Tractometer: Towards validation of tractography pipelines

Marc-Alexandre Côté, Gabriel Girard, Arnaud Boré, Eleftherios Garyfallidis, Jean-Christophe Houde, Maxime Descoteaux*

Sherbrooke Connectivity Imaging Laboratory (SCIL), Computer Science Department, Université de Sherbrooke, Sherbrooke, Québec, Canada

A R T I C L E   I N F O

Article history:
Available online 25 April 2013

Keywords:
Diffusion MRI
Tractography
Validation
Connectivity analysis

A B S T R A C T

We have developed the Tractometer: an online evaluation and validation system for tractography processing pipelines. One can now evaluate the results of more than 57,000 fiber tracking outputs using different acquisition settings (b-value, averaging), different local estimation techniques (tensor, q-ball, spherical deconvolution) and different tracking parameters (masking, seeding, maximum curvature, step size). At this stage, the system is solely based on a revised FiberCup analysis, but we hope that the community will get involved and provide us with new phantoms, new algorithms, third party libraries and new geometrical metrics, to name a few. We believe that the new connectivity analysis and tractography characteristics proposed can highlight limits of the algorithms and contribute in solving open questions in fiber tracking: from raw data to connectivity analysis. Overall, we show that (i) averaging improves quality of tractography, (ii) sharp angular ODF profiles helps tractography, (iii) seeding and multi-seeding has a large impact on tractography outputs and must be used with care, and (iv) deterministic tractography produces less invalid tracts which leads to better connectivity results than probabilistic tractography.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Diffusion MRI and fiber tractography have gained importance in the medical imaging community for the last decade. The neuroscience community often uses fiber tractography as a black box, and its limits are ignored by most. The diffusion community has done a good job in the last years to highlight the limitations of diffusion tensor imaging (DTI) (Descoteaux et al., in press) in crossings and high curvature areas. Thus, numerous new high angular resolution diffusion imaging (HARDI) tractography techniques have been proposed (Descoteaux, 2008; Seunarine and Alexander, 2009; Descoteaux et al., in press). Most of these HARDI techniques have been tested quantitatively on the sharpness of the peaks (angular resolution) and accuracy of the orientation distribution function (ODF) reconstruction based on simulated data. In neurosurgery, quantitative evaluation of fiber tracking results has been done by comparing tracts to electrocortical stimulation points (Kinoshita et al., 2005), or by measuring patient outcomes given the use of tractography for the resection of tumors (Wu et al., 2007; Spena et al., 2010). However, the comparison between tractography algorithms remains mainly qualitative for the most part. Most often, fiber tracking results are shown on major fiber bundles such as the corticospinal tract (CST), the corpus callosum (CC), superior longitudinal fasciculus (SLF), cingulum (Cg) and inter-hemispheric projections able to connect the left and right hemispheres. Several groups have studied the effect of interpolation (Tournier et al., 2012; Dyrby et al., 2011), step size (Tournier et al., 2011; Tournier et al., 2012; Sotiropoulos, 2010), stopping criteria (Girard and Descoteaux, 2012; Girard et al., 2012; Smith et al., 2012a; Smith et al., 2012b; Guevara et al., 2011; Garyfallidis, 2012), but mostly qualitatively on major fiber bundles.

Validation of fiber tractography remains an open question and a challenge on real data. Good reviews can be found in (Hubbard and Parker, 2009) and in (Fillard et al., 2011), where synthetic phantom data (Perrin et al., 2005; Fieremans et al., 2008; Pullens et al., 2010; Bach et al., 2011; Moussavi-Biugui et al., 2011), histological data (Leergaard et al., 2010; Anderson et al., 2006; Dauguet et al., 2007; Budde and Frank, 2012), biological ex vivo phantom data...
Campbell et al., 2006), and numerical simulation data (Hall and Alexander, 2009) are detailed.

Recently, the FiberCup phantom dataset (Fillard et al., 2011) was proposed to study the effect of the huge variety of diffusion models, tractography and combinations thereof. This phantom is publicly available (www.lnao.fr/spip.php?article112) and is now used by the community for quantitative evaluation of tracking algorithms (Pontabry and Rousseau, 2011; Reisert et al., 2011; Wilkins et al., 2012b; Wilkins et al., 2012a; Röttger et al., 2011; MomayyezSiahkal and Siddiqi, 2010). However, the original FiberCup contest does not well reflect brain connectivity analysis, especially in terms of seeding and performance evaluation. In our opinion, two important drawbacks of the FiberCup are the placement of the seed points given and the quantitative metric used to compare with ground truth. Only 16 seeds are given and these are close to boundaries and in the middle of structures. These 16 seeds result in 16 individual streamlines that are compared with the ground truth in terms of spatial, tangent to the tract and curvature distances. These measures are local and do not capture well the global connectivity profile of the tractography algorithm. Other problems are that each participant of (Fillard et al., 2011) performed his own analysis and that the implementations used are not available to the community.

In this paper, we propose a revised FiberCup analysis that is closer in spirit to brain connectivity analysis. In brain connectivity, the importance is connectivity. Does region A connect to B as expected? Does region A connect to unexpected regions of the brain? Therefore, instead of using local seeds and local point-by-point distances for evaluation, we propose a global view of the dataset and the fiber tractography streamline output. We developed an evaluation method to compare tractography pipeline streamline outputs and evaluate the number of found and not found fiber bundles, the proportion of streamlines part of existent and non-existent bundles, and the proportion of incomplete streamlines. Since these characteristics have a direct impact on connectivity analysis, having a tractography evaluation tool is crucial for human connectome studies.

