Tractometer: Online Evaluation System for Tractography

Marc-Alexandre Côté, Arnaud Boré, Gabriel Girard,
Jean-Christophe Houde, and Maxime Descoteaux

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

Abstract. We have developed a Tractometer: an online evaluation system for tractography processing pipelines. One can now evaluate the end effects on fiber tracts of different acquisition parameters (b-value, number of directions, denoising or not, averaging or not), different local estimation techniques (tensor, q-ball, spherical deconvolution, spherical wavelets) and to different tractography parameters (masking, seeding, stopping criteria). At this stage, the system is solely based on a revised FiberCup analysis, but we hope that the community gets involved and provides 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 elucidating the open questions in fiber tracking: from raw data to connectivity analysis.

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) in crossings and high curvature areas. Thus, numerous new high angular resolution diffusion imaging (HARDI) tractography techniques have been proposed. Several groups have studied the effect of interpolation, step size, stopping criteria, but mostly on toy simulated data or qualitatively [1] on large tracts or inter-hemispheric connections. Validation of fiber tractography remains an open question and a challenge on real data.

A first attempt has been done with the FiberCup phantom dataset [2], which was made public and is now used by the community for quantitative evaluation of tracking algorithms. However, in our opinion, two important drawbacks of the FiberCup are the seed points given and the quantitative metric used to compare with ground truth. Only 16 seeds are given here and there in the dataset, close to boundaries and in the middle of structures. These 16 seeds result in 16 individual tracts that are compared with the ground truth in terms of spatial, tangent to the tract and a curvature distance. These measures are local and do not capture well the global connectivity profile of the tractography algorithm. Other

N. Ayache et al. (Eds.): MICCAI 2012, Part I, LNCS 7510, pp. 698–705, 2012.

problems are that each participant of [2] 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 in closer 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 *tractogram* (the fiber tractography output). We developed a tractogram evaluation method to compare tracts and evaluate the number of found and not found fiber bundles, the proportion of tracts part of existent and non-existent bundles, and the proportion of incomplete tracts. Since these tractogram characteristics have a direct impact on connectivity analysis, having a tractogram evaluation tool is crucial in the era of the human connectome studies [3].

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 **(url: *scil.dinf.usherbrooke.ca/tractometer*)** to evaluate and rank pipelines. 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 tractogram. The user can then obtain a ranking against the current database of state-of-the-art techniques. Presently, in this database, we have $N = 1152$ different tractograms coming from our in-house tools, *MRtrix* [1] and *TrackVis* (`trackvis.org`) using one of the acquisitions or the averaged, local estimation techniques (tensor, q-ball, constant solid angle q-ball, spherical deconvolution and wavelets), masking from a complete phantom mask or regions of interest (ROIs), multiple seeding strategies and different stopping criteria.

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 [4] for the MICCAI FiberCup workshop held in 2009, which resulted in a group publication in [2]. In our opinion, the metrics proposed in [2] are too local and vulnerable to the seeds given and, as a result, do not capture the global *connectivity* behavior of the tractography technique. 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. 1b), or 2) From ROIs mimicking gray/white matter interface, as seen in Fig. 1c). Hence, the tractogram from a resulting tractography algorithm (Fig. 1e)) can be filtered by the ROIs at the end of bundles to quantify the global success (Fig. 1f)) and errors present in the tractogram.

Definitions and Rationale. We performed a survey with neurosurgeons and neurologists at our institute concerning true and false connections of tractograms.

