A Novel Clustering Algorithm For Application To Large Probabilistic Tractography Data Sets

R. E. Smith[1,2], J-D. Tournier[1,2], F. Calamante[1,2], and A. Connelly[1,2]
[1]Brain Research Institute, Florey Neuroscience Institutes (Austin), Heidelberg West, Victoria, Australia, [2]Department of Medicine, The University of Melbourne, Melbourne, Victoria, Australia

Introduction

The segmentation of brain white matter through clustering of tractography data is a problem that has attracted considerable attention in recent years. Approaches to date have succeeded in identifying the major white matter structures from diffusion tensor tractography, but suffer from a number of limitations, including the dependence upon definition of regions of interest[1], use of a similarity matrix which restricts the quantity of data which can be processed[1, 2], requirement of manual segmentation[3], or a coarse resolution of clustering due to the use of a feature space[2, 4, 5]. These become critical when dealing with probabilistic tractography, where very large data sets must be produced to accurately represent the underlying biological structure, particularly when applied to the whole brain. We present a novel algorithm, capable of automated clustering of very large probabilistic track data sets (demonstrated on 1,000,000 tracks) at any chosen cluster scale.

Methods

Our clustering algorithm consists of three stages. The first efficiently produces clusters from the fiber set using a data stream clustering strategy[6]; this consists of comparing each track in sequence to a set of exemplars, and producing a new exemplar whenever an incoming track is sufficiently unique. The second stage utilises a fast sparse implementation of the popular K-means algorithm, allowing clusters within a local neighbourhood to exchange tracks to reduce the global sum of squared error. The optional third stage merges neighbouring clusters for which the inter-cluster boundary is deemed arbitrary, based upon inter- and intra-cluster track distances. The second and third stages improve the reproducibility of the algorithm, as the results of data stream clustering are dependent upon the order in which the data are presented. The number of clusters is not set explicitly; rather it is a function of the selected scale of clustering, and the quality of discrimination between neighbouring clusters at that scale. The Hausdorff similarity metric[7] with upper threshold was selected to appropriately cluster tracks at a very fine resolution.

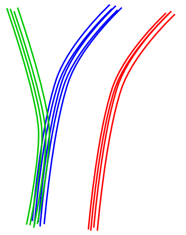

Figure 1: schematic illustration of 3 bundles identified by the first fine clustering stage. The green and blue clusters should be grouped as members of the same bundle, yet their Hausdorff distance is greater than that between the red and blue clusters.

An additional post-processing stage applies an agglomerative hierarchical clustering[8] approach to the resulting clusters to gradually identify larger structures within the brain. This approach provides logical sub-divisions of the primary white matter structures at a number of scales, down to cluster and even individual track level. A different similarity metric was used for the hierarchical clustering stage, since identifying clusters belonging to the same major pathways is a fundamentally different problem from the fine-scale clustering. This is illustrated in Figure 1: bundles which ought to be grouped together according to established human anatomy may not have high similarity according to the simple Hausdorff metric, and vice-versa. A novel metric was designed for this purpose, incorporating a number of measures including distance and directional coherence between tracks, as well as global information such as track density and continuity of track directionality between points.

Results

One million probabilistic streamlines were generated from DW data acquired on a 3T Siemens Trio from a healthy volunteer (2.3mm isotropic, 150 DWI directions, b=3000s/mm^2) using the MRtrix software package[10], with the fiber orientation distributions estimated using Constrained Spherical Deconvolution[11]. The clustering of these tracks (including hierarchical classification) was achieved in approximately 4 hours, running on a single 2.1GHz Intel processor with 4GB of RAM. With clustering performed using a Hausdorff threshold of 10mm, ~16,000 individual clusters (of 5 tracks or more) were identified. Figure 2 shows coronal, sagittal and transverse projections of a number of major fiber bundles identified through hierarchical classification – note that these bundles can be further sub-divided for visualisation down to the desired scope.

