FFClust: Fast fiber clustering for large tractography datasets for a detailed study of brain connectivity

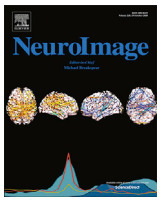
Andrea Vázquez, Narciso López-López, Alexis Sánchez, Josselin Houenou, Cyril Poupon, Jean-François Mangin, Cecilia Hernández, Pamela Guevara

NeuroImage 220 (2020) 117070

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage



FFClust: Fast fiber clustering for large tractography datasets for a detailed study of brain connectivity

Andrea Vázquez [b], Narciso López-López [b,c], Alexis Sánchez [b], Josselin Houenou [e,f,g,h], Cyril Poupon [e], Jean-François Mangin [e], Cecilia Hernández [b,d], Pamela Guevara [a,*]

[a] Universidad de Concepción, Department of Electrical Engineering, Concepción, Chile
[b] Universidad de Concepción, Department of Computer Science, Concepción, Chile
[c] Universidade da Coruña, Centro de investigación CITIC, A Coruña, Spain
[d] Center for Biotechnology and Bioengineering (CeBiB), Santiago, Chile
[e] Université Paris-Saclay, CEA, CNRS, Baobab, Neurospin, Gif-sur-Yvette, France
[f] INSERM U955 Unit, Mondor Institute for Biomedical Research, Team 15 "Translational Psychiatry", Créteil, France
[g] Fondation Fondamental, Créteil, France
[h] AP-HP, Department of Psychiatry and Addictology, Mondor University Hospitals, School of Medicine, DHU PePsy, Créteil, France

ARTICLE INFO

Keywords:
Fiber clustering
White matter bundle
Tractography

ABSTRACT

Automated methods that can identify white matter bundles from large tractography datasets have several applications in neuroscience research. In these applications, clustering algorithms have shown to play an important role in the analysis and visualization of white matter structure, generating useful data which can be the basis for further studies. This work proposes FFClust, an efficient fiber clustering method for large tractography datasets containing millions of fibers. Resulting clusters describe the whole set of main white matter fascicles present on an individual brain. The method aims to identify compact and homogeneous clusters, which enables several applications. In individuals, the clusters can be used to study the local connectivity in pathological brains, while at population level, the processing and analysis of reproducible bundles, and other post-processing algorithms can be carried out to study the brain connectivity and create new white matter bundle atlases. The proposed method was evaluated in terms of quality and execution time performance versus the state-of-the-art clustering techniques used in the area. Results show that FFClust is effective in the creation of compact clusters, with a low intra-cluster distance, while keeping a good quality Davies–Bouldin index, which is a metric that quantifies the quality of clustering approaches. Furthermore, it is about 8.6 times faster than the most efficient state-of-the-art method for one million fibers dataset. In addition, we show that FFClust is able to correctly identify atlas bundles connecting different brain regions, as an example of application and the utility of compact clusters.

1. Introduction

Diffusion Magnetic Resonance Imaging (dMRI) is an in-vivo and non-invasive technique that estimates the structure of white matter (WM) through the measurement of water molecules diffusion (Basser et al., 1994; Le Bihan and Iima, 2015). The main trajectories of WM can be reconstructed using tractography algorithms based on local orientation fields estimated from dMRI. The generated datasets consist of a 3D representation of the main WM fiber tracts (Basser et al., 2000). Streamline deterministic tractography follows the preferred direction of water diffusion in each voxel to reconstruct trajectories or lines represented by a sequence of point coordinates in 3D space. These lines are called "streamlines" or simply "fibers", though they do not represent real neural fibers but an estimation of the main trajectory of WM fascicles.

Tractography datasets contain fibers belonging to well known anatomical bundles, and also a set of bundles, mostly short association bundles, which have not yet completely described. They also have noisy fibers and artifacts, coming from dMRI intrinsic limitations and uncertainty, producing an incomplete reconstruction of the fibers (Maier-Hein et al., 1349). The application of clustering methods has helped to develop methods for the study of deep white matter bundles (DWM), in particular, the construction of deep white matter bundle atlases (O'Donnell and Westin, 2007; Guevara et al., 2012). More recently, the study of superficial white matter (SWM) or short association fiber has been carried out,

* Corresponding author.
E-mail address: pguevara@udec.cl (P. Guevara).

https://doi.org/10.1016/j.neuroimage.2020.117070
Received 30 December 2019; Received in revised form 19 March 2020; Accepted 16 June 2020
Available online 26 June 2020

with the study of reproducibility of these bundles and the construction of two SWM bundle atlases (Guevara et al., 2017; Román et al., 2017).

In general, there are two main strategies to study the brain connections given by tractograms. One is the segmentation based on anatomical Regions Of Interest (ROI) of the brain, which takes into account information on the morphology of the folding patterns of the cerebral cortex or other grey matter structures (Catani et al., 2002, 2012), to extract fibers connecting two regions. The second strategy is the clustering of fibers used to obtain bundles of similar fibers, considering their shape and position (O'Donnell et al., 2006; Guevara et al., 2011a).

Both strategies can be combined resulting in a hybrid approach that can improve the definition of anatomical bundles, since more information is included in the analysis (O'Donnell et al., 2013).

Typically, exploratory clustering methods find a great amount of bundles that characterize the structure of the white matter in its totality by using representatives such as clusters and centroids (Garyfallidis et al., 2012, 2016; Guevara et al., 2011a).

The method proposed by Guevara et al. (2011b) consists of several processing steps to subsequently subdivide the fibers into groups based on different criteria, like brain masks, voxel connectivity, fiber length and point-wise fiber distance. This method obtains compact and thin clusters that can be represented by a centroid. Another important work that performs clustering with large tractographies is QuickBundles (QB) (Garyfallidis et al., 2012). This is an unsupervised clustering algorithm that groups the fibers into clusters, without recalculating the clusters, like classical methods such as K-means. The algorithm uses a distance threshold to define whether a new fiber will be assigned to the closest cluster or will start a new cluster. It is based on the Minimum average Direct-Flip (MDF) fiber distance, although other fiber distance measures can be used. The clustering results of this method depend on the initial permutation of input fibers. This algorithm is very fast, taking about 30 min for a set of one million fibers.

Using a clustering method aids to process the tractography data, to subsequently apply other analyses on the resulting clusters. Example of analyses are the construction of WM bundle atlases (O'Donnell and Westin, 2007; Guevara et al., 2012, 2017; Román et al., 2017; Zhang et al., 2018a). Another application is the segmentation of bundles, also called virtual dissection, that seeks to label the anatomical bundles, already described by anatomists. Clustering-based segmentation methods use a fiber similarity or distance measures to group similar fibers along with anatomical information to identify known bundles. The algorithms embed anatomical knowledge commonly in the form of a bundle atlas or model (O'Donnell and Westin, 2007; Guevara et al., 2012; Ros et al., 2013; Jin et al., 2014; Yoo et al., 2015; Garyfallidis et al., 2018), or use a ROI atlas to guide the identification of anatomical bundles (Wassermann et al., 2010; Li et al., 2010; Chekir et al., 2014). Recently, methods using Deep Learning have been proposed for the segmentation of anatomical fascicles with promising results (Gupta et al., 2017, 2018; Wasserthal et al., 2018). Other applications are the study of reproducibility of white matter bundles (Guevara et al., 2017, 2020), the creation of diffusion-based cortex parcellations (Moreno-Dominguez et al., 2014; López-López et al., 2019) and the study of the human brain connectome (Zhang et al., 2018b). Furthermore, the segmentation of anatomical bundles have been extensively applied to perform clinical studies. These methods in general compare features extracted from the bundles, such as mean FA (Fractional Anisotropy) or other diffusion-based indices, bundle volume or bundle shape descriptors. For example, studies have been carried out to study bipolar disorder (Sarrazin et al., 2014), schizophrenia and autism spectrum disorder (Katz et al., 2016), parkinson (Cousineau et al., 2017) and major depressive disorder (Wu et al., 2018). Also, white matter fiber tracts can be identified in patients with brain tumors for neurosurgical planning (O'Donnell et al., 2017; Garyfallidis et al., 2018).

The size increment of tractography datasets from new high quality MRI databases, and various analysis that can take advantage of clustering results, has imposed the challenge to develop high quality and optimized fiber clustering methods.

This work proposes a new method for the clustering of large tractography datasets. The main goal is to develop an efficient clustering to group fibers into compact and regular clusters, representing the whole brain WM structure. The representation must be of good quality, i. e. the clusters must be compact along all the fibers, so that several analyses can be performed, using as input the resulting clusters or cluster centroids. A special interest is its use for the study of short association bundles and their segmentation, as well as of subdivisions of long anatomical bundles.

The proposed approach consists of four steps. The first step reduces data dimensionality by applying a partitioning clustering algorithm on fiber points instead of whole fibers. The second step groups fibers sharing the same cluster points into preliminary streamline clusters. Next, small preliminary streamline clusters are reassigned to larger clusters based on their direct or flipped distance. The last step builds compact clusters by merging candidate clusters based on a maximum Euclidean distance threshold using a graph representation of candidate cluster centroids. The experimental evaluation and comparison with the state-of-the-art shows that the proposed method is effective in the creation of compact clusters, with a low intra-cluster distance, while keeping an inter-cluster distance not excessively large, and it provides high performance. The proposed method is about 8.6 times faster than the state-of-the-art method, which enables a fast processing and visualization of main white matter fiber clusters. FFClust implementation is publicly available from https://github.com/andvazva/FFClust.

2. Materials and methods

2.1. Tractography dataset

We used the ARCHI database (Schmitt et al., 2012) that contains high quality MRI acquisition sequences of a 3T MRI scanner with an antenna of 12 channels (Siemens, Erlangen). The MRI protocol included the acquisition of T1 images at 1 mm isotropic spatial resolution using a MPRAGE sequence (Brant-Zawadzki et al., 1992) and the following parameters: 160 slices, $TH = 1.10mm$, $TE/TR = 2.98/2300ms$, $TI = 900ms$, deflection angle $FA = 9$; matrix $= 256x240$; $RBW = 240Hz/pixel$. The protocol also included a B0 fieldmap to correct for susceptibility artifacts, and a single-shell HARDI SS-EPI sequence (Mansfield, 1977) with an isotropic spatial resolution of 1.7x1.7x1.7 mm^3, with the following parameters: 60 optimized diffusion directions, $b = 1500s/mm^2$, 70 slices, $TH = 1.7mm$, $TE = 93ms$, $TR = 14,000ms$, $FA = 90$, matrix $= 128x128$, $RBW = 1502Hz/pixel$, echospacing $ES = 0.75ms$, partial Fourier factor $PF = 6/8$; GRAPPA $= 2$ (Griswold et al., 2002). All data was processed with the BrainVISA/Connectomist-2.0 software (Duclap et al., 2012).

