Primatologist: a modular segmentation pipeline for macaque brain morphometry[☆]

Yaël Balbastre[a,b,c], Denis Rivière[c,d], Nicolas Souedet[a,b], Clara Fischer[c,d], Anne-Sophie Hérard[a,b], Susannah Williams[a,b], Michel E. Vandenberghe[a,b], Julien Flament[b,e], Romina Aron-Badin[a,b], Philippe Hantraye[a,b,e], Jean-François Mangin[c,d], Thierry Delzescaux[a,b,f,*]

[a] *UMR9199, CNRS, CEA, Paris-Sud Univ., Univ. Paris-Saclay, Fontenay-aux-Roses, France*
[b] *MIRCen, Institut de biologie François Jacob, DRF, CEA, Fontenay-aux-Roses, France*
[c] *UNATI, NeuroSpin, Institut des sciences du vivant Frédéric Joliot, DRF, CEA, Univ. Paris-Saclay, Gif-sur-Yvette, France*
[d] *CATI Multicenter Neuroimaging Platform, France*
[e] *US27, INSERM, Fontenay-aux-Roses, France*
[f] *Sorbonne Universités, Université Pierre and Marie Curie, Paris, France*

Abstract

Because they bridge the genetic gap between rodents and humans, non-human primates (NHPs) play a major role in therapy development and evaluation for neurological disorders. However, translational research success from NHPs to patients requires an accurate phenotyping of the models. In patients, magnetic resonance imaging (MRI) combined with automated segmentation methods has offered the unique opportunity to assess *in vivo* brain morphological changes. Meanwhile, specific challenges caused by brain size and high field contrasts make existing algorithms hard to use routinely in NHPs. To tackle this issue, we propose a complete pipeline, Primatologist, for multi-region segmentation. Tissue segmentation is based on a modular statistical model that includes random field regularization, bias correction and denoising and is optimized by expectation-maximization. To deal with the broad variety of structures with different relaxing times at 7T, images are segmented into 17 anatomical classes, including subcortical regions. Pre-processing steps insure a good initialization of the parameters and thus the robustness of the pipeline. It is validated on 10 T2-weighted MRIs of healthy macaque brains. Classification scores are compared with those of a non-linear atlas registration, and the impact of each module on classification scores is thoroughly evaluated.

Keywords: MRI, Brain, Macaque, Segmentation, Expectation-Maximization, Primatologist

1. Introduction

Magnetic resonance imaging (MRI) is the modality of choice to investigate the human brain structure. It is non-invasive and offers the opportunity to image brain tissues *in vivo* at a millimetric resolution. With the advances in computer science, a number of automated methods have been developed to extract and analyze brain morphology and anatomy. When it comes to group comparisons, two strategies stand out: landmark-based and registration-based methods (Mangin et al., 2004b). Landmark-based methods consist in the segmentation in the subject's referential of well-defined anatomical regions such as the subcortical nuclei (Fischl et al., 2002; Patenaude et al., 2011; Visser et al., 2015), the hippocampus (Chupin et al., 2009) or the cortical sulci (Mangin et al., 2004a), while coordinate-based methods rely on the parametrization of the brain topography either through the registration of volumes (Ashburner, 2000) or surfaces (Fischl et al., 1999) towards a template, or through a theoretical coordinate system (Auzias et al., 2013; Régis et al., 2005; Talairach and Tournoux, 1988).

On a technical standpoint, most methods start with the classification of brain voxels into tissue classes, typically cerebrospinal fluid (CSF), gray matter (GM) and white matter (WM), sometimes into more precise anatomical regions. The first strategies proposed consisted in the segmentation of multi-contrast images based solely on the intensity histogram with supervised or unsupervised classification techniques (Vannier et al., 1988), with Gaussian mixture models (GMMs) showing the best efficacy. Liang et al. (1992) first proposed to fit such a parametric model to cerebral MRI data by expectation-maximization (EM). This framework was later extended with bias field estimation (Wells et al., 1996) and regularizing Markov random fields (MRFs) (Goldbach et al., 1991; Liang et al., 1994). To produce more robust segmentations in the case of low contrast images, Ashburner and Friston (1997) and Van Leemput et al. (1998) proposed to initialize the fitting process with tissue probability maps derived from a set of manually segmented images. However, to avoid segmenta-

[☆] ©2017. This manuscript version is made available under the CC-BY-NC-ND 4.0 license http://creativecommons.org/licenses/by-nc-nd/4.0
The final version is available at https://doi.org/10.1016/j.neuroimage.2017.09.00

[*] Corresponding author: Thierry Delzescaux (thierry.delzescaux@cea.fr)

tion bias, careful consideration should be given to the representativeness of the population within this set. Because GMMs are simple and extremely flexible, they are still the basis of some of the most frequently used MRI analysis tools such as Freesurfer (Fischl et al., 2002), FSL (Zhang et al., 2001) and the voxel-based morphometry toolbox of SPM (Ashburner and Friston, 1997, 2005). Another set of methods is based on the optimal fusion of multiple segmentations. Building on Warfield et al. (2004)'s STAPLE fusion method, Rohlfing et al. (2003) proposed a segmentation method solely based on the registration – eventually several registrations with varying parameters – of a representative collection of segmented images towards the target image space and the fusion of their labels. An accurate segmentation requires however a large number of registrations, and is thus computationally expensive. An extensive review of multi-atlas segmentation methods can be found in Iglesias and Sabuncu (2015).

Automated methods have easily found a field of application in neurodegenerative diseases. Indeed, cerebral atrophy is one of the hallmarks of these pathologies. Rosas et al. (2002) opened the way with the characterization of cortical thinning patterns in Huntington's disease (HD). Applications also include normal aging (Kochunov et al., 2005), Parkinson's disease (PD) (Lyoo et al., 2010; Pereira et al., 2012) and Alzheimer's disease (AD) (Chupin et al., 2009; Dickerson et al., 2009; Frisoni et al., 2007; Reiner et al., 2012). Supporting the use of MRI-based techniques, a strong correlation has been demonstrated between MRI-based volumetry and stereology-based neuron count in AD (Bobinski et al., 1999).

If numerous methods and software packages have been developed to study the human brain, preclinical image analysis has somehow been left aside, hampered by the small size of rodent and NHP brains and by the low availability of preclinical imaging systems. As a result, few morphological analyses of healthy NHP brains have been carried out: the majority were published later than 2012 whereas cortical thickness measures in patients were possible since 2002 (Rosas et al., 2002). Most of them took advantage of clinical pipelines such as FSL (Hopkins and Avants, 2013; Latzman et al., 2015; Liu et al., 2015; McLaren et al., 2010; Wey et al., 2013), Atropos (Hopkins and Avants, 2013), SPM (McLaren et al., 2009), FreeSurfer (Van Essen et al., 2012) or BrainVISA (Autrey et al., 2014; Bogart et al., 2014, 2012; Hopkins et al., 2010; Kochunov et al., 2010; Rogers et al., 2010) sometimes in combination with *ad hoc* NHP-specific treatments such as adapted skull stripping or intensity normalization. To ease the translational process, a pipeline dedicated to NHP MRI analysis should be made available.

Depending on the context, NHP imaging presents different challenges. Because the macaque brain is twenty times smaller than the human brain, a millimetric resolution is low and partial volume effect becomes a critical issue. Additionally, head muscles are prominent in NHPs and possess a T1 similar to that of brain tissue, hampering the skull stripping process. It is thus common to also acquire T2-weighted (T2w) images in which muscles are much more distinguishable from the brain. With high field preclinical systems, other challenges arise. At 7T, magnetization-prepared rapid gradient echo (MPRAGE) sequences usually used for T1-weighted (T1w) imaging are highly sensitive to B1 inhomogeneity (Seiger et al., 2015). Despite the development of a self-correcting T1w sequence (Marques et al., 2010; Van de Moortele et al., 2009), a common practice is to use fast spin echo T2w sequences that are less sensitive to magnetic field inhomogeneity. T2w images, however, show less contrast between gray and white structures. Additionally, different anatomical regions that would be considered WM in T1w images – cortical WM, corpus callosum and pallidum – present highly different T2w signal. The broad variety of T1w and T2w signals between regions had previously been described by Fischl et al. (2002). A generic segmentation pipeline should thus be sequence-independent and robust to magnetic field inhomogeneity, low resolution and low contrast.

As a result, we chose a statistical model, similar to that of Zhang et al. (2001), where major parameters are optimized by EM. It models intensity distribution as a mixture of Gaussians, with a MRF that integrates spatial dependency constraints. The MRF priors were obtained from the anatomical atlas published by Calabrese et al. (2015). That same atlas was registered towards the MRI space and derived into tissue priors. The bias field was estimated by low-pass filtering as in Wells et al. (1996) and Zhang et al. (2001). Additionally, we investigated the use of a denoising step, integrated to the EM scheme, that was never proposed before to the best of our knowledge. To make the whole process more robust, the EM segmentation was preceded by a first bias field estimation with BrainVISA bias correction tool (Mangin, 2000) and a new skull-stripping step robust to the problem caused by NHP head muscles was proposed. To better deal with registration errors, tissue priors were initialized from a fast 4-class GMM. The final segmentation can be used to compute volumes or to supervise PET analyses.

A complete automated pipeline was implemented in BrainVISA, a freely available image analysis software (`www.brainvisa.info`), and will be available with its next release (4.6). It performs tissue segmentation into multiple anatomical regions and is compatible with BrainVISA's sulci segmentation pipeline. It also takes advantage of BrainVISA's graphical interface, pipelining tools and parallel processing framework. It was validated on manually segmented T2w images acquired in 10 healthy macaques. Two impactful quantitative parameters that modulate the effects of the MRF and the atlas-derived priors were optimized and we took advantage of the modular structure of our pipeline to investigate the impact of its different parts (bias estimation, MRF, denoising) on the resulting segmentation.

2. Material and Methods

We tried to use common notations throughout this article with, notably, $\mathbf{y}$ naming observed intensities, $\mathbf{x}$ naming latent tissue classes and $\boldsymbol{\theta}$ naming model parameters. More details are given in Appendix A.

2.1. Statistical model

Statistical models of intensity are widely used for MRI segmentation because, in their context, Bayesian inference can be used to recover class labels based on both observed data and *a priori* knowledge. Such models have been extensively described before and "unified" models, which describe all parameters in the same statistical framework, have become increasingly popular. They present the advantage that parameters can then be optimized all together and that Bayesian priors can easily be included. Common quantitative parameters in MRI intensity models are, among other, Gaussian parameters, prior tissue probabilities and bias field.

Latent parameters can be optimized by searching for their maximum likelihood (ML) estimate, which makes sense if they are "subject-dependant", or by searching for their maximum *a posteriori* (MAP) estimate, which makes sense if their prior probability is known and thus if they are "population-dependant". If parameters are thought to be of very low-variance or if their optimization is considered intractable, they can be set to a pre-defined fixed value.

When the statistical model depends on hidden variables, class labels in our case, the EM algorithm (Dempster et al., 1977), whose general principles are recalled in Appendix B, can be used to find local ML estimates of the parameters. The algorithm alternates between approximating the posterior probability of the latent variable given parameter estimates and updating these parameter estimates.

2.1.1. Gaussian mixture model

Brain MR signal is usually described as a mixture of log-normal densities. In a theoretical case without partial volume effect, each voxel would belong to tissue class l that would generate a signal dependant on its intrinsic T1 or T2 under a log-normal law of parameters (μ_l, σ_l). The necessity to log-transform the data is justified by the exponential relationship that exists between intrinsic T1 or T2 values and the received MR signal at a given sequence parameter (inversion or echo time). From now on, we will always consider the log-transformed intensities $\mathbf{y} = \log(\mathbf{y}^{\mathrm{m}})$, where $\mathbf{y}^{\mathrm{m}}$ is the magnitude MR image.

Under this model, labels of different voxels are supposed to be independent from each other: X_i follows a discrete probability density $P(X_i = l) = A_{i,l}$ (*the prior*). The density function of Y_i is conditioned by X_i under the usual GMM and the full density function of Y_i is then:

$$p\left(Y_i = y_i\right) = \sum_{l \in \mathcal{L}} p\left(Y_i = y_i \mid X_i = l\right) P\left(X_i = l\right) \ . \qquad (1)$$

According to Bayes' rule, class probability conditioned by the observed intensity is:

$$P\left(X_i = l \mid Y_i = y_i\right) \propto g\left(y_i \mid \mu_l, \sigma_l\right) P\left(X_i = l\right) \ . \qquad (2)$$

If $\forall (i,j) \in \mathcal{I}$, $P(X_i = l) = P(X_j = l) = a_l$, the prior is said to be stationary and represents the proportion of voxels belonging to each class in the image. This assumption is often made in histogram fitting applications. In this work, we chose to incorporate prior knowledge about the possible location of the different tissue classes, as described by Van Leemput et al. (1999) and Ashburner and Friston (2005). Priors are then said to be non-stationary.

2.1.2. Markov random field

The fact that all voxels belonging to the same tissue class do not generate the exact same intensity μ_l can be explained in two non-exclusive ways: tissue type l can actually be composed of a variety of subtissues that possess a different but close T1 or T2, and noise inherent to the acquisition process can modify the intrinsic signal. The first explanation validates the finite mixture model of MR signal whereas the second hampers the segmentation process since it introduces outlier values, making voxels of tissue type l resemble another tissue type m.

The noise issue drove the use of MRFs in MR image segmentation. By modeling the tissue organization in the brain by an MRF, one can introduce prior probabilities on the existence of certain neighboring voxels organizations and bias the classification towards spatially homogeneous regions.

Within a random field, the variables X_i are not independent. However, with MRFs, dependence is restricted to connected variables with respect to a neighborhood system $\mathcal{N}$. Let $\mathcal{N} = (\mathcal{N}_i)_{i \in \mathcal{I}}$ such a neighborhood system with $\mathcal{N}_i$ the set of indices of $\mathcal{I}$ that are connected to index i:

$$P\left(x_i \mid \mathbf{x}_{\mathcal{I}\setminus\{i\}}\right) = P\left(x_i \mid \mathbf{x}_{\mathcal{N}_i}\right) \ . \qquad (3)$$

By taking advantage of the Hammersley-Clifford theorem (Hammersley and Clifford, 1971), it is common to formulate the MRF probability function in a logistic form factorized over the cliques of the lattice, *i.e.* the edges of the neighborhood. In this case:

$$P\left(x_i \mid \mathbf{x}_{\mathcal{N}_i}\right) \propto \exp\left(\beta \sum_{j \in \mathcal{N}_i} V(x_i, x_j)\right) \ , \qquad (4)$$

where β is a regularizing factor and $V(x_i{,}x_j)$ is a clique potential. A customary way to set cliques potentials, proposed by Besag (1986), is $V(l{,}m) = \delta_l{}^m$, where $\delta_l{}^m$ is the Kronecker product.

