Improved segmentation reproducibility in group tractography using a quantitative tract similarity measure

Jonathan D. Clayden[1], Mark E. Bastin[2] & Amos J. Storkey[3]

[1]Neuroinformatics Doctoral Training Centre, School of Informatics; [2]Medical and Radiological Sciences (Medical Physics); [3]Institute for Adaptive and Neural Computation, School of Informatics; University of Edinburgh, Edinburgh, UK

The field of tractography is rapidly developing, and many automatic or semiautomatic algorithms have now been devised to segment and visualize neural white matter fasciculi *in vivo*. However, these algorithms typically need to be given a starting location as input, and their output can be strongly dependent on the exact location of this "seed point". No robust method has yet been devised for placing these seed points so as to segment a comparable tract in a group of subjects. Here, we develop a measure of tract *similarity*, based on the shapes and lengths of the two tracts being compared, and apply it to the problem of consistent seed point placement and tract segmentation in group data. We demonstrate that using a single seed point transferred from standard space to each native space produces considerable variability in tractography output between scans. However, by seeding in a group of nearby candidate points and choosing the output with the greatest similarity to a reference tract chosen in advance—a method we refer to as *neighborhood tractography*—this variability can be significantly reduced.

Introduction

Diffusion magnetic resonance imaging (dMRI; see Basser *et al.*, 1994; Le Bihan, 2003) provides directional information about water self-diffusion in the brain. This information is particularly rich near to neural white matter fasciculi, since the self-diffusion is preferentially directed along rather than across these structures. The rapidly developing field of tractography consists of a growing number of voxel-level models of the dMRI signal data, and algorithms for integrating the results along fasciculi (*e.g.* Basser *et al.*, 2000; Behrens *et al.*, 2003b; Jansons & Alexander, 2003; Jones & Pierpaoli, 2005; Lazar & Alexander, 2005; Mori *et al.*, 1999; Parker *et al.*, 2002; Tournier *et al.*, 2004; Tuch *et al.*, 2002). These methods have been applied to neural connectivity analysis (*e.g.* Behrens *et al.*, 2003a; Huang *et al.*, 2005; Johansen-Berg *et al.*, 2005), as well as to the segmentation and visualization of white matter fasciculi (*e.g.* Jones *et al.*, 2005; Kanaan *et al.*, 2006; Pagani *et al.*, 2005; Zhang *et al.*, 2004). It is the segmentation application that we will be concentrating on here.

Tractography algorithms typically require as input a "seed point", a single voxel in the dMRI brain volume from which the algorithm begins its computation. The output of the algorithm is often highly sensitive to the specific location of this point. This sensitivity creates problems if one tries to segment comparable white matter structures in a group of brain volumes, because choosing congruent seed points is not straightforward. Placing the seed points manually (Ciccarelli *et al.*, 2005) is time consuming, and it has been shown that measurements of parameters such as mean fractional anisotropy (FA) or normalized volume in the resultant "tracts" (computed representations of the fasciculi of interest) can vary in value quite widely between observers, and particularly between scans (Ciccarelli *et al.*, 2003).

Relatively little work has been done to improve on manual seed point placement in group data. A two region of interest (2ROI) segmentation methodology, in which the tractography algorithm is instructed to ignore any pathways that do not pass through two predefined brain areas (Conturo *et al.*, 1999), has been applied to group tractography (Abe *et al.*, 2004); and Jones *et al.* (2002) developed a technique for averaging group dMRI data and performing tractography in the averaged brain volume. However, both of these methods have drawbacks. The strong *a priori* restriction that the 2ROI approach imposes upon the tractography algorithm may in some cases change the meaning of the output and make interpretation more complicated. The tensor averaging method by its nature discards individual anatomical variation, which may be crucial in clinical studies.

Rather than modifying tractography output, or the data from which it is generated, to suit a particular criterion, our aim in this work is to improve the reproducibility of tractography segmentation in group data by refining the input to the algorithm; *i.e.* the seed point. In order to eliminate observer subjectivity, the method should be automated.

A simple approach to automation is to place seed points in some standard space and then use standard image registration techniques to transfer them to each subject's native space (Clayden *et al.*, 2005). The aim of this approach, which we will refer to as the *registration method*, is to choose an equivalent seed point in each subject's brain volume, assuming that if this is done accurately then the same fasciculus will be segmented in each case. However, registration errors and anatomical variation between subjects make this assumption unsafe, limiting the usefulness of this approach. An alternative approach to the problem would be to choose, from a group of "candidate" seed points, that point which produces the best output. However, in this case one must quantitatively define what constitutes "good" or "correct" output.

In this study, we present a novel, quantitative tract similarity measure, based on the shape and length of the two tracts being compared. This measure is independent of the tractography algorithm being used to generate the tracts. In order to validate the measure, and demonstrate that it provides useful information, we use it to quantify similarity between independently generated comparable and disparate tracts in a group of volunteers. Finally, we apply the measure to the problem of consistent seed point placement across this subject group, and show that the set of tracts thus derived are more visually similar to one another than the set produced by the registration method.

Theory

Tracts and similarity

A calculation of similarity requires two tracts for comparison. We will assume that the tractography algorithm generating these tracts takes as input a single seed point, and produces voxelized, quantitative output. Hence we can define a tract r as

$$r = \{\mathbf{a}_r, \phi_r(\mathbf{x})\} \,, \tag{1}$$

where $\phi_r(\mathbf{x})$ is a discrete scalar field denoting the likelihood of a path from the seed point, $\mathbf{a}_r$, running through the voxel at location $\mathbf{x}$ in the native acquisition space of the subject. These two data elements are tied together because they represent both the input and output of the tractography algorithm. If $\mathbf{a}_r$ changes, then ϕ_r will change too.

The method by which the likelihood data are calculated from raw diffusion MR images will depend on the tractography algorithm and its underlying model of diffusion. For the purpose of illustration, we will briefly describe how this calculation was

performed for our data by the BEDPOST/ProbTrack algorithm (for full details, see Behrens *et al.*, 2003b). The diffusion weighted signal at each voxel, on the ith acquisition, is predicted by a partial volume model as a linear combination of an isotropic and an anisotropic component. That is,

$$A_i = A\left[(1-f)\exp(-b_i D) + f\exp(-b_i D \mathbf{r}_i^{\mathsf{T}} \mathsf{R} \mathbf{r}_i)\right] \ , \tag{2}$$

where A_i is the diffusion weighted signal, A is the signal without diffusion weighting, f is the anisotropic volume fraction, b_i and $\mathbf{r}_i$ are the b-value and gradient direction of the ith acquisition, D is the diffusivity and R is a matrix that encapsulates the directionality of the anisotropic component. Implicit in the latter matrix are the two Euler angles, $\{\theta, \phi\}$, that represent the direction of the underlying fiber tract. The distributions of the free parameters in this model, $\{A, D, f, \theta, \phi\}$, are estimated by BEDPOST using Markov Chain Monte Carlo sampling. Starting at the seed point, $\mathbf{a}_r$, ProbTrack then generates a large number of "probabilistic streamlines" by sampling from the local distribution over angles, moving a short distance in the sampled direction, and repeating until a termination criterion is met. Once this process is completed, the likelihoods $\phi_r(\mathbf{x})$ are given by the proportion of these streamlines that pass through the voxel at $\mathbf{x}$.