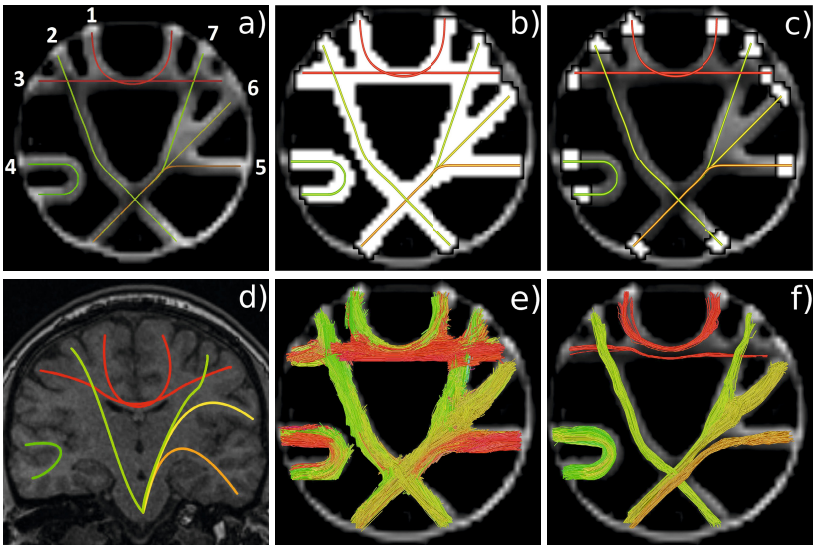

Fig. 1. a) FiberCup mimicking a brain in coronal slice d). b) Fibercup's white matter mask and c) regions of interest (ROIs) similar to white/gray matter interface on the cortex. e) Full tractogram (45,000 tracts) and f) tracts passing through ROIs of c).

We concluded that these terms were not the best choice for connectivity analysis purposes. Therefore, we define the following five new terms:

- *True Connections (TC)*: tracts connecting expected ROIs. This is illustrated by tracts in Fig. 1f). TC will be reported in percentage of true connections.
- *False but Plausible Connections (FPC)*: tracts connecting unexpected ROIs. These tracts are spatially coherent, have managed to connect ROIs, but do not agree with the ground truth (see Fig. 2a)). It is reported in %. According to our survey with clinicians, these are problematic tracts as they "look anatomically good" but are in fact non-existent from *a priori* knowledge.
- *Wrong Connections (WC)*: tracts that simply do not connect two ROIs. Depending on how the tractography algorithm handles stopping criteria, these tracts either stop prematurely due to stopping criteria or, most often, hit the boundaries of the tracking mask, as illustrated in Fig. 2b), c), e), f).
- *True Bundles (TB)*: bundle connecting expected ROIs. Figure 1a) shows the true bundles. TB is reported in bundle counts, from 0 to 7 for the FiberCup.
- *False but Plausible Bundles (FPB)*: bundle connecting unexpected ROIs. As TB, FPB is reported in bundle counts. Figure 2a) and d) show that bundles 1, 3 are mismatched, but look plausible, had we not known the ground truth.

Tractography. For this paper, we have focused only on the 3 mm isotropic, 64 directions, $b = 1500$ s/mm^2 dataset as 9/10 participants of [2] used this dataset). At the acquisition level, two measurements of the raw data are available. Hence, either one acquisition or an averaged were tested. For local estimation, diffusion

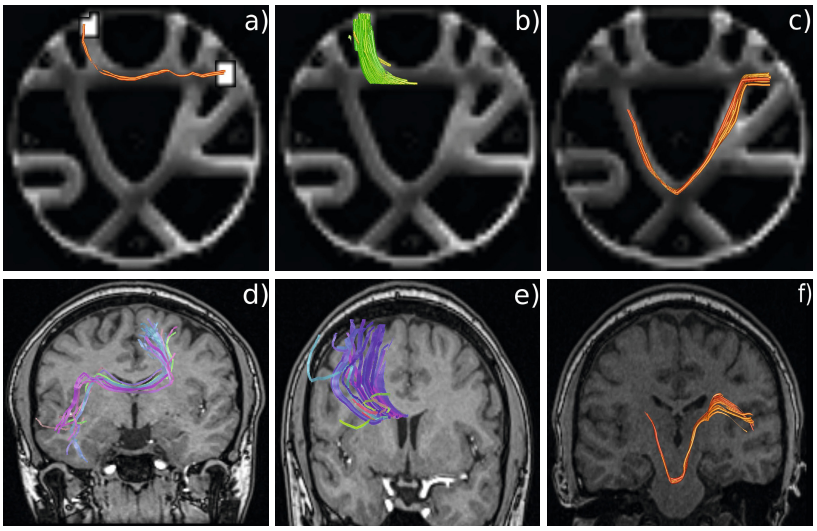

Fig. 2. a) A FPC between 2 ROIs of gray matter (FiberCup) and a real data analogy in d). In b), c), e) and f), we see WC due to many collisions with the tracking mask.