The issue with this formulation is that cliques potentials are arbitrarily set. In order to use learned clique priors, we have made a mean field approximation of the MRF conditional probability:

$$P\left(x_i \mid \mathbf{x}_{\mathcal{N}_i}\right) = \prod_{j \in \mathcal{N}_i} P(x_i \mid x_j) \ . \qquad (5)$$

We additionally state that clique prior probabilities are stationary and can thus be computed from a single reference segmentation. Because clique probabilities depend on the image resolution, we modulate reference clique priors C with a β parameter similar to that of equation (4). Additionally, to correct for anisotropy, we weighed each clique prior by its length $d(i,j)$ which is the distance between the centers of voxels i and j. Hence:

$$\forall i,j,\ P(X_i = l \mid X_j = m) = \frac{d(i,j)}{\sum_{k\in\mathcal{N}_i} d(i,k)} \left(C_{l,m}\right)^{\beta} \ . \quad (6)$$

In this framework, the posterior is expressed:

$$P(X_i = l \mid \mathbf{y}) \propto g(y_i \mid \mu_l, \sigma_l) P^{\mathrm{MRF}}(X_i = l \mid \mathbf{y}) \ , \quad (7)$$

where P^{MRF} incorporates the spatial dependency due to the MRF. The most common way to solve this problem is by iterated conditional modes (ICM), as described by Besag (1986). In this case, the MRF term is expressed:

$$P^{\mathrm{MRF}}(X_i = l \mid \mathbf{y}) = \lim_{k\to+\infty} P(X_i = l \mid \hat{\mathbf{x}}^{(k)}_{\mathcal{N}_i}) \ , \quad (8)$$

where $\hat{\mathbf{x}}^{(k)}$ is the current best estimate of $\mathbf{x}$. The initial estimate $\hat{\mathbf{x}}^{(0)}$ is often set to the MAP classification of the usual GMM:

$$\begin{aligned} \hat{x}^{(0)}_i &= \underset{l}{\mathrm{argmax}}\ P^{\mathrm{GMM}}(X_i = l \mid y_i) \\ &= \underset{l}{\mathrm{argmax}}\ g(y_i \mid \mu_l, \sigma_l) A_{i,l} \ . \end{aligned} \quad (9)$$

In the field of neuroimaging, this method is used in most MRF implementations (Fischl et al., 2002; Zhang et al., 2001). However, because it relies on a MAP estimate of class labels, it can tend to overly smooth the resulting segmentation in thin regions such as the CSF, especially when there is a strong partial volume effect. Consequently, we chose a different implementation, where class probabilities, in the MRF term, are approximated by their GMM form:

$$\begin{aligned} P^{\mathrm{MRF}}(X_i = l \mid \mathbf{y}) \propto \prod_{j\in\mathcal{N}_i} \sum_{m\in\mathcal{L}} & P(X_i = l \mid X_j = m) \\ & P^{\mathrm{GMM}}(X_j = m \mid y_j) \end{aligned} \quad (10)$$

2.1.3. Bias field

Measured MR signal is hampered by the inhomogeneity of the B0 (static) and B1 (transmission and reception) fields, especially with high intensity magnets (7T and more), that cause the presence of a slowly varying bias field in the MR image. This bias field is usually considered to be multiplicative in the measured signal space, and thus additive in the log-transformed intensity space. Consequently, the actual signal should be decomposed into its tissue component, for which the GMM is adequate, and its bias field component: $\mathbf{y} = \log(\mathbf{y}^{\mathrm{m}})$ - $\mathbf{b}$.

The bias field depends on a wide range of physical phenomena that are specific to the scanner, coil, sequence and subject and thus cannot be learned from a population of scans. Wells et al. (1996) and Van Leemput et al. (1999) have described two different ways to use the EM algorithm to estimate the bias field, based on its modelling either as a multi-dimensional, zero-centered Gaussian realization or as a grid of basis functions. We used the approach from Wells et al. (1996) since it is based on the EM algorithm and amounts to a simple low-pass filter that can be very effectively implemented as a separable, recursive Gaussian filter.

2.1.4. Noise

Noise in magnitude MR images is an additive feature that follows a Rice distribution, which can however be approximated by a Gaussian distribution in cases of high signal to noise ratio (SNR > 2) (Gudbjartsson and Patz, 1995). Here we will consider that noise mostly hampers the segmentation in tissue classes, and we will thus restrict ourselves to this latter case so that $\mathbf{y}^{\mathrm{m}} = \exp(\mathbf{y} + \mathbf{b}) + \mathbf{n}$.

While MRFs tackle the noise issue in the space of class probabilities, it could also be dealt with in the intensity space by estimating a denoised version of the MRI. Most denoising techniques consist in filtering the magnitude image with a kernel that would remove those spatially independent, zero-centered artifacts while keeping regions of true biological contrast preserved (Mohan et al., 2014). Some approaches tackle the issue in the K-space, but most end users only possess magnitude images, making those approaches unusable in the general case. Linear filters in the spatial domain show poor efficacy because they tend to smooth biological features such as frontiers between gray and white matters. The use of anisotropic filters can overcome this issue, but today's most popular methods rely on non-local means filters (Coupe et al., 2008; Manjón et al., 2008) which are very precise but bear heavy computational costs. Indeed, their complexity is $\mathcal{O}(KN^2)$, where K is the size of a neighboring window and N the number of voxels in the image, when more classic filtering techniques complexity is $\mathcal{O}(KN)$, where K is the size of the convolution kernel. Moreover, those techniques are purely based on information theory and cannot be easily included in a unified Bayesian model.

We propose to take advantage of class probability knowledge to filter the magnitude image while preserving interstructure contrast. To make the noise estimation process tractable, we suppose that the noise $\mathbf{n}$ is defined in a deterministic way when the magnitude image $\mathbf{y}^{\mathrm{m}}$ and class labels $\mathbf{x}$ are known. Indeed, inside each region, voxels are supposed to possess similar intensities ; consequently, there should exist a low-pass linear filter that yields a good approximation of the denoised intensity $\mathbf{y}^* = \mathbf{y}^{\mathrm{m}}$ - $\mathbf{n}$ if it was used only for intra-class smoothing. For each voxel i, let us call $\mathcal{N}_i$ the kernel domain and $\mathbf{w}$ its associated weights. The denoised signal can then be obtained by:

$$y^*_i = \frac{\sum_{j\in\mathcal{N}_i} \delta^{x_j}_{x_i} w_j y^{\mathrm{m}}_j}{\sum_{j\in\mathcal{N}_i} \delta^{x_j}_{x_i} w_j} \ , \quad (11)$$

where ${\delta_{x_i}}^{x_j}$ is the Kronecker product. When class probabilities are known, the expected value of $\mathbf{y}^*$ is then:

$$\mathbb{E}\left[y_i^*\right] = \sum_{l\in\mathcal{L}} P(X_i = l) \frac{\sum_{j\in\mathcal{N}_i} P(X_j = l) w_j y_j^{\mathrm{m}}}{\sum_{j\in\mathcal{N}_i} P(X_j = l) w_j} \ . \tag{12}$$

The noise image is then obtained through $\mathbf{n} = \mathbf{y}^{\mathrm{m}}$ - $\mathbf{y}^*$.

With this model, any smoothing kernel can be used and several could be studied. Here, we chose a very basic one based on a linear weighted moving average filter. Let $w_{\mathrm{lin}} = [0.25, 0.5, 0.25]$ the linear kernel, w is obtained by convolution: $w = w_x * w_y * w_z$.

2.2. Tissue and clique priors

The number of classes is an arbitrary choice, often set to 4 when non-stationary priors are used, as it is usually understood that the brain signal can be decomposed into GM, WM, CSF and background. This simplistic model is not always appropriate, especially in T2w images which generally show a broader range of intensities, with some anatomical regions such as the pallidum having a very singular signal. We thus chose to classify the brain into anatomical classes. Since many of these anatomical classes possess a similar intensity range, the use of non-stationary priors is necessary. Such priors are usually built from a population of ground truth segmentations registered into a common space (Evans et al., 1994). However, few digital atlases of the macaque brain are available compared to the human brain (see inline supplementary material 1) which led us to derive pseudo-probabilistic tissue maps from a hard label segmentation.

We chose to build those priors upon the atlas published by Calabrese et al. (2015) which consists in a high resolution *post mortem* T2w template with a parcellation into 241 regions. This labelling was provided with a hierarchy making it easy to extract a parcellation of any given complexity. Even though it was not probabilistic, we based our tissue map construction on this atlas. We first added two regions that were lacking from the atlas: CSF was obtained by a 1 mm dilation of the brain mask and corpus callosum was manually delineated from the template with Anatomist (`www.brainvisa.info`) and a Cintiq 24HD touchscreen (Wacom, Saitama, Japan). The Paxinos macaque atlas (Paxinos et al., 2008) was used as a reference. Based on the provided hierarchy, we then aggregated classes to keep only regions that made sense from the MR signal standpoint. The resulting atlas contained 18 labels (17 anatomical regions plus background) and 5 hierarchical levels, *intracranial* being the root node. The final hierarchy and the corresponding parcellation are depicted in figure 1. Clique priors can also be learned from a population of ground truth segmentations, as was done by Fischl et al. (2002) for their non-stationary clique priors. Here, we used stationary clique priors that were learned from the same ground truth segmentation. The modulating parameter β was optimized *a posteriori* based on segmentation results obtained with a validation database.

The model we presented is not fully unified beause the mapping of the atlas to the subject space was performed as a preprocessing step in a non-probabilistic way. An affine and a non-linear transformations were estimated by optimizing resemblance functions between the template and target MRI (Appendix C). Since both intensity correction and skull stripping have been shown to greatly improve the registration process (Acosta-Cabronero et al., 2008; Fein et al., 2006), our preprocessing pipeline starts with these two steps. An estimation of the bias field was performed with BrainVISA bias correction tool (Mangin, 2000), which principles are detailed in Appendix D. This first estimation was later used to initialize the EM-optimized bias field. A skull-stripping mask was then obtained with a combination of automated thresholding and morphomathematical operations described in Appendix E. Although the proposed method may seem quite basic, one should keep in mind that its sole purpose is to constrain and ease the registration process. As explained in Appendix C, the final registration step was performed on the non-stripped image, and the registered atlas was then used to perform a more robust skull-stripping: a mask obtained with a 2 voxels dilation of the registered atlas was systematically used to analyze the data in order to avoid unnecessary computations during EM optimization.

Registered atlas labels were resampled in the MRI space. Individual 3D volumes were created for each label. These volumes were then smoothed with a 3D Gaussian kernel (full width at half maximum = 3 voxels).

2.3. Mixture parameters initialization

Each class l is associated with two parameters characterizing its normal distribution: its mean μ_l and standard deviation σ_l. Because magnitude images are not quantitative, these parameters cannot be learned from a population of images, unless their intensities were previously homogenized. We have thus considered that these parameters were image-dependent and had to be optimized for each target image by EM.

Because the EM algorithm only ensures convergence towards a stationary point of the likelihood function, initialization of the Gaussian mixture is a defining step of the optimization process. If registered probability maps were used to initialize these parameters, overly smoothed tissue priors and registration errors could cause a bad evaluation of the Gaussian parameters, especially in small or thin regions such as the CSF. To make our process as robust as possible, we decided to first fit, without *a priori* and regularization, a 4-class Gaussian mixture to the log-transformed intensity histogram and use it to initialize the final 18-class mixture. The four naive classes were background, WM, GM and CSF.

First, the bias corrected MRI was log-transformed and its histogram was translated so that its maximum was matched to zero. A 4-class k-means clustering was performed with centroids initialized with values that were experimentally found to be in general close to the final

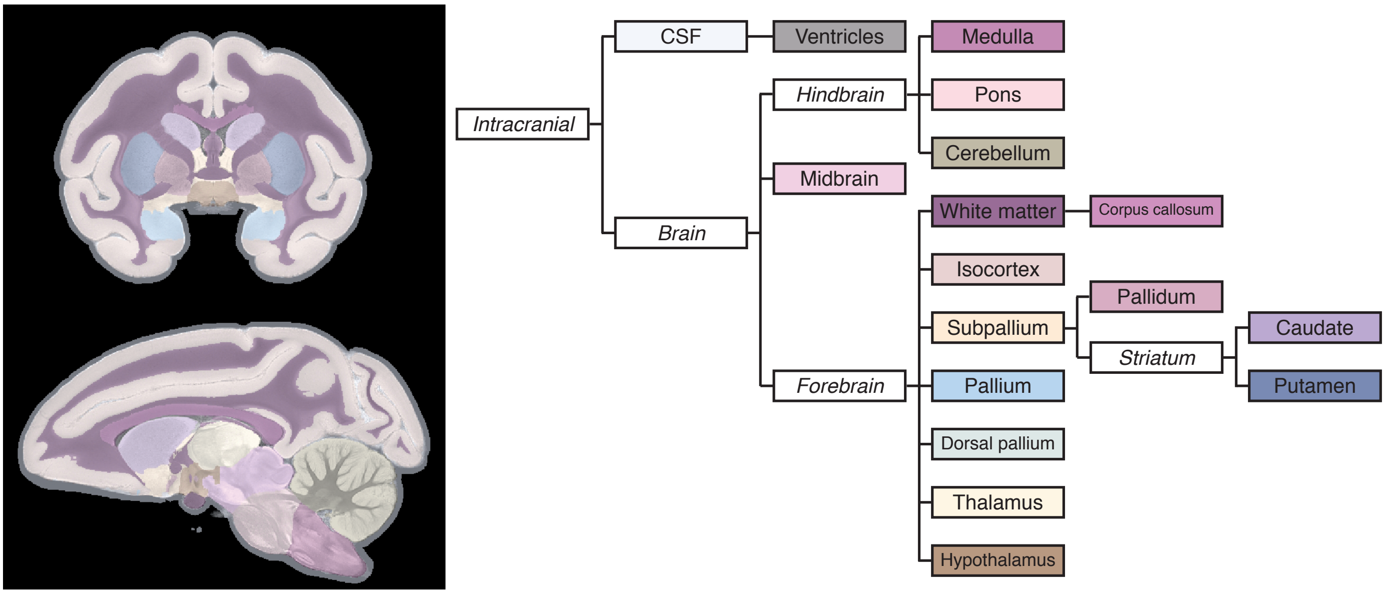


Figure 1: Simplified CIVM atlas and hierarchy. Labels are shown along with the T2w template in coronal (top left) and sagittal (bottom left) incidence. The corresponding hierarchy (right). Only labels associated with a color correspond to a class in the GMM. The other (in italics) are built by aggregation and used for multi-scale evaluation.

parameters (Figure 2). Means and standard deviations obtained from the resulting classification were used to initialize a 4-class Gaussian mixture that was then optimized by EM.

Previously, each atlas region had been classified as one of background, CSF, WM, GM or white-gray mixture (WGM). Our complete classification is shown inline supplementary table 1. Let $C(l)$ be the naive class corresponding to class l and $\boldsymbol{\theta}_4$ the parameters of the 4-class GMM. Posterior probabilities at each voxel were then computed according to:

$$P(X_i = l \mid y_i, \boldsymbol{\theta}_4) \propto p_{C(l)}(y_i \mid \boldsymbol{\theta}_4) P(X_i = l) \ , \tag{13}$$

with:

$$p_j(y \mid \boldsymbol{\theta}_4) = \alpha_j g(y \mid \mu_j, \sigma_j) \ , \tag{14}$$

and in the case where $C(l) = \mathrm{WGM}$:

$$p_j(y \mid \boldsymbol{\theta}_4) \propto \alpha_{\mathrm{W}} g_{\mathrm{W}}(y_i) + \alpha_{\mathrm{G}} g_{\mathrm{G}}(y_i) \ . \tag{15}$$

Finally, mixture parameters were initialized with their maximum likelihood estimate obtained from the above posterior probabilities.

