Update of the ISMRM 2015 Tractography Challenge: curated data and enhanced Tractometer scoring system

Emmanuelle Renauld[1], Antoine Théberge[1], Jean-Christophe Houde[2], Maxime Descoteaux[1]
[1]Sherbrooke Connectivity Imaging Laboratory (SCIL), Computer Sciences Department, Université de Sherbrooke
[2]Imeka Solutions Inc, Sherbrooke, QC, Canada

Introduction: The ISMRM 2015 challenge [1] was created to analyze the quality of bundles as whole entities, but the quality of individual streamlines composing them was not verified. Ground truth data contained short/long/looping/broken streamlines. The scoring system (the Tractometer [2]), was using Recobundles [3], which also contains limitations. Recobundles is influenced by the quality of the reference bundles, the order bundles are considered for scoring, and it relies on manually defined thresholds. Overall, although novel and useful at the time of the challenge, the data and scoring system led to sub-optimal bundle segmentation in some cases (see examples in Figures 1 and 2), leading to unstable overlap/overreach (OL/ORn) due to inclusion of outlier streamlines. Here, we propose an updated version: data without noisy streamlines, along with a more stable scoring system using carefully positioned anatomical ROIs. This phantom is the most widely used phantom in publications for tractography validation [4]. It is also one of the most used datasets as training / testing set in machine learning studies for tractography. In fact, it is virtually the only tractography dataset with human brain geometries offering a "ground truth". Machine learning trained on poorly curated data has poor chances of allowing good and generalizable results. We expect this new data to be an important upgrade for the tractography community.

Methods: It's important to note that the 2015 scores published publicly on the challenge website [2] contained some errors and were updated.
Updated data: Streamlines from the ground truth bundles were filtered to keep only those with length in the range 20-200mm. Streamlines presenting looping shapes or classified as outliers by the script scil_outlier_rejection.py [5] were discarded. Concerning bundles, the POPT, CST and FPT were merged together into a new bundle called corona radiata (Fig 1); the second tail of the ICP was judged non-anatomically verified and removed; streamlines from the ILF and OR were rejected based on manually drawn regions of interest (ROIs) to better separate the two bundles (Fig 2).
Alternate scoring technique: A new scoring technique was prepared, which relies on regions of interest (ROIs) (Fig 3) to segment bundles, instead of Recobundles. ROIs were created using a dilated mask of the ground truth bundles' endpoints. Manual modifications were applied by observing the variability of the initial challenge's submissions. Final ROIs are very large (Fig 3) but allow a good segmentation. Large inclusion masks were also created to exclude wrong path connections. The code also allows using other criteria, such as minimum or maximum lengths of each bundle, maximum angles, and maximum distance traveled in a given direction.

Results: New scoring of the initial 2015 submissions led, on average, to similar scores, but some bundles showed major differences. In particular, the CP was found in 22 submissions (instead of 2). Mean OL/ORn over all bundles were similar to initial scores, but some bundles scores were interestingly different, particularly the corona radiata (increase of 9% OL compared to the mean OL of the three initial bundles), the left optic radiation (increase of 9% OL), SCP (increase of 18% ORn), SLF (decrease of 10% ORn). Complete results were made available on the Tractometer website [2].

Discussion & conclusion: Although mean scores are similar over all teams/all bundles, some are now scored very differently. New suggested scoring has more chances of accurately representing the submission's data, both at the bundle level and at the streamline quality level. Updated data (no loops, no outliers, better bundle definitions) and scoring system is now offered [2, 5]. Cleaner data is important particularly for teams developing machine learning algorithms and using it as reference, as training data or as ground truth. Cleaner scoring system is important to accurately determine the best tractography algorithms, particularly when using test scores on machine learning experiments. With machine learning for tractography becoming increasingly investigated, and with the ISMRM 2015 data being one of the few "ground truth" dataset available and used a lot by that community, either as input, target or scoring reference, the quality of individual streamlines, and not only individual bundles, has become more critical. Our updated system and better GT data will help move the tractography field forward.