tensors were estimated using our in-house log-Euclidean implementation [5], the one from *TrackVis* or from *MRtrix* (public softwares). Otherwise, the diffusion orientation distribution function (ODF) from analytical q-ball of [6] (a-ODF) and normalized version with constant solid angle of [7] (csa-ODF) were implemented for several spherical harmonics (SH) order. Here, we report order 4 (r4) because it is the best for the FiberCup data. Next, a spherical wavelet (SW) decomposition based on [8] was performed at order 8. Moreover, spherical deconvolution (SD) techniques were tested at order 4 (r4), 6 (r6) and 8 (r8). We have an in-house implementation of [9] as well as the estimation of MRtrix to recover the fiber ODF (fODF). These local estimations are illustrated in Fig. 3.

As mentioned before, tracking is performed in a complete tracking mask shown in Fig. 1b), but seeding is either started from everywhere in that mask or from the ROIs (Fig. 1c)). Using these seeding options, we tested multiseeding with our implementation of [3] covering all possible maxima when seeding in crossing voxels, random multiseeding of *MRtrix* and the one from *TrackVis*. Finally, we used deterministic streamline option from *MRtrix*, *TrackVis* and our in-house tools (similar to [3]), as well as the probabilistic option of *MRtrix*. For DTI, tensorline, streamline, FACT and Runge-Kutta (rk) options of *TrackVis* were tested as well as our in-house tensorline and streamline algorithms.

Ranking system, database and website. To compute the TC, FPC, WC, TB and FPB from the tractograms, we use *MRtrix* (filter_tracks) to filter the resulting tracts of pipelines using different sets of ROIs. For each subset of tracts that corresponds to a specific connection between ROIs, the metrics are updated and sent to a database. Once all metrics are computed for all tractography pipelines,

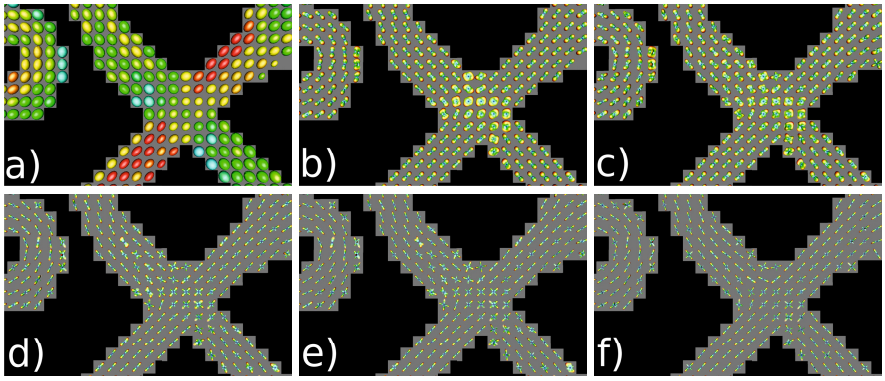

Fig. 3. Different local estimation methods provided in the Tractometer. a) DTI, b) a-ODF-r4, c) csa-ODF-r4, d) SD-r6, e) SD-r8 and f) SW-r8.

each once is ranked in 5 columns according to TC, FPC, WC, TB and FPB. The website allows one to submit his dataset(s) (modified diffusion dataset, fields of ODF, or tractogram) and a short description in order to compare it against what others have proposed. In the case of a diffusion dataset or fields of ODF submissions, our framework will automatically combine the submitted data with other pipelines already implemented in the database. Doing so, it will show the impacts of new contributions to tractography pipelines. 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.

3 Results and Discussion

Here is an overview of different results and messages that come out of the new analysis of the $N = 1152$ tractograms in the database. We advise the reader to have the new definitions of TC, FPC, WC, TB and FPB in mind before reading this section. Bold entries in tables represent "best" results. Firstly, not all methods are able to retrieve all the 7 out of 7 (7/7) TB.