2.4. EM optimization

Let us call $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \mathbf{b}, \mathbf{n})$ the model parameter vector. In practice, the first step consists in computing the hidden variable probabilities under the previous parameter estimate, $P(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta}^{(n)})$, that are stored in a matrix Z:

$$\forall\, i \in \mathcal{I},\ l \in \mathcal{L},\ Z_{i,l}^{(n)} = P(X_i = l \mid \mathbf{y}^{\mathrm{m}}, \boldsymbol{\theta}^{(n)}) \ . \tag{16}$$

Its elements values are obtained from the model's equations stated before, with:

$$\mathbf{y}^{(n)} = \log(\mathbf{y}^{\mathrm{m}} - \mathbf{n}^{(n)}) - \mathbf{b}^{(n)} \tag{17}$$

The second step consists in computing the parameters maximum likelihood estimates under these class probabilities, as described by equation (B.3). Because of the high dimensional nature of the parameter vector, an exact optimum cannot be obtained in practice. We will thus compute the optimal parameter for each "component" (Gaussian mixture, bias and noise) independently, with the other parameters supposed known. The optimum for the full vector will thus be approximated with acceptable precision:

1. Means and standard deviations estimation (Liang et al., 1992):

$$\mu_l^{(n+1)} = \frac{\sum_{i \in \mathcal{I}} Z_{i,l}^{(n)} y_i^{(n)}}{\sum_{i \in \mathcal{I}} Z_{i,l}^{(n)}} \ , \tag{18}$$

$$\left(\sigma_l^{(n+1)}\right)^2 = \frac{\sum_{i \in \mathcal{I}} Z_{i,l}^{(n)} (y_i^{(n)} - \mu_l^{(n+1)})^2}{\sum_{i \in \mathcal{I}} Z_{i,l}^{(n)}} \ . \tag{19}$$

2. Bias field estimation (Wells et al., 1996):
   (a) Computation of the residuals vector and the symmetric covariance matrix:

$$\bar{r}_i^{(n+1)} = \sum_{l \in \mathcal{L}} Z_{i,l}^{(n)} \frac{\log(y_i^{\mathrm{m}} - n_i^{(n)}) - \mu_l^{(n+1)}}{\left(\sigma_l^{(n+1)}\right)^2} \ , \tag{20}$$

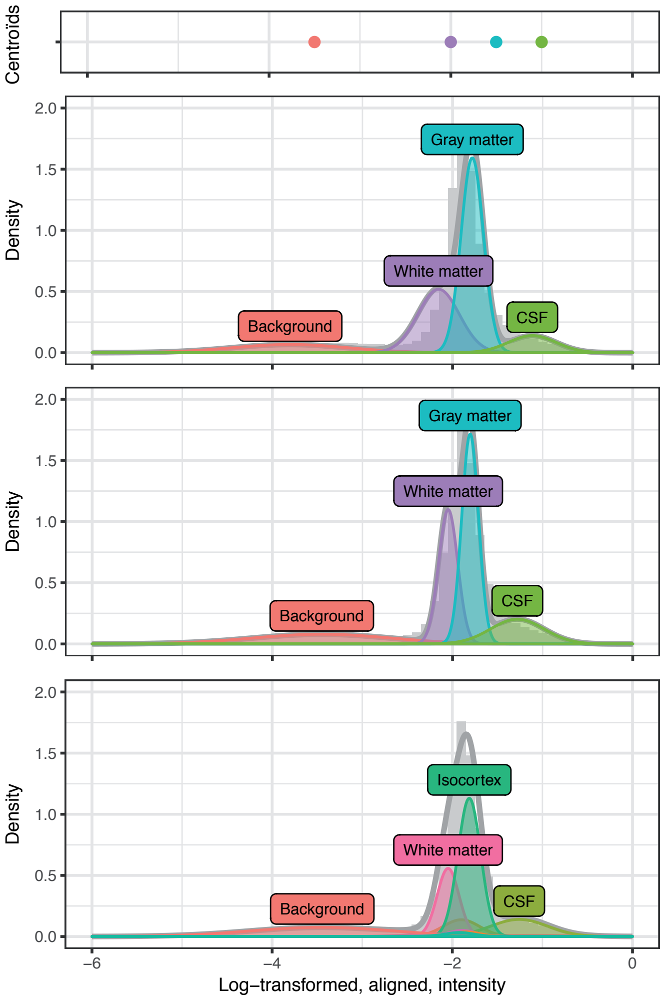


Figure 2: Initialization of the GMM parameters. First from top: centroids initialization. Second from top: 4-class GMM parameters after k-means. Third from top: 4-class GMM parameters after EM. Fourth from top: 18-class GMM parameters after conversion. In all panels, the image histogram is shown in light gray and the GMM PDF is plotted in dark gray.

$$\overline{\Sigma^{-1}}_{i,i}^{(n+1)} = \sum_{l\in\mathcal{L}} Z_{i,l}^{(n)} \left(\sigma_l^{(n+1)}\right)^{-2} \quad . \tag{21}$$

(b) Filtering:

$$b_i^{(n+1)} = \frac{[F\hat{\mathbf{r}}^{(n+1)}]_i}{[F\overline{\Sigma^{-1}}^{(n+1)\top}\mathbf{1}]_i} \quad . \tag{22}$$

3. Noise estimation:

$$n_i^{(n+1)} = y_i^{\mathrm{m}} - \sum_{l\in\mathcal{L}} Z_{i,l}^{(n)} \frac{\sum_{j\in\mathcal{N}_i} Z_{j,l}^{(n)} w_j y_j^{\mathrm{m}}}{\sum_{j\in\mathcal{N}_i} Z_{j,l}^{(n)} w_j} \quad . \tag{23}$$

The minimum number of EM iterations was set to 1 and the maximum number to 5. Optimization was stopped if the log-likelihood gain between two iterations was inferior to 0.01. Output images were the MAP classification, posterior probability maps for each class, the estimated bias field and the corrected and denoised MRI. The different components of the statistical model (bias estimation, denoising, MRF) were optional, enabling the user to deactivate each one if necessary.

2.5. Software implementation

All tools described in this paper were implemented in the BrainVISA framework, along with a dedicated pipeline and an input/output ontology, allowing fast and simple processing of user data (T1w or T2w macaque MRIs). The most demanding algorithms (registration, EM segmentation) were implemented in C++, as parts of AIMS which is BrainVISA's collection of image processing tools. Skull-stripping and fast GMM tools were implemented in Python, in the form of BrainVISA processes, with calls to low-level image processing libraries such as SciPy and pyAIMS (python bindings for the AIMS library).

2.6. Evaluation

Validation of the pipeline was performed by computing similarity scores between automated and manual segmentations performed on a set of MRIs acquired in 10 healthy macaques. This validation dataset was made publicly available concomitantly to this paper (Balbastre et al., 2017). MRIs were T2w and had a final lattice of 256 × 256 × 80 voxels of size 0.45 × 0.45 × 0.8 mm^3.

Let us recall that rather than manually segmenting all 80 coronal slices that constitute a MR volume, we decided to select a subset of sections in all three incidences. This choice was guided by the will to avoid any incidence-induced bias in the segmentation as well as lower the segmentation load. As a result, 7 coronal, 5 axial and 3 sagittal sections were selected so that all anatomical classes were found in all three incidence

To investigate the usefulness and influence of the different model parameters and components, automated segmentations characterized by varying parameter values and activation/deactivation of components were obtained and compared to the reference segmentations.

2.6.1. Hierarchical evaluation metric

Image segmentation can be seen as a classification problem, a domain where the F_1 score is a widely used metric. The F_1 score is exactly equivalent to the Dice coefficient (Dice, 1945), a more common designation in the field of image segmentation. However, this score was only defined for binary classifications, where observations can be separated between positives and negatives. In the case of multi-labels segmentation, it must be extended. We used the micro-averaged F_1 score, which can be obtained with a multi-labels definition of sets "positives" and "classified as positives". Details are provided in Balbastre et al. (2017).

In addition to the micro-F_1 score, the binary F_1 score was computed for each node of the atlas hierarchy. When optimizing parameters, decisions were made based on the micro-F_1 score.

2.6.2. Evaluated methods and statistical analysis

Because they were computed on a lattice, MRF clique priors depend on its resolution. In our case, the resolution of the atlas, in which priors were computed, differs from that of the target MRI. Consequently, we used a modulating parameter, β, similar to one classically used in Gibbs fields:

$$P(X_i = l \mid X_j = m) = \frac{(C_{l,m})^\beta}{\sum_{k\in\mathcal{N}_i} (C_{l,k})^\beta} , \qquad (24)$$

where C contains stationary clique priors computed from the reference atlas. The effect of this parameter is to make C's diagonal elements more or less influential, and thus to bias the segmentation towards more or less compact regions. We evaluated 12 different values for this parameter, ranging from 0.025 to 10, with otherwise all components activated and $\alpha = 1$.

Non-stationary priors also have a great influence on the segmentation, and registration errors may hamper it. We investigated the usefulness of a modulating parameter, α, that made those priors more or less equipossible:

$$P(X_i = l) = \alpha A_{i,l} + \frac{1-\alpha}{n} . \qquad (25)$$

We evaluated 10 different values, ranging from 0.1 to 1, with otherwise all components activated and β set to its previously optimized value.

We also investigated the influence of each component of the statistical model in the quality of the resulting segmentation. Each combination of activation-deactivation for the MRF, bias correction and denoising was tested, yielding 8 different combinations. One should keep in mind that, when the bias estimation component was deactivated, the bias field estimated in Appendix D was still used to correct the MRI. What was investigated here was the additional improvement brought by the statistical bias estimation. In order to analyze the influence of each component, a linear mixed-effects model (type III ANOVA) was used with activation of MRF, bias estimation and denoising as fixed factors and subjects as random factors.

Finally, optimized statistical segmentations were compared with those obtained by the sole non-linear atlas registration. Student's t-tests for paired measures were performed on a region-wise basis and p-values were corrected for multiple comparisons with Bonferroni's method.

All statistical analyses were performed in R (R Core Team, 2016), linear mixed-effects analysis was performed with the *nlme* package (Pinheiro et al., 2016) and graphs were generated with ggplot2 (Wickham, 2009). Data points are in general depicted as Tukey's boxplots that show the first, second (the median) and third quartiles. Upper and lower whiskers extend to the last values within the 1.5 interquartile range. Data points outside this range can be considered outliers according to Tukey's method.

3. Results

For the sake of clarity, in addition to the micro-F_1 score, only F_1 scores for regions CSF, isocortex and WM are depicted on graphs. Mean scores for all regions can be found in the associated tables. Let us recall that these are hierarchical regions ; consequentely, CSF includes both ventricular and external CSF and WM includes both cerebral WM and the corpus callosum.

3.1. Primatologist toolbox

The complete pipeline (Figure 3) is implemented in BrainVISA in the form of a toolbox that contains all individual processing tools, ready-to-use pipelines that allow easy processing of a batch of images, and database ontology that describes and organizes all input and output files. An example pipeline window is shown in figure 4.

BrainVISA is deeply intertwined with Soma-Workflow, a job distribution software that allows to speed-up the processing of image batches by running them in parallel on a multi-core workstation or on a computing cluster. For this paper, we were consequently able to process 10 images with 30 different combinations of parameters, which represent 300 pipeline runs.

Running Primatologist with a single 2.4 GHz processor and 2 GB of RAM on one of the images from our validation database lasts 1 hour and 15 minutes, with atlas registration being the most demanding step (affine transform estimation: 22 min, non-linear transform estimation: 44 min, transform application: 2 min, EM optimization: 3 min). If, instead of probabilizing a segmentation, an existing prior probability volume is used, the transform application step is more expensive (up to 16 minutes). In order to take advantage of multi-core workstations or computing clusters, some of the steps (transform application and sulci extraction) are parallelized.

All tools are shipped, along with dedicated documentation, in the Primatologist toolbox that will be made available with the next BrainVISA release (4.6). Sources of the C++ segmentation command and all python tools will be open.

3.2. Optimization of the MRF β parameter

F_1 scores obtained with different β values are depicted in inline supplementary figure 1 and complete results are given in inline supplementary table 2. A maximum common to 12 out of 22 regions (17 if we count co-maximums with a precision of 0.01) is found for $\beta = 0.25$ and most of the other regions possess a score close to their maximum at that value. Only CSF has maximum at a lower β value (0.1), but F_1 values are very similar for both parameters at 0.72 and 0.71. Interestingly, this optimum is very close to the ratio between the atlas and MRI resolutions, as:

$$\frac{0.15}{(0.45 \times 0.45 \times 0.8)^{1/3}} = 0.28 . \qquad (26)$$

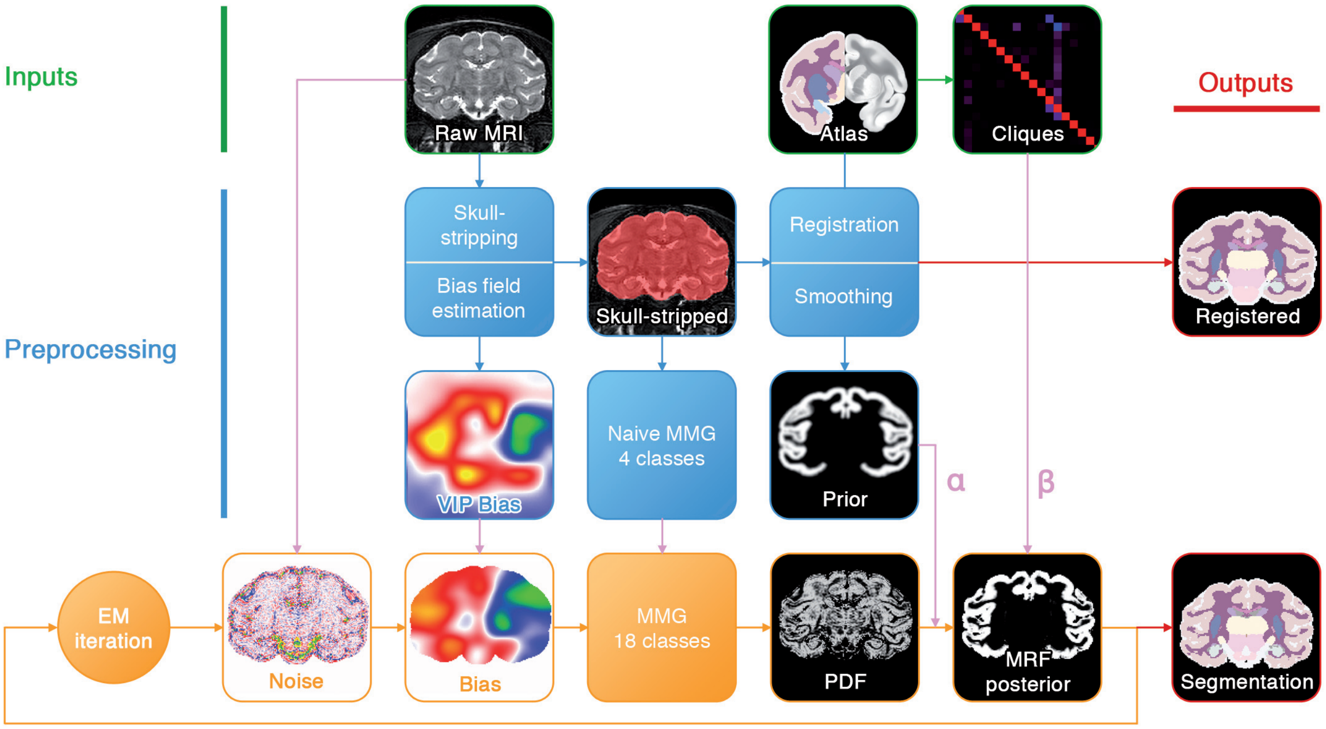


Figure 3: Segmentation workflow. Inputs are shown in the first line: only the original MRI, atlas (template and labels) and clique statistics are necessary. The second and third lines show preprocessing steps applied to initialize the statistical model. The fourth line shows the EM iterations and the resulting segmentation.

Qualitatively, most errors induced by an inadequate β value are found in the CSF for large values and in the WM for small values, as shown in inline supplementary figure 2.

For a given image resolution, micro-F_1 optimums are quite reproducible between subjects as $\beta = 0.25$ was an optimum for 8 out of 10 subjects. The other two had their optimum for $\beta = 0.1$, with almost identical micro-F_1 for both values of β ($\left|f_1^{\beta=0.1} - f_1^{\beta=0.25}\right| < 10^{-3}$).

3.3. Optimization of the prior α parameter

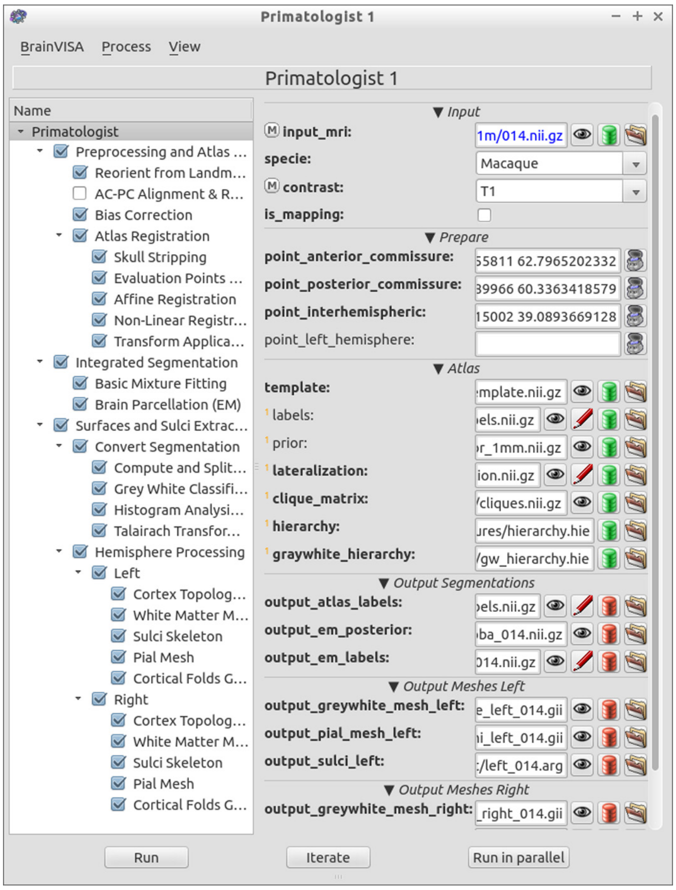
As shown in inline supplementary figure 3 and in inline supplementary table 3, no positive impact on the segmentation was brought by non-default values of the α parameter, apart from the CSF and medulla where a subtle optimum can be found for $\alpha = 0.9$.