The HARDI dataset was corrected for artifacts using a threefold strategy. First, signal dropouts and spikes are detected and corrected. Then, geometrical distortions induced by susceptibility effects are corrected using a further fieldmap calibration providing a non linear deformation to be applied along the phase axis. Finally, a joint correction for eddy currents and motion is applied to each diffusion-weighted volume. This processing corrects each diffusion direction using the rotation stemming from the rigid motion of each DW volume (Dubois et al., 2014). The diffusion-weighted (DW) processing pipeline includes the calculation of the analytic q-ball diffusion model (Descoteaux et al., 2007). Also, a robust brain white matter propagation mask is created, relying on a T1-weighted segmentation (Guevara et al., 2011c). Whole-brain streamline deterministic tractography was calculated with one seed per voxel, from all the voxels of the propagation mask, in forward and backward directions, with a tracking step of 0.2 mm and a maximum curvature angle of 30°. Resulting tractography datasets contain about one million fibers per subject. As a post-processing step, all the fibers were resampled using 21 equidistant points, as in (Guevara et al., 2011a, 2012).

2.2. Approach

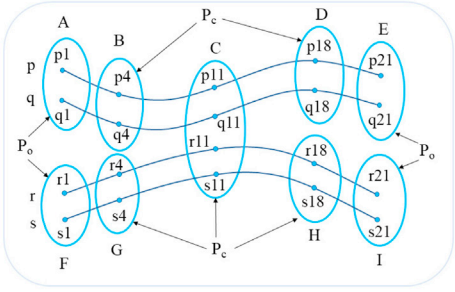
Let T_s be a tractography dataset of an individual subject consisting of a collection of fibers or streamlines, where each fiber is formed by 21 points in IR^3. The streamlines on a dataset are loaded into the main memory following the order of points calculated during the tracking. Hence, two orientations are possible: direct, or reverse (flipped), which must be considered in the streamline analysis. In this work, the similarity between fibers is defined by the maximum Euclidean distance between corresponding points (d_{ME}) (Guevara et al., 2011b, 2012, 2017; Román et al., 2017). It is a restrictive distance, since any relevant local difference between the fibers, on the fiber extremities and along all their shape, will be successfully captured.

We denote a fiber in direct order as a with 21 points in IR^3, that is each $a_i \in a$ is a 3D point, with coordinates x, y and z. In addition, we denote the fiber flipped representation as a^F, which contains 21 points in IR^3 in reverse order, that is, the first 3D point in a^F is the last in a, and so on. We denote $d_E(a_i, b_i)$ as the Euclidean distance between corresponding points a_i and b_i of fibers a and b. We assume a direct (d_E), as the maximum distance between any of the 21 points in fibers a and b, in direct order, and a flipped Euclidean distance (d_{EF}) with one fiber in inverse order. We consider the minimum of distances d_E and d_{EF}, denoted as d_{ME}, to measure the distance between streamlines a and b, as defined in Eq. (1).

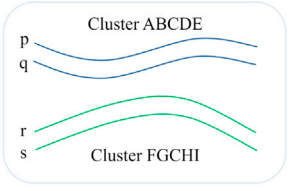
$$\begin{aligned} d_E(a_i, b_i) &= ||a_i - b_i|| = \sqrt{(a_{i_x} - b_{i_x})^2 + (a_{i_y} - b_{i_y})^2 + (a_{i_z} - b_{i_z})^2} \\ d_E(a, b) &= max_{i \in 21}(d_E(a_i, b_i)) \\ d_{EF} &= d_E(a, b^F) = d_E(a^F, b) \\ d_{ME} &= min(d_E(a, b), d_{EF}(a, b)) \end{aligned} \tag{1}$$

Hence, distance d_{ME} calculates the maximum Euclidean distance between corresponding points, taking into account the two possible fiber orientations.

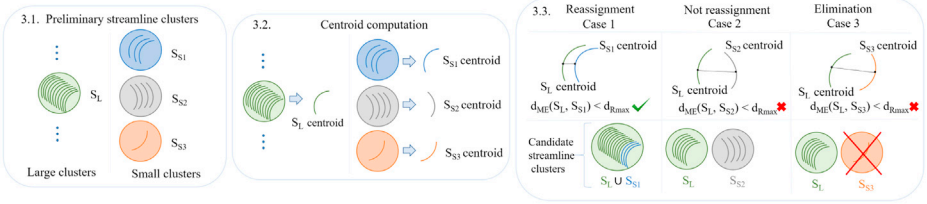
Our approach aims at improving the final clusters quality and the algorithm time complexity. A special interest is to keep a good similarity between fibers along all the fiber shape, in particular, on the extremities of the fibers. This feature is crucial for the study of superficial white matter, where short association fibers connect small grey matter regions, and a difference in the fiber end points for a group of fibers must lead to different clusters. To achieve this goal, the algorithm is based on a



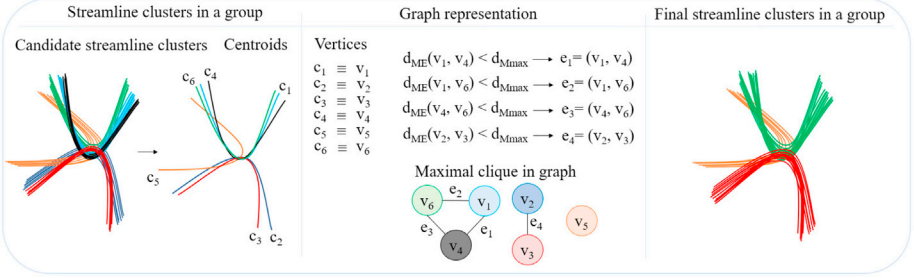
(a) STEP 1: Building point clusters.



(b) STEP 2: Generating preliminary streamline clusters.



(c) STEP 3: Reassigning small preliminary streamline clusters.



(d) STEP 4: Merging candidate streamline clusters.

Fig. 1. FFClust. (a) STEP 1: Building point clusters. MK is applied on the marked points. (b) STEP 2: Generating preliminary streamline clusters. Fibers sharing the

partition clustering applied separately to a subset of points along the fibers in parallel. Then, fibers with points sharing the same cluster points are merged. This strategy is more efficient than using the complete fiber data representation which requires more expensive fiber distance computation.

The algorithm proceeds in the following four steps: *(1)* building point clusters, *(2)* generating preliminary streamline clusters, *(3)* reassigning small preliminary streamline clusters and *(4)* merging candidate streamline clusters. Fig. 1 displays the complete workflow.

STEP 1 Building point clusters

This step aims at reducing the dimensionality of the input data by applying a partition clustering on a subset of streamline points. Applying the clustering locally on a subset of fiber points reduces the number of dimensions of the input elements and then the number of pairwise distance computations needed to form clusters. Distance computations are performed on three dimensions fiber points instead of fibers formed by 21 points, where each point has three dimensions. The method uses a subset of five points, including the two ending points (1, 21), the central point (11) and two intermediate points (4, 18). The decision of using these points is given because it has been shown that this sampling strategy is efficient and significant to estimate the maximum distance between fibers (Labra et al., 2017) and hence, to discriminate fiber differences.

The Minibatch K-means (Sculley, 2010) (MK) was chosen as a partition algorithm because it is known to provide good quality, and low time and space complexity. Moreover, given that the clustering algorithm is applied on each streamline point independently, the number of clusters do not need to be the same in all streamline points. In fact, the proposed algorithm uses different numbers of clusters for streamline ending points and central points. We denote the number of clusters for ending points as Kp_o, and the number of clusters for intermediate and central points as Kp_c. We apply the elbow method to find out the best number of clusters on each point (Kodinariya and Makwana, 2013).

Fig. 1. (a) illustrates with an example this STEP, using the five streamline points 1, 4, 11, 18, and 21. The intermediate and central points of the clusters are shown as p_c and the ending points as p_o. The central and intermediate point clusters are identified by the membership labels C, B, G, D, and H; and the ending points clusters by the labels A, F, E and I.

STEP 2 Generating preliminary streamline clusters

This step builds preliminary streamline clusters by grouping streamlines based on the membership labels of point clusters obtained in the previous step. A preliminary streamline cluster will contain all the fibers which points share the same point cluster labels. The method uses a dictionary data structure in which the key is the set of membership labels where each streamline point belongs to, and the value contains all fiber IDs that share the same set of point cluster membership labels. A preliminary streamline cluster contains all fiber IDs stored in the value associated to a key in the dictionary.

Fig. 1-(b) shows two streamline preliminary clusters, one is formed by streamlines (p, q) and the other by streamlines (r, s). As seen in Fig. 1-(a), points 1, 4, 11, 18 and 21 of fibers p and q belong to the same point clusters, then such streamline cluster is defined by the corresponding point cluster labels (A, B, C, D, E). In the same way the streamlines (r, s) formed a second preliminary streamline cluster identified by the point cluster labels (F, G, C, H, I).

STEP 3 Reassigning small preliminary streamline clusters

This step reassigns small preliminary streamline clusters that could be separated from large clusters in the previous steps. A small cluster is reassigned to the nearest large cluster, given a maximum distance threshold. In addition, with this processing, small noisy clusters are identified and discarded.

In order to do this, we first divide the preliminary clusters in two sets based on their number of streamlines. A set S_L contains all preliminary clusters with number of streamlines equal or greater than 6, and set S_S contains all preliminary clusters with 5 or fewer streamlines.

Centroids for each preliminary cluster in both sets are computed as the arithmetic mean of each streamline point. Then, a preliminary cluster in set S_S is reassigned to the closest preliminary cluster in set S_L, only if the distance between their centroids is below the threshold d_{Rmax}, that is, if $d_{ME}(a, b) < d_{Rmax}$ (see Eq. (1)).

Otherwise, corresponding clusters are not reassigned.

At the end of this step, if there still are preliminary clusters in set S_S containing one or two streamlines, these are considered noise and eliminated by default. Fig. 1-(c) shows preliminary streamline cluster set separation (Fig. 1-(c).3.1.), centroid computation (Fig. 1-(c).3.2.), reassignment, not reassignment and elimination cases (Fig. 1-(c).3.3). At the end of this step we obtain *candidate clusters*.