We will work on the principle that the characteristics of interest when comparing white matter tracts are length and shape. That is, if two tracts have the same shape *and* have the same length, then they are considered identical. For the purposes of comparison, we will make a distinction between "reference" and "candidate" tracts. There is no structural difference between the two, with both having the form given in Eq. (1), but similarity is always calculated for a candidate tract relative to a reference tract, rather than vice versa.

The following algorithm, which is based on a simplification and specialization of a general curve alignment algorithm (Sebastian *et al.*, 2003), provides sensitivity to the shapes of both the reference tract, r, and the candidate tract, c. Its output also depends on the length of the shorter of the two tracts. It moves along the two tracts simultaneously, voxel by voxel, finding a maximum likelihood pathway through the data, ϕ_r and ϕ_c, subject to certain path direction constraints. The output of the algorithm is a scalar value, $\sigma(r, c)$. The calculation is asymmetric, so that in general, $\sigma(r, c) \neq \sigma(c, r)$. The algorithm implicitly assumes that the seed points are equivalently located in the two tracts.

1. Initialize two sets of visited voxel locations, V_r and V_c, to the empty set.

2. Set tract pointers to the seed point location in each tract.

3. Add the current pointer position in the reference tract to the set V_r, and the position in the candidate tract to V_c.

4. Check the voxel values, from the field of connection likelihoods, ϕ_r, of the 26 voxels forming a cube around the current pointer location in the reference tract, and choose the largest valued neighboring voxel not in V_r. Note the step vector, $\mathbf{v}_r$, required to move to this new location.

5. Prohibiting movement at any angle greater than or equal to 90° from the chosen step direction in the reference tract, find the largest valued neighbor to the pointer in the candidate tract that is not in V_c. Note the step vector used here, $\mathbf{v}_c$.

6. Add the normalized inner product of the two step vectors to the result, $\sigma(r, c)$.

7. Move in the directions of the chosen steps and update the pointers in each tract.

8. Return to step 3, and repeat until there are no unvisited, nonzero voxels adjacent to one of the pointers. At this point, the algorithm has followed the reference tract to its end in one direction.

9. Return to step 2, and repeat until there are no unvisited, nonzero voxels adjacent to one of the starting points. The algorithm has now followed the reference tract to its end in all directions.

The normalized inner product calculated in step 6 is given by

$$\frac{\mathbf{v}_r \cdot \mathbf{v}_c}{\|\mathbf{v}_r\|\|\mathbf{v}_c\|} \ , \tag{3}$$

which is equivalent to the cosine of the angle between the two step vectors. The formulation of step 5 may seem to be excessively restrictive, but it simply ensures that the result is not undervalued due to the pointers drifting in opposite directions along the tract. This is an important issue because seed points are rarely placed at tract extremities, since such areas tend to be associated with high directional uncertainty, and so traversal away from the seed point can usually be in two, almost equally likely, directions. Note that there is no angle restriction in step 4.

The value of the σ function is translation invariant; but because we compare the local absolute directions of the tracts relative to the dMRI acquisition coordinate system, rather than curvature, it is not rotation invariant. This is desirable, since we do not want to produce spurious matches of potentially rotationally symmetric tracts such as the corpus callosum genu and splenium, or bilateral pairs.

Reduced tract

Tract data of the form given by Eq. (1) is not constrained to be a single voxel wide, and in general it will not be. As a result, the exact path taken through a reference tract can vary, and may be different during comparisons with different candidate tracts. This makes establishing an upper bound on the value of $\sigma(r, c)$ extremely difficult.

In order to alleviate this problem, we define a reduced version of the tract r to include that subset of the nonzero data in ϕ_r which is visited during the comparison of r with itself, a process that is illustrated, for a two dimensional case, in Fig. 1. Parts (a) and (c) of the figure represent two consecutive iterations of step 4 of the algorithm, and part (b) illustrates step 5. The shaded squares in the figure represent those voxels that the algorithm is allowed to move into, and the boxes with bold borders indicate visited voxels. After this calculation of $\sigma(r, r)$, the reduced tract, $\tilde{r}$, is defined as

$$\mathbf{a}_{\tilde{r}} = \mathbf{a}_r \qquad \phi_{\tilde{r}}(\mathbf{x}) = \begin{cases} \phi_r(\mathbf{x}) & \text{if } \mathbf{x} \in V_r \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where V_r is the set of visited voxel vectors calculated by the algorithm above. While r and $\tilde{r}$ are generally not identical, they are equivalent to the σ function in the sense that

$$\sigma(r, r) = \sigma(\tilde{r}, \tilde{r}) = \sigma(r, \tilde{r}) = \sigma(\tilde{r}, r) \ , \tag{5}$$

because all voxel locations whose data value is nonzero in r but not in $\tilde{r}$ are never visited. It must be remembered here that the tract data r includes the seed point, $\mathbf{a}_r$,