As for the MRF parameter, optimal α are quite robust. Seven out of 10 subjects had a maximal micro-F_1 for $\alpha = 1$. Two had their optimum at $\alpha = 0.9$ and one at $\alpha = 0.7$, with, in all cases, very close micro-F_1 ($\left|f_1^{\max} - f_1^{\alpha=1}\right| < 10^{-3}$).

3.4. Evaluation of the different model components

As shown in figure 5, micro-F_1 scores vary between 0.75 and 0.8 depending on the combination, with few outliers detected. Complete results are given in inline supplementary table 4. The optimum for 8 out of 22 regions (16 with a precision of 0.01) is found with a model that includes a MRF and bias estimations but no denoising (${f_1}^{\text{micro}} = 0.8$). The use of denoising slightly improves the segmentation for regions CSF, intracranial, forebrain, hypothalamus and cerebellum. Only regions midbrain and thalamus do not reach their maximum score with MRF activated. Regions midbrain, hindbrain, hypothalamus, corpus callosum, subpallium, striatum, caudate, pons, medulla and cerebellum did not reach their maximum with statistical bias estimation activated, even though one should note that scores had a low variability in these regions.

Results of the linear mixed-effects model are summarized in table 1. The use of such a model, instead of a type I ANOVA, was made necessary by the non-independence of the observations between groups (the same 10 images are processed with different parameters). Briefly, this model can be written the same way as an ANOVA, with an additional term that captures inter-subject variability. Explanatory factors are called “fixed effects” while subject-related factors are called “random effects”, with the latter supposed drawn from a zero-centered normal law. Such a model is fitted to the data by restricted maximum likelihood optimization, contrary to linear models that are usually fit by least squares minimization. Here, all fixed-effect factors are categorical variables taking only two values, 0 or 1. Fitted coefficients for all fixed-effect factors are given in the table, along with the significance level of the test on whether or not the coefficient is non-zero (the null hypothesis is “the coefficient value is zero”).



Primatologist 1

Name

- Primatologist
  - ☑ Preprocessing and Atlas ...
    - ☑ Reorient from Landm...
    - □ AC-PC Alignment & R...
    - ☑ Bias Correction
    - ☑ Atlas Registration
      - ☑ Skull Stripping
      - ☑ Evaluation Points ...
      - ☑ Affine Registration
      - ☑ Non-Linear Registr...
      - ☑ Transform Applica...
  - ☑ Integrated Segmentation
    - ☑ Basic Mixture Fitting
    - ☑ Brain Parcellation (EM)
  - ☑ Surfaces and Sulci Extrac...
    - ☑ Convert Segmentation
      - ☑ Compute and Split...
      - ☑ Grey White Classifi...
      - ☑ Histogram Analysi...
      - ☑ Talairach Transfor...
    - ☑ Hemisphere Processing
      - ☑ Left
        - ☑ Cortex Topolog...
        - ☑ White Matter M...
        - ☑ Sulci Skeleton
        - ☑ Pial Mesh
        - ☑ Cortical Folds G...
      - ☑ Right
        - ☑ Cortex Topolog...
        - ☑ White Matter M...
        - ☑ Sulci Skeleton
        - ☑ Pial Mesh
        - ☑ Cortical Folds G...

▼ *Input*

Ⓜ input_mri: 1m/014.nii.gz
specie: Macaque
Ⓜ contrast: T1
is_mapping: □

▼ *Prepare*

point_anterior_commissure: 55811 62.7965202332
point_posterior_commissure: 39966 60.3363418579
point_interhemispheric: 15002 39.0893669128
point_left_hemisphere:

▼ *Atlas*

template: mplate.nii.gz
[1] labels: els.nii.gz
[1] prior: r_1mm.nii.gz
[1] lateralization: ion.nii.gz
[1] clique_matrix: /cliques.nii.gz
[1] hierarchy: ures/hierarchy.hie
[1] graywhite_hierarchy: /gw_hierarchy.hie

▼ *Output Segmentations*

output_atlas_labels: els.nii.gz
output_em_posterior: ba_014.nii.gz
output_em_labels: 014.nii.gz

▼ *Output Meshes Left*

output_greywhite_mesh_left: e_left_014.gii
output_pial_mesh_left: i_left_014.gii
output_sulci_left: :/left_014.arg

▼ *Output Meshes Right*

output_greywhite_mesh_right: _right_014.gii

Figure 4: A Primatologist pipeline window in BrainVISA. The filling of all filenames and parameters is automated by the databasing and ontology system.

These results can be interpreted the same way as an ANOVA: the intercept coefficient corresponds to the expected micro-F_1 score when no component is activated ; each fixed-effect coefficient represents the expected change brought to the micro-F_1 score by activation of the corresponding component when all the other factors are considered equal ; interaction coefficients represent the expected change brought by a specific co-activation of components. Here, we find a statistically significant effect of the MRF and the bias estimation on the micro-F_1 as well as a negative interaction between the MRF and denoising components. The most impactful component is the MRF which, when activated, causes an expected 0.038 improvement of the micro-F_1 score. Qualitative results are shown in figure 6.

3.5. *Comparison between registration-based and EM segmentations*

F_1 scores at all nodes of the hierarchy for both the optimized EM segmentation and the registration-based segmentation are summarized in table 2, along with paired Student's t-tests results. Registration-based scores vary from 0.29 for the total CSF to 0.93 for the intracranial region. EM scores vary from 0.57 for the hypothalamus to 0.97 for the intracranial region. EM scores are higher than registration-based scores for all nodes except for the corpus callosum (0.61 *vs.* 0.63). A significant increase is found in total CSF, ventricles, isocortex, white matter, pallium, dorsal pallium, hypothalamus, pons, medulla, and cerebellum as well as in meta-regions such as hindbrain, midbrain, forebrain, brain and intracranial region. The micro-F_1 score is also significantly improved, with a 20% increase. The most massive improvements are found in total CSF and ventricles with respective increases of 137% and 89%. Qualitative results are shown in figure 7.

Table 1: Results of the linear mixed-effects model fitting of formula `f1 ~ Bias * MRF * Denoising + (1 | Subject)`. Significance levels: 0.1 (+) 0.05 (∗) 0.01 (∗∗) 0.001 (∗∗∗).

Factor	Value	p-value	
Intercept	*0.75*		
Bias (B)	0.011	5.1×10^{-5}	∗∗∗
MRF	0.038	$< 10^{-21}$	∗∗∗
Denoising (D)	0.002	0.42	
B ×MRF	0.001	0.78	
B ×D	-0.003	0.43	
MRF ×D	-0.012	2.6×10^{-3}	∗∗
B ×MRF ×D	-0.001	0.92	

4. Discussion

The main goal of this work was to develop an automated pipeline for Macaque brain MRI segmentation and to make it freely available. The proposed pipeline is based on a state-of-the-art statistical model, enabling sequence-independent segmentation and bias field estimation. Furthermore, a novel noise-estimation method is proposed and evaluated. Additionally, the use of cliques statistics, which is non-standard in the literature, makes the method more automatic and less reliant on an arbitrarily set parameter. This pipeline is implemented in BrainVISA, relies on its databasing system, and can be easily used thanks to its graphical interface, integrated viewers and comprehensive documentation.

Challenges arised from the low resolution of images, caused by macaques smaller brains, from the large variability of MR signals and from the substantial magnitude of the bias fields, both caused by the use of a 7T magnet. Images resolution mainly impacted the implementation of the MRF, making us diverge from common ICM optimization methods. Inter-region signal variability made us choose a segmentation into anatomical regions rather than into tissue classes (GM, WM, CSF) as more commonly done. However, no region probability maps of the macaque brain are freely available, and the absence of whole brain reference segmentation databases forbade us to build our own. Consequently, our pipeline relies on the registration and probabilization of a hard-label atlas. This incidentally forced us to use non-linear registration, which bears higher computational costs than affine registration.

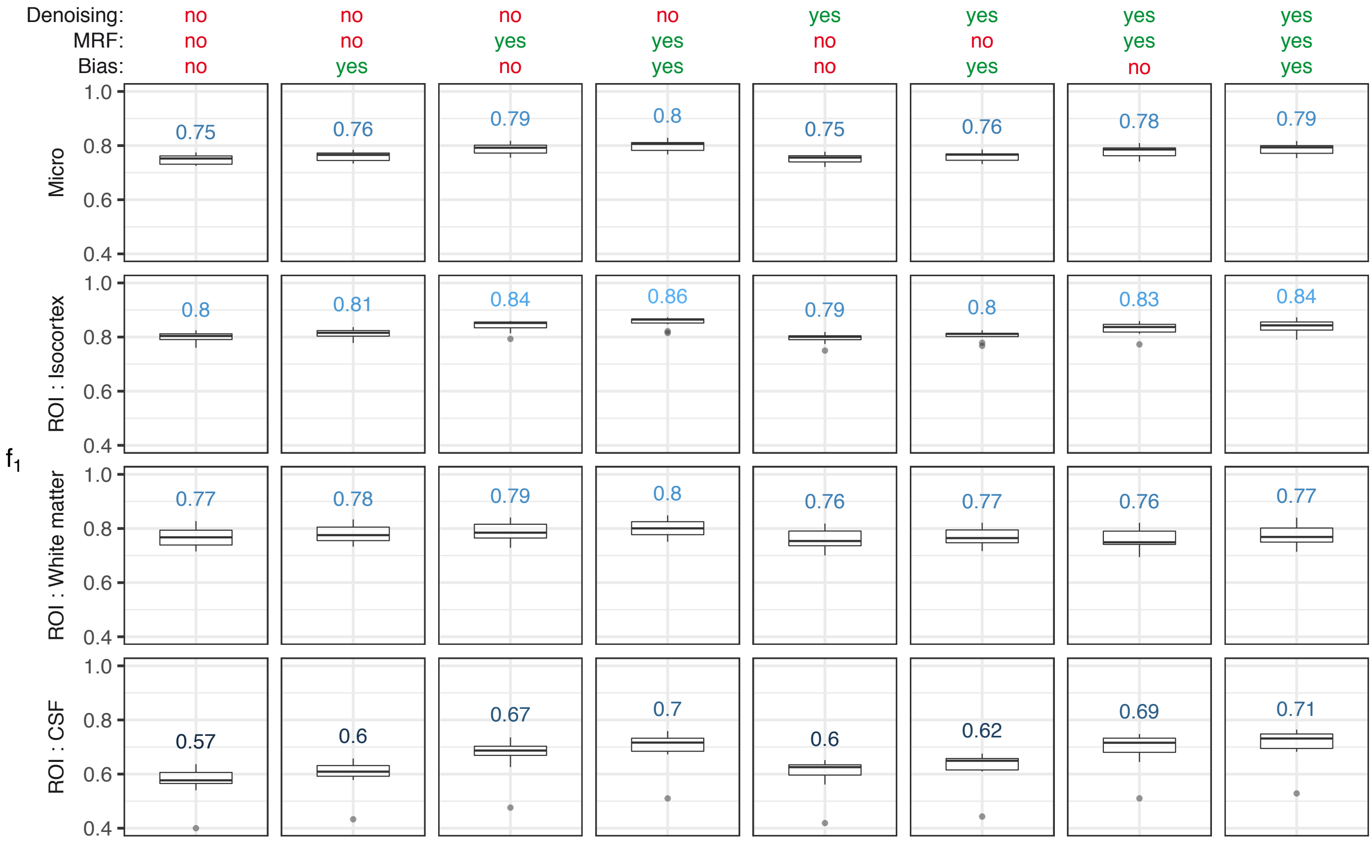


Figure 5: Effect of the activation of the different model components on the similarity score. For each tested combination, a Tukey's boxplot represents the different quartiles of the F_1 score for regions CSF, isocortex and white matter as well as those of the micro-F_1 score. The mean score is also indicated in blue.

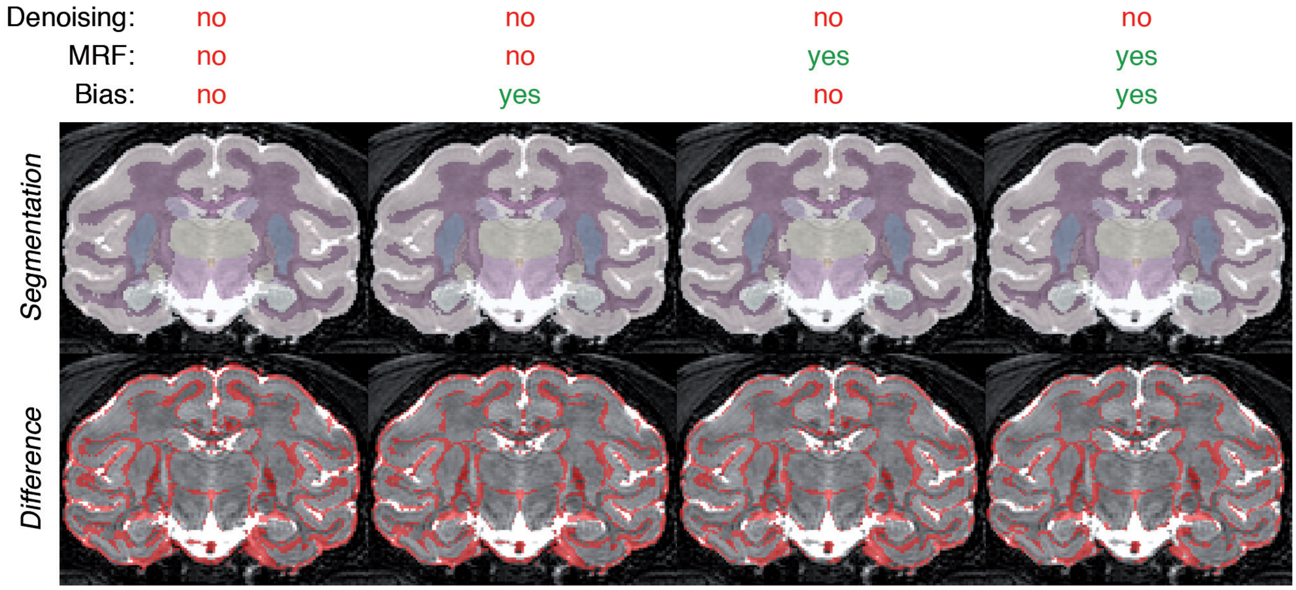


Figure 6: Resulting segmentation of an arbitrarily chosen subject for different activation combinations of MRF and bias estimation modules. The first row shows the segmentation while the second one shows in red voxels which classification differ from the manual segmentation of reference.