This paper is thus aimed at providing a framework to encourage the community to rigorously choose a tractography processing pipeline and report the known limitations of their technique. Therefore, in the rest of the paper, we describe our new online system to evaluate and rank pipelines (url: tractometer.org). At this stage, a user has the choice of providing 3 things to the system: (1) A diffusion dataset corrected with the user’s best algorithm, (2) a field of ODFs coming from the user’s best algorithm, or (3) a set of streamlines. The user can then obtain a ranking against the current database of state-of-the-art techniques. Presently, in this database, we have \( N = 57,096 \) different streamline outputs of tractography pipelines coming from our in-house tools, MRtrix (Tournier et al., 2012) and TrackVis (Diffusion Toolkit) (trackvis.org) using the different FiberCup acquisitions, local estimation techniques, and tracking parameters. A preliminary version of this paper appears in the MICCAI 2012 proceedings (Côté et al., 2012). Relative to that contribution, the current work adds 56,000 tractography pipelines to the testing framework, including new datasets of different b-values and new tractography algorithms. We also added new quantitative connectivity measures.

2. Materials and methods

2.1. Terminology

The diffusion MRI community needs to agree on definitions for fiber tracking. Unfortunately, there are many examples of confusing...
and contradictory definitions used in the literature and used by clinicians, researchers and laboratories. Here, we make several suggestions and define the terms used in this paper. We first separate our terms in three sections: (i) general terms, (ii) terms that concern anatomy and (iii) terms that concern imaging.

General. Fiber: a long and thin structure. Hence, fiber tracking is a general term that can be used in any field that reconstructs fibrous structures, such as hair fibers, celery fibers, muscle fibers, prostate fibers, brain fibers, etc.

Brain Anatomy. Axon: the long fiber-like part of a nerve cell along which impulses are conducted from the cell (μm scale). Tract: a group of neuronal axons in the central nervous system (mm scale). Fiber bundle: a group of fibers usually with an anatomical or functional meaning, e.g. arcuate fasciculus, fornix, etc. A fiber bundle for the area of brain anatomy is synonymous to a tract, also often called fiber tract. The term tract can be misleading when talking for example about the corticospinal tract. The corticospinal tract is not a single tract but a group of tracts.

Brain Imaging. Streamline: an imaginary line approximating the underlying fiber. Streamlines bundle: a group of streamlines with similar shape and spatial characteristics. These do not necessarily correspond to individual fiber bundles. Tractography: the process of generating streamlines of brain fibers. This is synonymous with fiber tracking in the central nervous system.

2.2. A revised FiberCup analysis

Fig. 1 illustrates the FiberCup dataset mimicking a coronal slice of the human brain. The phantom was built following the procedure of (Poupon et al., 2008; Poupon et al., 2010) for the MICCAI FiberCup workshop held in 2009, which resulted in a tractography comparison paper in (Fillard et al., 2011). Note that streamlines are colored using the rgb (red–green–blue) convention throughout the paper, i.e. red streamlines for left–right and green for top-down streamlines. Since the FiberCup is a 2D axial dataset, there are no blue streamlines coming out of the page. This also means that the connections mimicking the CST are actually green in the FiberCup dataset as opposed to the normal blue color in a classical brain rgb map.

In our opinion, the metrics proposed in (Fillard et al., 2011) are too local and vulnerable to the seeds given and, as a result, do not capture the global connectivity behavior of the fiber tracking algorithm. To better reflect brain connectivity studies, especially in terms of seeding and performance evaluation, we revisit the FiberCup analysis. The main difference is to consider two different starting configurations: (1) From a complete mask of the phantom mimicking a brain white matter (WM) mask, as seen in Fig. 7, or (2) from ROIs mimicking gray/white matter interfaces, as seen in Fig. 7. Hence, the output streamlines from a tractography algorithm, such as those shown in Fig. 2 (left), can be filtered by the ROIs at the end of bundles (Fig. 7, middle-left) to quantify the global success and errors present in tractography output. Hence, we propose new connectivity metrics highlighting the global errors and success rate of tractography pipelines.

2.2.1. Definitions and rationale

We performed a survey with neurosurgeons and neurologists at our institute concerning true and false connections, and false positives of streamlines. We concluded that these terms were not the best choices for connectivity analysis purposes. Therefore, we define the following six new terms:

- **Average Bundle Coverage (ABC)**: the proportion of a fiber bundle covered by streamlines. ABC is reported in % (percentage) and is the average of the number of voxels crossed by streamlines divided by the total number of voxels in the bundle. For example, in Fig. 2, bundles 4, 6 and 7 have a high ABC, whereas bundle 3 has a much lower ABC.
- **Valid Connections (VC)**: streamlines connecting expected ROIs and not exiting the expected fiber bundle mask. This is illustrated by streamlines in Fig. 2 (right). VC will be reported in % of valid connections.
- **Invalid Connections (IC)**: streamlines connecting unexpected ROIs or streamlines connecting expected ROIs but exiting the expected fiber bundle mask. These streamlines are spatially coherent, have managed to connect ROIs, but do not agree with the ground truth (see Fig. 3). IC are reported in %. According to our survey with clinicians, these are the most problematic streamlines as they “appear plausible” (because for example they connect gray matter regions) but are in fact non-existent from a priori anatomical knowledge.
- **No Connections (NC)**: streamlines that do not connect two ROIs. Depending on how fiber tracking handles stopping criteria, these streamlines either stop prematurely due to angular constraints or, most often, due to hitting the boundaries of the tracking mask, as illustrated in Fig. 4.
- **Valid Bundles (VB)**: bundles connecting expected ROIs. Fig. 1 (left) shows the valid bundles. VB is reported in bundle counts, from 0 to 7 for the FiberCup. For example, in Fig. 2, VB = 7.
- **Invalid Bundles (IB)**: bundle connecting unexpected ROIs. As VB, IB is reported in bundle counts and is similar to IC, but at the bundle scale. For example, this is shown in Fig. 3, where bundles 1, 3 are mismatched. However, note that they “look anatomically plausible”, had we not known the ground truth. In theory, there are a total of 39 possible IB.