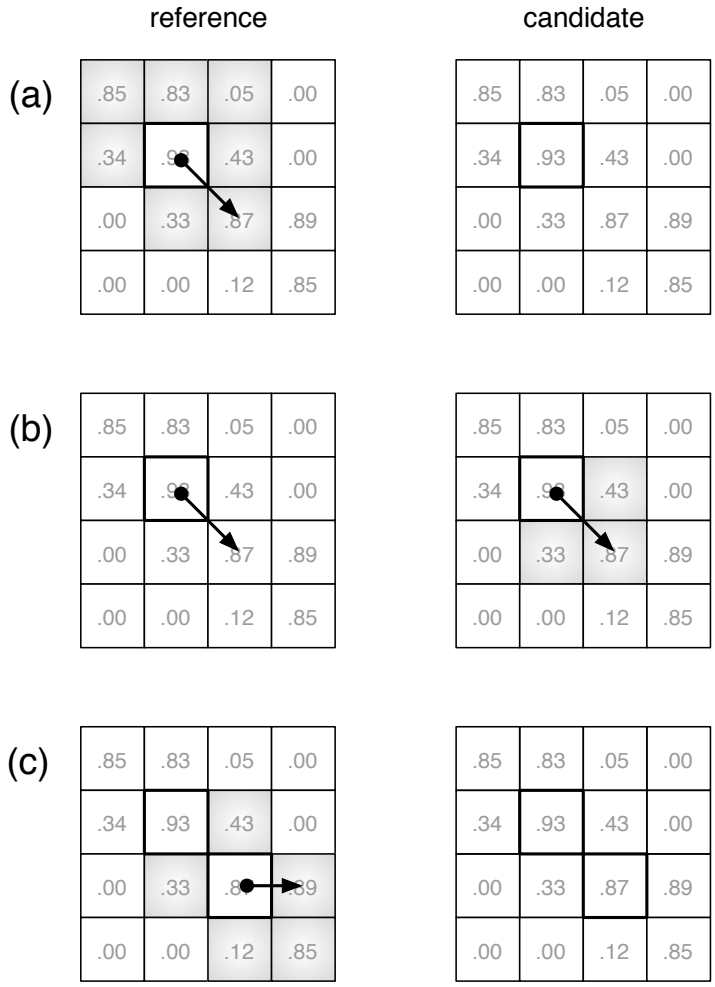

Figure 1: Two dimensional illustration of the shape similarity algorithm, as applied to two identical tracts. In (a), the boxes with bold borders represent the starting point, which has been marked visited. The shaded voxels in the reference tract indicate those nonzero, unvisited locations that the algorithm may legally move into, and the line represents the chosen step vector, from the current pointer location (circle head) to the next location (arrow head). In (b), movement in the candidate tract is restricted to those voxels whose angle from the chosen step direction in the reference tract is less than 90°. Since the voxel values are identical, the same direction is chosen. In (c), the next step in the reference tract cannot be back to the previous pointer location, since it is marked visited. In each diagram, numbers represent connection likelihood values at each voxel.

since this property will not hold if the same voxel data but different seed points were to be passed to the σ function.

When comparing a tract to itself the inner product calculated in step 6 of the algorithm will always be unity, and so the algorithm is merely counting the number of steps taken. Thus, the value of $\sigma(r, r)$ is exactly equal to the number of nonzero voxels in $\tilde{r}$ (excluding the seed point), and since each nonzero voxel can be visited at most once, producing a maximum contribution of one, we can establish the bounds

$$0 \leq \sigma(\tilde{r}, c) \leq \sigma(r, r) \quad \forall c \; . \tag{6}$$

The restriction that the pointer in the candidate tract can never move in a direction opposite to the reference tract ensures that all inner products are positive, and this fixes the lower bound in Eq. (6) at 0. Equivalently, $0 \leq \sigma(r, \tilde{c}) \leq \sigma(c, c)$ for any r.

Similarity measure

Using the tract comparison algorithm described above, we here develop measures of shape and length similarity, and then combine them together to form an overall similarity score.

We first approximate the length, L_r, of tract r as the number of voxels visited when it is compared to itself, excluding the seed point, which is given by

$$L_r \equiv \sigma(r, r) \; . \tag{7}$$

This length value is unchanged in the reduced tract, $\tilde{r}$, as shown by Eq. (5). Note that when comparing a tract to itself, shape is irrelevant because the local directionality of the reference and candidate tracts is always the same. If there are no nonzero voxels adjacent to the seed point, the data represents a "point tract", with length zero.

Given the definition of length in Eq. (7), and having calculated its value for the reference and candidate tracts, we establish the similarity of these two numbers using the symmetric normalized difference given by

$$S_1(r, c) = 1 - \left| \frac{L_r - L_c}{L_r + L_c} \right| = \frac{2 \cdot \min\{L_r, L_c\}}{L_r + L_c} = S_1(c, r) \; . \tag{8}$$

This measure has the value zero if either L_r or L_c is zero, and unity if the lengths are equal.

The other component of the similarity measure, the similarity in shape between the reference and candidate tracts, can be established using the asymmetric formulation

$$S_2(r, c) = \frac{\sigma(\tilde{r}, \tilde{c})}{\min\{L_r, L_c\}} \neq S_2(c, r) \; . \tag{9}$$

The denominator in Eq. (9) removes the length dependence of the σ function. The bounds on the σ function that were established above ensure that the value of Eq. (9) is always in the interval $[0, 1]$.

Finally, the two score components given by Eqs (8) and (9) are combined to form the overall similarity score,

$$S(r, c) \equiv \sqrt{S_1(r, c) \cdot S_2(r, c)} = \sqrt{\frac{2 \cdot \sigma(\tilde{r}, \tilde{c})}{L_r + L_c}} \; , \tag{10}$$

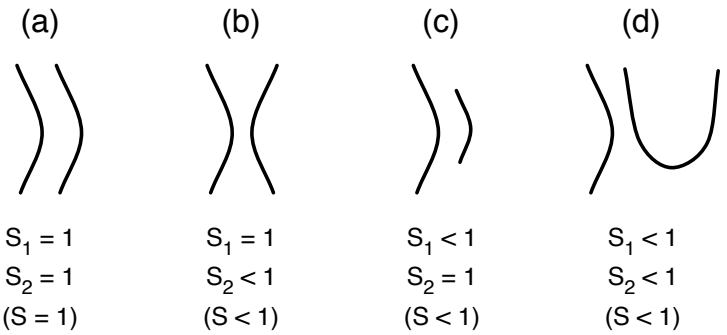

Figure 2: Qualitative demonstration of the effect on the two score components, S_1 (shape) and S_2 (length), of different types of relationship between the reference tract (fixed, and on the left in each case) and the candidate tract (variable, and on the right). The seed points are assumed to be in the centre of each tract throughout.

the geometric mean of the two components. A higher value of Eq. (10) indicates a better match, and a lower value indicates a worse match. The score will be 1 if r and c are the same tract. It will be 0 if either r or c is a point tract. The geometric mean lends a far stronger influence to very small values in one score component than does the arithmetic mean when finding the "average" similarity of c to r, and in particular, if *either* score component is 0 then the overall score is also 0. This formulation emphasizes that *both* length and curvature must be similar for the candidate tract to be considered a likely equivalent to the reference.