STEP 4 Merging candidate streamline clusters

This is a global refining step that aims to merge *candidate clusters* which still might be close based on a maximum distance parameter d_{Mmax}, in particular, clusters with flipped streamlines. This step first makes candidate cluster groups, where each group consists of clusters that share the same central point membership label obtained during the *STEP 1*. This makes groups that are close only by the central point. This processing is done only to avoid the pairwise comparison among all cluster centroids, however, this processing adds no error to the merging computation.

Then candidate cluster centroids in each group are merged based on the maximum distance parameter d_{Mmax}. If candidate cluster centroids are below d_{Mmax} for a maximum Euclidean direct or flipped distance, then such clusters are merged. To avoid the computation of all possible configurations of multiple candidate clusters that can be merged we formulate the problem using a graph representation and approximate the solution using a graph algorithm.

The graph representation considers that each candidate cluster centroid is a vertex, u, in an undirected graph, $G(u, v)$, and there is an edge, e, between two vertices, u and v, only if the cluster centroids they represent are below a maximum Euclidean distance threshold d_{Mmax}, that is, only if $d_{ME}(u, v) < d_{Mmax}$. Fig. 1-(d) shows an example where there are six candidate cluster centroids (c_1, c_2, c_3, c_4, c_5, c_6) represented with corresponding vertices (v_1, v_2, v_3, v_4, v_5, v_6). In this case, there are four edges (e_1, e_2, e_3, e_4), which exist only because the distance threshold is satisfied.

Then, a graph algorithm is applied to find groups of centroids, where each group contains all of its centroids close to each other. In a graph representation this is called a *clique*, where each group consists of vertices where all pair of vertices are connected by an edge. Specifically, the proposed method aims at finding *maximal cliques*, which are cliques that cannot grow by adding another vertex. Fig. 1-(d) shows three maximal cliques, one is formed by vertices (v_1, v_4, v_6), another is formed by vertices (v_2, v_3) and the other has vertex (v_5). Next, the method sorts all maximal cliques by decreasing size and merges all candidate clusters represented in cliques having at least two vertices, which clusters have not been previously merged. Note that a clique of size one means that the centroid is not close to any other centroid and then no merge should be performed. Given that a vertex in a clique represents a cluster centroid, and that the graph representation does not include the actual distance values among centroids, merging largest cliques first aims at joining more clusters that are close to each other based on the given threshold d_{Rmax}.

Fig. 1-(d) shows that candidate clusters represented by the centroids c_1, c_4 and c_6 are merged into the final *green cluster*, candidate clusters represented by the centroids c_2 and c_3 are merged into the final *red cluster* and candidate cluster represented by the centroid c_5 becomes the final *orange cluster*.

3. Results

This section describes the experimental evaluation and the results obtained in terms of clustering quality and performance of the method. It also performs a comparison analysis with the state-of-the-art techniques including the methods: Guevara (Guevara et al., 2011a), QuickBundles (Garyfallidis et al., 2012), and QuickBundlesX (Garyfallidis et al., 2016). In addition, it presents the results obtained for the segmentation of bundles based on a recent SWM atlas (Guevara et al., 2017). All results are obtained by using 50 subjects.

The method proposed by Guevara et al. (2011a) consists of several processing steps to subsequently subdivide the fibers into groups based on different criteria, like brain masks, fiber length, voxel-based connectivity, and point-wise fiber distance. The method provides high quality clusters, but it has about 13 configuration parameters and it is time consuming. QuickBundles (Garyfallidis et al., 2012) builds clusters using a greedy approach, and it uses a distance threshold to define whether a new fiber is assigned to the closest cluster or will start a new cluster. It is based on the Minimum average Direct-Flip (MDF) fiber distance, although other fiber distance measures can be used. QuickBundlesX (Garyfallidis et al., 2016) optimizes the QuickBundles algorithm using a tree data structure. More details in Supplementary file.

FFClust method was implemented in Python version 3.6 and in C language using compiler g++ version 7.4.0. The method supports sequential and parallel execution using OpenMP. All experiments were executed in a machine consisting of a 12-core Intel Core i7-8700K CPU with 3.70 GHz,680, and 32 GB of RAM, using Ubuntu 18.04.2 LTS with kernel 4.15.0–51 (64 bits).

The experiments were performed using deterministic tractography datasets from the ARCHI database (Schmitt et al., 2012), with one million of streamlines. To compare execution times, we use tractography datasets with resampled subjects from 330,000 to 2,729,000 streamlines.

3.1. Parameter configuration for quantitative analysis

The method has three configurable parameters. First, the number of clusters (Kp_c and Kp_o) for each of the five streamline points on which the MK algorithm is applied (*STEP 1*). Second, the maximum Euclidean distance threshold (d_{Rmax}) for the reassignment of small to large preliminary clusters (*STEP 3*). Third, the maximum Euclidean distance threshold (d_{Mmax}) for merging candidate clusters into final clusters in the last step (*STEP 4*).

3.1.1. Finding the number of fiber point clusters

First, the number of clusters for applying MK in *STEP 1* is determined using the Elbow method (Kodinariya and Makwana, 2013). To do this, the MK algorithm is executed for each of the five streamline points of a subject of one million fibers of a subject of the ARCHI database with 50, 150, 200, 250, 300, 350, 400 and 450 clusters. Fig. 2 displays the total intra-cluster variation or total within-cluster sum of squares (WCSS) for different number of clusters for the five streamline points. The elbow method shows the total WCSS as a function of the number of clusters. As the number of clusters increases the WCSS decreases, which indicates that clusters get more compact. The idea of the elbow method is to choose a number of clusters where the WCSS does not decrease much when using more clusters.

Fig. 2 shows that a good number of clusters for the intermediate and central points (Kp_c) is between 150 and 200, and for the ending points (Kp_o) is between 150 and 300. Finding the best values for the parameters d_{Rmax} and d_{Mmax} were considered using intra-cluster maximum distance. Such experiments are available in the Supplementary file.

3.1.2. Best configuration

The best parameter configuration for the proposed method consists of the number of clusters for the intermediate and central points $Kp_c = 200$ and the ending points $Kp_o = 300$ (*STEP 1*), and the value of 6 mm for the threshold distances d_{Rmax} in *STEP 3*, and d_{Mmax} in *STEP 4*.

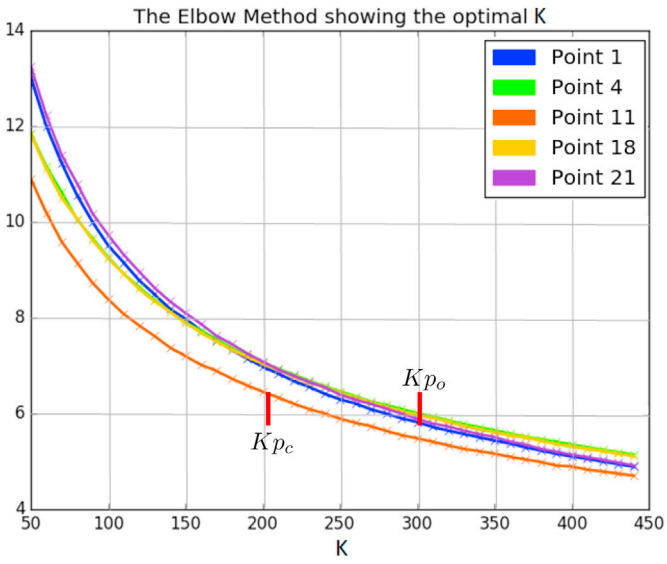


Fig. 2. Elbow method showing the optimal number of clusters K. The x-axis shows the number of clusters, y-axis shows the inertia. K's optimal values are located at the elbow of the line.

3.2. Comparison with the state-of-the-art methods

This section provides a quality comparison with the state-of-the-art methods including Guevara (Guevara et al., 2011a), QuickBundles (Garyfallidis et al., 2012), and QuickBundlesX (Garyfallidis et al., 2016). QuickBundlesX, is the successor of QuickBundles, and it performs clustering on streamlines by building a tree at different levels, with different distance thresholds. The experimental evaluation considers both the default and best parameter configurations for all methods. Specifically, *QBmdf6* refers to using QuickBundles using MDF distance of 6 mm; *QBmdf10* using MDF distance with 10 mm; *QBXmdf6* using QuickBundlesX with distance 6 mm; and *QBXmdf10*, using QuickBundlesX with distance 10 mm. Finally, *Guevara* refers to the method proposed in (Guevara et al., 2011a), which uses a maximum Euclidean distance of 10 mm.

The first evaluation, showed in Fig. 3, considers the quality of the clustering approaches, by computing the intra-cluster, inter-cluster maximum Euclidean distances and the cluster sizes obtained by all the alternatives. The comparison includes the error bars using 50 subjects, with tractography datasets of 1,045,676 fibers in average.

Fig. 3-(top) shows the number of clusters with corresponding intra-cluster distance, Fig. 3-(middle) shows the number of clusters with inter-cluster distance, and Fig. 3-(bottom) shows the number of clusters with cluster sizes. As observed, FFClust provides clusters with small intra-cluster distance, where all clusters have distances below 60 mm. QuickBundles with MDF of 6 mm also provides clusters with small intra-cluster distance, but the number of clusters with intra-cluster distance over 30 mm is greater than the number of clusters FFClust generates. All other methods produce considerable more clusters with intra-cluster distance greater than 45 mm. Note that the intra-cluster distance is measured with the maximum Euclidean distance between the corresponding points. This distance is more restrictive that MDF distance, based on the mean Euclidean distance, and employed by QB. Furthermore, the distance threshold used to compare and fuse the clusters in all the analyzed methods (QB, Guevara and FFClust) is applied to the cluster centroids. Hence, the distance between the fibers of the clusters can be higher than the threshold.