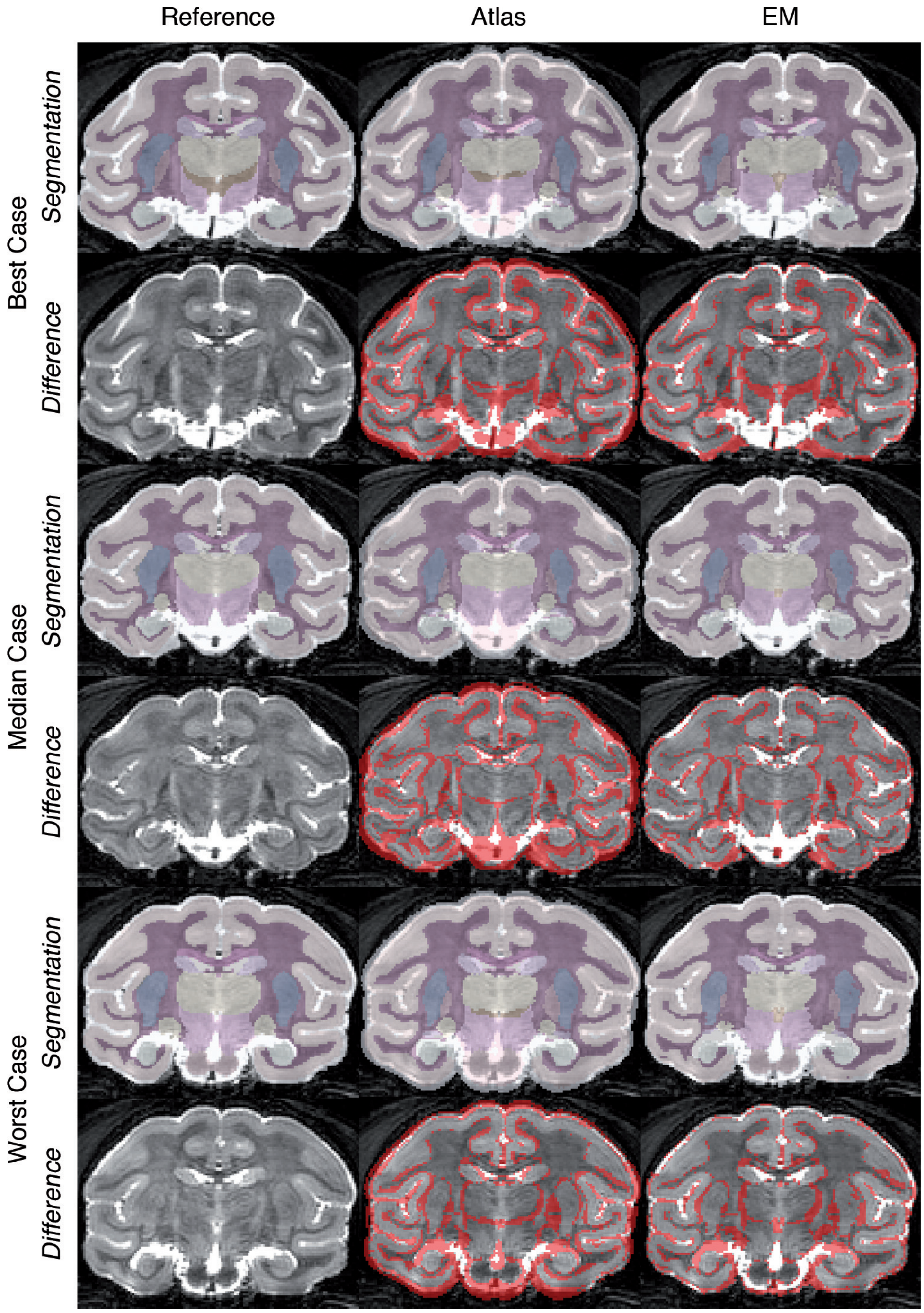


Figure 7: Resulting segmentation of subjects with the best, median and worst micro-F_1 scores obtained with the EM segmentation (right), compared with manual references (left) and segmentations obtained through a simple non-linear registration of the atlas (middle). Odd rows show the segmentation while the even rows show in red voxels which classification differ from the manual segmentation of reference.

Table 2: Difference in mean F_1 score between registration-based and EM segmentations. A paired Student's t-test was performed for each region as well as for the micro-F_1 score. Significance levels after Bonferroni correction: 0.004 (+) 0.002 ($*$) 0.0004 ($**$) 0.00004 ($***$).

ROI	${f_1}^{\mathrm{EM}}$		${f_1}^{\mathrm{atlas}}$	
${F_1}^{\mathrm{micro}}$	0.80	>	0.66	***
Intracranial	0.97	>	0.93	***
CSF	0.71	>	0.30	***
Ventricles	0.71	>	0.37	***
Brain	0.95	>	0.90	***
Forebrain	0.95	>	0.90	***
Midbrain	0.79	>	0.76	*
Hindbrain	0.92	>	0.87	***
Isocortex	0.84	>	0.75	***
Pallium	0.70	>	0.62	**
Dorsal Pallium	0.76	>	0.70	**
Thalamus	0.81	>	0.79	
Hypothalamus	0.54	>	0.47	*
White matter	0.77	>	0.70	***
Corpus callosum	0.61	<	0.63	
Subpallium	0.80	>	0.78	
Pallidum	0.70	>	0.68	
Striatum	0.80	>	0.78	
Caudate	0.74	>	0.72	
Putamen	0.84	>	0.83	
Pons	0.76	>	0.67	***
Medulla	0.76	>	0.72	**
Cerebellum	0.92	>	0.88	***

Rather than using null or random parameter initialization, we took advantage of non-integrative methods to approximate the bias field and mixture parameters. This allowed us to initialize these parameters close to the expected optimum, saving the EM algorithm from being stuck in spurious local optimums.

4.1. Quality of the segmentation

Brain morphology is much less variable in macaques than in humans. Consequently, we could expect a straightforward non-linear atlas registration to yield good segmentation results. This approach was incidentally used by Knickmeyer et al. (2010), Liu et al. (2015) and Scott et al. (2015) to segment cortical lobes and subcortical structures in MRIs of the developing macaque brain, following a procedure described by Styner et al. (2007). It should however be noted that, in all cases, when it comes to segmenting the isocortex, gray and white matters were previously separated with an EM approach.

However, our results show that, for healthy subjects, our approach yields better results than atlas registration in almost all regions, with very significant differences found in the CSF, white matter and isocortex, regions that are the most variable. Few statistically significant differences were found for subcortical regions, though different results would be expected when it comes to pathological cases, such as models of striatal atrophy, as we have described before (Balbastre et al., 2015). Significant improvements due to our approach were also found in macroscopic regions such as the whole brain, forebrain and hindbrain. The most striking differences were found in total CSF and ventricles which are thin and variable regions that are very difficult to correctly segment through registration.

These results are obtained on 2D reference sections, which is not standard. However, the F_1 score is not associated with any neighborhood structure and can be efficiently approximated from a dense sampling of points in the image. However, the fact that our sampling is performed sections-wise, and not completely randomly, might bias the estimation because not all regions are equally sampled. On the other hand, the fact that reference sections were manually segmented in different incidences reduces incidence-induced bias. In our case, the less represented class in the reference set is the hypothalamus (406 voxels ± 135). However, if the best automated segmentation is taken as an estimate of the total number of voxels, it corresponds to one of the highest sampling rates with 24% of all hypothalamus voxels that were manually segmented. Mean sampling rates vary between 15%, for the subpallium, and 25%, for the pallium.

4.2. Role of the Markov Random Field

Analysis of the influence of the different model components shows that the use of a MRF greatly improves the segmentation. Even the CSF gets a F_1 score close to its maximum with the MRF activated, even though its thinness could lead one to predict a loss of sensitivity due to the MRF in this structure.

Magnitude of clique prior probabilities is intrinsically linked to the image resolution and the size of structures. To understand this, let us focus on a binary volume containing a square object of dimensions $2n \times 2n$. The probability for a pixel to be black when one of its neighbors is white, $P(X_j = 0 \mid X_i = 1)$, is expressed as:

$$\frac{24 + 48\ (n-1) + 24\ (n-1)^2}{48 + 144\ (n-1) + 120\ (n-1)^2 + 48\ (n-1)^3}\ . \tag{27}$$

The evolution of $P(X_j = 0 \mid X_i = 1)$ and $P(X_j = 1 \mid X_i = 1)$ is depicted accordingly in figure 8. The square is supposed of constant size and n is then directly linked to the lattice resolution. One can then understand that the more resolved the lattice, the bigger the diagonal terms $(P(X_j = k \mid X_i = k))$. When stationary clique priors are learned from a reference segmentation, our study hints towards a linear connection between the modulating β parameter of the random field and the resolution ratio between reference and target images. However, this link was not thoroughly validated and could be the subject of a future investigation.

This also shows, as explained before by Morris et al. (1996), that clique priors are intrinsically linked to the size of regions through both their volume and surface (area and perimeter in 2D), even though differences diminish with

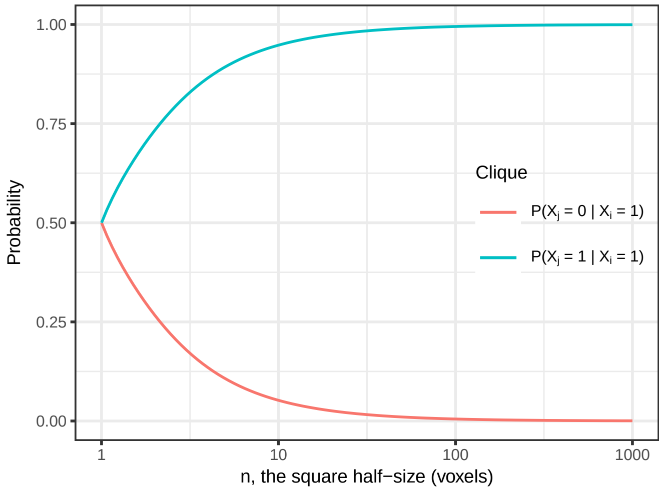


Figure 8: Evolution of the conditional probabilities of cliques (0 | 1) and (1 | 1) in the case of a square of increasing resolution n. The x axis is plotted on a logarithmic scale.

higher resolutions. The common way of setting identical potentials for all regions, through Potts model for example (Avants et al., 2011; Besag, 1986), might be imprecise in low resolution images.

Two potential improvements of our MRF model are possible:

1. The use of non-stationary clique priors learned from a set of reference segmentations, as described in Fischl et al. (2002), would allow for more precise prior information to be injected in the model. This requires, however, a large number of reference segmentations that are not yet available in our database.
2. Stationary clique priors could be set more freely if they were part of the statistical model parameters that are optimized by EM, as in Rajapakse and Kruggel (1998). They would still be initialized from the reference segmentation but could, this way, be refined to adapt better to the specificity of the target image. The drawback is that by enlarging the parameter space, we would increase the chance to get stuck in local likelihood maximums or to diverge greatly from the optimal solution.

4.3. Non-stationary prior modulation

As shown by the optimization of the α parameter, the method we chose to modulate the non-stationary priors seems to be poorly adequate. The issue raised by this formula is that all voxels are modulated the same way, without any consideration for their location, enabling regions to appear in voxels quite distant from their atlas location. Even though this was the intended behavior, in order to compensate the lack of freedom caused by a hard atlas compared to a probabilistic one, objective results show that drawbacks outweigh benefits. The size of the smoothing kernel appears as a better candidate for optimization, and this should be investigated in a near future.

Let us note that part of the prior could be freed and optimized by EM. This parametrization was described by Ashburner and Friston (2005) and is also implemented in Atropos (Avants et al., 2011). The prior is written:

$$P(X_i = l) \propto r_l A_{i,l} \ , \tag{28}$$

where r_l is the ratio between the global proportion of voxels of class l in the image and that same proportion in the atlas. This way, the segmentation is not biased by the proportion of voxels from each class in the atlas. The r_l can then be optimized by EM. Let us call ${\alpha_l}^{(n)}$ the estimated proportion of voxels of class l at iteration n and ${\alpha_l}^{\text{atlas}}$ the proportion of voxels of class l in the atlas:

$$\alpha_l^{\text{atlas}} \propto \sum_{i \in \mathcal{I}} A_{i,l} \ , \tag{29}$$

$$\alpha_l^{(n+1)} \propto \sum_{i \in \mathcal{I}} P(X_i = l \mid \mathbf{y}, \boldsymbol{\theta}^{(n)}) \ , \tag{30}$$

$$r_l^{(n+1)} = \frac{\alpha_l^{(n+1)}}{\alpha_l^{\text{atlas}}} \tag{31}$$

4.4. Improvement brought by a statistical estimate of the bias field

Several bias field estimation methods that do not rely on the estimation of class probabilities, being based either on information theory or non-explicit distribution modeling, have shown to perform extremely well (Mangin, 2000; Sled et al., 1998; Tustison et al., 2010). In our study, using images acquired at 7T, we showed that the inclusion of a bias estimation module in the EM optimization, similar to those described by Wells et al. (1996), Van Leemput et al. (1999) and Zhang et al. (2001), improved the resulting classification compared to an independent bias correction. Indeed, the knowledge of class probabilities allows a better estimate of the intensity distribution in the corrected image and thus a better estimate of the bias field. The maximum estimated bias field value in each image ranged from 1.12 to 1.48 (mean ± SD: 1.33 ± 0.11) and an illustrative example is provided in figure 9. No correlation was found between the maximum bias field value and the micro F_1 score (Pearson's product-moment correlation, $r = 0.17$, $p = 0.6$). Nonetheless, non-statistical estimates are still a good way to initialize the bias field before EM optimization.

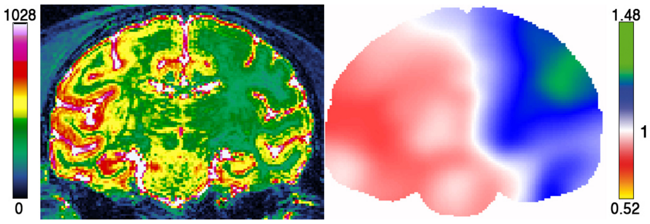


Figure 9: Color map of the most biased MRI of the validation database (left), along with the estimated bias field (right).

4.5. Statistical denoising

No statistical improvement of the segmentations was brought by the use of denoising, even when the MRF was deactivated. One should note, however, that magnetic noise was not overly present in our validation set. A positive impact was nonetheless expected, and its absence can be explained by our choice of kernel, which was very basic, as well as by the absence of posterior control on spatial dependency in the estimated noise image.

The analysis of the noise image in one representative subject (Figure 10) shows that the estimated noise distribution seems zero-centered, as expected, but suffers from an over-representation of high values, especially positive ones, which positively skews the distribution ($\gamma = 0.12$). A qualitative inspection of the noise image shows that extreme noise values seem to be located in the CSF which can be caused either by a non fully additive noise or by a bad estimation in thin regions.

Note also that, in this paper, we used a deterministic definition of noise. The presence of outliers could be reduced by searching for a MAP estimate in a Bayesian setting, where noise is supposed to be a realisation of a zero-centered Gaussian of variance σ^2. In this case, a ML estimate of σ could even be optimized by EM along with the other model parameters.

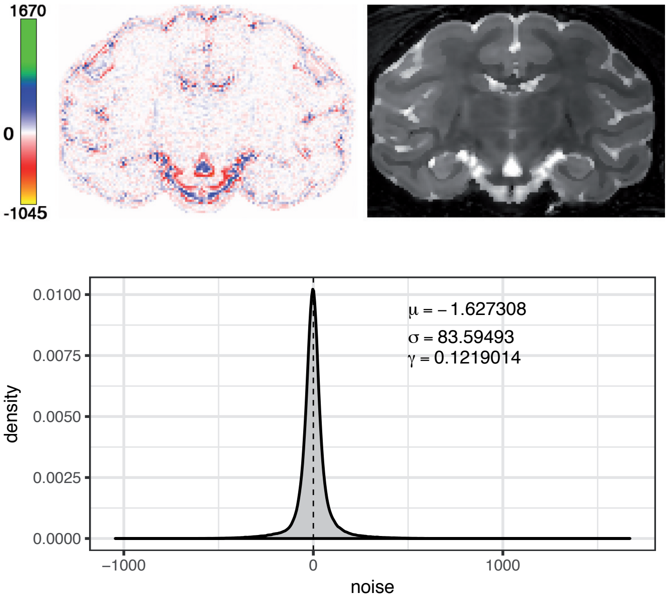


Figure 10: Color map of the estimated noise in a representative subject (top left). Corresponding noise-corrected MRI (top right). Density histogram of the estimated noise, with the distribution mean value μ, standard deviation σ and skewness γ (bottom).