2.3. Overview of tractography pipelines currently part of the Tractometer

**Acquisition.** There are three acquisitions with different b-values available in the current version of the Tractometer, corresponding to the b = 650 s/mm², b = 1500 s/mm² and b = 2000 s/mm² datasets shown in Fig. 5. For each acquisition, there were 64 uniformly distributed diffusion-weighted measurements and 1 b = 0 image, with two repetitions. Hence, for the rest of the paper, the first repetition is called Acquisition 1 (acq1), the second Acquisition 2 (acq2) and the average of the two Average (ave). The spatial resolution for all three datasets is 3 mm × 3 mm × 3 mm and 3 slices were acquired. Specific parameters are as follows: field of view 19.2 cm, matrix size 64 × 64, read bandwidth 1775 Hz/pixel, partial Fourier factor of 6/8, GRAPPA factor of 2, TR = 5 s, TE = 77/94/102 ms respectively for each b-value, and SNR of DWI was estimated to be 9.1/2.6/1.1 respectively (Fillard et al., 2011). The b = 0 image has SNR of approximately 15.8.

**Local reconstruction techniques.** There are several diffusion tensor estimations included in the Tractometer. Diffusion tensors were estimated using our in-house log-Euclidean implementation (Arsigny et al., 2006), the TrackVis implementation and the MRtrix implementation. Moreover, most existing local reconstruction techniques based on the spherical harmonics (SH) representation are currently implemented in the Tractometer. It is beyond the scope of this paper to present the mathematical details of all of these techniques. The reader is referred to the following reviews (Descoteaux, 2008; Seunarine and Alexander, 2009).

SH-based techniques included in the Tractometer are: the diffusion orientation distribution function (ODF) from the numerical Funk–Radon transform (Tuch, 2004) implementation of TrackVis (num-ODF), as well as the analytical q-ball imaging solution of (Descoteaux et al., 2007) (a-ODF) and normalized version with constant solid angle (csa) of (Aganj et al., 2010; Tristán-Vega and Aja-Fernández, 2010) (csa-ODF). a-ODF and csa-ODF are implemented...
for two SH orders (4 and 6) and all use a Laplace–Beltrami regularization with $k = 0.006$. For the rest of the paper, $r_4$, $r_6$ and $r_8$ represent maximum SH order 4, 6, and 8 respectively (Descoteaux et al., 2006).

Moreover, spherical deconvolution (SD) techniques were tested at order 6 and 8. We currently have two implementations in the Tractometer: our in-house implementation of (Descoteaux et al., 2009), based on diffusion ODF deconvolution to recover the fiber ODF (fODF), and the MRtrix implementation of (Tournier et al., 2007), based on the raw diffusion signal deconvolution to recover the fiber orientation transform (FOD). For both SD techniques, default parameters of the constrained regularization were used (Tournier et al., 2012), and the single fiber response function was estimated using voxels with FA value above 0.2 in the white matter mask seen in Fig. 7. Voxels with FA values above 0.2 are only present in the ‘U’ bundle 4 and the bottom part of bundle 2. A subset of these local estimations are illustrated in Fig. 6.

Tractography: seeding There are two seeding parameters in the Tractometer. First, whether or not we seed in the whole white matter (WM) or seed from regions of interest (ROI) corresponding to the gray/white interface (as seen in Fig. 7). These masks were manually drawn and are provided in the Data section of the Tractometer website (tractometer.org/data). Note that gray/white interface ROIs all have the same number of voxels in them (exactly 24 voxels).

At the voxel level, there are also different seeding strategies.

- **Our in-house strategy** randomly draws $N$ seeds in the voxel and ensures that all ODF maxima are used (illustrated in Fig. 8). This is to ensure that streamlines will be launched in all principal directions when there is an underlying crossing configuration. Hence, seeds are randomly placed spatially within the voxel, and the tracking direction is a random pick on the maxima of the interpolated ODF at that spatial point.
In the case of our in-house deterministic and probabilistic tractography algorithms (Girard and Descoteaux, 2012), seeds are also randomly placed in the voxel, but the first tracking direction is always picked according to the fODF distribution, as done by probabilistic tracking.

- The tractography pipelines using TrackVis also use multiseeding, which is available with the -rseed option.

- MRtrix also randomly distributes seeds in voxels but follows a different philosophy. The user can provide a required number of streamlines, \( N \), and MRtrix will do everything to reconstruct \( N \) streamlines by randomly putting seeds in the seeding mask provided. Hence, even if a user wants \( M \) seeds per voxel, MRtrix cannot guarantee that exactly \( M \) seeds will be used in each voxel. It can only approximately get \( M \) seeds per voxel if the user provides a desired number of streamlines \( N = VM \), where \( V \) is the number of voxels in the seeding mask.