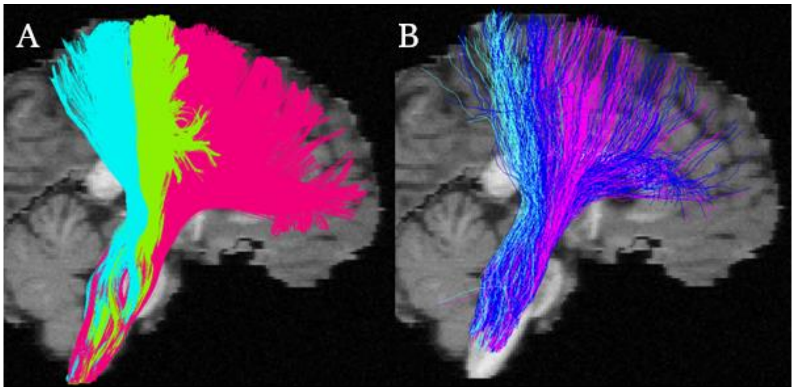

Fig 1. A. Ground truth POPT, CST, and FPT. B. Recovered bundles in team 16.4. In dark blue, additional streamlines found with ROI segmentation

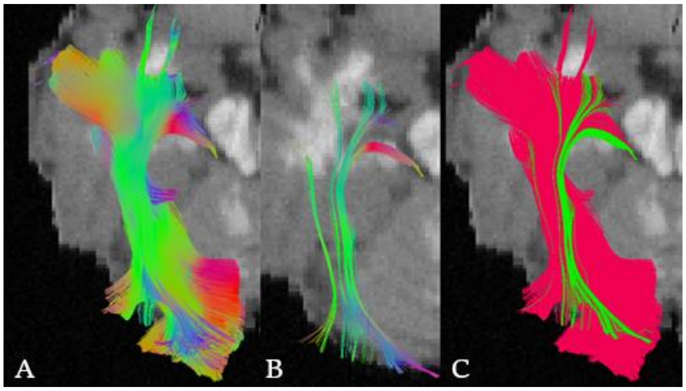

Fig 2. Left ILF (A) and left OR (B) recovered in team 1.3 with Recobundles are not adequately separated and both contain parts of the real ILF and real OR (C).

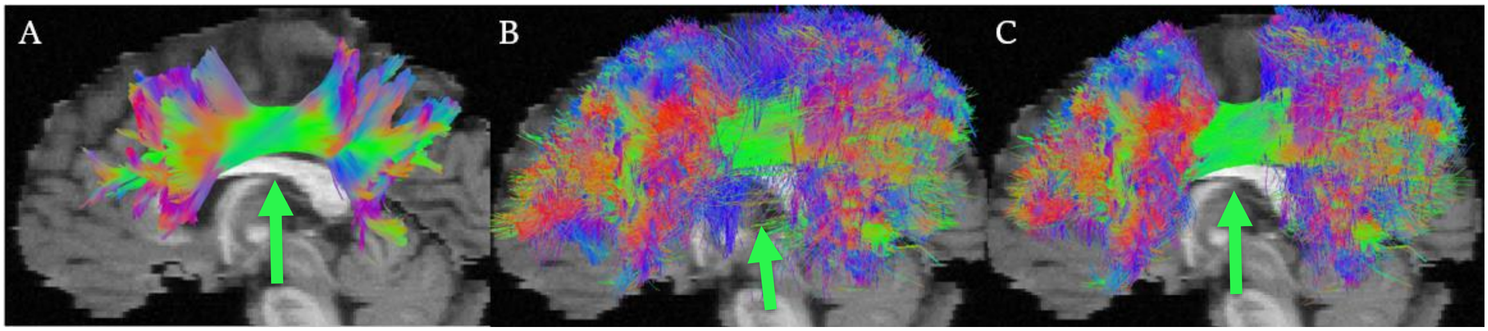

Fig 3. Ground truth SLF (A) and the union of all teams' recovered SLF, using either Recobundles (B) or ROI segmentation (C). In the newest segmentation, the shape of the bundle is better defined.

References:

[1] Maier-Hein, Klaus H., et al. Nat com 8.1 (2017): 1-13. [2] http://tractometer.org/ismrm_2015_challenge/ [3] Garyfallidis et al. Neuroimage, 2017
[4] Drobnjak, Ivana, et al. NeuroImage 245 (2021): 118704. [5] https://scilpy.readthedocs.io/