

This method may also have interesting applications beyond the extraction of spurious streamline data. For instance, the examination of streamline geometry could aid researchers in understanding how brain lesions affect the structure of white matter. Applications outside the field of neuroscience could include investigating path exploration of traffic or migration
patterns. As travelers explore new paths, the required radius of the population may expand to predict a change in behavior in a population there is a shift in the mean or median path. This is just one example of a system to which our algorithm could potentially be applied and indicates an opportunity for further exploration in the future.
References


REFERENCES


REFERENCES