4.6. Potential applications

This pipeline is an entry point for any morphometry analysis of the brain. Such analyses are particularly employed for studying brain development, brain aging and neurodegenerative diseases. In the case of brain development studies, the macaque is an extremely interesting model since ethical and practical constraints – movement artifacts, experimentation time, *etc.* – make difficult the constitution of healthy children cohorts. Contrary to children, young macaques can be anesthetized and have been a subject of study for many years (Knickmeyer et al., 2010; Liu et al., 2015; Malkova et al., 2006; Scott et al., 2015), whereas works to constitute the first human cohort started later (Almli et al., 2007). Macaques have also been used as models for studying neurodegenerative diseases such as HD (Ferrante et al., 1993) and PD (Porras et al., 2012) and to evaluate possible therapies with a translational aim (Jarraya et al., 2009). In this context, morphometry measures can be used as biomarkers of disease progression. In AD, NHPs are even more crucial as rodents need to be genetically humanized to mimic the pathology. Morphometry features are already widely used to study this disease in humans and they would be vital in macaques if emerging models (Forny-Germano et al., 2014) were to be used to evaluate promising therapies.

Another major application would be its use to supervise PET analyses and extract region-wise kinetics in group studies (Ballanger et al., 2013). The use of multi-compartments models, which need anatomical priors, have shown to more accurately quantify ligand binding (Ginovart et al., 2001). Moreover, automated ROIs are not expert-dependent and are less time-consuming than manual delineations.

4.7. Limitations and future directions

The most striking limitation of our approach is the lack of a true probabilistic atlas of the macaque brain. Such an atlas would capture accurately the inter-subject variability and would be obviously more informative than the smoothed atlas we currently use. Consequently, the development of macaque brain segmentation database would be a welcome project in our field. The *UNC-Wisconsin Rhesus Neurodevelopment Database* (Young et al., 2017) is an important first step in this direction. It provides, under a permissive free software license, longitudinal MRI scans acquired in dozens of macaques from age 0 to 3 years. They could be the building base for expert segmentations that would allow the construction of age-appropriate probabilistic atlases usable as priors in our software and others. The existence of such a database would also allow the use of non-stationary clique priors, which effect could be compared with that of stationary clique priors.

Let us also note that, besides being non probabilistic, the atlas we used was also based on images acquired *post mortem*, after extraction from the skull. It means that, in addition to global deformations inherent to the elastic nature of the tissues, no CSF is present and sulci and ventricles are partially closed (the banks of opposing gyri touch each other). The CSF prior is thus poorly informative and can even hamper the segmentation. Luckily, contrast between GM and CSF is one of the strongest in the brain and the GMM was generally able to compensate this poor prior. This problem could also be solved by using a truly probabilistic atlas built from images acquired *in vivo*. There are also differences in species (rhesus *vs.* cynomolgus) which add some slight imprecision to the segmentation. Once again, building more appropriate templates is the key.

When it comes to probabilistic atlases, two strategies stand out, depending on the use of affine or non-linear registration. The second one allows for more precise and strong priors but bears additional computational costs. If the additional precision it brings to the segmentation outweighs the computational time, it becomes beneficial. Another strategy relies on the construction of subject-specific prior maps from the registration of multiple atlases towards the subject space, rather than using a pre-computed probabilistic atlas. The computational cost is even higher, but generative multi-atlas models prove to bring additional precision to the segmentation (Iglesias et al., 2013).

Our scheme could also be extended to multi-channel images, since GMMs easily apply to multivariate data. Multi-contrast MRIs are extremely useful to segment subcortical structures (Visser et al., 2015). In NHPs, because of their head muscles that can be better separated in T2w images, such images would additionally help skull-stripping the brain.

Finally, our pipeline should be applied to other species by taking advantage of freely available atlases. In particular, the CIVM published atlases of the Mouse brain, which incidentally includes anatomical probability maps (Johnson et al., 2010), and of the Rat brain (Papp et al., 2014) that could be used with our pipeline. The challenge would be to adequately select GMM classes based on the contrasts that are distinguishable in MRIs. In particular, it could be necessary to use separate classes for gray and white regions of the hippocampus, since this structure takes a larger proportion of the brain in rodents than in primates. Obviously, because of their small brain size, partial volume issues and artifacts due to high fields would also be present. On the bright side, their cerebral morphology is much less complex and the cortex is not folded, leading to more compact regions that are easier to accurately segment than the thin gyri of the macaque brain.

5. Conclusion

In this paper, we present the first pipeline dedicated to anatomical segmentation of macaque brain MRIs. It allows computing of 17-region parcellations out of unstripped, uncorrected and unaligned T1w and T2w images. The segmentation process is based on a statistical model of intensity and is modular: the user may activate on demand spatial dependency priors (MRF), bias field estimation and noise estimation. Segmentations were validated on a database of manually segmented images acquired in healthy macaques. The combination of MRF and bias field estimation yielded the best results. Our software was implemented as an open source BrainVISA toolbox which will be made available with BrainVISA 4.6. This paper also provides results on statistical MRI models that are not restricted to macaque brain images. In particular, we show how the use of learned clique priors improve such models compared to Potts models. We also accurately describe how the EM bias field estimation improves segmentations compared to independent estimations. Thanks to the free availability of this pipeline, researchers working with NHPs will be able to extract new morphological features that could play a major role in the study of neurodegenerative disorders at the preclinical level.

Acknowledgments

This research was funded by a rolling grant from the CEA's Molecular Imaging Research Center (MIRCen) and was partly supported by the Translational Research Infrastructure for Biotherapies in Neuroscience (NeurATRIS, "Investissement d'Avenir", ANR-11-INBS-0011). We would like to thank Duke Center for In Vivo Microscopy (CIVM), a U.S.A. National Institutes of Health Biomedical Technology Resource Center (P41 EB015897), for kindly granting us permission to use their atlas in our work and software.

Appendix A. Mathematical notations

We use uppercase letters (X) for random variables, lowercase letters (x) for their realisations and bold letters ($\mathbf{x}$) for vectors. Uppercase P will denote probability density functions (PDFs) of discrete random variables while lowercase p denote PDFs of continuous random variables. We will generally write intensities y, labels x and model parameters θ. When no ambiguity exist, the expression $P(A = a \mid B = b)$ is abbreviated $P(a \mid b)$. When defining discrete PDFs, we may avoid writing the normalization term and will use the "proportional to" sign ($\propto$) instead. Additionally, because, in the case of MRI segmentation, intensities ($\mathbf{y}$) take value in a continuous space and hidden variables ($\mathbf{x}$) take value in a discrete space, we stick to this case in the all formulations.

Let the image domain $\mathcal{I} = [1, n]$ a set of indices with n the number of voxels. The multivariate label variable $\mathbf{X} = (X_i \ ; \ i \in \mathcal{I})$ takes values in $\mathcal{X} = \mathcal{L}^n$ with $\mathcal{L} = [1, N]$. The multivariate intensity variable $\mathbf{Y} = (Y_i \ ; \ i \in \mathcal{I})$ takes value in $\mathbf{R}^n$.

Appendix B. Expectation-Maximization

The EM algorithm is an iterative method to find local maximums of the likelihood function of a statistical model given a set of realizations $\mathbf{y}$, *i.e.*, estimate its maximum likelihood parameter $\hat{\boldsymbol{\theta}}$. To ease the analysis, the log-likelihood is more often used:

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta}\ ;\ \mathbf{y}) \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log p(\mathbf{y} \mid \boldsymbol{\theta})\ .\end{aligned} \tag{B.1}$$

When hidden variables play a role, it is easier to have them appear in the likelihood formulation:

$$\begin{aligned}p(\mathbf{y} \mid \boldsymbol{\theta}) &= \sum_{\mathbf{x}\in\mathcal{X}} p(\mathbf{y},\mathbf{x} \mid \boldsymbol{\theta}) \\ &= \sum_{\mathbf{x}\in\mathcal{X}} p(\mathbf{y} \mid \mathbf{x},\boldsymbol{\theta}) P(\mathbf{x} \mid \boldsymbol{\theta})\ .\end{aligned} \tag{B.2}$$

The EM algorithm states that, knowing an estimate $\boldsymbol{\theta}^{(n)}$ of $\hat{\boldsymbol{\theta}}$, a better estimate $\boldsymbol{\theta}^{(n+1)}$ is found by maximizing the following function Q over $\boldsymbol{\theta}$:

$$\begin{aligned}Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(n)}\right) &= \mathbb{E}\left[\log p(\mathbf{y},\mathbf{x} \mid \boldsymbol{\theta}) \mid \mathbf{y},\boldsymbol{\theta}^{(n)}\right] \\ &= \sum_{\mathbf{x}\in\mathcal{X}} P(\mathbf{x} \mid \mathbf{y},\boldsymbol{\theta}^{(n)}) \log p(\mathbf{y},\mathbf{x} \mid \boldsymbol{\theta})\ .\end{aligned} \tag{B.3}$$

This scheme insures that (1) $L\left(\boldsymbol{\theta}^{(n+1)}\right) \geqslant L\left(\boldsymbol{\theta}^{(n)}\right)$ and (2) the sequence $(\boldsymbol{\theta}^{(n)})$ converges at least towards a stationary point of L (McLachlan and Krishnan, 2008). For a short and clear overview of the EM algorithm, see Borman (2004).

Appendix C. Registration

For the affine registration step, we used our implementation of a strategy inspired by Thevenaz and Unser (1997). Briefly, mutual information (MI) between a reference (the target MRI) and a moving (the atlas template) image was maximized at several pyramid levels with a relaxed version of Newton's optimization scheme that included backtracking and fronttracking steps. MI gradient and Hessian were computed as in Thevenaz and Unser (1997). To successively free transform parameters, a translation was first computed, initializing a rigid transform that initialized an affine transform. Different resolution levels were computed by transforming the moving image into a L2 spline pyramid (Unser et al., 1993). To speed up the process, 20000 points were randomly selected in the reference brain mask for MI evaluation. This number does not depend on the image resolution: because of the low number of optimized parameters (12), additional points would only marginally improve the estimation. Sixty four bins were used to compute the joint histogram and optimization was stopped when the difference between consecutive MI values was inferior to 10^{-5}. The damping parameter was multiplied (resp. divided) by 2 for the fronttracking (resp. backtracking) operations. Translation was initialized by aligning intensity-weighted gravity centers.

Elastic registration was performed with our implementation of Mattes' Free Form Deformation (Mattes et al., 2003), which we have previously presented in the context of mouse atlas registration (Lebenberg et al., 2010). Spline grids of size 4, 6, 8 and 10 were successively optimized. Sixty four bins were used to compute the joint histogram and optimization was stopped when the relative MI gain between two successive iterations was inferior to 5 × 10^{-3}. In order to correct for eventual errors induced by a potential bad skull-stripping, the resulting 10 × 10 × 10 transform was used to initialize another optimization, this time with the non-stripped MR image as a target.

Appendix D. Initial bias correction

A multiplicative bias field was modeled with a grid of cubic B-splines whose coefficients were set by optimizing an objective function in a multi-resolution scheme by simulated annealing. The objective function contains energy terms that penalize entropy in the corrected image, spatial irregularity of the bias field and mean intensity difference between the original and corrected images. In the most recent version of this algorithm, the robustness of the process was increased by extracting WM ridge points characterized as voxels of high curvature in the intensity space (positive curvature in T1w images, negative curvature in T2w images). An energy term penalizing the entropy of the corrected image in the ridge was added to the objective function. Let $\mathbf{y}$ be the original image, $\mathbf{w}$ the WM points, $\mathbf{b}$ the multiplicative bias field and $\mathbf{c}$ its coefficients, the objective function was expressed as :

$$\begin{aligned}f(\mathbf{c}) = \;& K_e f_{\text{entropy}}(\mathbf{by}) \\ & + K_r f_{\text{regul}}(\mathbf{b}) \\ & + K_o f_{\text{offset}}(\mathbf{by},\mathbf{y}) \\ & + K_w f_{\text{entropy}}(\mathbf{by}(\mathbf{w}))\ ,\end{aligned} \tag{D.1}$$

with

$$f_{\text{regul}}(\mathbf{b}) = \sum_i \sum_{j\in\mathcal{N}_i} \log^2\left(\frac{c_i}{c_j}\right)\ , \tag{D.2}$$

$$f_{\text{offset}}(\mathbf{x},\mathbf{y}) = (\bar{\mathbf{x}} - \bar{\mathbf{y}})^2\ , \tag{D.3}$$

and f_{entropy} the entropy of the image histogram. We used the default parameter values set by the tool: $K_e = 1$, $K_r = 50$, $K_o = 0.5$, $K_w = 20$. Grid spacing was set at 16 mm. Bias field, original and corrected images were saved for further processing steps.

Appendix E. Skull-stripping

Because of their massive head muscles, which volume depends on age and species, skull-stripping is particularly challenging in NHPs (Maldjian et al., 2015). As stated earlier, T1w and T2w images present different challenges which led us to propose two adapted tools.

With T1w images acquired in humans, a common and simple way of performing skull-stripping is, after an automated thresholding of the MRI, to separate brain tissues from skin tissues with an erosion or radius r (E_r) and an extraction of the biggest connected component with respect to connectivity d (B_d), followed by a dilation (D_r) to recover lost tissues. Those operators are defined in inline supplementary material 2. In NHPs, however, the brain can be smaller than head muscles, which present a similar T1w signal. To get around this issue, we have introduced another relation to order components based on their compactness. Indeed, because of the hole left by the brain, the muscle component should be much less compact. In order to keep the algorithmic complexity low, we have defined compactness as the ratio between the volume of the component and that of its bounding box. Consequently, we note M_d the selection of the most compact component. A basic skull extraction function can then be written :

$$D_r \circ M_d \circ E_r \ , \tag{E.1}$$

where $\circ$ is the function composition operator. To increase robustness, we have added two steps. First, after component selection, its topology was corrected by filling its cavities. However, we did not deal with holes because their filling is too computationally expensive. Second, in order to allow the dilation to accurately recover lost brain tissues without spurious inclusions of skin tissues, it was obtained from a distance map computed in a space constrained by the first automated threshold, B. Let us note $D_{B,r}$ such a constrained dilation, the final skull stripping function can be written :

$$D_{B,r_2} \circ F_d \circ M_d \circ E_{r_1} \tag{E.2}$$

In T2w images, muscles present a hypo-intense signal compared to brain tissues, making them easily separable by thresholding. In our method, we classified bias corrected voxels into 6 classes by k-means clustering. The least intense class was considered as background (Bg), the second least intense as muscle (Mu), the following two as brain tissue (Ti) and the two most intense as CSF (Csf). The following processing was performed:

1. Raw $= D_r \circ F_d \circ B_d \circ E_r \circ F_d$ (Ti $\cup$ Csf)
2. Brain $= C_r \circ D_r \circ F_d \circ B_d \circ E_r \circ F_d$ (Ti $\cap$ Raw)

References

Acosta-Cabronero, J., Williams, G. B., Pereira, J. M., Pengas, G., Nestor, P. J., Feb. 2008. The impact of skull-stripping and radio-frequency bias correction on grey-matter segmentation for voxel-based morphometry. NeuroImage 39 (4), 1654–1665.

Almli, C., Rivkin, M., McKinstry, R., Mar. 2007. The NIH MRI study of normal brain development (Objective-2): Newborns, infants, toddlers, and preschoolers. NeuroImage 35 (1), 308–325.

Ashburner, J., 2000. Computational neuroanatomy. Ph.D. thesis, University College London, London, UK.

Ashburner, J., Friston, K., 1997. Multimodal image coregistration and partitioning - a unified framework. NeuroImage 6 (3), 209–217.

Ashburner, J., Friston, K. J., Jul. 2005. Unified segmentation. NeuroImage 26 (3), 839–51.

Autrey, M. M., Reamer, L. A., Mareno, M. C., Sherwood, C. C., Herndon, J. G., Preuss, T., Schapiro, S. J., Hopkins, W. D., Jun. 2014. Age-related effects in the neocortical organization of chimpanzees: Gray and white matter volume, cortical thickness, and gyrification. NeuroImage 101C, 59–67.

Auzias, G., Lefevre, J., Le Troter, A., Fischer, C., Perrot, M., Regis, J., Coulon, O., May 2013. Model-driven harmonic parameterization of the cortical surface: HIP-HOP. IEEE Trans. Med. Imaging 32 (5), 873–887.

Avants, B. B., Tustison, N. J., Wu, J., Cook, P. A., Gee, J. C., 2011. An open source multivariate framework for n-tissue segmentation with evaluation on public data. Neuroinformatics 9 (4), 381–400.