On the other hand, as seen in Fig. 3-(middle), the inter-cluster distance of FFClust is similar to the inter-cluster distance of *QBmdf6* and

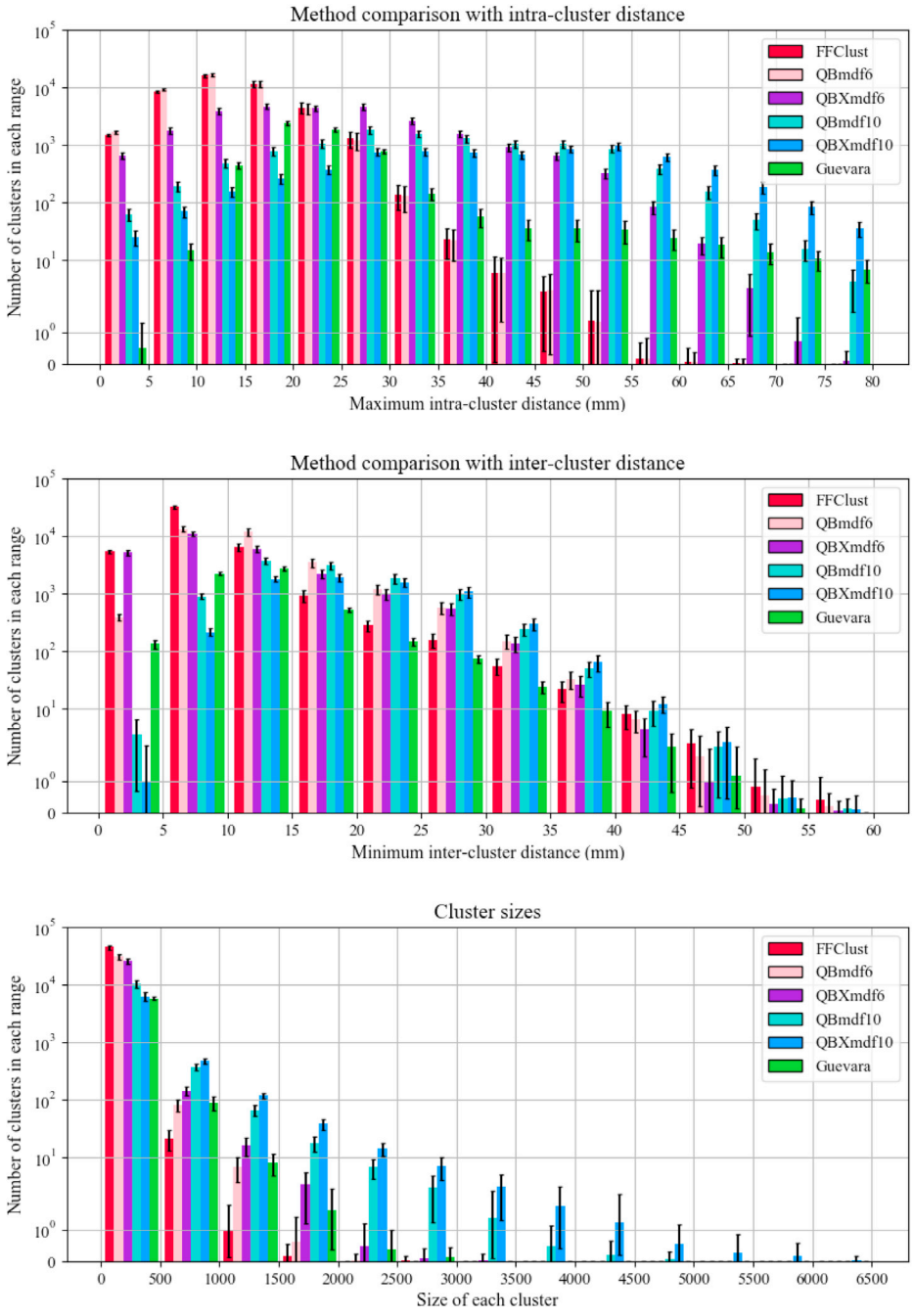


Fig. 3. Method comparison with intra-cluster, inter-cluster maximum distances, and cluster sizes. Error bars are computed by using 50 subjects with approximately one million fibers. Top figure shows intra-cluster maximum distance, middle figure shows inter-cluster maximum distance, and bottom figure shows cluster sizes.

QBXmdf6. The other methods have more clusters with a greater inter-cluster distance than FFClust, but as mentioned, they also have greater intra-cluster distances. Also, Fig. 3-(bottom) shows that FFClust and *Guevara* generates smaller clusters than the other methods, however, they are also able to find large clusters. The maximum cluster size is about 8,000 for FFClust and 9,000 for *Guevara*. In contrast *QBmdf10*, *QBXmd6*, and *QBXmdf10* have clusters of sizes over 15,000.

Note that Fig. 3 shows the results in logarithmic scale. The Supplementary file contains the same results in linear scale.

Both FFClust and *Guevara* methods eliminate fibers, but QuickBundles and QuickBundlesX do not. The Guevara method eliminates 37%, and FFClust eliminates 13% of the total number of fibers. Note that FFClust eliminates only small clusters containing 1 or 2 fibers.

Another experimental evaluation was performed to consider QuickBundles and QuickBundlesX using maximum Euclidean distance instead of the MDF. However, both methods become time consuming, taking days to complete on a dataset of 100,000 fibers. Experiments showed the quality of clusters, considering the same measures, i. e. intra-cluster and inter-cluster maximum distances, and cluster sizes, are similar to FFClust's. The Supplementary file contains the results for this experiment.

A third quality clustering evaluation uses the Davies–Bouldin (DB) index (Xu and Tian, 2015). The DB index is defined as the average similarity between each cluster with its most similar cluster, where similarity refers to the ratio of intra-cluster and inter-cluster distances. The DB index is computed by Equation (2), where n is the number of clusters; α_i and α_j are the average distances between all elements of cluster i and j respectively; c_i and c_j are the centroids of cluster i and j; and $d(c_i, c_j)$ is the

average distance between both centroids. A DB index has normalized values between 0 and 1, where a value closer to zero indicates a better separation between the clusters.

$$DB = \frac{1}{n}\sum_{i=1}^{n}\max_{i\neq j}\left(\frac{\alpha_i + \alpha_j}{d(c_i, c_j)}\right) \quad (2)$$

Fig. 4 displays the DB index for all methods, which shows that *QBmdf6* provides the best score, and FFClust is the second best method. Then, closely follows the Guevara method. The figure also shows that QB provides better DB score than QBX.

3.3. Qualitative analysis using segmentation

This section describes a qualitative analysis based on the segmentation of bundles using an atlas of superficial white matter bundles (Guevara et al., 2017). This analysis compares the segmentation results obtained by the state-of-the-art methods. The segmentation method selects and labels the closer cluster to each atlas bundle, based on a maximum Euclidean distance threshold (Labra et al., 2017; Vázquez et al., 2019).

Fig. 5 shows the segmentation comparison with the SWM atlas (Guevara et al., 2017), obtaining the closest clusters for each method using the four atlas bundles connecting the postcentral (PoC) and precentral (PrC) gyri. To obtain the closest clusters, the strategy uses a segmentation threshold distance of 6 mm, and then if a method does not find all bundles, it is increased to 8 mm. The Guevara and FFClust methods are able to identify the four bundles with a threshold distance of 6 mm, and QB alternatives are able to identify the four bundles using the segmentation threshold of 8 mm. Fig. 5 also shows an error percentage on the bottom right of each image. This error measures the percentage of fibers that are out of the correct regions. Since there is no ground truth, this is not a real measure of quality but provides an insight into the fibers that are included in the clusters but differ from the main fiber shape, with fibers that connect surrounding cortical regions. The Guevara method achieves the best streamline endpoint error (5.4%) and the second best is achieved by FFClust (6.2%), and QBmdf6 follows with 6.5%.

Fig. 6 shows the segmentation comparison with the SWM bundle atlas (Guevara et al., 2012) for the corticospinal tract (CST), the inferior fronto-occipital fasciculus (IFO), and the inferior longitudal fasciculus (IL). We can observe that all the methods are able to regroup the fibers belonging to these long bundles, with a variable number of clusters, depending on the threshold and the method itself.

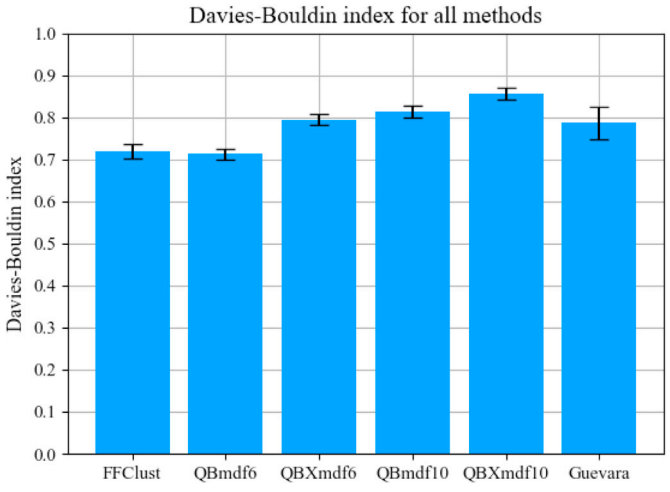


Fig. 4. Davies-Bouldin index for each method. X-axis contains each method. FFClust ($d_{Rmax} = 6mm$ and $d_{Mmax} = 6mm$). Y-axis shows the DB index, the closer to zero the better value.

3.4. Comparison with QuickBundles by qualitative analysis

This section evaluates the final clusters obtained in FFClust, *QBmdf6*, *QBmdf10*, and *Guevara* using a visual inspection to observe the quality of the final clusters. QBX is not considered because both the DB index and the segmentation evaluation show that QB achieves better quality than QBX.

