Automatic clustering and population analysis of white matter tracts using maximum density paths

Gautam Prasad a,b, Shantanu H. Joshi c, Neda Jahanshad a,b, Julio Villalon-Reina a,b, Iman Aganj g, Christophe Lenglet h, Guillermo Sapiro j,k, Katie L. McMahon k, Greig I. de Zubicaray l, Nicholas G. Martin m, Margaret J. Wright l,m, Arthur W. Toga a,b,d,e, Paul M. Thompson a,b,c,d,e,f

a Imaging Genetics Center, Institute for Neuroimaging & Informatics, University of Southern California, Los Angeles, CA, USA
b Laboratory of Neuro Imaging, Institute for Neuroimaging & Informatics, University of Southern California, Los Angeles, CA, USA
c Department of Neurology, University of California Los Angeles, CA, USA
d Dept. of Neonatology, Psychiatry, Engineering, Radiology, University of Southern California, Los Angeles, CA, USA
e Dept. of Ophthalmology, University of Southern California, Los Angeles, CA, USA
f Center for Magnetic Resonance Research, Department of Radiology, University of Minnesota, Minneapolis, MN, USA
g Dept. of Electrical and Computer Engineering, Computer Science, Duke University, NC, USA
h Dept. of Biomedical Engineering, Duke University, NC, USA
i Center for Advanced Imaging, University of Queensland, Brisbane, Australia
j School of Psychology, University of Queensland, Brisbane, Australia
k Department of Pediatrics, University of Southern California, Los Angeles, CA, USA
l Martinos Center for Biomedical Imaging, Radiology Department, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
m Department of Neurology, Psychiatry, Engineering, Radiology, University of Southern California, Los Angeles, CA, USA
n Laboratory of Neuro Imaging, Institute for Neuroimaging & Informatics, University of Southern California, Los Angeles, CA, USA

A B S T R A C T

We introduce a framework for population analysis of white matter tracts based on diffusion-weighted images of the brain. The framework enables extraction of fibers from high angular resolution diffusion images (HARDI); clustering of the fibers based partly on prior knowledge from an atlas; representation of the fiber bundles compactly using a path following points of highest density (maximum density path; MDP); and registration of these paths together using geodesic curve matching to find local correspondences across a population. We demonstrate our method on 4-Tesla HARDI scans from 565 young adults to compute localized statistics across 50 white matter tracts based on fractional anisotropy (FA). Experimental results show increased sensitivity in the determination of genetic influences on principal fiber tracts compared to the tract-based spatial statistics (TBSS) method. Our results show that the MDP representation reveals important parts of the white matter structure and considerably reduces the dimensionality over comparable fiber matching approaches.

© 2014 Elsevier Inc. All rights reserved.

Introduction

Diffusion weighted imaging (DWI) measures the directional diffusion of water through the brain in vivo. By following the dominant directions of diffusion across the brain, whole-brain tractography algorithms can reconstruct the brain’s major white matter pathways, extracting a vast number of fibers that are amenable to statistical analysis. We can then study these white matter regions in individuals and populations to better understand disease effects (Daianu et al., 2013; Jahanshad et al., 2012b; Takahashi et al., 2002), changes in brain microstructure and connectivity with age (Abe et al., 2002; Dennis et al., 2012), hemispheric differences (Jahanshad et al., 2010), sex differences (Peled et al., 1998), and genetic influences (Jahanshad et al., 2013a; Kochunov et al., 2010).

High angular resolution diffusion imaging (HARDI) enables a more accurate representation of fiber directions compared to the more standard single-tensor model (Basser and Pierpaoli, 1996). The single-tensor model does not account for fiber crossing or mixing, but the orientation distribution function (ODF) (Tuch, 2004) can be derived from HARDI images to discriminate multiple fibers with different orientations passing through a voxel (Leow et al., 2009; Zhan et al., 2010).
The large number of fibers generated by the tractography algorithms first needs to be clustered according to known anatomical pathways before comparing them across subjects. A wealth of clustering methods has been applied to tractography results including fuzzy clustering (Shimony et al., 2002), normalized cuts (Brun et al., 2004), k-means (O’Donnell and Westin, 2005), spectral clustering (O’Donnell et al., 2006), Dirichlet distributions (Maddah et al., 2008), hierarchical clustering (Visser et al., 2011), a Gaussian process framework (Wassermann et al., 2010b), and median filtering (Prasad et al., 2011a). Some of these methods readily benefit from prior anatomical information provided by an atlas of likely locations of the tracts in the brain (Yendiki et al., 2011), suggesting when to split or combine clusters to conform to known anatomy. In one approach (Jin et al., 2011a,b, 2013), several labeled atlases are deformed onto a fiber set extracted from a new subject, and a fiber matching and voting process are used to help decide the anatomical bundles to which the fibers belong.

