6

Neighbourhood tractography

If a breakdown of communication between regions of the brain is indeed a significant factor in cognitive impairments, as the disconnection hypothesis posits, then dMRI and tractography are surely apt for studying this kind of pathology. But for the potential of the technique to be fully realised, its ability to provide proxy measures for white matter integrity such as FA will not be sufficient—it is also important that robust methods exist to compare such measures between normal and abnormal conditions, and to spatially localise any reproducible differences whenever possible.

In this chapter, we review the methods that have been applied to the localised study of white matter with dMRI, and describe a novel and automated approach to the issue. Our method treats segmentation as an *a posteriori* tract matching problem. We define a reference tract in a single brain volume, and then use a tract similarity measure to select from a number of candidate tract segmentations based on their topological resemblance to the reference. We demonstrate that this approach improves the consistency of segmentation results in a group of healthy young volunteers, thus reducing the impact of one source of within-group variance on anisotropy measurements.

6.1 Group comparison in white matter

Approaches to the identification and localisation of systematic differences between the white matter of two or more populations fall into two broad categories. One can either use a technique that is itself capable of highlighting local regions where differences are focussed; or hypothesise where such regions may be, *a priori*, and then study those target areas specifically. Tractography, when applied as a segmentation technique, provides a white matter-specific tool for implementing the latter approach. However, we begin here by examining those methods which work with the whole brain.

Voxel-Based Morphometry (VBM) is a whole brain technique which was originally conceived to find areas of structural difference between groups (Wright *et al.*, 1995), but has since been generalised to voxelwise comparison of many types of medical image data (Ashburner & Friston, 2000; Good *et al.*, 2001). Uptake of the technique in the clinical dMRI literature has been significant over the past few years, particularly in the study of schizophrenia, whose effect on the brain is thought to be diffuse (e.g. Ardekani *et al.*, 2003; Burns *et al.*, 2003; Park *et al.*, 2004). In these applications, the analysis is typically performed on FA maps, and the comparison is between patient and control groups.

A typical VBM pipeline involves spatially normalising the set of images by transforming them into a common space, filtering the normalised images with a Gaussian smoothing kernel, and then performing a statistical comparison of intensity at each voxel. The smoothing process confers several benefits for both sensitivity and specificity: it improves the signal to noise ratio in the data; it makes correction for multiple comparisons less onerous[a]; and it helps to avoid

[a]Smoothing with a Gaussian kernel allows one to use the theory of Gaussian random fields to perform a multiple comparisons correction which is less conservative than a Bonferroni-type correction. However, the extent to which

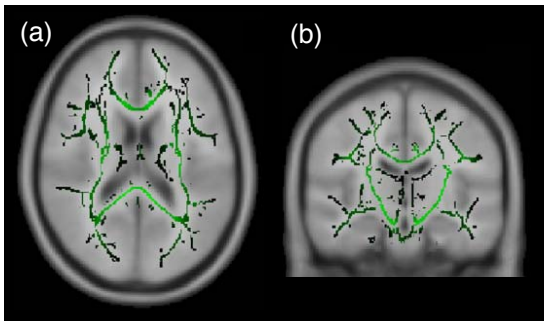

Figure 6.1: A typical FA skeleton created by the TBSS technique with data from 10 healthy subjects. The skeleton and underlying standard brain image are shown in axial **(a)** and coronal **(b)** planes. Reproduced from Smith *et al.* (2006).

false positives due to misregistration between images (Ashburner & Friston, 2001). However, the choice of filter size is a problematic issue. Ideally the full width at half maximum (FWHM) for the filter should be approximately equal to the size of features of interest, but the spatial extent of population differences in anisotropy is usually not known in advance; and no other principled method for choosing the filter size has been established for DMRI data. As a result, kernels with FWHMs of between 3 mm and 16 mm have been applied to FA maps, usually without any explicit justification; but it has been shown that the eventual conclusion drawn from a data set can depend directly on this choice (Jones *et al.*, 2005a). Moreover, even when smoothing is applied, spurious "results" due to registration errors persist to some extent. Unfortunately, these two drawbacks represent significant limitations to the method.

The difficulty of choosing a filter FWHM illustrates an important general point. If one is interested in contrasting one group of subjects against another, at whatever scale, one must compare like with like. Consistency is a crucial prerequisite. But for findings to be meaningful and reproducible, the techniques they employ must additionally be robust. A strong dependency of a result on the value of any methodological parameter is strongly undesirable, particularly if there is no principled way to choose that parameter.

Concern over the limitations of VBM, as well as its lack of specificity to diffusion data, motivated the recent development of another whole brain method called tract-based spatial statistics (TBSS; see Smith *et al.*, 2006), which is just beginning to be applied to clinical data (Anjari *et al.*, 2007; Kochunov *et al.*, 2007; Rouw & Scholte, 2007).

The key innovation of the TBSS method is its use of an anisotropy "skeleton"—a ridge of locally maximal FA running through the brain's white matter structures—to establish voxel homology for comparison (see Fig. 6.1). This approach allows one to perform voxelwise statistics without relying on the accuracy of image registration methods alone. By focussing on areas of the brain which can be confidently classified as white matter, the method additionally reduces the number of comparisons that need to be performed, thus improving the statistical power of any given data set. It is assumed that local variation in anisotropy *across* a tract is entirely attributable to partial volume effects.