The evaluation considered five cases: thinner clusters (50 thinnest clusters that present between 2 and 5 fibers), thicker clusters (50 clusters with the most fibers), short fiber clusters (200 short fiber clusters, between 30 and 60 mm), long fiber clusters (50 clusters with the longest fibers, starting at 80 mm), the least homogeneous clusters, i. e., clusters with the largest maximum intra-cluster distance, and the most similar clusters (94 clusters with the most similar ones among them). The following is a detailed description of each of the cases:

1. **Thinner clusters.** The thinnest clusters are those with the least number of streamlines. We considered such clusters are those having between 2 and 5 streamlines. Fig. 7 shows the results of FFClust, *QBmdf6*, *QBmdf10*, and *Guevara* with the 50 thinner clusters. Visual inspection shows that the methods provide similar clusters, however QB and *Guevara* clusters seem to be more scattered than clusters obtained by FFClust.
2. **Thicker clusters.** The thickest clusters are those having the largest sizes, that is, clusters with the largest number of streamlines. We consider the 50 thickest clusters for visual inspection. Fig. 8 shows the results for both algorithms. We provide each of the three views of the brain (coronal, axial and sagittal). We note that FFClust, *QBmdf6* and *Guevara* obtain clusters that look uniform. As for *QBmdf10*, we see that the clusters have less homogeneous and scattered ending points.
3. **Short fiber clusters.** We visualize short fiber clusters having fibers of length up to 60 mm. Fig. 9 shows the comparison of short fibers for the three algorithms. We observe that the quality of *QBmdf6*, FFClust and *Guevara* are very similar, but we see some clusters with more scattered ending points for QB. Again *QBmdf10* presents clusters too wide with frizzy ending points. Short fibers are usually analyzed based on the regions they connect (Guevara et al., 2020). Hence, having endpoints from different streamlines close to each other in a cluster will help to minimize the number of streamlines connecting neighboring regions. However, this does not ensure that all the cluster streamlines will land on the same anatomical structure.
4. **Long fiber clusters.** We denote long fiber clusters those clusters with fibers of length greater than 80 mm. To facilitate visual inspection we present the 50 longest clusters. Fig. 10 shows the comparison between FFClust, *QBmdf6*, *QBmdf10*, and *Guevara* for coronal, axial and sagittal views. As in the previous experiments, we observe that the clusters generated by FFClust, *QBmdf6* and *Guevara* are very similar and compact, whereas clusters for *QBmdf10* have frizzy ending point.
5. **Clusters with largest maximum intra-cluster distance.** This experiment studies the final clusters that have a maximum intra-cluster distance over 40 mm. FFClust obtains three clusters over this maximum distance, shown in Fig. 11-(left). *QBmdf6* obtains about 15 clusters over 50 mm, shown in Fig. 11-(middle left).

QBmdf10 and *Guevara* obtain about 140 over 50 mm, hence we display only the clusters with maximum intra-cluster distance over 70 mm (see Fig. 11-(middle right) and Fig. 11-(right)).

Fig. 11 shows that FFClust has only a few clusters with small number of streamlines and atypical forms. We suggest that they are noise in the tractography. Some clusters found in QB with intra-cluster distance were probably divided into several small clusters by FFClust. FFClust only discards small and isolated clusters, that are dissimilar to all the other large clusters. This is performed in its third step, which tries to reassign the small clusters (with five or fewer streamlines) to the largest clusters. All small clusters, with one or two streamlines that are not reassigned to a large cluster are eliminated. It should be noted that none of the compared

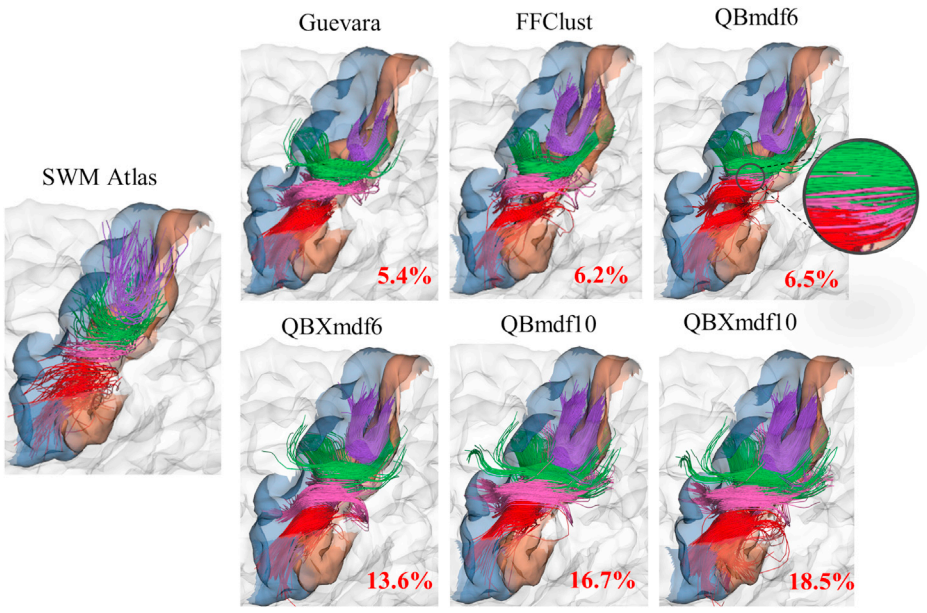


Fig. 5. Segmentation comparison with bundles of a SWM atlas (Guevara et al., 2017) connecting the postcentral (PoC) and precentral (PrC) gyri. Top image show the

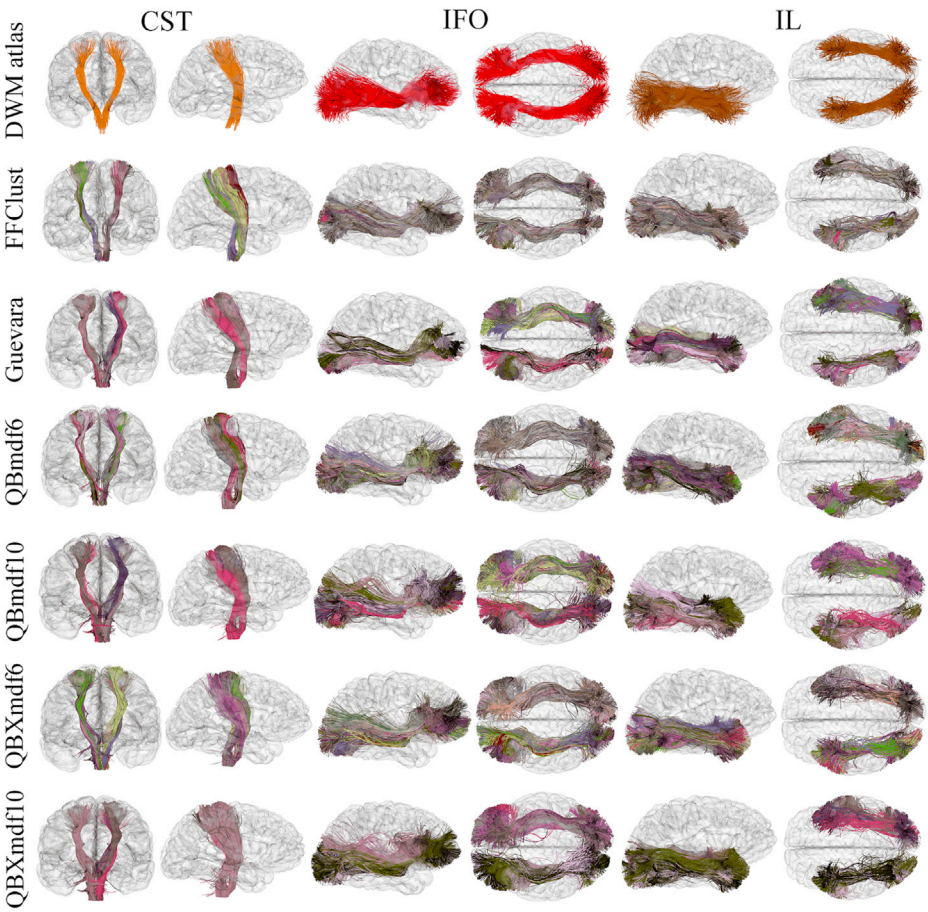


Fig. 6. Segmentation comparison based on a DWM bundle atlas (Guevara et al., 2012) for the corticospinal tract (CST), the inferior fronto-occipital fasciculus (IFO),

methods were designed to eliminate erroneous clusters of medium or large size. Further analyzes with additional information are required to perform this kind of filtering.

6. **The most similar clusters.** We also analyze the most similar clusters, i. e. those that most resemble each other for FFClust, *QBmdf6*, *QBmdf10*, and *Guevara*. To identify similar clusters, we used the 100 bundles of a SWM atlas (Guevara et al., 2017) as a reference. Those bundles where found to be the short association bundles most stable across subjects (Guevara et al., 2017). We identified the clusters that most resemble them, using a maximum threshold of 6 mm, obtaining 94 bundles for the four configurations. Fig. 12 shows the results of

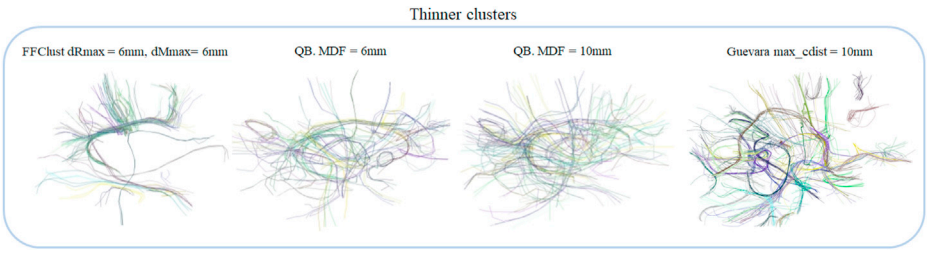


Fig. 7. Thinner cluster comparison. In the comparison we show FFClust with $Kp_c = 200$, $Kp_o = 300$, $d_{Rmax} = 6$ mm, $d_{Mmax} = 6$ mm, together with QB with $MDF = 6$ mm, $MDF = 10$ mm, and Guevara $max\_cdist = 10$ mm.

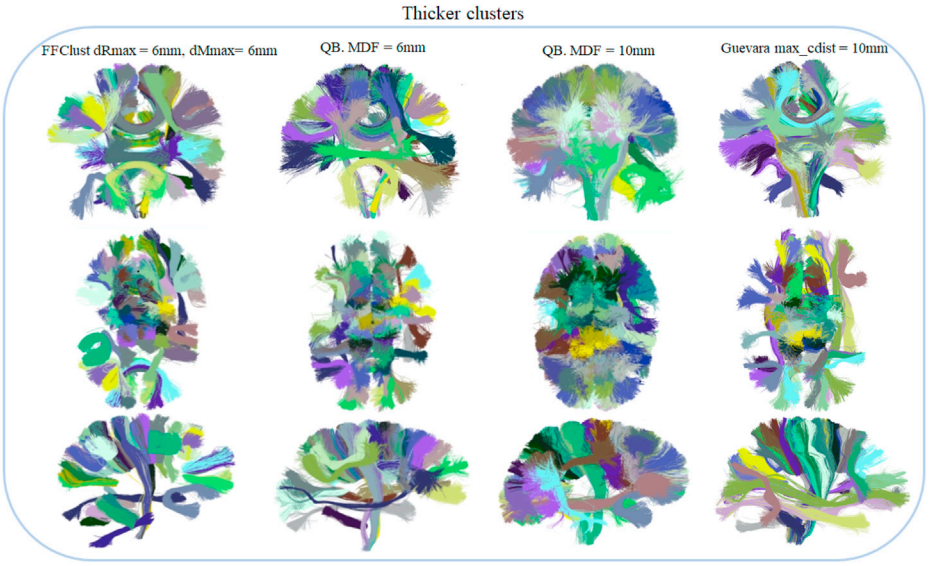


Fig. 8. Thicker cluster comparison. The comparison shows FFClust with $Kp_c = 200$, $Kp_o = 300$, $d_{Rmax} = 6$ mm, $d_{Mmax} = 6$ mm, together with QB with $MDF = 6$ mm, $MDF = 10$ mm, and Guevara $max\_cdist = 10$ mm. Display of coronal, axial and sagittal views.