Fig. 2 shows four examples of tract pairs and their associated score components. In each case the reference tract is on the left, and the seed points are assumed to be placed exactly in the middle of each tract. These are idealized, and continuous rather than voxelized, tract curves; but they illustrate how the two score components will be affected in various scenarios. In (a), the the candidate tract is identical to the reference tract. This is equivalent to the case in Fig. 1. In (b), the candidate is a reflected copy of the reference. Note that the shapes of this pair of curves are considered different. In (c), the candidate is a central segment from the reference, so the shape is considered identical, but the lengths differ. It should be noted that this case represents a truncation rather than a scaling of the reference tract, as the latter would not produce an S_2 score of 1. Finally, in (d), the tracts are different in both shape and length.

Methods

Image acquisition

Six normal volunteers (2 male, 4 female; mean age 27 ± 3.4 years) were recruited for this study. Each subject underwent a dMRI protocol on a GE Signa LX 1.5 T clinical scanner (GE Medical Systems, Milwaukee, WI, USA), equipped with a self-shielding gradient set (22 mT m^{-1} maximum gradient strength) and manufacturer supplied "bird-cage" quadrature head coil. The protocol used a single-shot spin-echo echo-planar imaging sequence with 51 noncollinear diffusion weighting gradient directions at a b-value of 1000 s mm^{-2}, and 3 T_2-weighted scans. 48 contiguous axial slice locations were imaged, with a field of view of 220×220 mm, and a slice thickness of 2.8 mm. The

acquisition matrix was 96×96 voxels in-plane, zero filled to 128×128. TR was 17 s per volume and TE was 94.3 ms.

In order to investigate the variation in similarity scores between acquisitions, 2 of the subjects were scanned twice, and 3 were scanned three times. Those subjects that went through the protocol three times were taken out of the scanner between the second and third acquisitions, and the slice locations were repositioned for the third acquisition without reference to those chosen for the first two.

Data processing

The data were initially preprocessed to remove skull data and eddy current induced distortion effects from the images, using FMRIB Software Library (FSL) tools (FMRIB, Oxford, UK). As mentioned above, the underlying tractography algorithm used in this study was the BEDPOST/ProbTrack algorithm (Behrens *et al.*, 2003b), which is also part of the FSL suite. It should be remembered that the BEDPOST/ProbTrack model of the dMRI signal is a partial volume model assuming a single anisotropic diffusion direction at each voxel, and as such the measure of anisotropy it uses is the anisotropic volume fraction (AVF), rather than the more common, diffusion tensor based fractional anisotropy (FA). However, the two measures are closely related.

The aim of our first experiment was to validate the similarity measure described above, by investigating whether the measure could differentiate between comparable and disparate tracts in the group of volunteers. A series of 8 seed points were placed in major white matter fasciculi on a Montréal Neurological Institute (MNI) standard brain (Evans *et al.*, 1993), and transferred to each subject's native space using the FLIRT registration algorithm (Jenkinson & Smith, 2001), with the MNI white matter map used as a weighting volume (Clayden *et al.*, 2005). The specific seed regions chosen were genu and splenium of corpus callosum (CC), right and left anterior limb of internal capsule (ALIC), right and left posterior limb of internal capsule (PLIC), and right and left sagittal stratum (SS). Whilst the accuracy of seed point placement using this registration method is limited, it provides an independent mechanism for generating groups of tracts that can be expected to be more or less similar to one another. The ProbTrack tractography algorithm was run with each of these points as a seed, and similarity scores were calculated for various tract pair permutations. Comparisons between equivalent seed regions on the left and right of a single brain volume (*e.g.* left ALIC versus right ALIC) were labeled "bilateral", and all other comparisons within a single volume (*e.g.* left ALIC versus right PLIC) were labeled "nonbilateral". Comparisons across subjects for a single seed region (*e.g.* left ALIC in subject 1 versus left ALIC in subject 2) were labeled "intersubject"; and additional similarity scores were calculated between 1st and 2nd scans ("inter-NEX") and 2nd and 3rd scans ("interscan"), where available, within each subject and seed region. We expect that similarity scores will be lowest for the nonbilateral comparisons, and highest for the interscan and inter-NEX cases where the two tracts are from the same seed region and same subject. For every pair of tracts thus compared, similarity scores were calculated using each in turn as the reference tract.

A second experiment was then performed, aimed at applying the similarity approach to the problem of improving the robustness of seed point placement across a group of scans. For each seed region, a representative reference tract was chosen. For each scan, a $7 \times 7 \times 7$ cube of voxels around, and including, the voxel suggested by the registration method (hereafter the "original" seed point) for each fasciculus of interest were used as seed points for the tractography algorithm, except where the

Subject	Scan 1	Scan 2 (inter-NEX)	Scan 3 (interscan)
1	(a)	(b)	(c)
2	(d)	(e)	
3	(f)	(g)	(h)
4	(i)	(j)	(k)
5	(l)	(m)	
6	(n)		

Table 1: Correspondence between the different scans and the subfigure labels used in Figs 4 and 5.

Seed	RM score mean	RM score s.d.	NT score mean	NT score s.d.
1 (CC genu)	0.488	0.056	0.597	0.018
2 (CC splenium)	0.354	0.106	0.542	0.031
3 (right ALIC)	0.529	0.098	0.651	0.026
4 (left ALIC)	0.463	0.099	0.644	0.027
5 (right SS)	0.329	0.220	0.680	0.023
6 (left SS)	0.365	0.077	0.516	0.024
7 (right PLIC)	0.405	0.096	0.570	0.030
8 (left PLIC)	0.444	0.054	0.594	0.025

Table 2: Mean and standard deviation of similarity scores for all tracts chosen by neighborhood tractography (NT) in each of the 8 seed regions, determined from the 6 volunteers (14 scans). The means and standard deviations for tracts chosen by the registration method (RM) are given for comparison.

voxel AVF was less than 0.2, an empirically chosen threshold used to avoid seeding in cerebrospinal fluid or gray matter. The tract with the highest similarity score when compared to the relevant reference tract was then selected as the "best" tract from each brain volume. We will refer to this technique as *neighborhood tractography*.

In all of the experiments described above, reference and candidate tract data (*i.e.* the fields ϕ_r and ϕ_c) were thresholded at the 1% level before similarity scores were calculated. This was done to avoid inclusion of very low confidence paths in the comparisons.