Given a suitably preprocessed DMRI data set, the full TBSS pipeline consists of the following four steps.

1. Perform nonlinear registration of each individual FA map to the most typical one—that is, the one requiring the smallest average displacement. Resample each registered map to a standard resolution of $1\times1\times1$ mm.

2. Average the registered maps and generate an FA skeleton for this average map, by finding the voxel with maximal FA along lines perpendicular to local tract directions. Threshold this skeleton at an FA value of around 0.2 to 0.3.

3. Search each individual FA map for locally maximal values in the same way, and project each separate skeleton onto the average one, thus establishing voxel homology.

4. Perform voxelwise comparisons within the skeleton.

the assumptions of this approach are met in FA images may be limited (for details see Jones *et al.*, 2005a).

Note that there is no direct spatial smoothing involved in this process—although the average FA map used for step 2 will be implicitly smoothed to some degree by registration inaccuracies—so the problem of choosing a filter width does not occur.

The capacity of whole brain analysis techniques like VBM and TBSS to obviate the need for predefined brain areas of interest can be an invaluable one, particularly for exploratory studies where detailed *a priori* information is simply not available. However, just as the increased specificity of TBSS provides gains in statistical power over VBM, the advantage of making more detailed hypotheses is that subtler effects can be found more easily. Hence, when prior information regarding the likely location of interesting effects is at hand, methods for studying a particular fasciculus come into their own.

6.2 Tract-specific comparison

Perhaps the simplest method for searching for tract-specific differences between populations involves manually superimposing regions of interest (ROIs) with fixed dimensions onto an MRI image with high grey matter–white matter contrast. Indices of white matter integrity such as FA can then be averaged within these regions and compared between the subject groups of interest. This practice was employed in many of the first clinical comparative studies that used DTI (e.g. Ellis *et al.*, 1999; Jones *et al.*, 1999), and it is still far from obsolete.

Manual placement of ROIs has several advantages, many of which stem from its simplicity. Computational run time is trivial. There is no complex relationship between the "original" (i.e. native space) FA data and those values which are used for comparison, as there is in the common-space comparisons of VBM and TBSS. Depending on the level of specificity of the hypothesis, it may not be necessary to find congruent regions in each brain—placing an ROI in the correct structure with high confidence is often sufficient. Smoothing and multiple comparisons correction are typically minimal or unnecessary; and averaging within each region reduces the effects of noise. On the other hand, the method is not without its limitations. The choice of ROI size is arbitrary—too large and partial volume effects will be considerable; too small and noise will be a problem—although the relationship between this choice and the results is likely to be less complex than that of VBM's filter width. Study of any given fasciculus is limited to a very small and arbitrary portion of its length, and the region of the tract in which the ROI is placed will tend to depend on where it is most easy to fit it. Consequently, some tracts may be wholly excluded from this type of study because they are never wide enough to receive an ROI of the chosen size.

Probably the biggest drawback of manual ROI-based comparison, though, is its subjectivity. Although it requires almost no computational effort, the number of man-hours required to place the required regions in a typical data set is considerable; and it is hard to justify the particular placement choices made by any given observer. The related approach of segmenting an entire white matter structure by hand is not only time consuming; it is also extremely difficult to do well, since tracts are three-dimensional and can have highly irregular shapes.

Tractography provides an alternative. It is both objective and specific to white matter; and it lends itself directly to the segmentation of whole tracts, thus minimising noise issues whilst still focussing on a single fasciculus. The complex shape of white matter structures is not a problem. Of course, the caveats that apply to fibre tracking in general also apply to this application, but as problems such as crossing fibre degeneracy are handled better by new algorithms, so segmentation accuracy can be expected to improve.

The problem is then one of choosing seed points. Since the output of tractography algorithms is usually highly sensitive to the particular choice of starting location, care must be taken to ensure that a study is truly comparing like with like. Indeed, however sophisticated tractography algorithms become, the question of how best to initialise the tract reconstruction process is likely to remain an important one, especially in segmentation applications where consistency is important. Nevertheless, relatively little work has been done to date towards a principled and practical approach to seed point placement. Seeds are sometimes placed by hand, but average FA measured in tracts segmented in this way has been shown to vary quite widely between observers, and particularly between scans, even for a single subject (Ciccarelli

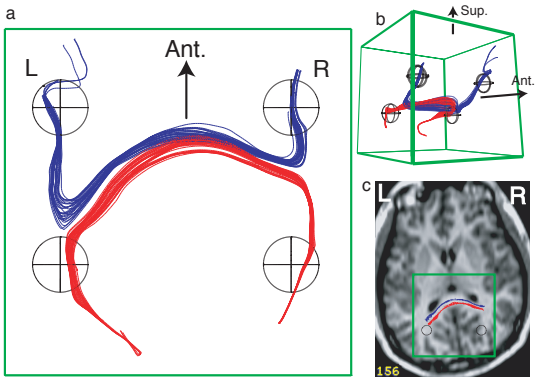

Figure 6.2: The two regions of interest constraint can separate disparate tracts which pass close to one another, in this case in the corpus callosum splenium. **(a,b)** Two and three-dimensional illustrations of the two tract trajectories, the former in axial projection. **(c)** The location of subfigure (a) in the brain. Reproduced from Conturo *et al.* (1999).

et al., 2003a). Some of this variation will be due to noise, but subjectivity remains a major confound if one is looking for group differences.