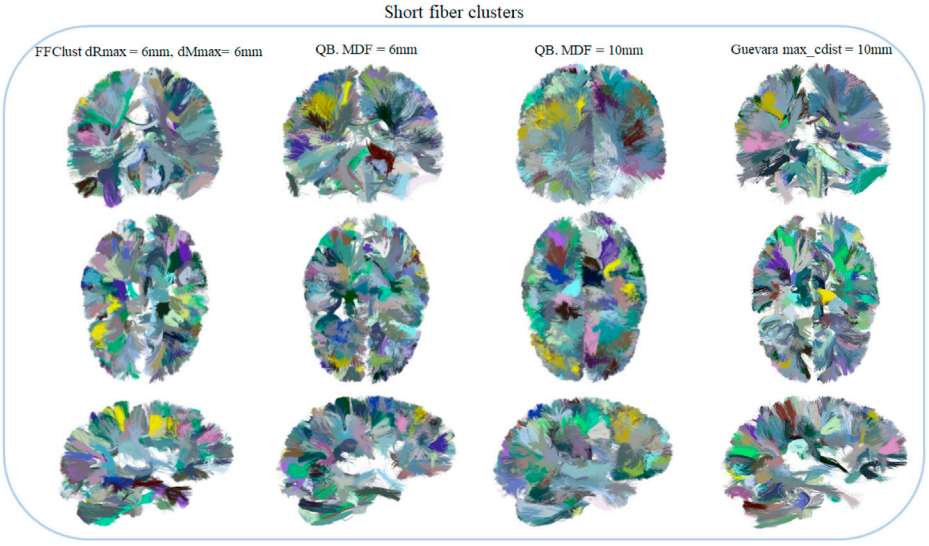


Fig. 9. Comparison of short fiber clusters. The comparison contains the 200 thickest clusters with fiber lengths between 30 mm and 60 mm. It shows FFClust with $Kp_c = 200$, $Kp_o = 300$, $d_{Rmax} = 6$ mm, $d_{Mmax} = 6$ mm, together with QB with $MDF = 6$ mm, $MDF = 10$ mm, and $Guevaramax\_cdist = 10$ mm. Display of coronal, axial and sagittal views.

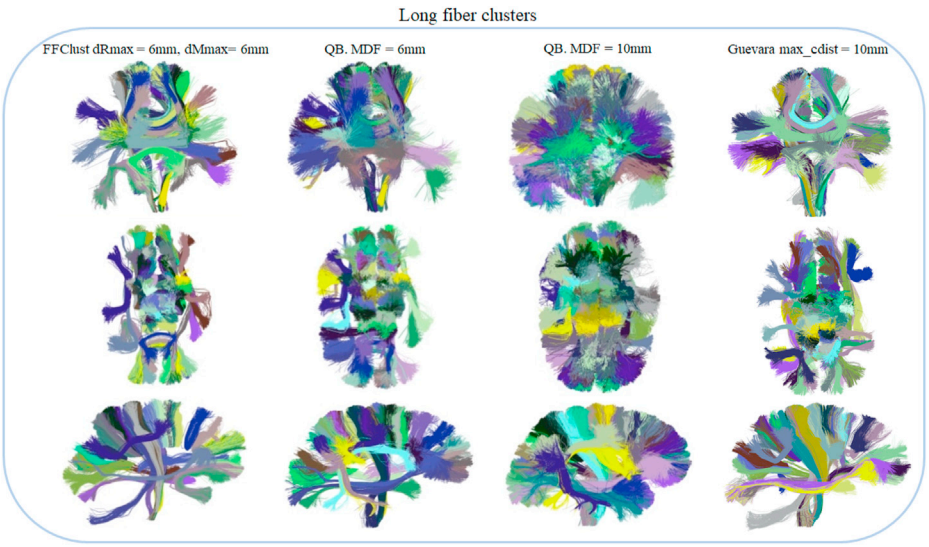


Fig. 10. Comparison of long fiber clusters. In the comparison appears the 50 thickest clusters with fibers longer than 60 mm long. We show FFClust with $Kp_c = 200$, $Kp_o = 300$, $d_{Rmax} = 6$ mm, $d_{Mmax} = 6$ mm, together with QB with $MDF = 6$ mm, and $MDF = 10$ mm, and Guevara $max\_cdist = 10$ mm. Display of coronal, axial and sagittal views.

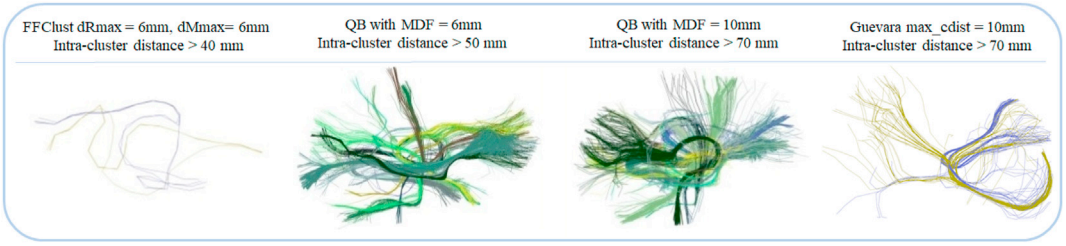


Fig. 11. Images of clusters with greater intra-cluster distance. Clusters with large intra-cluster distance, FFClust with $d_{Rmax} = 6$ mm and $d_{Mmax} = 6$ mm are shown with distance > 40 mm, QB with $MDF = 6$ mm with distance > 50 mm, QB with $MDF = 10$ mm with distance > 70 mm, and in Guevara with $max\_cdist = 10$ mm.

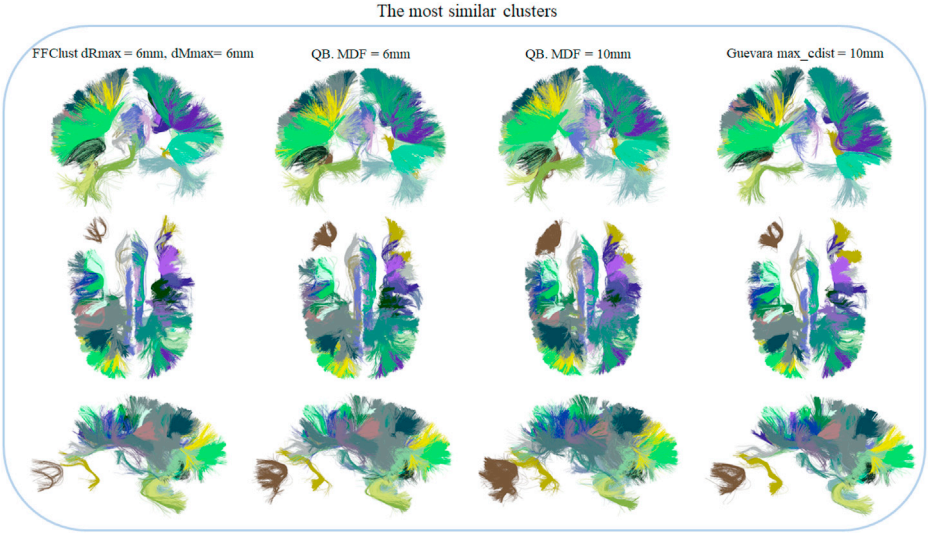


Fig. 12. Comparison of most similar clusters. In the comparison appears the 94 most similar clusters taking as a reference a SWM atlas (Guevara et al., 2017). We apply FFClust with $d_{Rmax} = 6$ mm and $d_{Mmax} = 6$ mm, together with QB with $MDF = 6$ mm, and $MDF = 10$ mm, and Guevara $max\_cdist = 10$ mm. Display of coronal, axial and sagittal views.

FFClust, *QBmdf6*, *QBmdf10*, and *Guevara* with the 94 most similar clusters. Those clusters appear very similar for all the methods.

3.5. Runtime comparison

This section evaluates the execution time of FFClust with state-of-the-art methods. FFClust-seq denotes the sequential version of FFClust, where the complete algorithm is executed using only one thread. FFClust-par denotes the parallel implementation of FFClust, where each of the steps of the algorithm is executed using five threads in the first step and 12 in the other steps. The methods used for comparison are: QuickBundles with MDF of 6 mm (*QBmdf6*), with MDF of 10 mm (*QBmdf10*), QuickBundlesX with MDF of 6 mm and 10 mm (QBX). However, we did not include the Guevara (Guevara et al., 2011a) method for this experiment because its execution times are longer than 2 h for any dataset, far exceeding the execution times of algorithms such as QB and FFClust, which were designed to be efficient. The evaluation was performed using subjects from the ARCHI database with a number of streamlines varying from 330,000 to 2,729,000.

Table 1 shows the execution times in seconds for all considered methods. Fig. 13 shows the execution times in logarithmic scale of the methods. FFClust-seq and FFClust-par provide the best execution times. It is at least an order of magnitude faster than QuickBundles. Also, we can see the trend of the QB algorithm, where the execution time increases when *MDF* is set to 6 mm, instead of using the QB default value of 10 mm.

3.5.1. Execution time complexity

As FFClust consists of four steps, its time complexity is based on each of the steps.

The time complexity of the *STEP 1: Building point clusters* is based on MK. Its upper bound is $\mathcal{O}(tKDN)$, where N is the number of elements, D is the dimensionality of the elements, K is the number of clusters and t is the number of iterations or until convergence. FFClust reduces the impact of the dimensionality parameter by defining the input elements as streamline points instead of complete streamlines. Then, the algorithm reduces the dimensionality from $D = 21 \times 3$ to $D = 3$. FFClust sequential implementation executes the MK algorithm on five streamline points, as FFClust uses five points. FFClust parallel implementation executes MK on each streamline point in parallel.

The *STEP 2: Generating preliminary streamline clusters* of the algorithm builds a dictionary data structure to map the clusters of points to streamline clusters. It is the fastest step of all and its complexity is $\mathcal{O}(N)$. This step is also parallelized in the FFClust parallel implementation.