Following clustering, several methods can be used for fiber bundle matching (Colby et al., 2011) using a parametric curve-based method to resample fibers in a bundle based on shared seed points and then compute correspondences from the resampling to create a representative path for an individual or group. A similar re-sampling approach is used in a method (Yeatman et al., 2012) that filters fiber bundles to match a probabilistic atlas. (Corouge et al., 2006) analyze fiber bundles by resampling and then aligning them across subjects using Procrustes analysis (Goodall, 1991) to generate a mean shape. (Roberts et al., 2005) apply a density measure derived from tractography results. Their measure (fiber density index; FDI) quantifies the average number of detected fiber paths passing through voxels in a ROI (Wassermann et al., 2010a) use Gaussian processes to create voxel-wise probability maps of white matter structure. The fiber locations in high density regions of the image space are used by O’Donnell et al. (2009) as a template to align other fibers and compute correspondences. Yushkevich et al. (2008) analyze white matter tracts using deformable geometric medial models that allow for integration of nearby tensor-based features to reduce the dimensionality and improve registration. (Patel et al., 2010) use a fast-marching algorithm to encapsulate white matter tracts in voxel-based boundaries, which are then matched using variational techniques.

In contrast to the parameterized methods mentioned above, white matter analysis can also be performed using a voxel-based approach. A popular method known as tract based spatial statistics (TBSS) (Smith et al., 2006), uses a skeletonized representation of white matter and uses non-linear registration for matching the skeletons. Although it is a very popular approach, TBSS does not explicitly represent tracts that would be recognized by anatomists, and therefore is not guaranteed to produce a consistent labeling of tracts from one brain to another (Schwarz et al., 2013). Although voxel-based methods can also be used to analyze DWI, they are often sensitive to the image registration (Tustison et al., 2012). Most existing white matter analysis techniques focus on non-linear registration of fractional anisotropy (FA) images as in TBSS (Smith et al., 2006) and voxel-based morphometry (VBM), which can be applied to DWI-derived maps such as FA (Jones et al., 2005). Other approaches that focus on diffusion tensor correspondences are usually based on a global image registration (Wang et al., 2011; Ye et al., 2009), but a high-dimensional registration of tensor fields may also be used, as can tensor-based statistics (Chiang et al., 2008; Lee et al., 2009; Lepore et al., 2008). Given the richness of information provided by tractography, it seems advantageous to directly study the fiber tract bundles rather than simply analyzing voxel-based representation.

**Approach**

Our work adopts a parameterized approach by refining the representation of white matter structure into compact and localized paths, represented as 3D curves. These paths represent the most influential regions in tractography and are used as compact dimensional representations of the fiber bundle. Our method uses an additional local registration of specific white matter regions to fix biases (Tustison et al., 2012) in voxel-based analysis and many of the problems of registration algorithms (Klein et al., 2009) that work on the entire image. Additionally, our approach may offer increased statistical power as it finds shape homologies across different white matter tracts.

**Materials and methods**

This section describes important steps starting with the extraction of fibers using HARDI tractography, clustering of fibers using a white matter ROI atlas, representation and matching of fiber bundles using MDPs, and finally, statistical analysis of MDPs in a population. The schematic pipeline outlining the extraction and representation of MDPs is shown in Fig. 1, whereas the workflow for statistical group analysis is shown in Fig. 2.

**HARDI tractography using the Hough transform**

We use a global tractography algorithm (Aganj et al., 2011) to extract fibers from HARDI images. The algorithm uses extensive information provided by HARDI at each voxel, parametrized by the orientation distribution function (ODF). Our tractography method selects fibers in the diffusion image space by generating scores for all possible curves at a seed point. These curves are parameterized using 2nd order polynomials. An additional parameter controls the maximum expected curve length and is set to a value representing the largest dimension of the volume. In practice, the...
Fig. 1. Schematic of the pipeline for extraction, clustering, and representation of maximum density paths (MDPs) for a single subject. The first panel shows the fibers from our HARDI global tractography method. The co-registered region of interest (ROI) atlas is used to select a fiber bundle representing a particular white matter tract. The resulting fiber bundle is then converted to a volumetric density image, which is transformed into a graph. Selected seed points in the image form the nodes of the graph, that are used to compute maximum density paths. The maximum density path compactly summarizes a given white matter structure and enables specific matching of these regions across subjects using curve registration.

Fig. 2. Schematic of the pipeline for performing statistical analysis of populations of MDPs. The first panel represents MDPs for a population of subjects for a given white matter tract. A nonlinear iterative method that uses geodesic curve registrations is used to compute an average MDP representing the mean shape of the population. A correspondence is established for all subject MDPs via the average MDP. Fractional anisotropy (FA) from each subject is resampled for the corresponding points and compared across the population.

Please cite this article as: Prasad, G., et al., Automatic clustering and population analysis of white matter tracts using maximum density paths, Neuroimage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.04.033
First, seed points are generated using a prior probability based on FA from the single-tensor model of diffusion (Basser and Pierpaoli, 1996), defined as

$$\sqrt{\frac{1}{2} \left( (\lambda_1 - \lambda_2)^2 + (\lambda_2 - \lambda_3)^2 + (\lambda_3 - \lambda_1)^2 \right)} \sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}$$

(1)

where \(\lambda_1, \lambda_2, \) and \(\lambda_3\) are the eigenvalues of the diffusion tensor. These seed points are used to generate curves that receive a score estimating the probability of their existence, computed from the voxels the curve passes through.