Seed points can be placed in some standard space and then transferred to each subject's native space using an image registration algorithm. The assumption is that if the transformation between spaces is accurate then the seed points are congruous, and so the same fasciculus will be segmented in each brain volume. There is a very real risk, however, that registration errors and anatomical variation between subjects will make this assumption unsafe. Performance can be improved by weighting the registration cost function so as to maximise the importance of aligning white matter regions well, but we will see that this approach, which we will refer to as the *registration method*, still has problems. On the other hand, it has the advantage of being semiautomatic, and so reducing the effect of observer bias—although the choice of registration algorithm and its parameters will of course affect the outcome.

Choosing a single seed point *a priori* is not the only way to use tractography for white matter segmentation. One could instead seed at a number of voxels, which raises the questions of which seeds in particular to use and what to do with the multiple tracts that result.

There are several answers to these questions available in the literature, and the spectrum of supported responses is still tending to enlarge rather than shrink. Probably the most well established method is to constrain the tractography algorithm so that all reconstructed pathways must pass through two or more "waypoint" ROIs (Conturo *et al.*, 1999). The use of this constraint is illustrated in Fig. 6.2. Seeding near the middle of the corpus callosum splenium, in this case, may produce a streamline that projects posteriorly into occipital cortex (coloured red) or one that projects anteriorly (coloured blue). Since the pathways run very close to one another, the reconstruction will be highly sensitive to the exact location of the seed point. However, if one seeds in a number of locations in the splenium and then retains only those streamlines which pass through the two posterior ROIs, it is assured that the anterior projection will be ignored. Although these ROIs are often placed by hand or using registration-based transformation, they can often be far wider than the tract of interest—unlike regions used directly for segmentation—and their exact placement may therefore not be crucial. This kind of method is applicable to any streamline-based tractography algorithm, deterministic or probabilistic; and has been applied to group tractography (Abe *et al.*, 2004). There are some limitations, however, which we will discuss later in the chapter.

Another alternative is to seed throughout the brain. The advantages and disadvantages of this are obvious: on the one hand, choosing seed points is no longer an issue; on the other, a large proportion of the results are irrelevant to the study of any given fasciculus. A two ROI constraint can be applied, or one can use clustering techniques to divide up a brainful of streamlines into related bundles. Various distance metrics have been proposed for streamline clustering (Brun *et al.*, 2004; Corouge *et al.*, 2006; Maddah *et al.*, 2005; O'Donnell & Westin, 2005), but one has still to identify the bundle or bundles of interest, and the general approach is not directly applicable to probabilistic tract representations.

The remainder of this chapter is dedicated to describing a novel perspective on the seeding problem, and applying it to some real-world data. We then compare the new method, which we refer to as *neighbourhood tractography* (NT), with region of interest-based alternatives.

6.3 Similarity and matching

Rather than modify tractography output to suit a particular criterion, the aim of the following work is to improve the consistency of tractography-based segmentation in group data by refining the initialisation of the algorithm; i.e. the seed point. In order to eliminate observer subjectivity, the method is to be automated. Our approach is to choose, from a group of "candidate" seed points, that point which produces the best output. In order to quantitatively define what constitutes "good" or "correct" output, we develop a novel tract similarity measure, based on the shape and length of two tracts being compared (first described in Clayden *et al.*, 2006a). 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 define a series of reference tracts, and 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 (cf. Clayden *et al.*, 2006b, from which Figs 6.3–6.8 are taken).

Since there is a diverse array of tractography algorithms available, and studies may wish to use different algorithms depending on the nature of the problem or hypothesis that they are working on, it is desirable that the process of tract matching be as independent as possible of the choice of algorithm. However, different algorithms produce different tract representations, as we saw in chapter 5, which creates a problem when we want to compare them using a single method. The solution is to use a common representation for all tracts, for the purposes of matching only. The "field of connection likelihoods" representation of a tract that is natural for probabilistic algorithms such as BEDPOST/ProbTrack can easily be generated by spatial discretisation of a deterministic streamline, so for the purposes of our tract similarity measure, we will assume that the tractography algorithm takes as input a single seed point, and produces voxelised, quantitative output. Hence we can define a tract r as the ordered pair

$$r = (\mathbf{a}_r, \phi_r(\mathbf{x})) \,, \tag{6.1}$$

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.

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. (6.1), 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 tacitly assumes that the seed points are equivalently located in the two tracts.

1. Initialise 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 neighbouring voxel not in V_r. Note the step vector, $\mathbf{v}_r$, required to move to this new location.