Balbastre, Y., Rivière, D., Souedet, N., Fischer, C., Hérard, A.-S., Williams, S., Vandenberghe, M. E., Flament, J., Aron-Badin, R., Hantraye, P., Mangin, J.-F., Delzescaux, T., 2017. A validation dataset for Macaque brain MRI segmentation. Data Brief (submitted).

Balbastre, Y., Vandenberghe, M. E., Flament, J., Hérard, A.-S., Gipchtein, P., Williams, S., Souedet, N., Guillermier, M., Bugi, A., Perrier, A. L., Aron-Badin, R., Hantraye, P., Mangin, J.-F., Delzescaux, T., Oct. 2015. An original approach for personalized parcellation of macaque MR brain images: Application to caudate volume estimation in a model of Huntington's disease. In: Program No. 42.09. 2015 Neuroscience Meeting Planner. Society for Neuroscience, Chicago, IL, USA.

Ballanger, B., Tremblay, L., Sgambato-Faure, V., Beaudoin-Gobert, M., Lavenne, F., Le Bars, D., Costes, N., Aug. 2013. A multi-atlas based method for automated anatomical Macaca fascicularis brain MRI segmentation and PET kinetic extraction. NeuroImage 77, 26–43.

Besag, J., 1986. On the statistical analysis of dirty pictures. J. R. Stat. Soc. 48 (3), 259–302.

Bobinski, M., de Leon, M., Wegiel, J., DeSanti, S., Convit, A., Saint Louis, L., Rusinek, H., Wisniewski, H., Dec. 1999. The histological validation of post mortem magnetic resonance imaging-determined hippocampal volume in Alzheimer's disease. Neuroscience 95 (3), 721–725.

Bogart, S. L., Bennett, A. J., Schapiro, S. J., a Reamer, L., Hopkins, W. D., Mar. 2014. Different early rearing experiences have long-term effects on cortical organization in captive chimpanzees (Pan troglodytes). Dev. Sci. 17 (2), 161–74.

Bogart, S. L., Mangin, J.-F., Schapiro, S. J., Reamer, L., Bennett, A. J., Pierre, P. J., Hopkins, W. D., Jul. 2012. Cortical sulci asymmetries in chimpanzees and macaques: A new look at an old idea. NeuroImage 61 (3), 533–41.

Borman, S., 2004. The expectation maximization algorithm a short tutorial.

Calabrese, E., Badea, A., Coe, C. L., Lubach, G. R., Shi, Y., Styner, M. A., Allan Johnson, G., 2015. A diffusion tensor MRI atlas of the postmortem rhesus macaque brain. NeuroImage.

Chupin, M., Gérardin, E., Cuingnet, R., Boutet, C., Lemieux, L., Lehéricy, S., Benali, H., Garnero, L., Colliot, O., 2009. Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI. Hippocampus 19 (6), 579–587.

Coupe, P., Yger, P., Prima, S., Hellier, P., Kervrann, C., Barillot, C., Apr. 2008. An optimized blockwise nonlocal means denoising filter for 3-D magnetic resonance images. IEEE Trans. Med. Imaging 27 (4), 425–441.

Dempster, A. A. P., Laird, N. M. N., Rubin, D. D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B 39 (1), 1–38.

Dice, L. R., Jul. 1945. Measures of the amount of ecologic association between species. Ecology 26 (3), 297–302.

Dickerson, B. C., Bakkour, A., Salat, D. H., Feczko, E., Pacheco, J., Greve, D. N., Grodstein, F., Wright, C. I., Blacker, D., Rosas, H. D., Sperling, R. A., Atri, A., Growdon, J. H., Hyman, B. T., Morris, J. C., Fischl, B., Buckner, R. L., Mar. 2009. The cortical signature of Alzheimer's disease: Regionally specific cortical thinning relates to symptom severity in very mild to mild AD dementia and is detectable in asymptomatic amyloid-positive individuals. Cereb. Cortex 19 (3), 497–510.

Evans, A. C., Kamber, M., Collins, D. L., MacDonald, D., 1994. An MRI-based probabilistic atlas of neuroanatomy. In: Magnetic Resonance Scanning and Epilepsy. Springer US, Boston, MA, pp. 263–274.

Fein, G., Landman, B., Tran, H., Barakos, J., Moon, K., Di Sclafani, V., Shumway, R., May 2006. Statistical parametric mapping of brain morphology: Sensitivity is dramatically increased by using brain-extracted images as inputs. NeuroImage 30 (4), 1187–1195.

Ferrante, R. J., Kowall, N. W., Cipolloni, P. B., Storey, E., Beal, M. F., Jan. 1993. Excitotoxin lesions in primates as a model for Huntington's disease: Histopathologic and neurochemical characterization. Exp. Neurol. 119 (1), 46–71.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A. M., Jan. 2002. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. Neuron 33 (3), 341–355.

Fischl, B., Sereno, M. I., Tootell, R. B. H., Dale, A. M., 1999. High-resolution intersubject averaging and a coordinate system for the cortical surface. Hum. Brain Mapp. 8, 272–284.

Forny-Germano, L., Lyra e Silva, N. M., Batista, A. F., Brito-Moreira, J., Gralle, M., Boehnke, S. E., Coe, B. C., Lablans, A., Marques, S. A., Martinez, A. M., Klein, W. L., Houzel, J. C., Ferreira, S. T., Munoz, D. P., De Felice, F. G., 2014. Alzheimer's disease-like pathology induced by amyloid-β oligomers in nonhuman primates. J. Neurosci. 34 (41), 13629–13643.

Frisoni, G. B., Pievani, M., Testa, C., Sabattoli, F., Bresciani, L., Bonetti, M., Beltramello, A., Hayashi, K. M., Toga, A. W., Thompson, P. M., Mar. 2007. The topography of grey matter involvement in early and late onset Alzheimer's disease. Brain 130 (3), 720–730.

Ginovart, N., Wilson, A. A., Meyer, J. H., Hussey, D., Houle, S., Nov. 2001. Positron Emission Tomography Quantification of [11C]-DASB Binding to the Human Serotonin Transporter: Modeling Strategies. J. Cereb. Blood Flow Metab. 21 (11), 1342–1353.

Goldbach, M., Menhardt, W., Stevens, J., 1991. Multispectral tissue characterization in magnetic resonance imaging using bayesian estimation and markov random fields. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Vol. 13. IEEE, pp. 62–63.

Gudbjartsson, H., Patz, S., Dec. 1995. The Rician distribution of noisy MRI data. Magn. Reson. Med. 34 (6), 910–4.

Hammersley, J. M., Clifford, P., 1971. Markov fields on finite graphs and lattices.

Hopkins, W. D., Avants, B. B., Mar. 2013. Regional and hemispheric variation in cortical thickness in chimpanzees (Pan troglodytes). J. Neurosci. 33 (12), 5241–8.

Hopkins, W. D., Coulon, O., Mangin, J., Dec. 2010. Observer-independent characterization of sulcal landmarks and depth asymmetry in the central sulcus of the chimpanzee brain. Neuroscience 171 (2), 544–51.

Iglesias, J. E., Sabuncu, M. R., Aug. 2015. Multi-atlas segmentation of biomedical images: A survey. Med. Image Anal. 24 (1), 205–219.

Iglesias, J. E., Sabuncu, M. R., Van Leemput, K., Dec. 2013. A unified framework for cross-modality multi-atlas segmentation of brain MRI. Med. Image Anal. 17 (8), 1181–1191.

Jarraya, B., Boulet, S., Ralph, G. S., Jan, C., Bonvento, G., Azzouz, M., Miskin, J. E., Shin, M., Delzescaux, T., Drouot, X., Hérard, A.-S., Day, D. M., Brouillet, E., Kingsman, S. M., Hantraye, P., Mitrophanous, K. A., Mazarakis, N. D., Palfi, S., Oct. 2009. Dopamine gene therapy for Parkinson's disease in a nonhuman primate without associated dyskinesia. Sci. Transl. Med. 1 (2), 2ra4–2ra4.

Johnson, G. A., Badea, A., Brandenburg, J., Cofer, G., Fubara, B., Liu, S., Nissanov, J., Nov. 2010. Waxholm space: An image-based reference for coordinating mouse brain research. NeuroImage 53 (2), 365–72.

Knickmeyer, R. C., Styner, M., Short, S. J., Lubach, G. R., Kang, C., Hamer, R., Coe, C. L., Gilmore, J. H., May 2010. Maturational trajectories of cortical brain development through the pubertal transition: Unique species and sex differences in the monkey revealed through structural magnetic resonance imaging. Cereb. Cortex 20 (5), 1053–1063.

Kochunov, P., Glahn, D. C., Fox, P. T., Lancaster, J. L., Saleem, K., Shelledy, W., Zilles, K., Thompson, P. M., Coulon, O., Mangin, J. F., Blangero, J., Rogers, J., Nov. 2010. Genetics of primary cerebral gyrification: Heritability of length, depth and area of primary sulci in an extended pedigree of Papio baboons. NeuroImage 53 (3), 1126–34.

Kochunov, P., Mangin, J.-F., Coyle, T., Lancaster, J., Thompson, P., Rivière, D., Cointepas, Y., Régis, J., Schlosser, A., Royall, D. R., Zilles, K., Mazziotta, J., Toga, A., Fox, P. T., Nov. 2005. Age-related morphology trends of cortical sulci. Hum. Brain Mapp. 26 (3), 210–20.

Latzman, R. D., Hecht, L. K., Freeman, H. D., Schapiro, S. J., Hopkins, W. D., Aug. 2015. Neuroanatomical correlates of personality in chimpanzees (Pan troglodytes): Associations between personality and frontal cortex. NeuroImage 123, 63–71.

Lebenberg, J., Hérard, A., Dubois, A., Dauguet, J., Frouin, V., Dhenain, M., Hantraye, P., Delzescaux, T., Jul. 2010. Validation of MRI-based 3D digital atlas registration with histological and autoradiographic volumes: An anatomofunctional transgenic mouse brain imaging study. NeuroImage 51 (3), 1037–46.

Liang, Z., Jaszczak, R. J., Coleman, R. E., 1992. Parameter estimation of finite mixtures using the EM algorithm and information criteria with application to medical image processing. IEEE Trans. Nucl. Sci. 39 (4), 1126–1133.

Liang, Z., Macfall, J. R., Harrington, D. P., 1994. Parameter estimation and tissue segmentation from multispectral MR images. IEEE Trans. Med. Imaging 13 (3), 441–449.

Liu, C., Tian, X., Liu, H., Mo, Y., Bai, F., Zhao, X., Ma, Y., Wang, J., Feb. 2015. Rhesus monkey brain development during late infancy and the effect of phencyclidine: A longitudinal MRI and DTI study. NeuroImage 107C, 65–75.

Lyoo, C. H., Ryu, Y. H., Lee, M. S., Mar. 2010. Topographical distribution of cerebral cortical thinning in patients with mild Parkinson's disease without dementia. Mov. Disord. Off. J. Mov. Disord. Soc. 25 (4), 496–9.

Maldjian, J. A., Shively, C. A., Nader, M. A., Friedman, D. P., Whitlow, C. T., Dec. 2015. Multi-atlas library for eliminating normalization failures in non-human primates. Neuroinformatics, 1–8.

Malkova, L., Heuer, E., Saunders, R. C., 2006. Longitudinal magnetic resonance imaging study of rhesus monkey brain development. Eur. J. Neurosci. 24 (11), 3204–3212.

Mangin, J., Rivière, D., Cachia, A., Duchesnay, E., Cointepas, Y., Papadopoulos-Orfanos, D., Scifo, P., Ochiai, T., Brunelle, F., Régis, J., Jan. 2004a. A framework to study the cortical folding patterns. NeuroImage 23 Suppl 1, S129–38.

Mangin, J., Rivière, D., Coulon, O., Poupon, C., Cachia, A., Cointepas, Y., Poline, J.-B., Le Bihan, D., Régis, J., Papadopoulos-Orfanos, D., Feb. 2004b. Coordinate-based versus structural approaches to brain image analysis. Artif. Intell. Med. 30 (2), 177–97.

Mangin, J.-F., 2000. Entropy minimization for automatic correction of intensity nonuniformity. In: Proceedings IEEE Workshop on Mathematical Methods in Biomedical Image Analysis. MMBIA-2000 (Cat. No.PR00737). Vol. 00. IEEE Comput. Soc, pp. 162–169.

Manjón, J. V., Carbonell-Caballero, J., Lull, J. J., García-Martí, G., Martí-Bonmatí, L., Robles, M., Aug. 2008. MRI denoising using Non-Local Means. Medical Image Analysis 12 (4), 514–523.

Marques, J. P., Kober, T., Krueger, G., van der Zwaag, W., Van de Moortele, P.-F., Gruetter, R., Jan. 2010. MP2RAGE, a self bias-field corrected sequence for improved segmentation and T1-mapping at high field. NeuroImage 49 (2), 1271–1281.

Mattes, D., Haynor, D. R., Vesselle, H., Lewellen, T. K., Eubank, W., Jan. 2003. PET-CT image registration in the chest using free-form deformations. IEEE Trans. Med. Imaging 22 (1), 120–8.

McLachlan, G. J., Krishnan, T. T., 2008. The EM Algorithm and Extensions. Wiley-Interscience.

McLaren, D. G., Kosmatka, K. J., Kastman, E. K., Bendlin, B. B., Johnson, S. C., Mar. 2010. Rhesus macaque brain morphometry: A methodological comparison of voxel-wise approaches. Methods San Diego Calif 50 (3), 157–65.

McLaren, D. G., Kosmatka, K. J., Oakes, T. R., Kroenke, C. D., Kohama, S. G., Matochik, J. A., Ingram, D. K., Johnson, S. C., Mar. 2009. A population-average MRI-based atlas collection of the rhesus macaque. NeuroImage 45 (1), 52–9.

Mohan, J., Krishnaveni, V., Guo, Y., Jan. 2014. A survey on the magnetic resonance image denoising methods. Biomedical Signal Processing and Control 9, 56–69.

Morris, R. D., Descombes, X., Zerubia, J., Sep. 1996. The Ising/Potts model is not well suited to segmentation tasks. In: 1996 IEEE Digital Signal Processing Workshop Proceedings. pp. 263–266.

Papp, E. A., Leergaard, T. B., Calabrese, E., Johnson, G. A., Bjaalie, J. G., Aug. 2014. Waxholm Space atlas of the Sprague Dawley rat brain. NeuroImage 97, 374–86.

Patenaude, B., Smith, S. M., Kennedy, D. N., Jenkinson, M., Jun. 2011. A Bayesian model of shape and appearance for subcortical brain segmentation. NeuroImage 56 (3), 907–22.

Paxinos, G., Huang, X.-F., Petrides, M., Toga, A. W., 2008. The Rhesus Monkey Brain in Stereotaxic Coordinates, 2nd Edition. Academic Press, San Diego.

Pereira, J. B., Ibarretxe-Bilbao, N., Marti, M.-J., Compta, Y., Junqué, C., Bargallo, N., Tolosa, E., Nov. 2012. Assessment of cortical degeneration in patients with Parkinson's disease by voxel-based morphometry, cortical folding, and cortical thickness. Hum. Brain Mapp. 33 (11), 2521–34.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., R Core Team, 2016. Nlme: Linear and nonlinear mixed effects models.

Porras, G., Li, Q., Bezard, E., Jan. 2012. Modeling Parkinson's Disease in Primates: The MPTP Model. Cold Spring Harb Perspect Med 2 (3), a009308.

R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rajapakse, J. C., Kruggel, F., 1998. Segmentation of MR images with intensity inhomogeneities. Image Vis. Comput. 16 (3), 165–180.

Régis, J., Mangin, J.-F., Ochiai, T., Frouin, V., Rivière, D., Cachia, A., Tamura, M., Samson, Y., 2005. "Sulcal root" generic model: A hypothesis to overcome the variability of the human cortex folding patterns. Neurol. Med. Chir. (Tokyo) 45 (1), 1–17.

Reiner, P., Jouvent, E., Duchesnay, E., Cuingnet, R., Mangin, J.-F., Chabriat, H., Jan. 2012. Sulcal span in Azheimer's disease, amnestic mild cognitive impairment, and healthy controls. J. Alzheimers Dis. 29 (3), 605–13.