The ODFs at each voxel from our HARDI images were computed using a normalized and dimensionless estimator derived from Q-ball imaging (QBI) (Aganj et al., 2010). This method uses the Jacobian factor \(\rho^2\) for the constant solid angle (CSA) ODF as

$$\frac{1}{4\pi} \int_{\mathbb{S}^2} \ln \left( -\ln \left( \frac{S(\hat{u})}{S_0} \right) \right) \, d\hat{u}.$$

(2)

In this equation, \(S(\hat{u})\) is the diffusion signal, \(S_0\) is the baseline image, \(\text{FRT}\) is the Funk–Radon transform, and \(\nabla^2_{\hat{u}}\) is the Laplace–Beltrami operator. This estimate outperforms the original definition (Tuch, 2004) with superior resolution for detecting multiple fiber orientations (Aganj et al., 2010; Descoteaux and Bore, 2012; Fritzsche et al., 2010; Ghosh et al., 2013). The Hough transform tractography chooses fibers from all possible curves generated in the image at a certain space and parameter resolution. These curves are parameterized by their arc length, \(s\), ranging in value from \(L_\text{min}\) to \(L_\text{max}\) and approximated using simple polynomials. The scores for each possible curve, \(s\), derive from the ODF and FA values

$$\int_{\mathbb{L}} \left( \log \left( \text{ODF}_{\hat{x}(s)}(t(s)) \cdot \text{FA}(\hat{x}(s)) \right) + \lambda \right) \, ds,$$

(3)

where \(\text{ODF}_{\hat{x}(s)}(t(s))\) is the value from the ODF at the 3D location \(\hat{x}(s)\) and direction specified by the unit tangent vector \(t(s)\). The method scores as many fibers as possible arising from a seed point and uses the voting process provided by the Hough transform to select the best fitting curve. These filtered curves comprise the final set of fibers produced by the method for a single subject, which we refer to as \(F\).

The method is probabilistic in its selection of a fiber at a certain seed point but does not generate volumetric data giving a distribution of fibers in the white matter. It chooses seed points (voxel locations) randomly throughout the white matter tissue with a probability proportional to their fractional anisotropy. Once a seed point is chosen, the algorithm scores all possible fibers that pass through this point. The number of fibers is restricted by the parameterization and range of the variables involved, but it is close to one million candidate fibers. For each fiber this score represents the probability of that fiber existing and is constructed by integrating the orientation distribution function over the span of the fiber combined with the probability of the corresponding seed point. The method then uses the Hough transform to select the fiber with the highest probability or highest score as the final fiber for that seed point.

The tractography algorithm was run on each subject image and generated around 10,000 fibers (Fig. 1 shows a representative example with our data), which represents a good balance between computational efficiency and sampling enough of the image space (Prasad et al., 2013c). We subsequently clustered these fibers using a white matter atlas.

Fiber clustering with using white matter ROI atlas

Fibers extracted using the Hough transform-based tractography method are clustered using a ROI atlas to incorporate prior anatomical information. We use the Johns Hopkins University (JHU) atlas (Wakana et al., 2004), which delineates 50 white matter regions of interest (ROI). This ROI atlas is first registered to our subject space using an affine transform provided by FMRIB’s Linear Image Registration Tool (FLIRT) (Jenkinson and Smith, 2001). This is then followed by a nonlinear transform from the Automatic Registration Toolbox (ART) (Ardekani et al., 2005; Klein et al., 2009) to refine the registration further.

We then cluster the fiber bundles by measuring the intersection of fibers with the ROI atlas as follows. A fiber intersection score is computed by counting the number of ROI voxels that intersect with the fiber tract. This score is used to select fibers that belong to an ROI and thus to a white matter tract. Spuriously intersecting fibers are eliminated by applying an experimentally determined threshold that is dependent on the number and type of fibers obtained from the tractography algorithm. Formally, if \(F\) is the set of fibers for a subject and \(r\) is a specific white matter ROI label in the atlas, then the subset of selected fibers in a bundle is given by,

$$B = \{ f \in F : \int A(s, r) \, ds > t \}$$

(4)

where \(t\) is the intersection threshold and \(A\) is an indicator function defined to be

$$A(s, r) = \begin{cases} 1 & \text{if } s \in \text{region } r \\ 0 & \text{otherwise}. \end{cases}$$

(5)

Bundle representation using the maximum density path

The fiber bundle \(B\) representing a white matter tract is reduced to a compact representation also referred to as the maximum density path as follows. We first compute a density volume of our fiber bundle to characterize our search space, and denote it as

$$I_d(\bar{x}) = \sum_{b \in B} Q(b, \bar{x}).$$

(6)

where \(\bar{x}\) represents a 3D voxel location and \(Q\) is the indicator function defined as

$$Q(b, \bar{x}) = \begin{cases} 1 & \text{if } b \text{ intersects } \bar{x} \\ 0 & \text{otherwise}. \end{cases}$$

(7)

This value specifies the number of fibers that intersects each voxel. We then construct a graph that represents the voxel-wise fiber density in our fiber bundle. The above voxels are used as nodes in a graph, \(G = (N, E)\) (a set of nodes and undirected edges connecting them) with those of positive value connected to their 26 neighbors by undirected edges. In our formulation, the weight of an edge between nodes \(i\) and \(j\) is calculated as the sum of the voxel densities it connected as

$$I_d(\bar{x}(i)) + I_d(\bar{x}(j)).$$

(8)

with \(I_d(\bar{x}(k))\) as the density for the voxel location \(\bar{x}(k)\) corresponding to node \(k\). These edge weights are then modified by subtracting each from the maximum initial edge cost, \(\epsilon_{\text{pre}}\), such that edges in high density regions have weights close to zero. These edge weights are designed to allow the shortest path algorithm to go through edges in high density regions.