Tractography: masking. In the current version of the Tractometer, complete white matter masking is used, using the mask shown in Fig. 7, which was manually segmented. Hence, the tractography process is always stopped when stepping outside this binary tracking mask.

Tractography: angular constraint, radius of curvature and step size. During the tractography procedure, discrete steps (\( s \) in Fig. 9) are taken to estimate streamlines through the white matter. This step varies from one publication to the other (Tournier et al., 2012; Tournier et al., 2011; Röttger et al., 2011; Hagmann et al., 2003; Parker et al., 2003; Descoteaux et al., 2009; Garyfallidis, 2012; Lazar et al., 2003), and some algorithms have no step size (Mori et al., 1999). Too large of a step has the risk of stepping outside a bundle and into another one. Too small of a step size enhances the risk of deviating from the tract trajectory and increases computational burden. In the current Tractometer, we vary step size with the following four increments: 0.3 mm, 0.6 mm, 1.0 mm and 3.0 mm.

Moreover, depending on the tractography implementation, there is often a maximum allowed angle of curvature between two consecutive directions defined by the maximum aperture angle (\( \theta \) in Fig. 9), or the radius of curvature (\( R \) in Fig. 9). In the current Tractometer, we vary \( R \) in the same four increments as the step size. The mathematical relation between \( R, s \), and \( \theta \) is: 
\[
\theta = \min(2 \arcsin(s/(2R)), 90) \in [0, 90]\]
(Tournier et al., 2012), which results in constraints shown in Table 1.

Finally, we do not impose a maximum or minimum length for streamlines to perform a fair comparison between methods.

Tractography: deterministic and probabilistic. For HARDI local reconstruction techniques, we use the deterministic streamline implementations of TrackVis (odf_tracker command), MRtrix (streamtrack SD_STREAM) (Tournier et al., 2012) and two in-house...
implementations. The first (Det. in-house 1 in the tables) follows the local model maxima closest to its incoming direction (Descoteaux et al., 2009). The second technique always follows the direction associated to the maximum value on the sphere within the aperture angle permitted. Moreover, we use the probabilistic streamline implementations of MRtrix (streamtrack SD_PROB) (Tournier et al., 2012) and our in-house implementation (Girard and Descoteaux, 2012; Girard et al., 2012). For DTI, tensorline (Weinstein et al., 1999), FACT (Mori et al., 1999) and Runge–Kutta (rk) (Basser et al., 2000) implementations of TrackVis (dti_tracker commands) were tested as well as our in-house implementation of the tensor advection (TEND) algorithm (Lazar et al., 2003).

**Resulting tractography pipelines.** The combinatorial number of possible tractograms quickly explodes if we start modifying all
possible parameters. Overall, we have $N = 57,096$ different tractography pipelines. Table 2 illustrates the different steps of the tractography pipeline that were combined.

**Ranking system, database and website.** An automated system is used to create the pipelines (using the different parameters values combination) to launch all the computations needed to obtain the final streamlines sets. Once the streamlines sets are computed, the system uses the `track_counts` function of Dipy (Garyfallidis et al., 2011) to filter those streamlines, using the different sets of ROIs. Each ROIs combination can result in a subset of streamlines that correspond to a specific connection between ROIs. For each of those subsets, the system computes the metrics and updates the database. Once all metrics are computed for all tractography pipelines, each pipeline can be ranked according to its VC, IC, NC, VB, IB and ABC values.

The final results are accessible on the Tractometer website (tractometer.org). Users can search, filter and sort the different pipelines according to the different metrics. This way, people will be able to find the best methods and parameters values for their specific needs.

The website also allows users to submit datasets in three categories:

1. A modified diffusion-weighted dataset.
2. A field of ODFs. The field of ODFs must be submitted as a 3D dataset with discrete spherical function values at every voxel. The discretization of the sphere corresponds to a fifth order tessellation of the icosahedron having 724 points on the hemisphere. This tessellation is dense and has an approximate 4 degrees angle between each sample on the sphere.
3. A dataset of streamlines. The Tractometer currently supports the .trk (TrackVis), .tck (MRtrix), .bundledata (brainvisa.info), and fib (binary and ASCII VTK) streamlines formats.

Hence, a user working on artefact correction or denoising techniques can submit in category (1), a user working on local reconstruction techniques can submit in category (2), or one can submit the streamlines resulting from a new tractography algorithm in category (3). The user must also include a short description of the methods and algorithms used for the creation of this dataset, in order to compare it against what others have proposed. In the case of a diffusion dataset or a field of ODFs submission, our framework will automatically generate the streamlines resulting from this new dataset combined with the tractography pipelines already implemented in the system. Doing so, it will show the impact of new contributions on the final tractography results. As the website grows, more features will be added like new algorithms, the possibility to submit third party libraries, new geometrical metrics and other phantoms, to name a few.

**Fig. 8.** Multiple seeding scenarios in a crossing configuration. Our in-house seeding technique insures that all peaks of a multiple maxima angular profile are covered by the seeding.

