A scalable approach to streamline tractography clustering

E. Visser[1,2], E. Nijhuis[1,3], and M. P. Zwiers[1,2]

[1]Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, Netherlands, [2]Department of Psychiatry, Radboud University Nijmegen Medical Centre, Nijmegen, Netherlands, [3]Department of Technical Medicine, University of Twente, Enschede, Netherlands

Introduction

Diffusion tractography has become an important method for assessing white matter (WM) structure. Depending on the application, the large number of streamlines typically produced by the tracking algorithm may make the dataset difficult to handle. As there will be many streamlines corresponding to each anatomical WM tract, it is often advantageous to group them into clusters in which all streamlines correspond to the same anatomical tract. In group studies, identifying tracts across subjects may be an additional objective. A common problem with clustering methods is their ability to scale, especially if data from multiple subjects is analysed. We present an approach based on repeated clustering of subsets that is conceptually transparent and that scales well to large tractography datasets and subject groups.

Clustering

The distance between streamlines *A* and *B* is defined as

$$D_{AB} = \min\left(\sum_{i=1}^{N_p}\left\|a_i - b_i\right\|, \sum_{i=1}^{N_p}\left\|a_i - b_{N_p-i}\right\|\right)$$

where a_i and b_i are the coordinates of the points of *A* and *B* respectively. In words, the distance between two tracts is defined as the sum of the Euclidean distances between their points. In general, two streamlines will have different numbers of points. Therefore, prior to clustering, all tracts are resampled to consist of N_p=25 points each. As there are two possible combinations of both streamlines' indexing directions, the minimum of the corresponding distances is used. For a dataset containing N_s streamlines, *D* will be an N_s-by-N_s matrix. This matrix can be used to run clustering on, but as its size scales with N_s^2, it is clear that this approach is not suitable for large (high-resolution) datasets. Practically, the upper limit on the number of tracts would probably lie below N_s=50,000, at which point the size of *D* would be almost 10 GB if it were stored as single precision numbers.

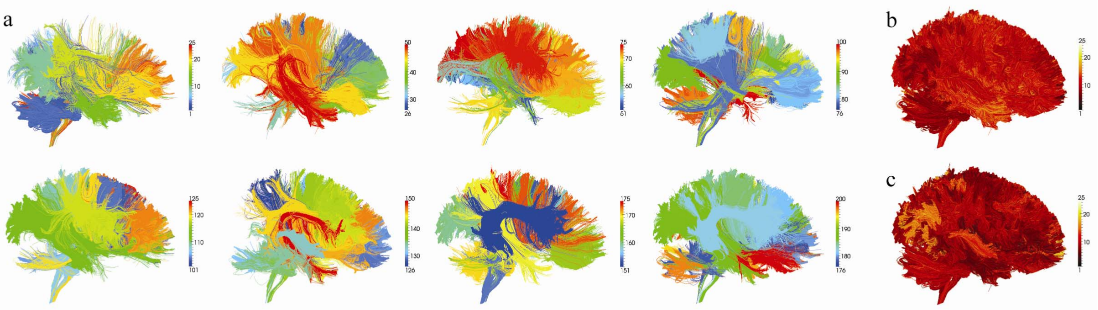

Figure 1. a) Clustering results on a single subject; N_c=200, N_p=25. Clusters were split into 8 groups for visualisation. b) Number of different clusters each streamline was assigned to across permutations. c) Number of times each streamline was assigned to its final cluster.

We solve the size problem by taking a number N_p of random permutations of all streamlines and dividing them into subsets of predetermined size N_s', which we chose to be 10,000. On every subset of every permutation, hierarchical clustering with complete linkage as implemented in MATLAB is run. As the number of clusters N_c that we are looking for is small (≤300) compared to the size of the subsets, it seems reasonable to expect clustering to be consistent to a large degree between subsets. This was confirmed by analysis of intermediate results. Cluster label correspondence between the subsets in a permutation and between permutations is determined using the distance D_{ab} between the mean tracts of the clusters. In the final step, each tract is assigned the label that occurred most often over all permutations, exploiting the consistency of clustering between subsets.