We use Dijkstra’s algorithm (Dijkstra, 1959) to compute the path through this graph following the nodes with highest density. Dijkstra’s algorithm is a graph search method that finds the shortest
path from a source node to every other node. However, the number of
nodes in the graph may be large and when the algorithm is used for a
single destination node, it may be stopped once that path is found. To
find the shortest path to represent a white matter region, we require
the graph to have start and end nodes to constrain the path to a specific
region of the graph. These nodes are specified by an expert in the ROI
atlas. The ROI points for the start and end locations may not always
 correspond to the positive density values derived from our bundle.
Thus, we find the closest voxels in the density volume at the corre-
sponding start and end nodes in subsequent computation with
Euclidean distance used as the metric.

In our implementation, we used Dijkstra’s algorithm (Dijkstra, 1959)
to find only the shortest path between the start and the end nodes
selected in the graph. If Dijkstra’s algorithm is unable to find the path
between the start and end nodes our method automatically identifies
this situation and takes steps to remedy the graph and finds the shortest
path. The algorithm will be incapable of finding a connection between
the two nodes if the structure of the graph is such that there are no
edges from the subgraph containing the start node with the subgraph
containing the end node. This can be caused by scanner artifacts or
suboptimal solutions due to the fiber tractography algorithm. In this
situation, we add extra edges and nodes to the graph so all voxels within
our ROI are fully connected with their neighbors. The edges are weighted
by the largest edge cost in the current graph. This allows gaps between
the start and end nodes to be filled in and use as few edges as possible
in regions with unknown data. The nodes in the path correspond to a
set of voxel locations in our image space. We smooth the path so it is
better conditioned for subsequent processing. We convolve the 3D
coordinates of our path with a Gaussian kernel to achieve this, though
fitting these points to a spline would also have sufficed. A summary of
these steps is presented in Algorithm 1. We represent the maximum
density path

by the coordinate function of the parameterized curve, and denote it to
be β such that β: [0, 1] → ℝ3. Figure 3 shows an example of a maximum
density path computed for a fiber bundle. Additionally, we show the
density and the FA images that correspond to the fibers. For comparison,
we also show a representative mean fiber by applying Procrustes anal-
ysis to align the fibers in the bundle to their mean. We then compute
a new mean (shown in blue) of the fibers after alignment. This example
shows that even if the bundle includes a few spurious fibers, it can dra-
stically change the appearance of the mean fiber derived from Procrustes
analysis, while the MDP remains stable.

Figure 5 shows an example of MDPs for a population of subjects
overlaid on each other. Some of these paths are short because the
 corresponding regions of interest in the white matter atlas are small.

This means the seed points specified in the atlas are not very far apart
and even if the fibers are much larger they are summarized by the
structure within the white matter region and points with the highest
density. An alternative could be to use a probabilistic white matter
atlas and threshold the regions so they encompass a larger fraction of
the fiber lengths in the white matter region.

**Shape analysis of maximum density paths**

This section outlines the method for shape representation and ana-
lysis for maximum density paths. The maximum density paths denoting
tracts are modeled as continuous open curves in ℝ3 but they can also be
considered as points in an infinite-dimensional space of curves. This
space is induced by a suitable Riemannian metric defined on its tangent
space. Shape matching between MDPs is enabled by measuring shortest
length paths, also known as geodesics connecting two shapes in the
shape space. The geodesic not only measures the length of the path
and quantifies the geometric distance between two shapes, but also
represents an optimal geometric deformation that highlights the ana-
tomical differences between the shapes. Additionally, geodesics are an
important ingredient for constructing intrinsic population averages for
shapes — an essential step in statistical analysis of shapes.

**Representation of MDPs**

We represent the shape of the coordinate function of the MDP
using a vector-valued function (Joshi et al., 2007a,b; Srivastava et al.,
2011) as

\[
q(s) = \frac{\beta(s)}{\sqrt{\|\beta'(s)\|}} \in \mathbb{R}^3. 
\]  

(9)

Here, \(s \in [0, 1]\), \(\|\cdot\| \equiv \sqrt{(\cdot, \cdot)_{\mathbb{R}^3}}\), and \((\cdot, \cdot)_{\mathbb{R}^3}\) is the standard Euclidean
inner product in \(\mathbb{R}^3\). Our goal is to achieve an elastic shape matching
between MDPs. We would also prefer that the shape matching is
invariant with respect to the orientation and scale of the MDPs. Owing
to the derivative, the function \(q(s)\) is invariant to translation and uniform scaling. The elasticity
of the representation is due to the presence of the square-root in the
denominator that allows the \(q\) function to have arbitrary speeds. To
define a metric on the space \(Q\), we first define its tangent space which
is the set of all tangent vectors orthogonal to \(Q\). Formally, the tangent
space of \(Q\) is given by \(T_Q Q = \{v = (w_1, w_2, \ldots), (w_1): [0, 1] \to \mathbb{R}^3, \}
\(w_i \in [0, 1]\) such that \(\int_0^1 (w(s), q(s))_{\mathbb{R}^3} ds = 0\), where \(n = 3\). Here
each \(w_i\) represents a tangent vector in the tangent space of \(Q\). Now,
the metric on the tangent space \(T_Q Q\) is defined as follows. Given a curve \(q \in Q\), and the first order perturbations of \(q\) given by \(u, v \in T_Q Q\), respectively, the inner product between the tangent
curves \(u, v \in T_Q q\) at \(q\) is defined as