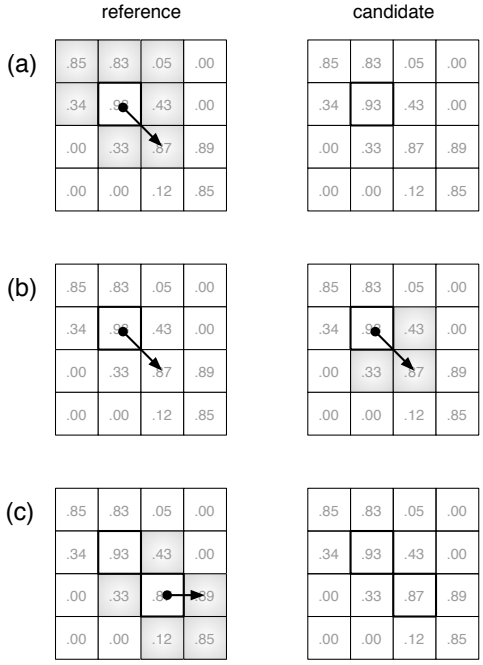

Figure 6.3: Two-dimensional illustration of the shape similarity algorithm, as applied to two identical tracts. In **(a)**, the boxes with bold borders represent the starting point, which has been marked visited. The shaded voxels in the reference tract indicate those nonzero, unvisited locations that 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.

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 neighbour 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 normalised 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 normalised inner product calculated in step 6 is given by

$$\frac{\mathbf{v}_r \cdot \mathbf{v}_c}{\|\mathbf{v}_r\| \|\mathbf{v}_c\|}, \tag{6.2}$$

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 between rotationally symmetric tracts such as the corpus callosum genu and splenium, or bilateral pairs.

6.3.1 The reduced tract

Tract data of the form given by Eq. (6.1) are not constrained to be a single voxel wide, and in general they will not be. Moreover, since the algorithm ceases stepping through the data when either tract terminates, 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. 6.3. 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

$$\mathbf{a}_{\tilde{r}} = \mathbf{a}_r \qquad \phi_{\tilde{r}}(\mathbf{x}) = \begin{cases} \phi_r(\mathbf{x}) & \text{if } \mathbf{x} \in V_r \\ 0 & \text{otherwise,} \end{cases} \tag{6.3}$$

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

$$\sigma(r,r) = \sigma(\tilde{r},\tilde{r}) = \sigma(r,\tilde{r}) = \sigma(\tilde{r},r)\,, \tag{6.4}$$

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, $\mathbf{a}_r$, 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 score contribution of one, we can establish the bounds

$$0 \le \sigma(\tilde{r},c) \le \sigma(r,r) \quad \forall c\,. \tag{6.5}$$

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. (6.5) at 0. Equivalently, $0 \le \sigma(r,\tilde{c}) \le \sigma(c,c)$ for any r.

6.3.2 A similarity measure

Using the tract comparison algorithm described above, we now 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

$$L_r \equiv \sigma(r,r)\,. \tag{6.6}$$

This length value is unchanged in the reduced tract, $\tilde{r}$, as shown by Eq. (6.4). 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. (6.6), and having calculated its value for the reference and candidate tracts, we establish the similarity of these two numbers using the symmetric normalised difference given by

$$S_1(r,c) = 1 - \left|\frac{L_r - L_c}{L_r + L_c}\right| = \frac{2\cdot\min\{L_r,L_c\}}{L_r + L_c} = S_1(c,r)\,. \tag{6.7}$$

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

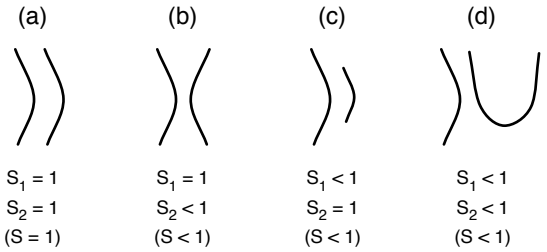

Figure 6.4: Qualitative demonstration of the effect on the two score components, S_1 (length) and S_2 (shape), of different types of relationship between the reference tract (fixed, and on the left in each case) and the candidate tract (variable, and on the right). The seed points are assumed to be in the centre of each tract throughout.

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

$$S_2(r,c) = \frac{\sigma(\tilde{r},\tilde{c})}{\min\{L_r, L_c\}} \neq S_2(c,r)\ . \tag{6.8}$$

The denominator in Eq. (6.8) removes the length dependence of the σ function. The bounds on the σ function that were established above ensure that the value of Eq. (6.8) is always in the interval $[0,1]$.

Finally, the two score components given by Eqs (6.7) and (6.8) are combined to form the overall similarity score,

$$S(r,c) \equiv \sqrt{S_1(r,c)\cdot S_2(r,c)} = \sqrt{\frac{2\cdot\sigma(\tilde{r},\tilde{c})}{L_r + L_c}}\ , \tag{6.9}$$

the geometric mean of the two components. A higher value of Eq. (6.9) 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 emphasises that *both* length and curvature must be similar for the candidate tract to be considered a likely equivalent to the reference.

Fig. 6.4 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 idealised, and continuous rather than voxelised, tract curves; but they illustrate how the two score components will be affected in various scenarios. In (a), the candidate tract is identical to the reference tract. This is equivalent to the case in Fig. 6.3. In (b), the candidate is a reflected copy of the reference. Note that the shapes of these two 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.

6.4 Validation and application

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, consisting of 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.