Results

Fig. 3 shows the results of the first experiment as a box-and-whisker plot. The mean ($\pm$ standard deviation) similarity score for each group of tract comparisons was 0.14 ($\pm$0.13) for nonbilateral, 0.31 ($\pm$0.13) for bilateral, 0.38 ($\pm$0.12) for intersubject, 0.47 ($\pm$0.09) for interscan, and 0.46 ($\pm$0.12) for inter-NEX. Two sample, two tailed t-tests showed significant differences between nonbilateral and bilateral scores ($P < 10^{-9}$), between bilateral and intersubject scores ($P = 0.005$), and between intersubject and interscan scores ($P < 10^{-6}$). There was no significant difference between interscan and inter-NEX similarity scores ($P = 0.89$).

Results from the second experiment are shown visually in Figs 4 and 5. The correspondence between the letters labeling each subfigure and the different scans is shown in Table 1. Fig. 4 shows the tract fields produced by seeding ProbTrack at the original

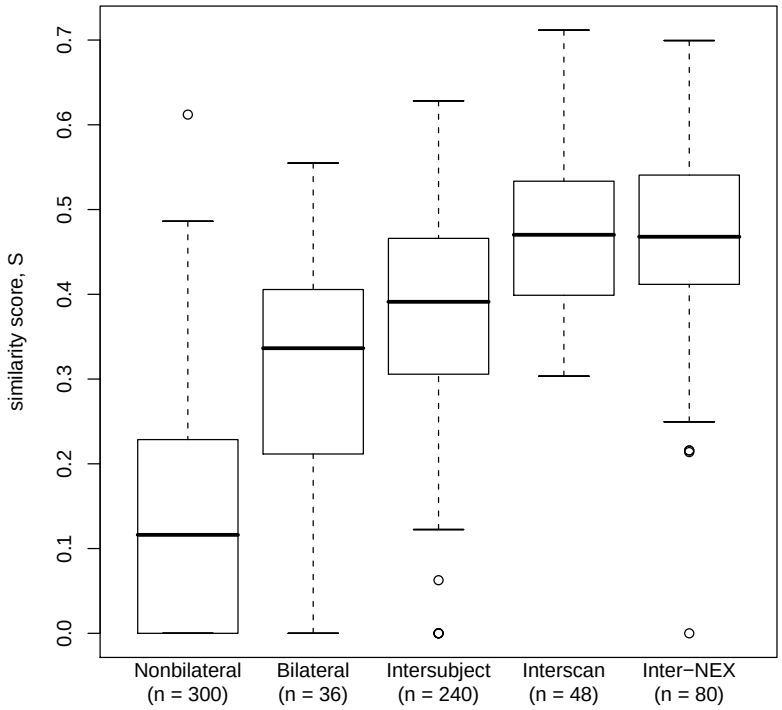

Figure 3: Box-and-whisker plot showing the range of similarity scores for the five different categories of comparison in the first experiment. The thick horizontal line across each box represents the median, the box shows the interquartile range, the whiskers show the extent of the bulk of the data, and circles show outliers more than 1.5 interquartile ranges from the box. The n values indicate the number of scores making up the data for each plot. The data demonstrate appropriate score increases across the different test conditions, suggesting that the score provides meaningful and useful information.

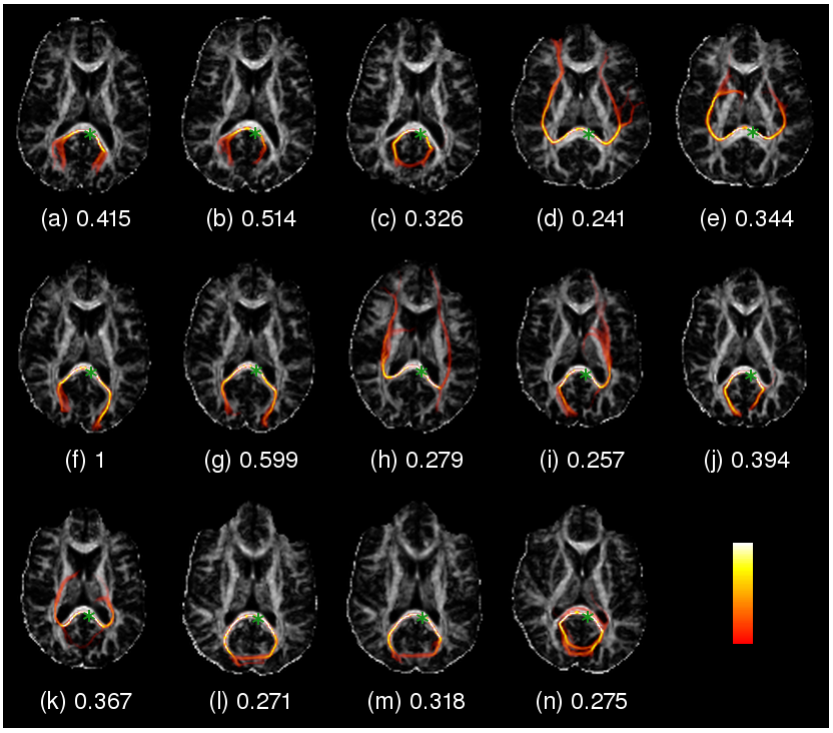

Figure 4: Two dimensional axial projections of the tracts generated by the ProbTrack algorithm using the original seed points chosen by the registration method, overlaid on AVF maps of the slice in plane with the seed in each case. White indicates high AVF and black low. In the tracts, yellow indicates high likelihood of connection to the seed point, and red low. The green stars indicate the seed point locations. The similarity score to the reference tract (f) is shown in each case.