\[
\langle u, v \rangle = \int_0^1 (u(s), v(s))_{\mathbb{R}^3} ds. 
\]  

(10)

Now given two shapes \(q_1\) and \(q_2\), the translation and scale-invariant
shape distance between them is simply found by measuring the length of
the geodesic, or the great circle connecting them on the sphere. Thus

Please cite this article as: Prasad, G., et al. Automatic clustering and population analysis of white matter tracts using maximum density paths, Neuroimage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.04.033
given a tangent vector \( f \in T_{q_1}(\mathcal{Q}) \) in the direction of \( q_2 \) given by \( f = q_2 - q_1 \), the geodesic (Joshi et al., 2007a) on \( \mathcal{Q} \) between the two points \( q_1, q_2 \in \mathcal{Q} \) along \( f \), for an infinitesimal time \( t \) is given by

\[
\chi_t(q_1; f) = \cos \left( t \cos^{-1}(q_1, q_2) \right) q_1 + \sin \left( t \cos^{-1}(q_1, q_2) \right) f.
\]

Then the geodesic distance between the two shapes \( q_1 \) and \( q_2 \) is given by

\[
d(q_1, q_2) = \int_0^1 \sqrt{(\dot{\chi}_t)^2} \, dt.
\]

The geodesic distance (Joshi et al., 2007a) given in Eq. (12) is only invariant to translation and scale. To make it invariant to rotations, we consider the shortest distance

\[
d_r(q_1, q_2) = \arg \min_{\gamma \in SO(3)} d(q_1, \hat{O}_2 q_2).
\]

Eq. (13) can be minimized either using gradient descent over the tangent space of \( SO(3) \) or by using singular value decomposition (Rohlf and Slice, 1990). In this paper, we find the rotation invariant distance as

\[
d_r(q_1, q_2) = d(q_1, \hat{O}_2 q_2),
\]

where \( \hat{O}_2 = ADB^T = \int_0^{\pi} q_1(s)q_2(s)^T \, ds \), \( A \) and \( B \) are left and right unitary matrices, and \( D \) is a matrix given by \( D = \begin{bmatrix} 1 & 0 \\ 0 & |A||B| \end{bmatrix} \). Finally, since we are representing MDPs by parameterized curves, we would like the shape matching to be invariant to reparameterizations. Following (Joshi et al., 2007a), we denote the reparameterization of a MDP curve using a group action by a diffeomorphism \( \gamma \), given by \( q \cdot \gamma = \sqrt{\gamma}q(\gamma) \). Then the optimal reparameterization \( \hat{\gamma} \) is approximated by the minimizer

\[
\hat{\gamma} = \arg \min_{\gamma} \left( \int_0^{2\pi} \left[ |q_1 - \tilde{q}_2 \cdot \gamma|^2 + \|	ilde{q}_2 - q_1 \cdot \gamma^{-1}|^2 \right] \, ds \right),
\]

Please cite this article as: Prasad, G., et al., Automatic clustering and population analysis of white matter tracts using maximum density paths, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.04.033
where $\bar{q}_2 = \hat{O}_4 q_2$. In this paper, we use dynamic programming to obtain the solution to Eq. (15). The fully elastic, pose, scale, and reparameterization invariant distance between MDPs is given by

$$d_e(q_1, q_2) = d_e(q_1, q_2; \gamma) = d\left(q_1, \left(\hat{O}_4 q_2\right) \cdot \gamma\right).$$

(16)

The optimal geodesic path can also be denoted by a one-parameter flow $\Psi$ and the tangent vector $\alpha_t$, such that

$$\Psi : \Psi_0(q_1, \alpha_t) = q_1, \Psi_1(q_1, \alpha_t) = \left(\hat{O}_4 q_2\right) \cdot \gamma.$$  

(17)

The optimal tangent vector can then be written from Eq. (17) as

$$\hat{\alpha}_t = \Psi^{-1}_1(q_1, \left(\hat{O}_4 q_2\right) \cdot \gamma).$$

(18)

**Statistical analysis of MDPs across a population**

To evaluate group-level effects of sex, age, disease or even genetic influences on the MDP representations of white matter tracts, we need a suitable mechanism for performing statistical analysis. As MDPs are represented by parameterized functions defined on a shape space, one natural approach is to use the inherent non-linearity of the shape space, and define appropriate statistical measures under the Riemannian metric in Eq. (10). This approach is also called an intrinsic statistical analysis and leads to the definition of the Karcher mean (also known as the Fréchet mean) (Joshi et al., 2013; Le, 1995; Srivastava et al., 2005) in the shape space of all MDPs. Given a collection of MDP shapes $\{q_i\}, i = 1, \ldots, N$, the Karcher mean is defined as

$$\mu = \arg\min_{\bar{q}} \sum_{i=1}^{N} d_e(q_i, \bar{q})^2.$$  

(19)