**Fig. 9.** Deterministic (left) and probabilistic (right) tractography illustrated with angular constraint ($\theta$), radius of curvature ($R$) (the curvature is $1/R$) and step size ($s$), over FODF glyphs. The mathematical relation between $R$, $s$, and $\theta$ is: $\theta = \min(2\arcsin(s/(2R)), 90) \in [0, 90]$ (Tournier et al., 2012). Deterministic tracking picks the principal direction of the FODF closest to its incoming direction whereas probabilistic streamline tracking randomly picks according to the FODF distribution.
3. Results

Here is an overview of different results and messages that come out of the new analysis of the $N = 57,096$ tractography streamline outputs in the database. We advise the reader to have Fig. 1 with the ground truth bundle's ID number nearby while reading this section as well as the new definitions of VC (valid connections), IC (invalid connections), NC (no connections), VB (valid bundles), IB (invalid bundles) and ABC (average bundle coverage) in mind. There are many ways to query the database and many possible views and messages that one can extract from it. Here, we decide to start from a very general to a very specific view.

A good tractography pipeline should have a high VC percentage, a high ABC percentage and 7/7 VB found. Table 3 gives a general and per bundle view comparing the success rates of the DTI, HARDI deterministic (HARDI-Det) and HARDI probabilistic (HARDI-Prob) tractography pipelines. Note that out of the 57,096 different pipelines, 7550 pipelines do not recover at least one VC (i.e. 49,546 pipelines have at least 1 VC). For the rest of the tables, we choose to remove these pipelines from statistics because they arise from an aberrant combination of parameters selection in the tractography process. We exclude them to focus on results and messages arising from successful tractography pipelines. Moreover, from Table 3, we note that not all methods are able to retrieve all the 7 out of 7 (7/7) VB. A total of 6,360 out of 57,096 pipelines recover 7/7 VB. This is encouraging and means there is a large number of different parameters that produce good streamlines outputs which find the 7 VB. In fact, 15%, 13% and 1% of HARDI-Det, HARDI-Prob and DTI pipelines that have at least one VC, also find 7/7 VB. For DTI, it is already a surprise that 58 out of 5065 pipelines recover 7/7 VB, given the difficulty of the fiber crossing configurations of the FiberCup. Most of these 58 DTI pipelines (56/58) come from the TrackVis implementation of the tensorline algorithm (none of the MRtrix and in-house DTI tracking implementations recover 7/7 VB).

In terms of per bundle statistics, we also note from Table 3 that results for DTI and HARDI-Det pipelines show similar results. One must keep in mind that these statistics are computed using pipelines with at least 1 VC, meaning that HARDI-Det statistics are normalized by a much higher pipeline count (PC). The pipelines from the HARDI-Prob family show, on average, a lower bundle recovery rate and a higher ABC. The standard deviation on ABC for DTI, HARDI-Det and HARDI-Prob pipelines is in the interval $[13.1\%, 20.6\%]$, $[14.7\%, 19.2\%]$ and $[23.9\%, 29.1\%]$, respectively. Even with this high standard deviation, the results show that probabilistic pipelines are better at recovering the full fiber bundles. The standard deviation table corresponding to Table 3 is provided in the supplementary materials. For the rest of the results, we choose to compute statistics only on tractography pipelines with 7/7 VB.

Acquisition. Table 4 shows the effect of the different acquisition options on the tractography results. We first note that connectivity metrics are improved when using the averaged acquisition. We also see that acq1 is systematically better than acq2. Best scores are always obtained with b-value 1500 s/mm$^2$. Moreover, given its low SNR, it is not surprising that the b-value 2000 s/mm$^2$ has the worst connectivity scores. Also, as expected, b-value 650 s/mm$^2$ has a much lower count of 7/7 VB. It has a high SNR, but the local angular profiles reconstructed from it do not properly recover crossing configurations, even with HARDI techniques. However, the averaged b-value 650 s/mm$^2$ acquisition also results in the best ABC. This is surely due to the over-smoothed local angular ODF profiles that are reconstructed. Finally, note that in the best case, we only have 12.7% valid connections. This is quite low and reflects the aggressiveness of the stopping criteria in the tractography pipelines. Based on Table 4, we compute the rest of the statistics only using the averaged acquisition and b-value 1500 s/mm$^2$.

By doing this, we gradually zoom into the best tractography pipelines. The standard deviation table corresponding to Table 4 is provided in the supplementary materials.

Local reconstruction techniques. Table 5 shows the effect of the different local reconstruction techniques. Sharp angular distributions have more success at recovering 7/7 VB (highest count), starting with FODs from MRtrix, followed by our in-house FODs and the smoother csA-ODF, a-ODF and num-ODF reconstruction techniques. Again, in terms of connectivity metrics, sharp angular profiles (FODs and FODFs) show the highest VC and lowest IC and IB. However, this is at the price of having more NC. DTI has an IB count of 25, HARDI-Det has an IB count of approximately 17.5 for smooth profiles (num-ODF, a-ODF and csA-ODF) and approximately 12.5 for sharper profiles (FOD, fODF). This is approximately 5 IB less for the sharper angular profiles, which is considerable. Moreover, HARDI-Prob has an IB count of approximately 22 for sharp models, which is also considerable. Also note that no probabilistic pipeline successfully recovered 7/7 VB for the smoother ODF profiles. Finally, as expected, FODs/FOD combined with probabilistic tracking have the best ABC. At this point, we set the local reconstruction techniques to FOD-r6,r8 and FODF-r6,r8 for the rest of the analysis. There was no significant difference between maximal SH order 4, 6, and 8 reconstructions (shown Table 4 of the supplementary materials).