The data were initially preprocessed to remove skull data and eddy current induced distortion effects from the images, using FSL tools. The underlying tractography algorithm used in this study was the BEDPOST/ProbTrack algorithm (Behrens *et al.*, 2003b). It should be remembered that the BEDPOST model of the DMRI signal is a partial volume model assuming a single anisotropic diffusion direction at each voxel, and 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 may be 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"[b]) 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 from a single scan. For each other 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 voxel AVF was less than 0.2, an empirically chosen threshold used to avoid seeding in cerebrospinal fluid or grey 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 refer to this technique as neighbourhood 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.

Fig. 6.5 shows the results of the first experiment as a box-and-whisker plot. The mean (± one standard deviation) similarity score for each group of tract comparisons was 0.14 (±0.13) for nonbilateral, 0.31 (±0.13) for bilateral, 0.38 (±0.12) for intersubject, 0.47 (±0.09) for interscan, and 0.46 (±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 6.6 and 6.7. The correspondence between the letters labeling each subfigure and the different scans is shown in Table 6.1. Fig. 6.6 shows the tract fields produced by seeding ProbTrack at the original seed point in splenium of corpus callosum, and thresholding the results at the 1% level. This seed region

[b]The acronym NEX, for "number of excitations", is commonly used to denote the number of times an imaging sequence was applied to the subject.

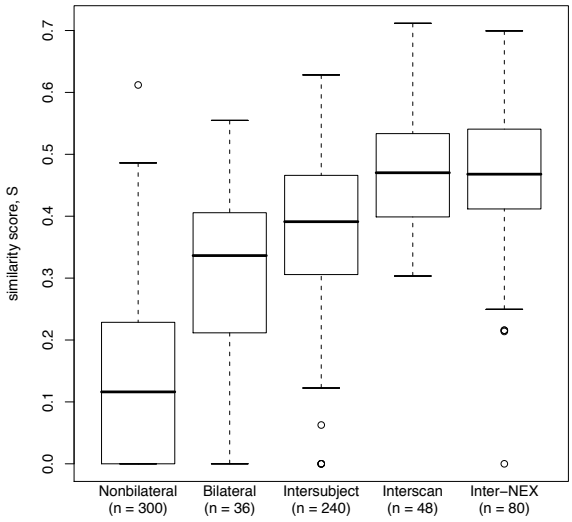

Figure 6.5: Box-and-whisker plot showing the range of similarity scores for the five different categories of comparison in the first experiment. The thick horizontal line across each box represents the median, the box shows the interquartile range, the whiskers show the extent of the bulk of the data, 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.

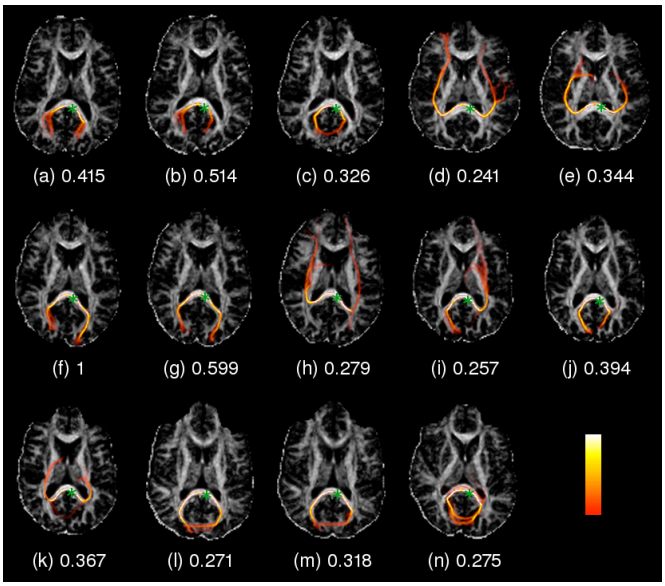

Figure 6.6: Two-dimensional axial projections of the tracts generated by the ProbTrack algorithm using the original seed points chosen by the registration method, overlaid on AVF maps of the slice in plane with the seed in each case. White indicates high AVF and black low. In the tracts, yellow indicates 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.

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 6.1: Correspondence between the different scans and the subfigure labels used in Figs 6.6 and 6.7.

Seed	MNI point	RM score mean	RM score s.d.	NT score mean	NT score s.d.
CC genu	(8,22,14)	0.488	0.056	0.597	0.018
CC splen.	(−6,40,14)	0.354	0.106	0.542	0.031
right ALIC	(18,10,6)	0.529	0.098	0.651	0.026
left ALIC	(−16,10,6)	0.463	0.099	0.644	0.027
right SS	(36,−54,10)	0.329	0.220	0.680	0.023
left SS	(−36,−54,10)	0.365	0.077	0.516	0.024
right PLIC	(22,−14,10)	0.405	0.096	0.570	0.030
left PLIC	(−22,−14,10)	0.444	0.054	0.594	0.025

Table 6.2: Mean and standard deviation of similarity scores for all tracts chosen by neighbourhood 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. The position of the seed point in MNI standard space is given in millimetres.