The computation of the Karcher mean involves computing geodesics at each step iteratively and proceeds as follows. For the first iteration, an extrinsic mean (Euclidean average) is computed and projected on the shape space. This is assumed to be the current estimate of the Karcher mean. For the subsequent iterations, geodesics are computed between all the individual shapes in the population to the current estimate. The tangent vectors $\left\{\hat{\alpha}_i, i = 1, \ldots, N\right\}$ are then computed as a result of minimizing Eq. (16) and averaged together. A geodesic flow is then constructed using Eq. (17) to yield a new estimate of the mean shape. This procedure is repeated until the geodesic variance given by the sum of the squared geodesic distances to the mean shape is minimized, and the mean converges. The Karcher mean completely relies on the geometry of the shape space and is useful in computing intrinsic statistical estimates such as covariances of MDPs. Additionally, the geodesics produce correspondences, making it easier to compare white matter measures projected on the MDPs across a population. This is also useful.
for studying differences in disease, sex, aging, or even heritability in a population.

Results

Experiments

We show experimental results on a dataset of \(N = 565\) young adults, including healthy twins and their siblings. The participants were scanned with a 4-Tesla Bruker Medspec MRI scanner, collecting 3D 105-gradient high angular resolution diffusion images (HARDI) and standard structural T1-weighted magnetic resonance images (MRI).

The images consisted of 55 slices, 2-mm thick, with a 1.79 × 1.79 mm\(^2\) in-plane resolution. For each person, we collected 94 diffusion-weighted images (\(b = 1159\) s/mm\(^2\)) using a uniform distribution of gradient directions on the hemisphere. We also collected 11 \(b_0\) (non-diffusion encoding) images and corrected all images for eddy current distortions and motion with FSL (www.fmrib.ox.ac.uk/fsl). Our cohort consisted of 367 women and 198 men, ranging from 20 to 29 years of age. Study participants gave informed consent; institutional ethics committees at the Queensland Institute of Medical Research, the University of Queensland, the Wesley Hospital, and at UCLA approved the study.

For each T1-weighted image, we manually removed non-brain tissue and registered it to the Colin27 (Holmes et al., 1998) high resolution brain template using a 9-parameter transformation. These tissue and registered it to the Colin27 (Holmes et al., 1998) high resolution brain template using a 9-parameter transformation. We also registered the JHU ROI atlas to the MDT for the MDP analysis. Our ROI consisted of 367 women and 198 men, ranging from 20 to 29 years of age. Study participants gave informed consent; institutional ethics committees at the Queensland Institute of Medical Research, the University of Queensland, the Wesley Hospital, and at UCLA approved the study.

We examined the reliability of the MDP construction procedure by analyzing subjects with repeat scans. Twenty-three subjects in the total population used in our analysis had repeat scans, which were used to test the stability of MDP construction across the two acquisitions.

We used the MDP algorithm to find corresponding points along each MDP and used paired-sample \(t\)-tests to study if the FA values in these white matter tract representations were significantly different.

Fig. 4 shows the collection of MDPs with 46 curves in each white matter region from the ROI atlas. Each of the 23 pairs is colored randomly with the two MDPs in a single pair having matching coloring. We corrected for multiple comparisons using the false discover rate (FDR) (Benjamini and Hochberg, 1995) at the 0.05 level and none of the values were significantly different between scans. This provides an indication of the stability of MDP representation and may help support a more meaningful interpretation of the subsequent statistical analyses.

Genetic effects on white matter morphology using MDPs

The twin cohort in the data is made up of monozygotic (MZ) and dizygotic (DZ) pairs, allowing us to estimate the relative contributions of additive genetic factors (A), shared environment (C), and unshared or unique environment (E) to the measures derived from the scans — in our case, FA along the MDPs. This standard "A/C/E" model describes the FA at each point on the MDP as a combination of latent variables, \(FA = aA + cC + eE\). In this formulation the total variance is \(\text{var}(FA) = a^2 + c^2 + e^2\) with \(\text{var}(A) = a^2\), \(\text{var}(C) = c^2\), and \(\text{var}(E) = e^2\). We are able to estimate the unobserved factor loadings as there is a difference in the theoretical covariance of FA for a MZ twin pair, \(a^2 + c^2\), and for a DZ twin pair, \(1/2)d^2 + c^2\), which we solved using a maximum likelihood fitting (Neale and Cardon, 1992) that estimated the parameters of the model. These methods are detailed in (Chiang et al., 2009).

Several studies (Chiang et al., 2011; Jahanshad et al., 2013b; Patel et al., 2010; Thompson et al., 2001) have shown evidence for heritability of the white matter structure in the brain. Here, we use heritability as a metric to understand how well MDPs were able to model and capture the underlying morphology of the white matter structure in our data. If the representation is able to effectively pick up heritability effects then our hypothesis is that the MDP matching across subjects reflects the underlying anatomical homology, and that the MDP model is better able to describe white matter brain structure.