TB (%)	0/7	1/7	2/7	3/7	4/7	5/7	6/7	7/7
DTI	0.0	0.7	6.2	11.8	54.9	19.4	6.9	0.0
a-ODF	0.0	0.0	0.0	0.0	4.2	12.5	37.5	45.8
csa-ODF	0.0	4.2	6.2	2.1	10.4	16.7	31.2	29.2
SD	0.0	2.1	0.0	0.0	2.1	8.3	16.7	**70.8**
SW	0.0	0.0	4.2	0.0	4.2	16.7	20.8	54.2
Total	0.0	1.4	3.3	2.8	15.1	14.7	22.6	40.0

Sharp angular distributions have more success at recovering 7/7 TB, from SD, SW and then csa-ODFs and a-ODFs. DTI never recovers 7/7 TB. Secondly, it may seem easy to obtain TB from the FiberCup dataset. However, we note that

there is only a small percentage of tracts that actually result in a TC. The results show that less than 38.2% of recovered tracts are TC! Hence, the major part of fiber tracts are either FPC or WC, as seen in the next table:

	TC (%)	FPC (%)	WC (%)	TB	FPB
Min	0.2	0.0	57.6	1.0	0.0
Max	38.2	15.2	98.4	7.0	22.0
μ	14.6	3.8	81.6	5.5	9.2
σ	7.8	3.2	9.2	1.6	4.5

Overall, between 57%-98% of tracts are WC and the remaining tracts are FPC (0% to 15%). These FPC account for 0 to 22 FPB out of the possible 39 FPB.

Thirdly, as one would expect, each TB of the FiberCup has a different success rate depending on its complexity, as seen in the following table:

No. Bundle / Local Estimation	1	2	3	4	5	6	7
DTI	81.9	8.3	34.7	83.3	82.6	92.4	23.6
a-ODF	87.5	70.8	45.8	**100.0**	91.7	91.7	52.1
csa-ODF	**100.0**	70.8	75.0	**100.0**	**100.0**	**100.0**	79.2
SD	93.8	85.4	**84.9**	99.0	97.4	97.9	**89.6**
SW	79.2	**91.7**	79.2	**100.0**	95.8	**100.0**	66.7

In fact, bundles 2 and 7 are harder to track than their homologue bundle in the brain because they go through several complex fiber crossing regions, whereas bundles 1, 4, 5, and 6 are most often tracked successfully by HARDI techniques.

Next, the following table shows that the more tracking seeds used, the more success the pipeline has in reaching the target ROI and thus increasing the % of pipelines scoring 7/7 TB. However, this also increases the chance to find FPB.

Mean (%) / Nb. Seeds	TC	FPC	WC	TB	FPB	(7/7)
1	14.2	**3.7**	82.1	4.7	**5.4**	15.7
9	14.6	3.8	81.6	5.6	9.3	**42.6**
17	**14.8**	3.8	**81.4**	5.8	10.6	50.0
33	**14.8**	3.9	**81.4**	**5.9**	11.5	52.8

Hence, one can see that TC, FPC and WC remain constant as multiseeding increases. Meaning that one must be careful when using aggressive multiseeding.

The next table compares *MRtrix*'s deterministic and probabilistic tracking based on a field of fODFs. All techniques recover, on average, the same number of TB and make low FPC. However, we note that the percentage of TC and WC seem to advantage deterministic tracking (i.e. better TC while keeping the WC and FPB lower). This highlights a limitation of probabilistic tracking, in that it explores the whole shape of the fODF but may take wrong turns and follow peaks that are part of different fiber bundles. This is something one should consider and further explore, especially in large-scale connectivity analysis studies.

MRtrix Pipelines		TC (%)	FPC (%)	WC (%)	TB	FPB
Local Estimation	Tracking					
SD-r6	Deterministic	**18.3**	3.2	**78.5**	**6.8**	10.1
SD-r6	Probabilistic	5.5	2.0	92.5	6.7	17.6
SD-r8	Deterministic	14.9	**1.6**	83.5	6.7	**8.0**
SD-r8	Probabilistic	7.0	**1.6**	91.4	**6.8**	16.4

Finally, in the following table, we explore the *decoupling* of fODF estimation from the tracking algorithm itself, between our in-house tools and *MRtrix*:

Pipelines		TC (%)	FPC (%)	WC (%)	TB	FPB
Local Estimation	Tracking (Det.)					
csa-ODF(in-house)	in-house	16.3	1.9	81.8	6.2	8.2
SD-r6 (in-house)	in-house	14.1	1.4	84.6	6.2	8.7
SW-r8 (in-house)	in-house	15.1	1.5	83.4	6.1	10.2
SD-r6 (MRtrix)	MRtrix	18.3	3.2	78.5	**6.8**	10.1
ODF-rk2 (TrackVis)	TrackVis	3.0	**0.1**	96.9	5.3	**3.4**
SD-r6(MRtrix)	in-house	**24.5**	3.4	**72.1**	6.5	10.5