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 (cf. Fig. 6.2). Fig. 6.7 shows the tracts chosen by the neighbourhood 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. 6.7, similarity scores are necessarily greater than or equal to the corresponding score in Fig. 6.6, 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. 6.8 shows examples of "reduced" reference and candidate tracts, in the sense described in §6.3.1. It shows how the reduction affects the tracts. In this case, the reference tract is simply slightly narrower than its unreduced equivalent, Fig. 6.6(f). The candidate tract is truncated at the edge of the splenium, where the unreduced version, Fig. 6.6(h), had an ambiguous branch.

The mean and standard deviation of the similarity scores for the tracts chosen before and after applying neighbourhood tractography for each seed region, across all subjects and acquisitions, are given in Table 6.2. The figures for the "best" tracts—as chosen by neighbourhood tractography—represent narrow and seed-specific score distributions, whose coefficients of variation (CVs, the standard deviations divided by the means) 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 CVs are in the range 11.5–66.9%.

In the second experiment, 63% of seed points chosen by neighbourhood tractography were not more than 2 voxels from the original seed point in any direction. This proportion is high enough to suggest that a $7\times7\times7$ search neighbourhood is generally sufficiently large. Run time for the method might be expected to increase with the cube of the neighbourhood width, although in practice it may be that more of the extra seed points are rejected by the anisotropy threshold, and so the scale factor may be a little less.

6.5 How many seeds?

To date, it has not been explicitly shown that tractography-based segmentation using a single seed point cannot yield consistent results, and we see no reason that this should be the case, particularly as the sophistication of tractography algorithms continues to improve. However it is certainly true, as we saw in §6.2, that choosing a single point so as to obtain useful results is hard. We have described here a method which emulates a process for selecting an appropriate seed that might be used by a human observer: the expected topology of the tract is clearly defined—in this case in terms of a reference tract—and then we try seeding at several plausible locations until a good match is found. Unlike the human observer, though, the algorithm is completely consistent in its assessment of candidate tracts and has no difficulties working in three dimensions. The selection process is also far faster using the algorithm: comparing each candidate tract with the reference tract takes only a second or so on a typical workstation.

The relationship between the neighbourhood tractography method, as we have presented it here, and region of interest-based methods is simple to explain. Let us assume that a plausible but suboptimal seed point has been selected in the native space of the subject using the registration method. This seed point could then be used directly for tractography, or one could grow an ROI around it and seed at every point therein, possibly subject to an anisotropy threshold. Neighbourhood tractography then performs all-but-one rejection of these seed points based on *a posteriori* tract similarity, whereas a multiple ROI method would combine results from the whole seed region, subject to the waypoint constraint on individual streamlines. Another option is to simply retain all the results with no constraints: this is a single ROI method. Selection between these strategies is therefore partly a question of deciding how many seed points should be used to generate the final tract representation.

It should be noted that the two ROI constraint, as demonstrated by Fig. 6.2, is effectively a modification of the tractography algorithm that it is applied to. A deterministic streamlining algorithm, thus modified, will return either a streamline passing through the seed point and the waypoint regions, or nothing. In the probabilistic case, the effect is to add an extra conditional dependency on the ROI locations to the connection likelihoods, making it somewhat more difficult to interpret the results.

Fig. 6.9 shows the results of applying the different strategies in a single scan, corresponding to Fig. 6.6(c), using 5000 streamline samples per seed point in each case. If the reference tract in Fig. 6.6(f) is taken to represent the pathway we are attempting to find in this scan, then Fig. 6.9(a) surely shows poor correspondence. Its projection into cortex is further anterior than it ought to be, and the reconstructed tract appears to cross the interhemispheric plane posterior to the splenium, which is definitely nonphysical. By constrast, tract (b) shows neither of these problems. It can be seen in the unthresholded version of this tract, (f), that a very small number of probabilistic streamlines (less than 1%) do project anteriorly from the splenium, but the main trajectory of the tract is consistent with the reference.

Full seeding in the $7 \times 7 \times 7$ voxel neighbourhood yields tract (g), which is very widely spread out and heavily affected by thresholding—compare tract (c). These two effects are related. A very large number of probabilistic streamlines are generated in this case (570,000 in total), but since they are very widely spread out no more than 80,710 pass through any single voxel in the brain. As a result, only 5.2% of nonzero voxels from (g) survive the threshold, compared to 21.5% from (f). Worse, the threshold scales with the volume of the neighbourhood, but the visit counts at each voxel do not keep pace (see Fig. 6.10), making the application of this kind of threshold undesirably sensitive to neighbourhood size. This effect is even more pronounced if an anisotropy threshold is not used to cull unpromising seeds from the seeding region. It seems, then, that tract (c) should be treated with caution, while (g) is too nonspecific to be of much use. Nevertheless, by carefully placing a smaller ROI well within the tract, a single seeding region can be practical (Kanaan *et al.*, 2006).