We fitted the "A/C/E" model to the FA values on the skeleton that fell within the ROI atlas. In our experiments we compared the full "A/C/E" model to the simpler formulation with two variables using minus two times the log-likelihood ratio, which approximately follows a \(\chi^2\) distribution, meaning that \(P > 0.05\) indicates a good fit. We found that the shared environment term (C component) did not have a significant fit for either method, so we used a simplified "A/E" model instead. This model selection procedure and selecting the "A/E" model instead of the more complicated "A/C/E" is widely used (Geschwind et al., 2002) and common with real data (Baaré et al., 2001). In the "A/E" model, \(a^2\) and \(e^2\) were estimated from the FA values and used in the A/C/E model to assess the heritability effects.
represents the proportion of the variance due to additive genetic factors and the other parameter in the model, $e^2$, represents the proportion of variance that is due to environmental factors including measurement error. Thus we model the components of FA variance as simply $\text{var}(\text{FA}) = a^2 + e^2$ and since we are interested in the relative proportions of variance captured by each component we can normalize this equation by $\text{var}(\text{FA})$ to interpret their relationship as $1 = a^2 + e^2$ or $1 - \frac{a^2}{e^2} = \frac{e^2}{\text{var}(\text{FA})}$. The goal here is to use the model that best fits the data to understand the genetic and environmental contributions to the variance. Maximizing heritable estimates in this case may imply minimizing measurement error and therefore may represent a stronger metric for measuring white matter microstructure. In addition, these highly heritable regions present good candidates for genetic associations and could be useful for cutting down on the dimensionality of the image for these types of analyses.

We also compared our method of analyzing white matter bundles using MDPs and analyzing the white matter skeletons from TBSS (Fig. 7) in our subjects. We used the genetic contribution due to FA using both methods to compare heritability. We found that the density and FA images smoothed with an isotropic Gaussian filter with full-width at half maximum (FWHM) = 3 mm produced higher $a^2$ values (Chiang et al., 2012). We restricted the analysis to 238 (48 monozygotic and 71 dizygotic) of the 565 twins because of issues with the nonlinear registration from TBSS in the omitted subjects. We computed genetic associations with mixed-model regression (Aulchenko et al., 2007; Jahanshad et al., 2012a) along the MDPs using the genes NTRK1, CLU, and COMT. We found NTRK1 passed a local false discovery rate (FDR) threshold (for a single MDP) in 20 regions across our white matter atlas represented as MDPs. In addition, we found CLU passed local FDR at the anterior limb of internal capsule right, posterior limb of internal capsule right, and anterior corona radiata left, and COMT passed at the sagittal stratum right (including inferior longitudinal fasciculus and inferior fronto-occipital fasciculus). When we used a global FDR, by combining the 67 MDPs into one large MDP, NTRK1 was the only SNP that survived in 600 of the total 1897 points in the global MDP. The results from NTRK1 and CLU agree with earlier studies of this dataset using voxel-based maps (Braskie et al., 2011, 2012).

### Discussion and conclusion

We have presented a method for extracting, representing, and analyzing the geometry of white matter bundles using maximum density paths. Our method enables population analysis of diffusion-weighted images without relying exclusively on global registration of the images into the same space. Image registration is performed only once to transform the ROI white matter atlas to the subject space for the purpose of initializing the seed points for clustering fibers from tractography. Density image volumes are computed from the fiber bundles, and MDPs are constructed using Dijkstra’s algorithm by imposing a graph structure on the images. The shapes of MDPs are then brought into correspondence through geodesic curve registration, allowing us to focus specifically on the white matter region we are interested in without involving the rest of the image. Further, our method introduces a way to perform localized statistical analysis of white matter tracts. The MDPs, use the start and end points from major

---

Fig. 7. Our twin data contained monozygotic (MZ) twins that share 100% of their genetic material and dizygotic twins (DZ) that share 50% of their genetic material. This structure in our data allows us to use structural equation models, particularly the A/E model, to estimate the amount of variance in a phenotype (in our case the white matter structure) that is due to genetic effects (heritability), or to unique environment factors and measurement error. In this figure we show the genetic contribution to white matter structure using our maximum density (MDP) path representation method compared with the white matter skeleton from tract-based spatial statistics (TBSS). A high value (red) means a large fraction of the variance in white matter structure is determined by genetic effects while a low value (blue) means the variance in structure was accounted for more by the environment. Since this is the proportion of variance accounted for by heritability, an analogous figure of the environmental contributions would involve simply reversing the color bar. Previous studies have shown high heritability of white matter tissue and we used the fraction of genetic determination as the metric to evaluate how well our MDP representation summarized the white matter structure in our data. The MDP method may have a better ability to pick up on the heritability because of the curve registration that is computed for each white matter region individually, which improves coherence of homologous points across subjects. We used a subset of 238 subjects for this analysis (48 MZ pairs, 71 DZ pairs). The top panel shows the MDPs from 67 white matter regions with slices of the atlas overlaid. The TBSS results show orthogonal slices of the TBSS skeleton overlaid on the white matter atlas.