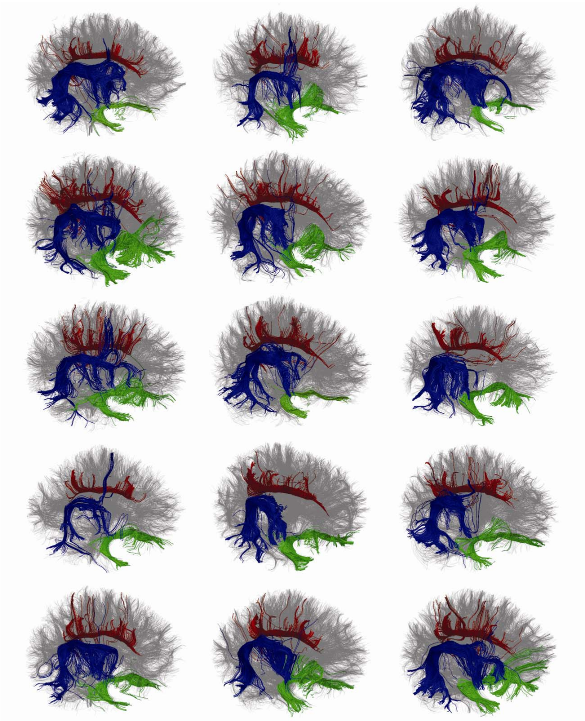

Figure 2. Group clustering results; N_c=300, N_p=100. Three unilateral clusters were selected in a single subject (cingulum, arcuate and uncinate fasciculi) and displayed for all 15 individual subjects.

Methods

Diffusion weighted imaging (DWI) datasets were acquired of the brains of 15 healthy volunteers on a Siemens Trio 3T system with a 32-channel head coil. The diffusion scheme consisted of 61 non-collinear directions at *b*=1000 s/mm^2 and 7 volumes with *b*=0. Resolution was 2.0x2.0x2.0 mm^3, with TR=8300 ms, TE=95 ms, matrix size 110 x 110 x 64 and GRAPPA acceleration factor 2. Diffusion analysis was performed using the Camino package [1]. Principal diffusion directions were obtained from q-ball reconstructions with spherical harmonics up to fourth order [2]. Deterministic tractography was run with a fractional anisotropy (FA) threshold of 0.15. All voxels with FA above 0.25 were used as seeds and streamlines whose length was shorter than 25 mm were discarded. This produced around 50,000 streamlines per subject. In one subject, tractography was run again with eight seeds per voxel, separated at 1 mm distance from each other, producing just over 375,000 streamlines. All streamlines were nonlinearly registered to a group template that was created using the FNIRT program that is included in FSL [3].

Results and discussion

Figure 1a shows the results obtained by applying the algorithm to the single subject data. The clusters are well-separated, confirming that clustering is indeed largely consistent across subsets. This conclusion is reinforced by figure 1b and c, which show that although most streamlines have been assigned to a relatively large number of different clusters, many have been assigned often to the final cluster. The differences between bundles in figure 1c are marked, but this is not unexpected as the anatomical separation between WM bundles differs greatly. In order to find consistent clusters across subjects, the registered streamline sets of different subjects were concatenated and clustering was run on the resultant dataset, containing approximately 850,000 streamlines. Results are comparable to those obtained in a single subject, but clusters are now consistent over all subjects. Example clusters are shown in figure 2. A key advantage of the method described here is its ability to scale to large datasets. It is conceivable that datasets would be much larger than those used here, but this should not pose a problem as execution time scales linearly with the number of streamlines N_s and working memory usage depends only on the subset size N_s'.

References

[1] Cook et al., Proc. ISMRM 2006. [2] Tuch, MRM 52(2004):1358-1372. [3] Smith et al., Neuroimage 23(2004):S208-S219.