Tractography parameters: step size, curvature, seeding. Considering tractography pipelines that recover 7/7 VB with the averaged acquisition at b-value 1500 s/mm$^2$ and the FOD-r6,r8 or FODF-r6,r8, we do not find a trend for best combination of step size, curvature and seeding strategy to obtain high VC, low IB and high ABC. The step size, the curvature and the seeding strategy seem to be very dependent on the choice of the tractography algorithm. Hence, averaging results from many different tractography pipelines does not provide any interesting observation or recommendation (shown in Tables 5 and 6 of the supplementary materials for completeness). Overall, higher ABC is generally obtained with a seeding strategy using more than 1 seed, but is similar for all multiseeding tested (9, 17, and 33). VC is also generally slightly higher on average using a single seed. Finally, in general, the number of IB is lower when ROI seeding and a single seed is used.

Best of. We finish by reporting an overview of best of tractography pipelines sorted in terms of maximum VC, minimum IB with
maximum VC, minimum NC and maximum ABC. As seen in Table 6, we report the best of each family of tractography pipelines and indicate their corresponding connectivity metrics and rank. The line corresponding to the overall best tractography pipeline is always indicated with a gray background and the resulting valid connection streamlines are also illustrated in Fig. 10. For comparison, we also illustrate VC of several worst tractography pipelines in Fig. 10.

1. In terms of maximum VC, the best deterministic tractography pipeline recovers nearly 25% VC, the best probabilistic pipeline 15% and the best DTI pipeline 10%. In general, the best of deterministic pipelines have much higher VC percentages than probabilistic pipelines. When looking at the top 100 pipelines, most combinations of step size, curvature and seeding strategy appear. The ave-b1500 dataset is the only one used in the top 100 pipelines and the FOD-r6,r8 are the most common local techniques. Note that the csa-ODF-r4,r6 come up only once in the top 100 pipelines. Finally, we note that the overall best pipeline in terms of maximum VC is the in-house 2 tractography algorithm, with the ave-b1500, FOD-r8 from MRtrix, a full white matter seeding with 1 seed per voxel, step size of 1 mm and radius of curvature of 3 mm.

2. In terms of minimum IB and maximum VC, the best pipeline produces 15.6% of VC, whereas the VC decreases by 4.7% for the pipeline coming at the second rank. The best pipeline is from our in-house 1 tractography algorithm, with ave-b1500 dataset, FOD-r6 from MRtrix, seeding only from ROIs with 1 seed per voxel, a step size of 0.3 mm and radius of curvature 0.6 mm. We also note that the top 8 pipelines produce 0 IB!

3. In terms of minimum NC, the best pipeline produces 60.4% of NC and comes from the deterministic MRtrix tractography algorithm with the ave-b1500 dataset, our in-house csa-ODF-r4 profiles, seeding only from ROIs with 1 seed per voxel, a step size of 3 mm and radius of curvature 1 mm. We also note a wide range of ratio VC/(VC + IC), which is in the interval [36.6%, 78.8%] in the top 100 pipelines. All HARDI local techniques appear in...
the top 100, including the smoother ones (num-ODF, a-ODF and csa-ODF). Overall, the ave-b1500 is the most appearing dataset, the most common step size is greater or equal to 1 mm, but mainly \( s = 3 \) mm, and the radius of curvature is most often \( R \geq 1 \) mm. Finally, all combinations of seeding and multiseeding strategies appear in the top 100.

4. In terms of maximum ABC, the best ABC percentage is obtained using probabilistic tracking. In fact, the top 100 pipelines are exclusively probabilistic ones. The best pipeline obtains an ABC of 91.8% but, at the same time, has a low VC percentage and the highest number of IB, with 33 invalid bundles! In fact, the IB is always in the interval \([23, 32]\), which is a huge number of invalid bundles. The first deterministic algorithm is ranked 120. We also note that DTI produces poor ABC results. Overall, in the top 100, all local techniques are FOD-r6,r8 and most pipelines use the WM seeding with a multiseeding of 33 seeds per voxel. Finally, in the top 100, acq1, acq2 and ave for \( b\)-value 1500 s/mm\(^2\) are equally common, the step size is most often smaller or equal to 1 mm and the radius of curvature is larger or equal to 1 mm.

4. Discussion

It is important to first note that results reported in this paper have changed compared to our MICCAI paper (Côté et al., 2012), which only had 1,152 tractography pipelines. Moreover, in the current version of the Tractometer, we ensure that all ROIs have the same number of voxels. We encourage the community to query the database differently and extract other messages that we might have omitted in this paper.

In general, a conclusion based on techniques available in the database is that the current best tractography pipeline configuration for optimal trade-off between VC, IC, NC, VB, IB and ABC is using the averaged dataset with \( b\)-value 1500 s/mm\(^2\), a sharp local estimation from spherical deconvolution and a deterministic tracking algorithm handling fiber crossings such as our in-house or MRtrix implementations. However, it is not straightforward to have general conclusions for the step size, curvature, seeding strategies. We now discuss each part of the processing pipeline in turn.