Finally, tracts (d) and (h) show the result of using two constraint ROIs in addition to the seeding ROI. In order to yield a tract as similar as possible to (b), we used information from the latter to place the constraint regions. The centres of the waypoint ROIs in the left and right hemispheres were placed at the locations with the greatest voxel value, in (b), within a plane normal to the anterior–posterior axis that is shown in green. The size of the constraint

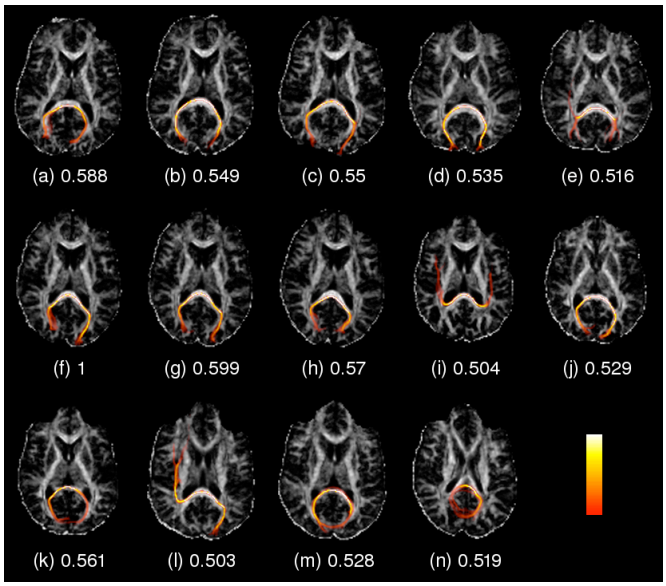

Figure 6.7: Projections of the tracts chosen as the "best" (highest similarity to the reference tract), using a 7×7×7 seeding neighbourhood around the original seed point. Individual similarity scores are also shown. Tract **(f)** is the reference tract.

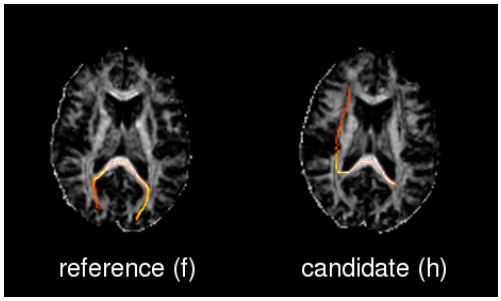

Figure 6.8: Examples of reduced reference and candidate tracts, produced from two of the unreduced tracts shown in Fig. 6.6.

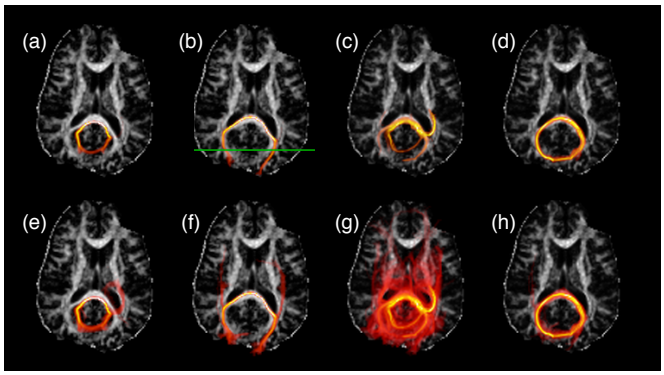

Figure 6.9: Illustrative results using various seeding strategies for the corpus callosum splenium in a single subject, ignoring potential seed points with AVF of less than 0.2 throughout. **(a,e)** Single seed point placed using the registration method. **(b,f)** Single seed point chosen using neighbourhood tractography. **(c,g)** Full seeding throughout the neighbourhood, with additive combination of results. **(d,h)** Full seeding, but with additional bilateral waypoint ROIs placed posterior to the splenium. **(a–d)** are thresholded at the 1% level; **(e–h)** are unthresholded. The colour scale is not consistent in meaning between subfigures.

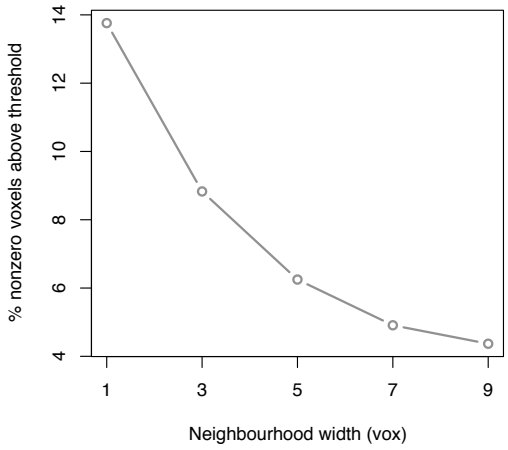

Figure 6.10: The effect of thresholding on connection likelihood when using a single ROI strategy. The impact of the threshold, expressed as a fixed proportion of the total number of streamlines initiated in each case, is more significant as the seeding ROI increases in size.

ROIs was $7\times7\times7$ voxels, as with the seeding region. The effect of thresholding in this case is far more modest than without the constraints, but again the segmentation appears to be passing between hemispheres twice, suggesting that the use of two waypoint regions may not be sufficient to ensure a plausible segmentation. One can attempt to rectify this by adding more constraints—in a recent reproducibility study, up to five constraint ROIs of four different types were used to segment each of the fasciculi of interest (Heiervang *et al.*, 2006)—but all such restrictions are simultaneously absolute and independent of the actual data, and so risk undermining the very advantages of using tractography for segmentation. The logical limit of the process of adding constraints, after all, is the case in which tracts are simply outlined by hand.