The time complexity of the *STEP 3: Reassigning small preliminary streamline clusters* is $\mathcal{O}(|S_s| \times |S_l|)$, where both $|S_s| \ll N$ and $|S_l| \ll N$ because they are centroids not fibers. The parallel version of this step is described by Vazquez et al. (Vázquez et al., 2019) and it is included in FFClust parallel implementation.

Finally, last *STEP 4: Merging candidate streamline clusters* time complexity is determined by the maximal clique algorithm. Although the problem is NP-hard, when using parameterized complexity in sparse graphs, algorithms can be near linear (Eppstein and Strash, 2011). This step is also parallelized in the FFClust parallel implementation.

Table 1
Execution times in seconds for FFClust-par, FFClust-seq, QBmdf6, QBmdf10 and QBX, varying the number of streamlines in the range of 330,000 and 2,729,000.

Total fibers	FFClust-par	FFClust-seq	QBmdf6	QBmdf10	QBX
	time (s)	time (s)	time (s)	time (s)	time (s)
330K	9.92	28.13	334.03	108.77	32.24
659K	24.61	70.09	3,594.02	1,031.37	217.41
955K	45.96	119.84	8,032.76	2,318.02	399.61
1296K	76.82	181.08	13,404.17	4,186.45	604.14
1634K	94.75	250.13	21,813.78	7,253.64	850.67
1945K	125.91	317.59	30,394.23	10,002.94	1,125.90
2338K	193.58	460.76	50,946.22	14,312.24	1,540.61
2729K	264.30	623.13	71,243.69	19,861.15	2,194.00

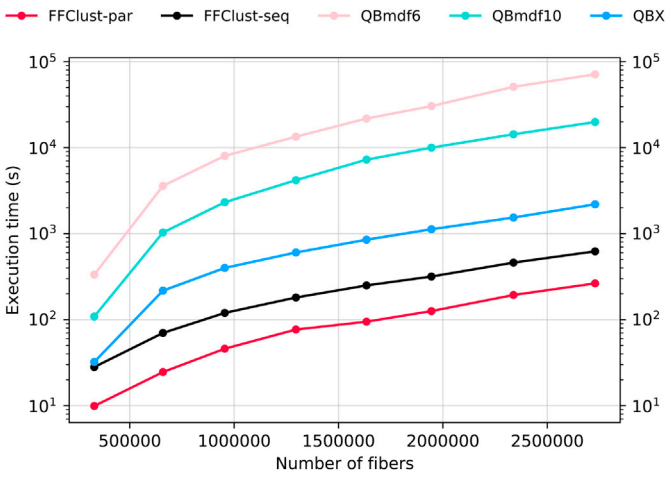


Fig. 13. Execution times for FFClust-par, FFClust-seq, QBmdf6, QBmdf10 and QBX, varying the number of streamlines in the range of 330,000 and 2,729,000.

4. Discussion and conclusions

We propose FFClust, a new fast clustering algorithm for large whole-brain tractography datasets of the brain's white matter. We compare our clustering results with the state-of-the-art clustering using the QuickBundlesX (QBX) (Garyfallidis et al., 2016), QuickBundles (QB) (Garyfallidis et al., 2012) and Guevara (Guevara et al., 2011b) methods.

After tuning the parameters of all methods, our experimental evaluation shows that FFClust identifies homogeneous clusters with a moderate maximum intra-cluster Euclidean distance and still it is able to find large clusters. Using the DB index as a metric of clustering quality, we found that QB, using MDF of 6 mm, is the best and FFClust is the second best, whereas using QB with the default distance MDF of 10 mm its DB index is less competitive. Based on the DB index, QBX does not improve the quality of QB.

We also compare the resulting clusters using as reference bundles connecting the postcentral (PoC) and precentral (PrC) gyri of a superficial WM bundle atlas. The results show that only the Guevara method and FFClust are able to find all bundles with a small error, i. e. with fewer fibers connecting surrounding regions. On the other hand, QB and QBX are able to identify the bundles, but with higher error. This analysis was performed to evaluate the potential applications of FFClust. It was designed to create compact clusters, with the purpose to be used in applications like bundle segmentation (Guevara et al., 2012; Labra et al., 2017; Vázquez et al., 2019) and inter-subject analyses for the creation of WM bundle atlases (Guevara et al., 2012, 2017; Román et al., 2017; Zhang et al., 2018a) and connectivity-based parcellations (Moreno-Dominguez et al., 2014; López-López et al., 2019). However, for some applications, bigger clusters would be more suitable. For example, if the main goal is the segmentation of large anatomical bundles, large clusters would be more useful, or easier to handle. On the other hand, if the clusters will be used for the study of short association bundles, small clusters are more suitable, since large clusters could connect neighboring anatomical regions. Another application is the diffusion-based parcellation, where, the size of the clusters depends on the size desired for the final parcels (or the number of parcels). Hence, the utility of each method must be evaluated by the user, in function of the particular requirements of the analysis to be performed.

Another advantage of FFClust, in comparison with the state-of-the-art methods, is the improvement in execution time. FFClust is at least an order of magnitude faster than QB. For instance, with a subject of 1 million of fibers, the sequential version of FFClust takes 1.99 min and its

parallel implementation takes 45 s. QB, on the other hand, takes 2.2 h using its best quality configuration (*MDF* of 6 mm). Also, at the expense of decreasing quality for some applications, using QB with its default value for *MDF* of 10 mm still takes 38 min, and its optimized version, QBX takes 6.6 min, which makes FFClust parallel implementation at least 50.4 times faster than QB and 8.6 times faster than QBX.

In summary, in addition to its reduced computation time, FFClust presents the advantage of producing good quality clusters, with a compact shape and without frizzy ending points. This feature will enable a more detailed study of brain connectivity, in particular, short association fibers, and could enable the development of diffusion-weighted parcellations. We notice that FFClust provides similar results to the Guevara method. In particular, they achieve a similar DB index score, and both are able to identify all bundles in the segmentation application. Moreover, both provide the lowest error percentages in the quality of such identified bundles. However, FFClust is more simple than the Guevara method, requires fewer parameters and it is faster. Then, we suggest that FFClust can be used in similar applications where the Guevara has been successfully used (Guevara et al., 2012, 2017).

Finally, it is important to note that since FFClust has four steps, as future work we plan to improve the execution times for the slower stages. This could be useful for integrating the clustering algorithm with visualization applications to enable the quick exploration and other post-processing analyses of the structure of the white matter for one or multiple subjects.

CRediT authorship contribution statement

Andrea Vázquez: Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization. **Narciso López-López:** Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization. **Alexis Sánchez:** Conceptualization, Methodology, Software. **Josselin Houenou:** Resources, Writing - original draft. **Cyril Poupon:** Resources, Writing - original draft. **Jean-François Mangin:** Resources, Writing - original draft. **Cecilia Hernández:** Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review & editing, Visualization, Supervision. **Pamela Guevara:** Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgements

This work has received funding by the ANID PFCHA/DOCTORADO NACIONAL/2016–21160342, ANID FONDECYT 1190701, ANID PIA/Anillo de Investigación en Ciencia y Tecnología ACT172121, ANID-Basal Project FB0008 (AC3E), ANID-Basal Project FB0001 (CeBiB), and by the European Unions Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 690941. This work was also partially funded by the Human Brain Project, funded from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreements N° 945539 (HBP SGA3), No. 785907 (HBP SGA2) and No:604102 (HBP SGA1).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2020.117070.