4.1. Acquisition

Limitations of the FiberCup dataset. The FiberCup phantom dataset is not perfect. The anisotropy of the phantom is quite low and thus privileges a certain class of techniques based on an angular distribution content of the DWI data. Raw signal modeling-based approaches are disadvantaged (such as implemented in FSL) and not well suited for the FiberCup dataset, as shown in (Fillard et al., 2011). The FiberCup dataset only provides three \( b\)-value acquisitions, with the \( b\)-value 2000 s/mm\(^2\) too noisy to be useful
and the \( b \)-value 650 s/mm\(^2\) resulting in over-smoothed local ODF profiles challenged in the crossing configurations. It is important to have richer \( b \)-value phantoms in a pursuit of having an optimal \( b \)-value acquisition for HARDI tractography.

Moreover, the FiberCup phantom does not provide a real 3D space example and is limited to the 2D plane. Other phantoms should take into account more complex bundles. Or, ideally, new techniques should be developed to compare streamlines bundles within a brain, as attempted in (Baumgartner et al., 2012). Of course, one has to be careful not to develop or tune his tractography algorithm to best perform solely on the FiberCup dataset. Since our preliminary version of this paper (Côté et al., 2012), several groups have contacted us to add new phantom datasets, such as Fieremans et al., 2008 (url: www.nitrc.org/projects/diffusion-data), to the Tractometer. We are currently investigating these new phantoms.

**Averaging.** It is reassuring and expected (Table 4) that averaging the two repetitions (acq1 and acq2) helps improve the connectivity metrics. However, the reader must keep in mind that this doubles the acquisition time, which is often not possible in clinical applications. This motivates the development of state-of-the-art denoising and noise correction techniques (Descoteaux et al., 2008; Coupé et al., 2010; Aja-Fernández et al., 2011; Brion et al., 2012). We are also currently considering providing masks and ROIs in a upscaled 128x128x6 space, so that groups developing new super-resolution techniques (Nedjati-Gilani et al., 2008; Scherrer et al., 2012) or using upsampling (Dyrby et al., 2011; Raffelt et al., 2012; Smith et al., 2012a) can use the Tractometer to evaluate their techniques.

**Direction undersampling.** Finally, groups working on novel compressed sensing and undersampling acquisitions with limited number of gradient directions can use the Tractometer to compare their novel techniques. The diffusion community needs a HARDI reconstruction acquisition scheme that is similar to a DTI-like scheme. Hence, it is important to seek high quality fiber crossing tractography results obtained from DTI-like acquisitions (approximately 12–30 directions).

### 4.2. Local estimation techniques

First, sharp angular distributions (FODs and fODFs) had more success than smoother ODFs (num-ODF, a-ODF, csa-ODF). Hence, in general, spherical deconvolution performed better than numerical and analytical ODFs. Depending on the acquisition used and its noise level, direct effects of local estimation techniques could also be seen in the resulting streamlines. Therefore, as the Tractometer grows, it can serve to further explore the effect of SH maximum order, \( \alpha \)-regularization parameters, response function estimation for deconvolution, amongst others. Finally, some SH-based techniques have not been implemented, such as spherical ridgelets (Michailovich and Rathi, 2009), spherical wavelets (Kezele et al., 2010), and the diffusion orientation transform (DOT) (Özarslan et al., 2006), but should soon appear in the Tractometer as we include more libraries, such as Dipy (Garyfallidis et al., 2011) and Camino (Cook et al., 2006). Similarly, we would like to compare multi-compartment parametric models like ball & sticks (Behrens et al., 2007) or multi-tensor (Kreher et al., 2005).
4.3. Tractography parameters

Step size. In the current literature, a general rule of thumb for choosing the step size is to set the step size as half the spatial acquisition grid, although (Tournier et al., 2012) recently suggested 1/10 the spatial acquisition grid. Note that some tractography techniques do not require any explicit step size, such as FACT (Mori et al., 1999) implemented in DTI studio (www.dtistudio.org), which steps according to the voxel size, and should be considered in future comparisons. The Tractometer has a large number of pipelines that produce 7/7 VB with good valid/invalid connection trade-off and acceptable average bundle coverage, using all possible step sizes tested. The step size clearly depends on other parameters and the implementation of the tractography algorithm as such. Optimal step size selection remains an open question.

Interpolation. In the current tractography techniques provided in the Tractometer, interpolation techniques have not been thoroughly studied. During the fiber tracking process, one wants to combine local neighborhood information. For this purpose, interpolation is used. Runge–Kutta interpolation is available in TrackVis when using the tensor model, but most other tractography techniques implemented use tri-linear interpolation on the field of ODFs. Should this interpolation take place on the original DWI data, on the field of DTs/ODFs, or simply on the principal directions extracted? When done on the field of principal directions, voxels at the boundary of the tracking mask are often problematic, and it is thus commonly performed on the field of DT/ODFs. However, interpolation certainly remains an open question and can considerably increase computation time of tractography pipelines. It also needs to be further addressed in the future.

Masking. Defining a good tracking mask is crucial for tractography algorithms (Guevara et al., 2011; Girard and Descoteaux, 2012). The novice tractography user soon realizes that if tracking is done on the whole brain, tracts can be found (i) passing through the ventricles; (ii) in between left and right hemispheres, through the inter-hemispheric space; and (iii) in between cortical regions, through sulci, among other physically impossible regions. Often, a classical approach is to define the mask using a thresholded fractional anisotropy (FA) map, often thresholded at values between 0.1 and 0.2. FA is fundamentally flawed in crossing and curving voxels and thus produces masks containing holes. In the Tractometer, the complete mask shown in Fig. 7 is used. We found that this is an aggressive stopping criteria (Girard and Descoteaux, 2012), leading to a large number of NC (more than 60% of streamlines), as seen in Fig. 4, in most tractography algorithms. Recent state-of-the-art techniques take into account information about the anatomy (Iturria-Medina et al., 2008; Girard and Descoteaux, 2012; Girard et al., 2012; Bloy et al., 2012; Smith et al., 2012a; Smith et al., 2012b) to improve the behavior of tractography. These techniques stress the importance for tractography algorithms not to stop in cerebral spinal fluid (CSF), not to stop in the white matter and terminate in gray matter or out of the brainstem (Girard and Descoteaux, 2012; Smith et al., 2012a). Finding the optimal boundary and masking conditions for tractography algorithms remains an open question.