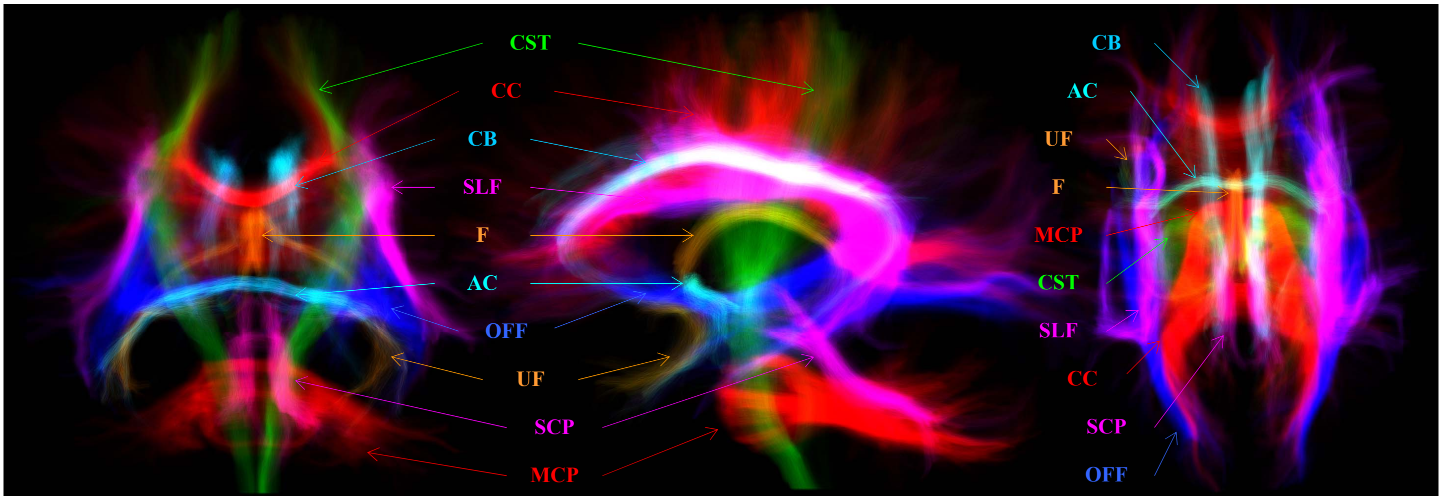

Figure 2: coronal (left), sagittal (middle) and axial (right) projections of a subset of the major bundles identified through hierarchical classification. The tracks are colour-coded with respect to their parent bundle (AC: anterior commissure; CB: cingulum bundle; CC: corpus callosum; CST: corticospinal tracts; F: fornix; MCP: middle cerebellar peduncle; OFF: occipitofrontal fasciculus; SCP: superior cerebellar peduncle; SLF: superior longitudinal fasciculus; UF: uncinate fasciculus). The tracks are displayed using transparency for ease of visualisation.

Discussion

We have described a novel clustering algorithm that can perform clustering on large probabilistic tractography data sets, permitting automated identification of the major white matter structures of the brain. Our proposed algorithm is fast, capable of processing very large data sets of 1,000,000 tracks (which has not been achievable using previous methods) in ~ 4 hours, making it applicable for use across a number of potential neuroscientific applications. The clustering results naturally depend on the quality of tractography; it is therefore expected to benefit from future advances in fiber-tracking algorithms.

References

[1] Leemans et al., ISMRM 17: 856 (2009) [2] Klein et al., SPIE 6509 (2007) [3] Hagler et al., HBM 30:1535-1547 (2009) [4] Klein et al., SPIE 6918 (2008)
[5] Liang et al., CIBCB 09: 292-297 (2009) [6] Guha et al., IEEE TKDE 15: 515-528 (2003) [7] Maddah, M., PhD thesis, CSAIL, MIT, 2008
[8] Jain et al., CSUR 31:3:264-323 (1999) [9] Hubert et al., JoC 2: 193-218 (1985) [10] MRtrix, http://www.brain.org.au/software/
[11] Tournier et al., NeuroImage 35: 1459-1472 (2007)