References

Basser, P.J., Mattiello, J., LeBihan, D., 1994. Estimation of the effective self-diffusion tensor from the nmr spin echo. J. Magn. Reson., Ser. B 103 (3), 247–254.
Basser, P.J., Pajevic, S., Pierpaoli, C., Duda, J., Aldroubi, A., 2000. In vivo fiber tractography using DT-MRI data. Magn. Reson. Med. 44 (4), 625–632.
Brant-Zawadzki, M., Gillan, G.D., Nitz, W.R., 1992. MP RAGE: a three-dimensional, t1-weighted, gradient-echo sequence–initial experience in the brain. Radiology 182 (3), 769–775. https://doi.org/10.1148/radiology.182.3.1535892.
Catani, M., Howard, R.J., Pajevic, S., Jones, D.K., 2002. Virtual in vivo interactive dissection of white matter fasciculi in the human brain. Neuroimage 17 (1), 77–94.
Catani, M., Dell'Acqua, F., Vergani, F., Malik, F., Hodge, H., Roy, P., Valabregue, R., De Schotten, M.T., 2012. Short frontal lobe connections of the human brain. Cortex 48 (2), 273–291.
Chekir, A., Descoteaux, M., Garyfallidis, E., Côté, M.-A., Boumghar, F.O., 2014. A hybrid approach for optimal automatic segmentation of White Matter tracts in HARDI. In: Biomedical Engineering and Sciences (IECBES), 2014 IEEE Conference on. IEEE, pp. 177–180.
Cousineau, M., Jodoin, P.-M., Garyfallidis, E., Côté, M.-A., Morency, F.C., Rozanski, V., Grand'Maison, M., Bedell, B.J., Descoteaux, M., 2017. A test-retest study on Parkinson's PPMI dataset yields statistically significant white matter fascicles. Neuroimage: Clin. 16, 222–233. https://doi.org/10.1016/j.nicl.2017.07.020.
Descoteaux, M., Angelino, E., Fitzgibbons, S., Deriche, R., 2007. Regularized, fast, and robust analytical Q-ball imaging. Magn. Reson. Med. 58 (3), 497–510.
Dubois, J., Kulikova, S., Hertz-Pannier, L., Mangin, J.-F., Dehaene-Lambertz, G., Poupon, C., 2014. Correction strategy for diffusion-weighted images corrupted with motion: application to the dti evaluation of infants' white matter. Magn. Reson. Imag. 32 (8), 981–992.
Duclap, D., Lebois, A., Schmitt, B., Riff, O., Guevara, P., Marrakchi-Kacem, L., Brion, V., Poupon, F., Mangin, J., Poupon, C., 2012. Connectomist-2.0: a novel diffusion analysis toolbox for BrainVISA. In: Proceedings of the 29th ESMRMB Meeting, vol. 842.
Eppstein, D., Strash, D., 2011. Listing all maximal cliques in large sparse real-world graphs. In: International Symposium on Experimental Algorithms. Springer, pp. 364–375.
Garyfallidis, E., Brett, M., Correia, M.M., Williams, G.B., Nimmo-Smith, I., 2012. Quickbundles, a method for tractography simplification. Front. Neurosci. 6, 175.
Garyfallidis, E., Côté, M.-A., Rheault, F., Descoteaux, M., 2016. QuickBundlesX: sequential clustering of millions of streamlines in multiple levels of detail at record execution time. In: Proceeding of International Society of Magnetic Resonance in Medicine (ISMRM), vol. 2016. May. Singa pore.
Garyfallidis, E., Côté, M.-A., Rheault, F., Sidhu, J., Hau, J., Petit, L., Fortin, D., Cunanne, S., Descoteaux, M., 2018. Recognition of white matter bundles using local and global streamline-based registration and clustering. Neuroimage 170, 283–295. https://doi.org/10.1016/j.neuroimage.2017.07.015.
Griswold, M.A., Jakob, P.M., Heidemann, R.M., Nittka, M., Jellus, V., Wang, J., Kiefer, B., Haase, A., 2002. Generalized autocalibrating partially parallel acquisitions (GRAPPA). Magn. Reson. Med. 47 (6), 1202–1210. https://doi.org/10.1002/mrm.10171.
Guevara, P., Poupon, C., Rivière, D., Cointepas, Y., Descoteaux, M., Thirion, B., Mangin, J.-F., 2011. Robust clustering of massive tractography datasets. Neuroimage 54 (3), 1975–1993.
Guevara, P., Poupon, C., Rivière, D., Cointepas, Y., Descoteaux, M., Thirion, B., Mangin, J.-F., 2011. Robust clustering of massive tractography datasets. Neuroimage 54 (3), 1975–1993.
Guevara, P., Duclap, D., Marrakchi-Kacem, L., Rivière, D., Cointepas, Y., Poupon, C., Mangin, J., 2011. Accurate tractography propagation mask using T1-weighted data rather than FA. In: Proceedings of the International Society of Magnetic Resonance in Medicine, p. 2018.
Guevara, P., Duclap, D., Poupon, C., Marrakchi-Kacem, L., Fillard, P., Le Bihan, D., Leboyer, M., Houenou, J., Mangin, J.-F., 2012. Automatic fiber bundle segmentation in massive tractography datasets using a multi-subject bundle atlas. Neuroimage 61 (4), 1083–1099.
Guevara, M., Román, C., Houenou, J., Duclap, D., Poupon, C., Mangin, J.F., Guevara, P., 2017. Reproducibility of superficial white matter tracts using diffusion-weighted imaging tractography. Neuroimage 147, 703–725.
Guevara, M., Guevara, P., Román, C., Mangin, J.-F., 2020. Superficial white matter: a review on the dMRI analysis methods and applications. Neuroimage 212, 116673. https://doi.org/10.1016/j.neuroimage.2020.116673.
Gupta, V., Thomopoulos, S.I., Rashid, F.M., Thompson, P.M., 2017. FiberNET: an ensemble deep learning framework for clustering white matter fibers. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2017. Springer International Publishing, pp. 548–555. https://doi.org/10.1007/978-3-319-66182-7_63.
Gupta, V., Thomopoulos, S.I., Corbin, C.K., Rashid, F., Thompson, P.M., 2018. FIBERNET 2.0: an automatic neural network based tool for clustering white matter fibers in the brain. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE. https://doi.org/10.1109/isbi.2018.8363672.
Jin, Y., Shi, Y., Zhan, L., Gutman, B.A., de Zubicaray, G.I., McMahon, K.L., Wright, M.J., Toga, A.W., Thompson, P.M., 2014. Automatic clustering of white matter fibers in brain diffusion MRI with an application to genetics. Neuroimage 100, 75–90.
Katz, J., d'Albis, M.-A., Boisgontier, J., Poupon, C., Mangin, J.-F., Guevara, P., Duclap, D., Hamdani, N., Petit, J., Monnet, D., Corvoisier, P.L., Leboyer, M., Delorme, R., Houenou, J., 2016. Similar white matter but opposite grey matter changes in schizophrenia and high-functioning autism. Acta Psychiatr. Scand. 134 (1), 31–39. https://doi.org/10.1111/acps.12579.
Kodinariya, T.M., Makwana, P.R., 2013. Review on determining number of cluster in K-means clustering. Int. J. 1 (6), 90–95.
Labra, N., Guevara, P., Duclap, D., Houenou, J., Poupon, C., Mangin, J.-F., Figueroa, M., 2017. Fast automatic segmentation of white matter streamlines based on a multi-subject bundle atlas. Neuroinformatics 15 (1), 71–86.

Le Bihan, D., Iima, M., 2015. Diffusion magnetic resonance imaging: what water tells us about biological tissues. PLoS Biol. 13 (7), e1002203.
Li, H., Xue, Z., Guo, L., Liu, T., Hunter, J., Wong, S.T., 2010. A hybrid approach to automatic clustering of white matter fibers. Neuroimage 49 (2), 1249–1258.
López-López, N., Vázquez, A., Poupon, C., Mangin, J.-F., Guevara, P., 2019. Cortical surface parcellation based on intra-subject white matter fiber clustering. In: 2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON). IEEE, pp. 1–6.
K. H. Maier-Hein, P. F. Neher, J.-C. Houde, et al., The challenge of mapping the human connectome based on diffusion tractography., Nat. Commun. 8 (1349). doi:10.1038/s41467-017-01285-x.
Mansfield, P., 1977. Multi-planar image formation using NMR spin echoes. J. Phys. C Solid State Phys. 10 (3), L55–L58. https://doi.org/10.1088/0022-3719/10/3/004.
Moreno-Dominguez, D., Anwander, A., Knösche, T.R., 2014. A hierarchical method for whole-brain connectivity-based parcellation. Hum. Brain Mapp. 35 (10), 5000–5025.
O'Donnell, L., Westin, C.-F., 2007. Automatic tractography segmentation using a high-dimensional white matter atlas. IEEE Trans. Med. Imag. 26 (11), 1562–1575.
O'Donnell, L.J., Kubicki, M., Shenton, M.E., Dreusicke, M.H., Grimson, W.E.L., Westin, C.F., 2006. A method for clustering white matter fiber tracts. AJNR Am. J. Neuroradiol. 27 (5), 1032–1036.
O'Donnell, L.J., Golby, A.J., Westin, C.-F., 2013. Fiber clustering versus the parcellation-based connectome. Neuroimage 80, 283–289.
O'Donnell, L.J., Suter, Y., Rigolo, L., Kahali, P., Zhang, F., Norton, I., Albi, A., Olubiyi, O., Meola, A., Essayed, W.I., Unadkat, P., Ciris, P.A., Wells, W.M., Rathi, Y., Westin, C.-F., Golby, A.J., 2017. Automated white matter fiber tract identification in patients with brain tumors. Neuroimage: Clin. 13, 138–153. https://doi.org/10.1016/j.nicl.2016.11.023.
Román, C., Guevara, M., Valenzuela, R., Figueroa, M., Houenou, J., Duclap, D., Poupon, C., Mangin, J.-F., Guevara, P., 2017. Clustering of whole-brain white matter short association bundles using HARDI data. Front. Neuroinf. 11, 73.
Ros, C., Güllmar, D., Stenzel, M., Mentzel, H.-J., Reichenbach, J.R., 2013. Atlas-guided cluster analysis of large tractography datasets. PloS One 8 (12), e83847.
Sarrazin, S., Poupon, C., Linke, J., Wessa, M., Phillips, M., Delavest, M., Versace, A., Almeida, J., Guevara, P., Duclap, D., Duchesnay, E., Mangin, J.-F., Dudal, K.L., Daban, C., Hamdani, N., D'Albis, M.-A., Leboyer, M., Houenou, J., 2014. A multicenter tractography study of deep white matter tracts in bipolar I disorder. JAMA Psychiatr. 71 (4), 388. https://doi.org/10.1001/jamapsychiatry.2013.4513.
Schmitt, B., Lebois, A., Duclap, D., Guevara, P., Poupon, F., Rivière, D., Cointepas, Y., LeBihan, D., Mangin, J., Poupon, C., 2012. Connect/archi: an open database to infer atlases of the human brain connectivity. ESMRMB 272, 2012.
Sculley, D., 2010. Web-scale k-means clustering. In: Proceedings of the 19th International Conference on World Wide Web. ACM, pp. 1177–1178.
Vázquez, A., López-López, N., Labra, N., Figueroa, M., Poupon, C., Mangin, J.-F., Hernández, C., Guevara, P., 2019. Parallel optimization of fiber bundle segmentation for massive tractography datasets. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE, pp. 178–181.
Wassermann, D., Bloy, L., Kanterakis, E., Verma, R., Deriche, R., 2010. Unsupervised white matter fiber clustering and tract probability map generation: applications of a Gaussian process framework for white matter fibers. Neuroimage 51 (1), 228–241.
Wasserthal, J., Neher, P., Maier-Hein, K.H., 2018. TractSeg - fast and accurate white matter tract segmentation. Neuroimage 183, 239–253. https://doi.org/10.1016/j.neuroimage.2018.07.070.
Wu, Y., Zhang, F., Makris, N., Ning, Y., Norton, I., She, S., Peng, H., Rathi, Y., Feng, Y., Wu, H., O'Donnell, L.J., 2018. Investigation into local white matter abnormality in emotional processing and sensorimotor areas using an automatically annotated fiber clustering in major depressive disorder. Neuroimage 181, 16–29. https://doi.org/10.1016/j.neuroimage.2018.06.019.
Xu, D., Tian, Y., 2015. A comprehensive survey of clustering algorithms. Ann. Data Sci. 2 (2), 165–193.
Yoo, S.W., Guevara, P., Jeong, Y., Yoo, K., Shin, J.S., Mangin, J.-F., Seong, J.-K., 2015. An example-based multi-atlas approach to automatic labeling of white matter tracts. PloS One 10 (7), e0133337. https://doi.org/10.1371/journal.pone.0133337.
Zhang, F., Wu, Y., Norton, I., Rigolo, L., Rathi, Y., Makris, N., O'Donnell, L.J., 2018. An anatomically curated fiber clustering white matter atlas for consistent white matter tract parcellation across the lifespan. Neuroimage 179, 429–447.
Zhang, Z., Descoteaux, M., Zhang, J., Girard, G., Chamberland, M., Dunson, D., Srivastava, A., Zhu, H., 2018. Mapping population-based structural connectomes. Neuroimage 172, 130–145. https://doi.org/10.1016/j.neuroimage.2017.12.064.