---

Please cite this article as: Prasad, G., et al., Automatic clustering and population analysis of white matter tracts using maximum density paths, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2014.04.033

---

G. Prasad et al. / NeuroImage xxx (2014) xxx–xxx
white matter pathways, and provide a compact representation so that
white matter tracts depend on an external parameter. These tools provide the foun-
dation for any study of white matter tracts or any type of population
analysis using diffusion weighted images. The complete procedure is
available as an end-to-end computational pipeline for white matter
tract-based analysis of diffusion images.

In the examples presented here, we used a global Hough transform
method for tractography, but the MDP representation is general enough
to be used with any type of tractography method. Our method relies on
density images from tractography, which could be computed using
streamlines (Basser et al., 2000), a deflection based algorithm (Lazar
et al., 2003), or any of the recent deterministic and probabilistic
methods (Zhan et al., 2013). In our case, we chose a tractography
method that could benefit from the information rich HARDI data, but
depending on the resources and data available, researchers may prefer
to use fibers from diffusion tensor imaging or diffusion spectrum
imaging (Wedeen et al., 2005) based algorithms. The graph-based rep-

representation for the fiber density volume led to conveniently incorporate

density information in the structure, and further led to an efficient
solution provided by Dijkstra’s algorithm. However we could have
formulated the problem using snakes (Kass et al., 1988) or splines
(Park and Lee, 2007) as well. Alternatively, other density representa-
tions such as those using surfaces (Zijdenbos et al., 1994) or using the
volumetric segmentations (Kubicki et al., 2005) directly would have
introduced a host of issues with registration and subsequent analysis
of correspondences. We chose to use an ROI atlas to select fibers for
analysis and representation with the MDP though alternative
approaches may work without relying on the registration of the atlas
into the image space. Future work could incorporate automatic clustering
of tractography fibers using approaches such as a hierarchical Dirichlet
processes mixture model (Wang et al., 2011), a voxel based approach
(Guevara et al., 2011), a spectral approach (Guevara et al., 2011), or
even shape clustering (Joshi and Srivastava, 2003; Joshi et al., 2004).

Hierarchical approaches may enable a user to specify the resolution of
MDPs in the brain tissue. As an alternative to FA, any other type of
statistics on the density paths could be used instead. We could compute
mean diffusivity (Le Bihan et al., 2001), generalized FA (Barmoutsis
et al., 2009), or the tensor distribution function and interpolate them
along each MDP. Our white matter analysis framework could even be
scored by their capacity (Prasad et al., 2013b) and used as measures of
connectivity to complement (Prasad et al., 2013a) and optimize our
representation of brain connectivity networks (Prasad et al., 2014).

Preliminary studies have used MDPs to study sex differences (Prasad
et al., 2011b), Alzheimer’s disease (Ni et al., 2012, 2014), 22q11.2
deletion syndrome (Villalon-Reina et al., 2012), and depression (Sacchet
et al., 2009), or the tensor distribution function and interpolate them
architectural intelligence (J. Neurosci. 32, 2212–2224).

Chiang, M., Barysheva, M., McMahan, K., de Zubicaray, G., Martin,
N., Wright, I., Thompson, P., 2008. Fluid registration of diffusion tensor imaging using

Chiang, M., Barysheva, M., Shattuck, D., Lee, A., Madsen, S., Avedissian, C., Klunder, A.,
algorithms for any study of white matter tracts or any type of population


Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and

Braskie, M., Jahanshad, N., Stein, J., Barysheva, M., McMahan, K., de Zubicaray, G., Martin,
31, 6570–6577.

Braskie, M., Jahanshad, N., Stein, J., Barysheva, M., Johnson, K., McMahan, K., de Zubicaray,
G., Martin, N., Wright, I., Ringman, J., et al., 2012. Relationship of a variant in the


Chiang, M., Leow, A., Dutton, R., Barysheva, M., Rose, S., McMahan, K., de Zubicaray, G.,
Toga, A., Thompson, P., 2008. Fluid registration of diffusion tensor imaging using

Chiang, M., Barysheva, M., Shattuck, D., Lee, A., Madsen, S., Avedissian, C., Klunder, A.,
Toga, A., McMahan, K., de Zubicaray, G., Martin, N., Hill, I., Toga, A., Wright, I.,
Thompson, P., 2011. Genetics of white matter development: a DTI study of 705

Chiang, M., Barysheva, M., McMahan, K., de Zubicaray, G., Johnson, K., Montgomery, G.,
Martin, N., Toga, A., Wright, I., Shapshak, P., et al., 2012. Gene network effects on
brain microstructure and intellectual performance identified in 472 twins. NeuroImage
72, 8732–8745.

Cichon, S., Cook, N., Daly, M., Farace, S., Gejman, P., Keleso, I., Lehoor, T., Levinson, D.,
Moran, A., Sklar, P., et al., 2009. Genomewide association studies: history, rationale, and

analysis allows for enhanced connectivity analysis. NeuroImage 59 (4), 3272–3282.


Breakdown of brain connectivity between normal aging and Alzheimer’s disease: a
structural k-core network analysis. Brain Connect. 3, 112–120.

Dennis, E., Jahanshad, N., McMahan, K., de Zubicaray, G., Martin, N., Hickie, I., Toga, A.,
Wright, M., Thompson, P., 2012. Development of brain structural connectivity
networks.