It is interesting to note that even though the neighbourhood tractography result was used to inform the placement of waypoint ROIs, the result is very different in shape. The pathway segmented in (b) is present in (h), of course, but the projection is so "unlikely", given the seed mask and waypoint constraints, that its end points in occipital cortex are absent from the thresholded version. Instead, the tract projection in (d) represents the most visited pathway from (g) that passes through the waypoints. It is very difficult to decide which of these results is more anatomically correct. Validation of tractography output is a complex issue in its own right, as we discussed in §5.5. However, we would certainly claim that the tracts shown in Fig. 6.7 are more similar to the reference tract—which is chosen for illustration—than those in Fig. 6.6. If the nature of the reference tract were later to be found to be inappropriate, it could be updated and neighbourhood tractography repeated without change. The method would then find the best match to the new reference tract.

Although we have focussed on probabilistic tractography in this comparison, our observations are just as valid in the deterministic—or maximum likelihood—case, where only one streamline is generated per seed point. However, a connection likelihood threshold would not then be relevant.

Multiple seed points for tractography need not necessarily be adjacent to one another in a neighbourhood. We have described, in separate work, an iterative approach in which each seed point is chosen from the tractography output for the previous seed such that it is as far as possible from its predecessor whilst still having high likelihood of connection to it (Clayden *et al.*, 2005). This can help the segmentation to recover, to some extent, from a poor starting seed point; but the choice of each seed point is not very strongly principled in that case, and the method is likely to be rather too scattershot for general purpose use.

Ultimately, the question of which seeding strategy generates the most useful segmentation of the fasciculus of interest will depend on the application. The all-but-one rejection approach of neighbourhood tractography offers the greatest specificity whilst providing a strong *preference*—as opposed to an enforced constraint—for tracts with topologies similar to the reference

tract. ROI methods, meanwhile, offer to segment the full width of the fasiculus if they are given a large enough seeding region. The latter may be desirable, or it may be problematic due to the larger expected role of partial volume effects. Of course, these two techniques are not mutually exclusive, and it may be that in some cases a combination is the most successful.

6.6 Evaluation of the similarity measure

While tract shape has been studied before (Batchelor *et al.*, 2006; Corouge *et al.*, 2004; Ding *et al.*, 2003), previous work has been aimed at modelling individual tracts, rather than doing pairwise similarity scoring. The kinds of tract characteristics that these previous studies have worked with, such as curvature and torsion, could in principle be applied to the tract matching problem; but as far as we are aware, the work described above represents the first actual attempt at using a quantitative tract similarity measure to improve segmentation consistency.

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. Behaviour 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 also 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.

The narrowness of the score distributions for each seed point—as shown in Table 6.2—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.

A major advantage of the NT approach 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. However, we have made no alterations or corrections for factors such as natural variation in brain size and shape, or head rotation; even for the purposes of tract comparison. In fact, a correction based on a transformation of the candidate tract into the space of the reference tract would have problems of its own, since interpolating the tract data could alter its structure in undesirable ways. For example, local duplication of voxel values—which would arise from a nearest neighbour interpolation scheme—would be strongly suboptimal for our similarity algorithm. The difficulty with registration also makes simple field-based distance measures such as the sum of squared differences between voxels highly problematic for comparing tracts.

Since differences in head rotation and head size between scans will have a complex, non-linear effect on the similarity measure, and may affect different tracts differently, it is not straightforward to establish the impact of these variates, nor 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.

The similarity measure described above aims to be relatively simple whilst capturing important characteristics of the two tracts that we wish to compare. This simplicity aids portability. Whilst probabilistic tractography algorithms tend to produce tract data of the form given by Eq. (6.1), 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.

There is an obvious limitation of comparing shape at the voxel scale, which is that voxel sizes vary between data sets and have no intrinsic physiological significance at all. Moreover, since voxels are often not equal in width in all dimensions, a step of "one voxel" may represent a different real-world distance depending on the orientation of the step.

The main weakness of the similarity measure presented here is that the termination criterion in step 8 of the algorithm (see page 66) 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. 6.7 than in Fig. 6.6. 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. Indeed it is plausible, even likely, that in some cases better matching tracts than those selected by this similarity measure were available but were underscored and therefore disregarded. This substantial issue could perhaps be alleviated by biasing the algorithm in favour 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. A different approach may be preferable.

6.7 The next step

This chapter began with a look at how segmentation and comparative analysis in white matter can be approached using either a tract-specific or a whole brain starting point, the choice between which should depend on the specificity of one's hypothesis. We saw that segmentation using tractography is typically subjected to constraints based on a number of "waypoint" regions of interest, which are described in advance. This multiple ROI approach is one way to incorporate prior knowledge into the fibre tracking process, but it represents a hard constraint which complicates the interpretation of the resulting tract. The novelty of the neighbourhood tractography method, as an alternative approach, is that it allows prior information to be introduced in the form of a reference tract, but rather than constraining tractography directly, the reference is used to select from a number of candidate segmentations generated with different seed point initialisations.

We have demonstrated here that NT can work as intended, although we have also identified some shortcomings in the particular similarity measure that we used. As well as considering how these issues might be overcome, it is important to investigate how successful the method is when applied to data of clinical interest, where there may be more confounding factors than are present in scans from healthy young volunteers. The latter will be the focus of the next chapter.