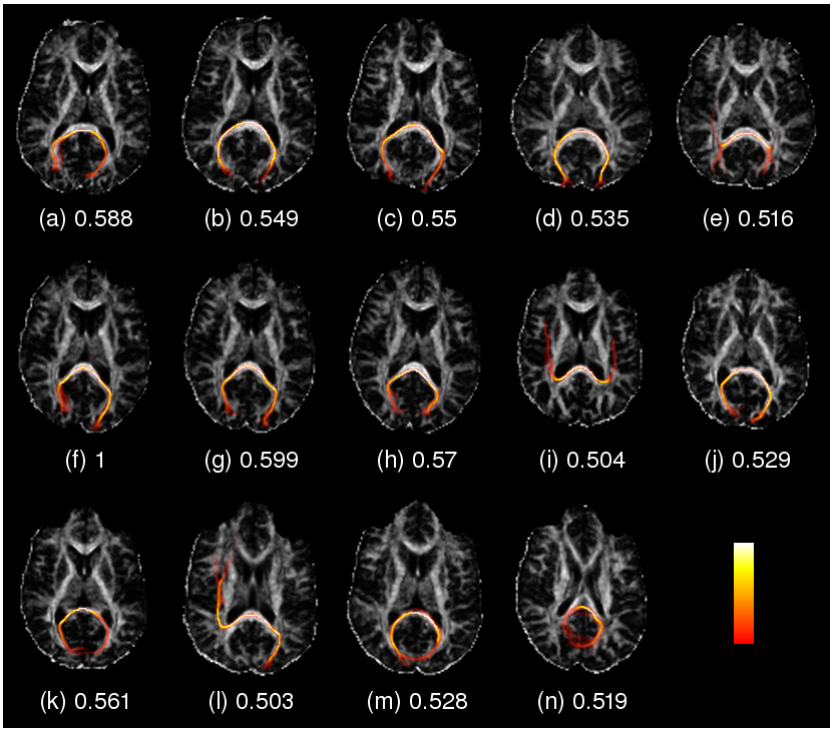

Figure 5: Projections of the tracts chosen as the "best" (highest similarity to the reference tract), using a $7 \times 7 \times 7$ seeding neighborhood around the original seed point. Individual similarity scores are also shown. Tract (f) is the reference tract.

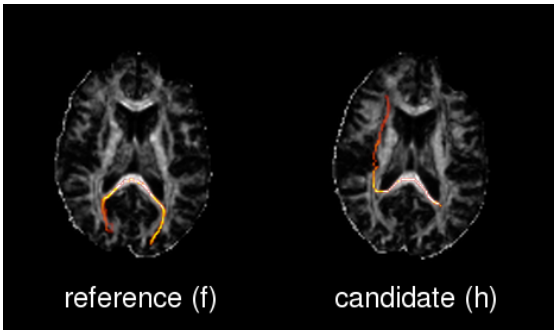

Figure 6: Examples of reduced reference and candidate tracts, produced from two of the unreduced tracts shown in Fig. 4.

seed point in splenium of corpus callosum, and thresholding the results at the 1% level. This seed region was chosen as the example because considerable variation in tract shape can be seen across the group: the resultant tracts demonstrate pathways running anterior (d, e, h, k), posterior (a–c, f, g, j, l–n) or both (i) from the edges of the corpus callosum itself. Fig. 5 shows the tracts chosen by the neighborhood tractography approach, after the same 1% threshold has been applied. Both figures also show the similarity scores associated with each tract, using (f), which is the same in both cases, as the reference tract. In Fig. 5, similarity scores are necessarily greater than or equal to the corresponding score in Fig. 4, and only two tracts (i, l) remain that do not project in the posterior direction from the corpus callosum. These two tracts have the two lowest similarity scores in the figure. Tract (g), which has the highest score apart from the reference tract, is found in the same subject as the reference tract, so the fasciculus it represents is identical.

Fig. 6 shows examples of "reduced" reference and candidate tracts, in the sense described in the Theory. It shows how the reduction affects the tracts. In this case, the reference tract is simply slightly narrower than its unreduced equivalent, Fig. 4(f). The candidate tract is truncated at the edge of the splenium, where the unreduced version, Fig. 4(h), had an ambiguous branch.

The mean and standard deviation of the similarity scores for the tracts chosen before and after applying neighborhood tractography for each seed region, across all subjects and acquisitions, are given in Table 2. The figures for the "best" tracts (as chosen by neighborhood tractography) represent narrow and seed specific score distributions, whose coefficients of variation are in the range 3.0–5.7%. By comparison, the original scores, generated by the registration method, are invariably lower with wider standard deviations. Their coefficients of variation are in the range 11.5–66.9%.

Discussion

While tract shape has been studied before (Corouge *et al.*, 2004), previous work has been aimed at modeling individual tracts, rather than doing pairwise similarity scoring. As far as we are aware, the present study represents the first attempt at using a quantitative tract similarity measure to improve segmentation reproducibility.

The results from our first experiment provide evidence that the similarity measure described above produces higher scores for a single seed region across a range of healthy subjects, than it does for a range of seed regions within a single subject, as demonstrated by higher intersubject than bilateral and nonbilateral similarity scores. Behavior of this nature is clearly crucial for any tract similarity measure that is intended to be used as a basis for the identification of comparable tracts across a group of subjects. It is not surprising to find that comparisons between bilateral seed regions (such as left versus right ALIC) produce generally higher scores than other comparisons (such as left ALIC versus right PLIC), since comparable white matter fasciculi in the two hemispheres can be expected to have similar lengths and related shapes. Nevertheless, even the bilateral scores are significantly smaller than the intersubject scores.

The finding that interscan and inter-NEX scores are indistinguishable is an interesting one. It suggests that repositioning of the slice positions introduces no consistent bias to the results of the similarity measure, demonstrating a useful robustness to subtle changes in the slice locations. It is additionally reassuring to see that both these sets of scores are significantly higher than the intersubject scores, since the underlying fasciculi are *the same* across acquisitions, rather than merely comparable as they are in

the intersubject case.

Our findings from the second experiment are also encouraging. The results in Fig. 4 show the inadequacy of the registration method: despite comparable seed point locations in each brain volume, the shapes of the resultant tracts are highly variable. By contrast, the tracts shown in Fig. 5 are generally more similar to the reference tract (f) than those in Fig. 4, at least in terms of the qualitative observation that fewer of them project anteriorly from the splenium of corpus callosum. In addition, those tracts which clearly do project along very dissimilar pathways to the reference tract (i, l) have similarity scores lower than the remaining 12 tracts. Tract (g) in Fig. 5, which appears most similar in shape to the reference tract, appropriately generates the largest similarity score.

The narrowness of the score distributions for each seed point (see Table 2) seems to indicate that the scoring algorithm is quite strongly influenced by the nature of the reference tract. This may be because the part of each tract near the seed point in each direction is relatively reproducible, whereas the spatially uncertain regions near the ends of tracts are very unlikely to produce a perfect match with the reference tract. The combination of these two factors may effectively impose reference tract specific upper and lower score bounds.

We do not claim that the tracts shown in Fig. 5 are anatomically correct. Validation of tractography output is a complex issue in its own right (Mori & van Zijl, 2002), since there are no other established techniques for studying white matter structures *in vivo*. We do, however, claim that these tracts are more similar to the reference tract (which is chosen for illustration) than those shown in Fig. 4. If the nature of the reference tract were later to be found to be inappropriate, it could be updated and neighborhood tractography repeated without change. The method would then find the best match to the new reference tract.

