SUPPLEMENTAL DIGITAL CONTENT 1

METHODS

Tensor Model and Tractography

Quality assurance of the acquired data was conducted to detect artifacts and outliers, followed by DWI de-noising using Slicer[36] and brain extraction using FSL[26]. Tensors were fitted to the DWI data using multivariate linear fitting[37] by in-house software. The WM fiber pathways were generated by the standard DTI tractography method (FACT) as implemented by TrackVis[25], with default parameters and by seeding from the entire WM region. In order to calculate the connectivity profiles of fibers, we used the probtrackx utility of the FSL software[26], again with default parameters.

The proposed methodology requires two steps to be performed for each participant, namely connectivity analysis and identification of the WM tracts. The automated identification of the entire set of WM tracts takes only a few minutes to run on a PC. However, generation of the connectivity profiles of fibers using the probabilistic tractography may take several hours based on the currently available implementations (FSL software[26]). Thus, further improvements should be considered to speed up the entire process, possibly by employing a faster probabilistic tractography method.

Methodological Details

Here, we present the proposed tract extraction framework by first describing the representation of fibers. Then, we demonstrate how a fiber bundle atlas can be constructed based on the Mixture of Multinomials (MMN) clustering model. Finally, an adaptive MMN is introduced which incorporates the generated atlas as a prior for the clustering of a new subject, so as to automatically establish correspondence between bundles of different subjects.

A similar approach was proposed before for healthy cases, using the Mixture of Gaussians Model (MGM).[18,19] The main technical difference between this work and the previous one is the way fiber bundles are represented in the model. Previously, fiber bundles were represented by Gaussian distributions, each parameterized by a mean vector and covariance matrix. MGM model poses several difficulties when the variation of bundles increases due to the distortion of white matter fibers by edema and mass effect. Specifically, the possible singularity of the

covariance matrix due to high dimensionality (95 in our case) hinders successful atlas generation, and thereby automated extraction of bundles in a test subject. Thus, here we implement a more stable model based on MMN that is not affected by dimensionality, unlike MGM. In MMN, each fiber bundle is represented by a multinomial distribution, encoding the probabilities that fibers tend to connect gray matter regions.

The connectivity profile of a fiber is defined as a collection of connectivity profile of voxels along the fiber.[18] Given a parcellation of the brain into K cortical regions $\{G_k\}$, the M dimensional vector $\boldsymbol{u}(x)$ for a voxel x consists of connection probabilities, $\boldsymbol{u}(x) \equiv [\, freq(G_1|x)\,, freq(G_2|x)\,, \dots, freq(G_K|x)\,]$, each corresponding to a connection to a specific region G_k ie, the number of fibers passing through the voxel x and connecting to region G_k. Then, a fiber is represented by a matrix with the vectors $\boldsymbol{u}(x)$ as its rows or columns.

Instead of working directly with matrices, we average over the voxels along a fiber to obtain a compact representation. Such an approach eliminates any need to define a metric for matrices of varying sizes due to different number of voxels of fibers. Finally, each fiber is represented by a single vector $\boldsymbol{f} \equiv [\, f_1\,, f_2\,, \dots, f_K]$, where f_k is calculated by averaging frequencies, $freq(G_k|x)$, over voxels along the fiber.[18]

One important issue with fiber clustering is the fact that correspondence might not be easy to establish between the resulting fiber bundles of different subjects when they are clustered individually. To assure the correspondence among subjects, we assume that each subject is an independent observation from an underlying common bundle model that is characterized by a fiber bundle *atlas*.[19] The atlas can be constructed by clustering fibers of a single or multiple training subjects. When using multiple subjects, their fibers can be combined easily since the proposed representation of fibers is invariant to spatial image coordinates.

We use MMN model for clustering, which has been used for document clustering in the past. Each fiber is assumed to be drawn from a multinomial distribution, $\boldsymbol{f} \sim MN(\boldsymbol{\beta})$. The probability mass function of a multinomial distribution is $p(\boldsymbol{f}|\boldsymbol{\beta}) = \frac{n!}{f_1! \dots f_K!} \beta_1^{f_1} \dots \beta_K^{f_K}$, where $n = \sum_k^K f_k$. Each element $\beta_k \geq 0$ gives the probability of being connected to a region G_k, where $\sum_k^K \beta_k = 1$. Fibers of the whole brain are assumed to be drawn from a mixture of M multinomial distributions, with the final likelihood of N fibers is $p(\boldsymbol{F}|\boldsymbol{\lambda}, \mathbf{B}) = \prod_i^N \sum_j^M \lambda_j \, p(\boldsymbol{f}_i|\boldsymbol{\beta}_j)$, where λ_j is the weight of j^{th} multinomial distribution. Given a set of fibers, the parameters $\boldsymbol{\lambda}$ and $\mathbf{B}$ can be

inferred by using the Expectation Maximization method. In the expectation step, we estimate the membership possibility γ_i^j of the fiber i to the j^{th} cluster by

$$\gamma_i^j = \frac{p(\boldsymbol{f}_i|\boldsymbol{\beta}_j)}{\sum_v^M \lambda_v\, p(\boldsymbol{f}_i|\boldsymbol{\beta}_v)}\,. \tag{1}$$

Then in the maximization step, we estimate the unknown parameters as

$$\beta_{jk} = \frac{\sum_i^N \gamma_i^j\, f_{ik}}{\sum_i^N \gamma_i^j\, z_i}\,;\quad z_i = \sum_k^K f_{ik}\,, \tag{2}$$

$$\lambda_j = \frac{\sum_i^N \gamma_i^j}{N}\,. \tag{3}$$

With a random initial guess on parameters, these two steps are repeated until convergence. Finally, the atlas is characterized by the defined mixture model, with each multinomial distribution corresponding to a fiber bundle. The resulting clusters are visually inspected by an expert to label them with white matter (WM) structures that they belong to.

Once the atlas is generated, it is used as a prior model for clustering fibers of a new subject. The adaptive clustering incorporates the generated atlas as a set of Dirichlet priors for the parameter set of the new MMN that is run for a test subject. For each multinomial distribution in the new model, we define a Dirichlet prior, $Dir(\boldsymbol{\alpha}_j)$ over the parameter $\boldsymbol{\beta}_j$, where $\boldsymbol{\alpha}_j$, is calculated by scaling the corresponding parameter $\widehat{\boldsymbol{\beta}}_j$ of the atlas, $\boldsymbol{\alpha}_i = c\widehat{\boldsymbol{\beta}}_j$. Under these settings, the Maximum a Posteriori (MAP) estimate of the parameter $\boldsymbol{\beta}$, given an observation (fiber) $\boldsymbol{f}$ is

$$\beta_k = \frac{f_k + \alpha_k - 1}{z + \sum_v^K (\alpha_k - 1)}\,. \tag{4}$$

Then, for clustering new subjects using the adaptive clustering scheme, the maximization step (2) is modified as

$$\beta_{jk} = \frac{\sum_i^N \gamma_i^j\, f_{ik} + \alpha_{jk} - 1}{\sum_i^N \gamma_i^j\, z_i + \alpha_j}\,;\quad z_i = \sum_k^K f_{ik}\;;\; \alpha_j = \sum_k^K \alpha_{jk} - 1\,. \tag{5}$$

In above formulation, the atlas introduces some pseudo counts for each cortical region. This means that one can adjust the compliance of a new subject with the atlas by changing the magnitudes of elements of $\boldsymbol{\alpha}_j$ while keeping their proportions fixed.