As seen in this table, *MRtrix* has, on average, the best success at recovering 7/7 TB. However, we also see that the best results *TrackVis*, based a Runge-Kutta interpolation of the qball ODF field, has a very low FPC and FPB. Moreover, we finally see that fODFs from *MRtrix* combined with our tracking has the best compromise between TC and WC. We believe that this is especially owed to the seeding strategy that differs between the two tracking techniques. Carefully covering all multiple orientations in a seed voxel when there are many (as done in our in-house tools and as opposed to random seeding done in *MRtrix*), achieves more TC and less WC. This needs to be further investigated and motivates the fact of *decoupling* tracking steps (seeding, masking, interpolation and stopping).

A conclusion based on available techniques in the database is that the current best tractography pipeline configuration for optimal trade-off between TC, FPC, WC, TB and FPB is using the averaged dataset, a sharp local estimation (e.g. spherical deconvolution), a multiseeding strategy, is starting from the ROIs and uses a tracking algorithm handling fiber crossings such as [3] and *MRtrix*.

Limitations. The current Tractometer system has limitations. Currently, only a single dataset has been included in the database. Of course, one has to be careful not to develop or tune his fiber tractography algorithm to best perform solely on the FiberCup dataset. As mentioned in [2], this phantom data is not perfect and can privilege a certain class of techniques such as streamline techniques based on an angular distribution content of the DW-MRI data. Raw signal modeling-based approaches are disadvantaged (such as implemented in FSL) and not well suited for the FiberCup [2]. 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 tracts bundles within a brain. Finally, we have currently

focused on the global TC, FPC, WC, TB and FPB metrics but other geometrical metrics could be incorporated to get a more complete ranking system.

4 Conclusion

We have developed a new Tractometer to evaluate tractography pipelines. Overall, we have shown that *MRtrix* and our in-house tools based on SD currently provide the best rankings. *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 and DTI. Of course, there are limitations to this system 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 tracts and you will be compared and ranked against other state-of-the-art techniques!

References

1. Tournier, J.D., Calamante, F., Connelly, A.: MRtrix: Diffusion tractography in crossing fiber regions. International Journal of Imaging Systems and Technology 22(1), 53–66 (2012)
2. Fillard, P., Descoteaux, M., Goh, A., Gouttard, S., Jeurissen, B., Malcolm, J., Ramirez-Manzanares, A., Reisert, M., Sakaie, K., Tensaouti, F., Yo, T., Mangin, J.F., Poupon, C.: Quantitative evaluation of 10 tractography algorithms on a realistic diffusion mr phantom. NeuroImage 56(1), 220–234 (2011)
3. Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C.J., Wedeen, V.J., Sporns, O.: Mapping the structural core of human cerebral cortex. PLoS Biology 6(7), e159 (2008)
4. Poupon, C., Rieul, B., Kezele, I., Perrin, M., Poupon, F., Mangin, J.F.: New diffusion phantoms dedicated to the study and validation of hardi models. Magnetic Resonance in Medicine 60, 1276–1283 (2008)
5. Arsigny, V., Fillard, P., Pennec, X., Ayache, N.: Log-euclidean metrics for fast and simple calculus on diffusion tensors. Magnetic Resonance in Medicine 56(2), 411–421 (2006)
6. Descoteaux, M., Angelino, E., Fitzgibbons, S., Deriche, R.: Regularized, fast, and robust analytical q-ball imaging. Magnetic Resonance in Medicine 58(3), 497–510 (2007)
7. Tristán-Vega, A., Aja-Fernández, S.: Dwi filtering using joint information for dti and hardi. Medical Image Analysis 14(2), 205–218 (2010)
8. Kezele, I., Descoteaux, M., Poupon, C., Poupon, F., Mangin, J.F.: Spherical wavelet transform for odf sharpening. Medical Image Analysis 14(3), 332–342 (2010)
9. Tournier, J.D., Calamante, F., Connelly, A.: Robust determination of the fibre orientation distribution in diffusion mri: Non-negativity constrained super-resolved spherical deconvolution. NeuroImage 35(4), 1459–1472 (2007)