After the fact examination of the locations of the "best" voxels relative to the original seed points showed that 63% of seed points chosen by neighborhood tractography are not more than 2 voxels from the original seed point in any direction. While this proportion is high enough to suggest that a $7 \times 7 \times 7$ search cube is generally sufficiently large, it may be that using a larger search space would have improved the results of our application experiment in certain cases. However, for larger search spaces it would probably be necessary to use more complex heuristics for culling seed points than the simple AVF threshold used here, so as to keep run times reasonable.

A major advantage of the approach taken in this work, when compared to group data averaging (Jones *et al.*, 2002), is that no spatial manipulation of each individual brain volume is required before tractography can be performed, and so potentially interesting anatomical variation across the group need not be averaged away or otherwise distorted. For this reason, we have made no alterations or corrections for factors such as natural variation in brain size and shape, or head rotation. These factors will have some effect on the absolute similarity scores, but they will affect all candidate tracts, and the neighborhood tractography method is only interested in *relative* similarity, compared to other candidates. A correction based on a transformation of the candidate tract into the space of the reference tract would, in any case, have problems of its own, since interpolating the tract data could alter its structure in undesirable ways. For example, local duplication of voxel values would be strongly suboptimal for our similarity algorithm.

Since differences in head rotation and head size between scans will have a complex, nonlinear effect on the similarity measure, and may affect different tracts differently, it is not straightforward to establish the impact of these variates, or to recommend upper bounds on acceptable rotations or scalings. Moreover, working with simulated

data would add another image processing step, which may be a source of variance, and would introduce similar interpolation issues to a correction. However, interscan rotations for single subjects are present in our data set. Linear registrations between pairs of T_2-weighted images suggest that the median rotation between a subject's first scan and their third was 1.5° (4.3° about the left–right axis, 0.6° about the anterior–posterior, and 1.1° about the superior–inferior). Hence, some variance due to rotation is incorporated into the results from our first and second experiments; but it should be remembered that in the first experiment, inter-NEX and interscan scores were statistically indistinguishable, despite much smaller rotations in the former case (median of 0.3°), suggesting a certain robustness to such effects.

Unlike neighborhood tractography as presented here, the 2ROI method (Conturo *et al.*, 1999) is a modification of tractography algorithms themselves. Neither method is explicitly dependent on the particular algorithm in use, but the 2ROI approach changes the meaning of the tractography output. Specifically, a probabilistic algorithm will no longer indicate absolute connection likelihood to the seed point. The data will instead represent some information about the relative likelihoods of different connection routes between the two regions of interest. By contrast, whilst neighborhood tractography provides a preference for certain shapes and lengths of output, it does not change the tract data at all. Where appropriate, neighborhood tractography and the 2ROI method could be complementary.

Some tractography studies have used a volume of interest (VOI) approach, in which the tractography algorithm is seeded from a cluster of voxels and the results combined together (*e.g.* Kanaan *et al.*, 2006; Pagani *et al.*, 2005). This method is related to neighborhood tractography, because both approaches seed in a region rather than a single point, but while the VOI method retains all of the resulting data, neighborhood tractography retains only the output from a single chosen seed point. The potential advantage of the all-but-one rejection performed by neighborhood tractography is that the specificity of single seed point tractography remains—allowing, in principle, for the study of white matter structures only a single voxel wide—whilst undesirable sensitivity that hinders more straightforward single seed methods is reduced. It is not clear to what extent it is reasonable, in general, to treat the sum of the outputs from each point in a VOI as a single white matter structure.

The similarity measure described above aims to be relatively simple whilst capturing important characteristics of the two tracts that we wish to compare. It would be possible to make minor specializations or refinements to the algorithm that calculates the σ function without changing the overall framework, provided the bounds on the score, S, remain. This simplicity aids portability. Whilst probabilistic tractography algorithms tend to produce tract data of the form given by Eq. (1), some other approaches, particularly streamline based algorithms, instead produce a single line of infinitesimal thickness through the seed point. In these cases, the principle of our similarity calculation would still be applicable, and in fact the method would become even simpler because there would no longer be any need to produce a reduced tract.

The main weakness of the similarity measure presented here is that the termination criterion in step 8 of the algorithm (see Methods) can be met prematurely if a local "loop" of relatively high valued voxels is encountered. This leads to underscoring or false negatives, and is likely to be at least a contributor to the problem of narrow score distributions for a particular reference tract, and the reason that tract (l) is *less* visually similar to (f) in Fig. 5 than in Fig. 4. That result, when taken in context with the rest of the data, suggests that while a high score seems to indicate a good match between tracts, a low score may not reliably indicate a bad match. The underscoring

problem could perhaps be alleviated by biasing the algorithm in favor of continuing in the same direction as its previous step, and introducing some fuzziness into the choice of local maximum voxel in step 4 of the algorithm. However, these changes would render the algorithm nondeterministic, and care would have to be taken to ensure that the maximum and minimum scores remain tractable.

We do not claim that our similarity measure is optimal, or that tract shape and length are the only characteristics of interest, but our results demonstrate that dMRI data does provide enough information to meaningfully compare tractography output algorithmically. The measure presented here is not directly based on any explicit model of intersubject tract variability, so a constructive future direction for development would be to construct such a model in a formal probabilistic framework, and derive a more rigorous similarity measure from that model.

The effect of the neighborhood tractography approach will necessarily depend on the underlying data, the tractography algorithm and the similarity measure being used, so further investigation will be required to find the best combination of these factors. Nevertheless, we believe that approaches to group tractography based on an attempt to pinpoint a single, "equivalent" seed point in each brain volume face almost unassailable difficulties; whereas after the fact comparisons of candidate tracts with a reference tract in the manner described in this study could provide a truly robust foundation for automated, reproducible tract segmentation in group dMRI data. With further improvement to similarity measures, it may ultimately be possible to produce a standard set of reference tracts, much as standard brain images are produced now, and to use these to reliably segment specific fasciculi with tractography in a group of brain volumes. Furthermore, a similarity measure with sensitivity to tract shape, such as ours, could be applied to quantify white matter distortion effects in pathologies such as brain tumors.

Acknowledgments

The authors would like to thank Dr Paul Armitage for his comments on this work. JDC is supported by an EPSRC/MRC studentship via the Neuroinformatics Doctoral Training Centre at the University of Edinburgh. All MR scanning was performed at the SHEFC Brain Imaging Research Centre for Scotland (HTTP://WWW.DCN.ED.AC.UK/BIC).

References

Abe O., Yamada H., Masutani Y., Aoki S., Kunimatsu A., Yamasue H., Fukuda R., Kasai K., Hayashi N., Masumoto T., Mori H., Soma T. & Ohtomo K. (2004). Amyotrophic lateral sclerosis: diffusion tensor tractography and voxel-based analysis. *NMR in Biomedicine* **17**(6):411–416.