Seeding. In general, using both complete and ROI seeding, 1 seed per voxel produces the best trade-off between valid and invalid connections. However, using only 1 seed per voxel suffers from a low ABC. The more seeds per voxel, the more IC but the better the ABC. In the end, the best ABC is obtained from complete seeding and 33 seeds per voxel but at the price of having more than 24 invalid bundles, which is a huge problem. Hence, the Tractometer currently suggests that both ROI and complete seeding are appropriate seeding strategies and should be used with care to control the proportion of valid/invalid connections. Complete fiber tractography (from all white matter voxels) should probably be done with a low seeding number per voxel to avoid invalid connections as much as possible. On the other hand, if extra information (anatomical, functional, amongst others) is available and can be used to filter out invalid connections, a large number of seeds per voxel can be used to obtain a good average bundle coverage.

Seeding and masking certainly have a huge impact on the fiber count and the density of fiber bundles. In a sense, “easier” bundles such as bundles 4 and 6 (those that have less fiber crossing configurations along their path), end up artificially over-estimated, as reported in (Smith et al., 2012b), with a much larger number of streamlines, compared to more “difficult” bundles such as bundles 2 and 3. Hence, as mentioned in (Jones et al., 2013), fiber count should be avoided. In (Smith et al., 2012b), it is advised to correct for this fiber density bias using a new filtering approach and by seeding only from the gray/white matter interface. Seeding and masking clearly remain open questions for the tractography community.

4.4. Deterministic and probabilistic tractography

Fiber tracking is an intuitive process, but one with many underlying parameters mentioned in the previous paragraphs. Streamlines are recovered by propagating through a field of DTs, ODFs, fODFs, FODs, or any other pre-computed set of directions. What directions to follow? How to follow them? How to numerically extract them? These questions are crucial in tractography and remain open. Many families of algorithms exist, just to name a few: deterministic, probabilistic, geodesic, and global algorithms. We are pursuing an extension for connectivity matrices provided by other probabilistic, geodesic (Péchaud et al., 2009; Sepasian et al., 2009), global tractography algorithms (Jbabdi et al., 2007; Kreher et al., 2008; Fillard et al., 2009; Reisert et al., 2011) or any other fiber tracking approach.

The Tractometer is currently developed for streamline-based deterministic and probabilistic algorithms. Deterministic techniques have a lower ABC percentage than probabilistic ones, but have a much better valid/invalid connection trade-off. Probabilistic tractography techniques are surely disadvantaged by the limitations of the FiberCup datasets but in any case, they are, in nature, much more aggressive in the way they randomly explore tracking directions from the field of fiber ODFs provided to the fiber tracking algorithm. Hence, probabilistic pipelines have a tendency to produce a large number of no and invalid connections. As for multiseeding strategies, we believe they should be used with care and only if extra information is available to filter out invalid connections.

5. Conclusion

We have developed the Tractometer, an online system for the evaluation of tractography pipelines. Overall, we have shown that MRtrix and our in-house tools based on spherical deconvolution currently provide the best rankings in terms of valid/invalid connections. TrackVis is based on q-ball ODFs or DTI, and thus, does not perform as well, just as our in-house tracking based on q-ball ODFs and DTI. Our in-house tracking will be available shortly in Dipy (Garyfallidis et al., 2011). In summary, the most important messages that come out of the Tractometer in its current state are: (i) Averaging improves results. (ii) Sharp angular ODF profiles help the tractography algorithm. (iii) The more you add seeds, the more invalid connections (IC) you get, and one should be careful about randomly seeding a large number of seeds in the complete white matter. (iv) Probabilistic streamline tracking has highest ABC at the price of severely increasing NC and IC. (v) Deterministic streamline techniques...
show the best valid/invalid connection trade-off in our connectivity metrics. The (iv)–(v) messages are contrary to the popular belief that probabilistic tractography gives, in general, better results than deterministic tractography. We have shown that this is not the case when using the FiberCup phantom. Hence, probabilistic tractography algorithms should be used with care, especially in global connectomics studies.

Of course, these are only the first steps of the Tractometer, but we believe that, as the community contributes to the system with more, and better phantoms, this new system can have a positive impact on the dMRI community. Just as the machine learning and computer vision communities have used benchmarks to move forward in algorithm design and evaluation, the dMRI community needs to do the same to answer open questions. Only then can new tractography algorithms be compared to the state-of-the-art and their contributions quantified.

Send us your corrected raw diffusion data, your ODFs or your fiber streamlines at tractometer.org and you will be compared and ranked against other state-of-the-art techniques!

Acknowledgments

We would like to thank Compute Canada for access to High Performance Computing (HPC) resources and thank NeuroSpin for the public FiberCup dataset (url: www.hnaf.fr/spip.php?rubrique79).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.media.2013.03.009.

References