Basser P., Mattiello J. & Le Bihan D. (1994). Estimation of the effective self-diffusion tensor from the NMR spin echo. *Journal of Magnetic Resonance, Series B* **103**(3):247–254.

Basser P., Pajevic S., Pierpaoli C., Duda J. & Aldroubi A. (2000). In vivo fiber tractography using DT-MRI data. *Magnetic Resonance in Medicine* **44**(4):625–632.

Behrens T., Johansen-Berg H., Woolrich M., Smith S., Wheeler-Kingshott C., Boulby P., Barker G., Sillery E., Sheehan K., Ciccarelli O., Thompson A., Brady J. & Matthews P. (2003a). Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging. *Nature Neuroscience* **6**(7):750–757.

Behrens T., Woolrich M., Jenkinson M., Johansen-Berg H., Nunes R., Clare S., Matthews P., Brady J. & Smith S. (2003b). Characterization and propagation of uncertainty in diffusion-weighted MR imaging. *Magnetic Resonance in Medicine* **50**(5):1077–1088.

Ciccarelli O., Parker G., Toosy A., Wheeler-Kingshott C., Barker G., Boulby P., Miller D. & Thompson A. (2003). From diffusion tractography to quantitative white matter tract measures: a reproducibility study. *NeuroImage* **18**(2):348–359.

Ciccarelli O., Toosy A., Hickman S., Parker G., Wheeler-Kingshott C., Miller D. & Thompson A. (2005). Optic radiation changes after optic neuritis detected by tractography-based group mapping. *Human Brain Mapping* **25**(3):308–316.

Clayden J., Marjoram D., Bastin M., Johnstone E. & Lawrie S. (2005). Towards an automated method for white matter integrity comparison between populations. In *Proceedings of the ESMRMB 22nd Annual Meeting*, 508. European Society for Magnetic Resonance in Medicine and Biology.

Conturo T., Lori N., Cull T., Akbudak E., Snyder A., Shimony J., McKinstry R., Burton H. & Raichle M. (1999). Tracking neuronal fiber pathways in the living human brain. *Proceedings of the National Academy of Sciences of the United States of America* **96**(18):10,422–10,427.

Corouge I., Gouttard S. & Gerig G. (2004). *A Statistical Shape Model of Individual Fiber Tracts Extracted from Diffusion Tensor MRI*, vol. 3217 of *Lecture Notes in Computer Science*. Springer-Verlag.

Evans A., Collins D., Mills S., Brown E., Kelly R. & Peters T. (1993). 3D statistical neuroanatomical models from 305 MRI volumes. In *Nuclear Science Symposium and Medical Imaging Conference*, vol. 3, pp. 1813–1817. IEEE.

Huang H., Zhang J., Jiang H., Wakana S., Poetscher L., Miller M., van Zijl P., Hillis A., Wytik R. & Mori S. (2005). DTI tractography based parcellation of white matter: application to the mid-sagittal morphology of corpus callosum. *NeuroImage* **26**(1):195–205.

Jansons K. & Alexander D. (2003). Persistent angular structure: new insights from diffusion magnetic resonance imaging data. *Inverse Problems* **19**(5):1031–1046.

Jenkinson M. & Smith S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis* **5**(2):143–156.

Johansen-Berg H., Behrens T., Sillery E., Ciccarelli O., Thompson A., Smith S. & Matthews P. (2005). Functional-anatomical validation and individual variation of diffusion tractography-based segmentation of the human thalamus. *Cerebral Cortex* **15**(1):31–39.

Jones D., Griffin L., Alexander D., Catani M., Horsfield M., Howard R. & Williams S. (2002). Spatial normalization and averaging of diffusion tensor MRI data sets. *NeuroImage* **17**(2):592–617.

Jones D. & Pierpaoli C. (2005). Confidence mapping in diffusion tensor magnetic resonance imaging tractography using a bootstrap approach. *Magnetic Resonance in Medicine* **53**(5):1143–1149.

Jones D., Travis A., Eden G., Pierpaoli C. & Basser P. (2005). PASTA: Pointwise assessment of streamline tractography attributes. *Magnetic Resonance in Medicine* **53**(6):1462–1467.

Kanaan R., Shergill S., Barker G., Catani M., Ng V., Howard R., McGuire P. & Jones D. (2006). Tract-specific anisotropy measurements in diffusion tensor imaging. *Psychiatry Research: Neuroimaging* **146**(1):73–82.

Lazar M. & Alexander A. (2005). Bootstrap white matter tractography (BOOT-TRAC). *NeuroImage* **24**(2):524–532.

Le Bihan D. (2003). Looking into the functional architecture of the brain with diffusion MRI. *Nature Reviews Neuroscience* **4**(6):469–480.

Mori S., Crain B., Chacko V. & van Zijl P. (1999). Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging. *Annals of Neurology* **45**(2):265–269.

Mori S. & van Zijl P. (2002). Fiber tracking: principles and strategies - a technical review. *NMR in Biomedicine* **15**:468–480.

Pagani E., Filippi M., Rocca M. & Horsfield M. (2005). A method for obtaining tract-specific diffusion tensor MRI measurements in the presence of disease: application to patients with clinically isolated syndromes suggestive of multiple sclerosis. *NeuroImage* **26**(1):258–265.

Parker G., Stephan K., Barker G., Rowe J., MacManus D., Wheeler-Kingshott C., Ciccarelli O., Passingham R., Spinks R., Lemon R. & Turner R. (2002). Initial demonstration of in vivo tracing of axonal projections in the macaque brain and comparison with the human brain using diffusion tensor imaging and fast marching tractography. *NeuroImage* **15**(4):797–809.

Sebastian T., Klein P. & Kimia B. (2003). On aligning curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**(1):116–125.

Tournier J.D., Calamante F., Gadian D. & Connelly A. (2004). Direct estimation of the fiber orientation density function from diffusion-weighted MRI data using spherical deconvolution. *NeuroImage* **23**(3):1176–1185.

Tuch D., Reese T., Wiegell M., Makris N., Belliveau J. & Wedeen V. (2002). High angular resolution diffusion imaging reveals intravoxel white matter fiber heterogeneity. *Magnetic Resonance in Medicine* **48**(4):577–582.

Zhang S., Bastin M., Laidlaw D., Sinha S., Armitage P. & Deisboeck T. (2004). Visualization and analysis of white matter structural asymmetry in diffusion tensor MRI data. *Magnetic Resonance in Medicine* **51**(1):140–147.