Rogers, J., Kochunov, P., Zilles, K., Shelledy, W., Lancaster, J., Thompson, P., Duggirala, R., Blangero, J., Fox, P. T., Glahn, D. C., Nov. 2010. On the genetic architecture of cortical folding and brain volume in primates. NeuroImage 53 (3), 1103–8.

Rohlfing, T., Russakoff, D. B., Maurer, C. R., Jul. 2003. Expectation maximization strategies for multi-atlas multi-label segmentation. In: Information Processing in Medical Imaging. Vol. 2732 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, Ambleside, UK, pp. 210–21.

Rosas, H. D., Liu, A. K., Hersch, S., Glessner, M., Ferrante, R. J., Salat, D. H., van der Kouwe, A., Jenkins, B. G., Dale, A. M., Fischl, B., Mar. 2002. Regional and progressive thinning of the cortical ribbon in Huntington's disease. Neurology 58 (5), 695–701.

Scott, J. A., Grayson, D., Fletcher, E., Lee, A., Bauman, M. D., Schumann, C. M., Buonocore, M. H., Amaral, D. G., 2015. Longitudinal analysis of the developing rhesus monkey brain using magnetic resonance imaging: Birth to adulthood. Brain Struct. Funct.

Seiger, R., Hahn, A., Hummer, A., Kranz, G. S., Ganger, S., Küblböck, M., Kraus, C., Sladky, R., Kasper, S., Windischberger, C., Lanzenberger, R., Mar. 2015. Voxel-based morphometry at ultra-high fields. A comparison of 7T and 3T MRI data. NeuroImage 113, 207–216.

Sled, J. G., Zijdenbos, A. P., Evans, A. C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging 17 (1), 87–97.

Styner, M., Knickmeyer, R., Joshi, S., Coe, C., Short, S. J., Gilmore, J., Mar. 2007. Automatic brain segmentation in rhesus monkeys. In: Pluim, J. P. W., Reinhardt, J. M. (Eds.), Medical Imaging 2007: Image Processing. Vol. 6512 of SPIE Proceedings. SPIE, pp. 65122L–65122L–8.

Talairach, J., Tournoux, P., 1988. Co-Planar Stereotaxic Atlas of the Human Brain. Thieme, New York.

Thevenaz, P., Unser, M., Oct. 1997. Spline pyramids for intermodal image registration using mutual information. In: Wavelet Applications in Signal and Image V. Vol. 3169 of SPIE Proceedings. SPIE, San Diego, CA, USA, pp. 236–247.

Tustison, N., Avants, B., Cook, P., Zheng, Y., Egan, A., Yushkevich, P., Gee, J., 2010. N4ITK: Improved N3 bias correction. IEEE Trans. Med. Imaging 29 (6), 1310–1320.

Unser, M., Aldroubi, A., Eden, M., 1993. B-spline signal processing. II. Efficiency design and applications. In: IEEE Transactions on Signal Processing. Vol. 41. pp. 834–848.

Van de Moortele, P.-F., Auerbach, E. J., Olman, C., Yacoub, E., Uğurbil, K., Moeller, S., Jun. 2009. T1 weighted brain images at 7 Tesla unbiased for Proton Density, T2* contrast and RF coil receive B1 sensitivity with simultaneous vessel visualization. NeuroImage 46 (2), 432–446.

Van Essen, D. C., Glasser, M. F., Dierker, D. L., Harwell, J., Oct. 2012. Cortical parcellations of the macaque monkey analyzed on surface-based atlases. Cereb. Cortex 22 (10), 2227–40.

Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., Oct. 1998. Automatic segmentation of brain tissues and MR bias field correction using a digital brain atlas. In: Wells, W. M., Colchester, A., Delp, S. (Eds.), Medical Image Computing and Computer-Assisted Intervention — MICCAI'98. Vol. 1496 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, Cambridge, MA, USA, pp. 1222–1229.

Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., Oct. 1999. Automated model-based tissue classification of MR images of the brain. IEEE Trans. Med. Imaging 18 (10), 897–908.

Vannier, M., Speidel, C., Rickman, D., Schertz, L., Baker, L., Hildebolt, C., Offutt, C., Balko, J., Butterfield, R., Gado, M., 1988. Validation of magnetic resonance imaging (MRI) multispectral tissue classification. In: 9th International Conference on Pattern Recognition. Vol. 2. IEEE, pp. 1182–1186.

Visser, E., Keuken, M. C., Douaud, G., Gaura, V., Bachoud-Levi, A.-C., Remy, P., Forstmann, B. U., Jenkinson, M., Oct. 2015. Automatic segmentation of the striatum and globus pallidus using MIST: Multimodal image segmentation tool. NeuroImage 125, 479–497.

Warfield, S. K., Zou, K. H., Wells, W. M., 2004. Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging 23 (7), 903–921.

Wells, W. M., Grimson, W. E. L., Kikinis, R., Jolesz, F. A., 1996. Adaptive segmentation of MRI data. IEEE Trans. Med. Imaging 15 (4), 429–42.

Wey, H.-Y., a Phillips, K., McKay, D. R., Laird, A. R., Kochunov, P., Davis, M. D., Glahn, D. C., Duong, T. Q., Fox, P. T., Aug. 2013. Multi-region hemispheric specialization differentiates human from nonhuman primate brain function. Brain Struct. Funct.

Wickham, H., 2009. Ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Young, J. T., Shi, Y., Niethammer, M., Grauer, M., Coe, C. L., Lubach, G. R., Budin, F., Knickmeyer, R. C., Alexander, A. L., Styner, M. A., 2017. The UNC-Wisconsin Rhesus Macaque Neurodevelopment Database: A Structural MRI and DTI Database of Early Postnatal Development. Front. Neurosci. 11.

Zhang, Y., Brady, M., Smith, S., Jan. 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging 20 (1), 45–57.

Inline Supplementary Material 1: Digital atlases of the Macaque brain.

Styner et al. (2007), McLaren et al. (2009) and Rohlfing et al. (2012) published *in vivo* T1w templates with probability maps for GM, WM and CSF. Styner et al. (2007) also provided a lobar parcellation into 13 regions and a subcortical parcellation comprising 4 regions and Rohlfing et al. (2012) provided a 502-region atlas obtained by non-linear registration from the BrainInfo macaque atlas (Dubach and Bowden, 2009). *In vivo* templates and brain parcellations were also proposed by Wisco et al. (2008) (T2w template, 14 regions), Frey et al. (2011) (T1w template, 255 regions obtained by non-linear registration from the Paxinos macaque atlas (Paxinos et al., 2008)) and Ballanger et al. (2013) (T1w template, 42 regions) but without probability maps. Recently, Shi et al. (2017) published age appropriate templates and GM, WM, CSF and subcortical probability maps. Those templates and maps were constructed using non-linear registration. The parcellation from Styner et al. (2007) was also propagated towards each template.

Supplementary references

Ballanger, B., Tremblay, L., Sgambato-Faure, V., Beaudoin-Gobert, M., Lavenne, F., Le Bars, D., Costes, N., Aug. 2013. A multi-atlas based method for automated anatomical Macaca fascicularis brain MRI segmentation and PET kinetic extraction. NeuroImage 77, 26–43.

Dubach, M. F., Bowden, D. M., Oct. 2009. BrainInfo online 3d macaque brain atlas: A database in the shape of a brain. In: Program No. 199.5. 2009 Neuroscience Meeting Planner. Society for Neuroscience, Chicago, IL, USA.

Frey, S., Pandya, D. N., Chakravarty, M. M., Bailey, L., Petrides, M., Collins, D. L., Apr. 2011. An MRI based average macaque monkey stereotaxic atlas and space (MNI monkey space). NeuroImage 55 (4), 1435–42.

McLaren, D. G., Kosmatka, K. J., Oakes, T. R., Kroenke, C. D., Kohama, S. G., Matochik, J. A., Ingram, D. K., Johnson, S. C., Mar. 2009. A population-average MRI-based atlas collection of the rhesus macaque. NeuroImage 45 (1), 52–9.

Paxinos, G., Huang, X.-F., Petrides, M., Toga, A. W., 2008. The Rhesus Monkey Brain in Stereotaxic Coordinates, 2nd Edition. Academic Press, San Diego.

Rohlfing, T., Kroenke, C. D., Sullivan, E. V., Dubach, M. F., Bowden, D. M., a Grant, K., Pfefferbaum, A., Jan. 2012. The INIA19 template and NeuroMaps atlas for primate brain image parcellation and spatial normalization. Front. Neuroinformatics 6, 27.

Shi, Y., Budin, F., Yapuncich, E., Rumple, A., Young, J. T., Payne, C., Zhang, X., Hu, X., Godfrey, J., Howell, B., Sanchez, M. M., Styner, M. A., 2017. UNC-Emory Infant Atlases for Macaque Brain Image Analysis: Postnatal Brain Development through 12 Months. Front. Neurosci. 10.

Styner, M., Knickmeyer, R., Joshi, S., Coe, C., Short, S. J., Gilmore, J., Mar. 2007. Automatic brain segmentation in rhesus monkeys. In: Pluim, J. P. W., Reinhardt, J. M. (Eds.), Medical Imaging 2007: Image Processing. Vol. 6512 of SPIE Proceedings. SPIE, pp. 65122L–65122L–8.

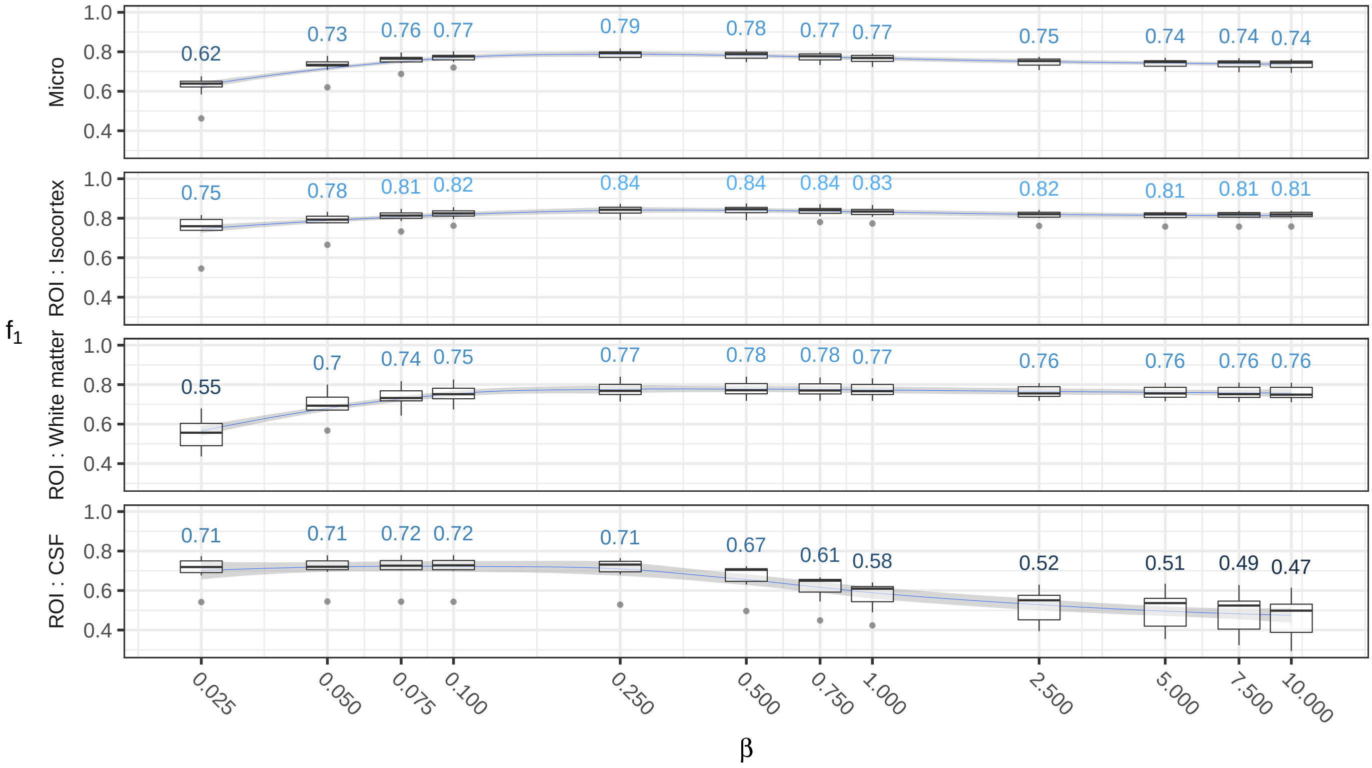
Wisco, J. J., Rosene, D. L., Killiany, R. J., Moss, M. B., Warfield, S. K., Egorova, S., Wu, Y., Liptak, Z., Warner, J., Guttmann, C. R. G., Oct. 2008. A rhesus monkey reference label atlas for template driven segmentation. J. Med. Primatol. 37 (5), 250–60.

Inline Supplementary Material 2: Basic morphomathematical operations.

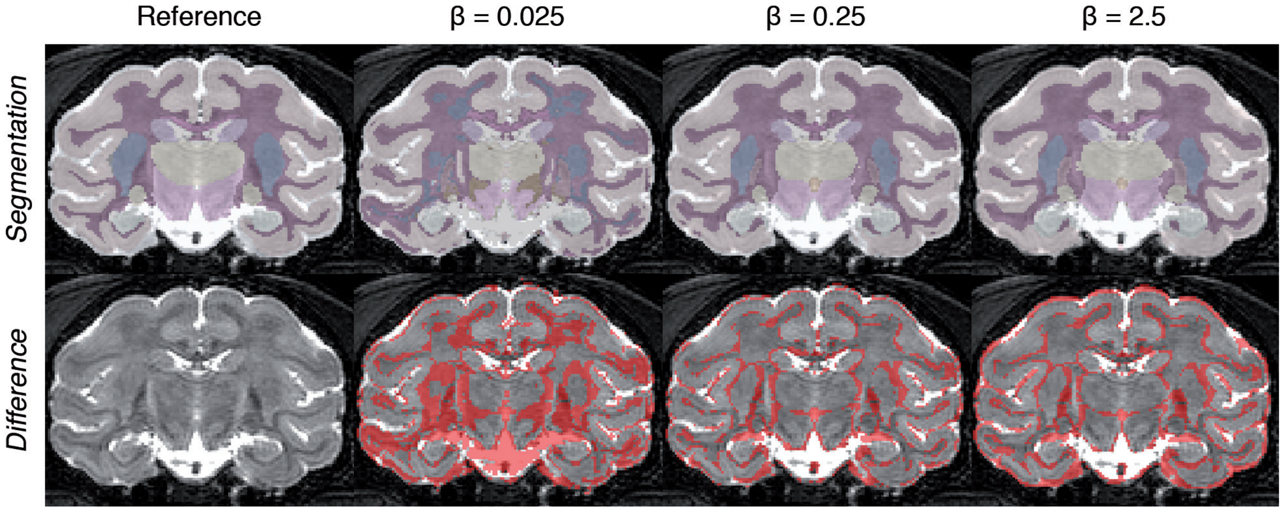
Let us describe basic morphomathematical operations that can be applied to a binary volume B. An erosion of radius r (E_r) is equivalent to thresholding a distance map to the background voxels of B at distance r. Dilation (D_r) is the opposite operation: $D_r(B) = \overline{E_r(\bar{B})}$, with $\bar{\cdot}$ the binary operation that inverts foreground and background. An opening is the concatenation of an erosion and a dilation ($O_r = D_r \circ E_r$) while a closing is that of a dilation and an erosion ($C_r = E_r \circ D_r$). We will note B_d the selection of the biggest connected component of B under connectivity order d, and F_d the filling of all cavities of B (cavities are background components that are not connected to the image domain boundaries).

Inline Supplementary Table 1: Classification of the 18 atlas classes into Background, CSF, WM, GM and WGM.

Background
Background
White Matter
Pallidum
White matter
Corpus callosum
Gray Matter
Subpallium
Pallium
Isocortex
White-Gray Mixture
Thalamus
Hypothalamus
Caudate nucleus
Putamen
Dorsal pallium
Midbrain
Medulla
Pons
Cerebellum
CSF
CSF
Ventricles



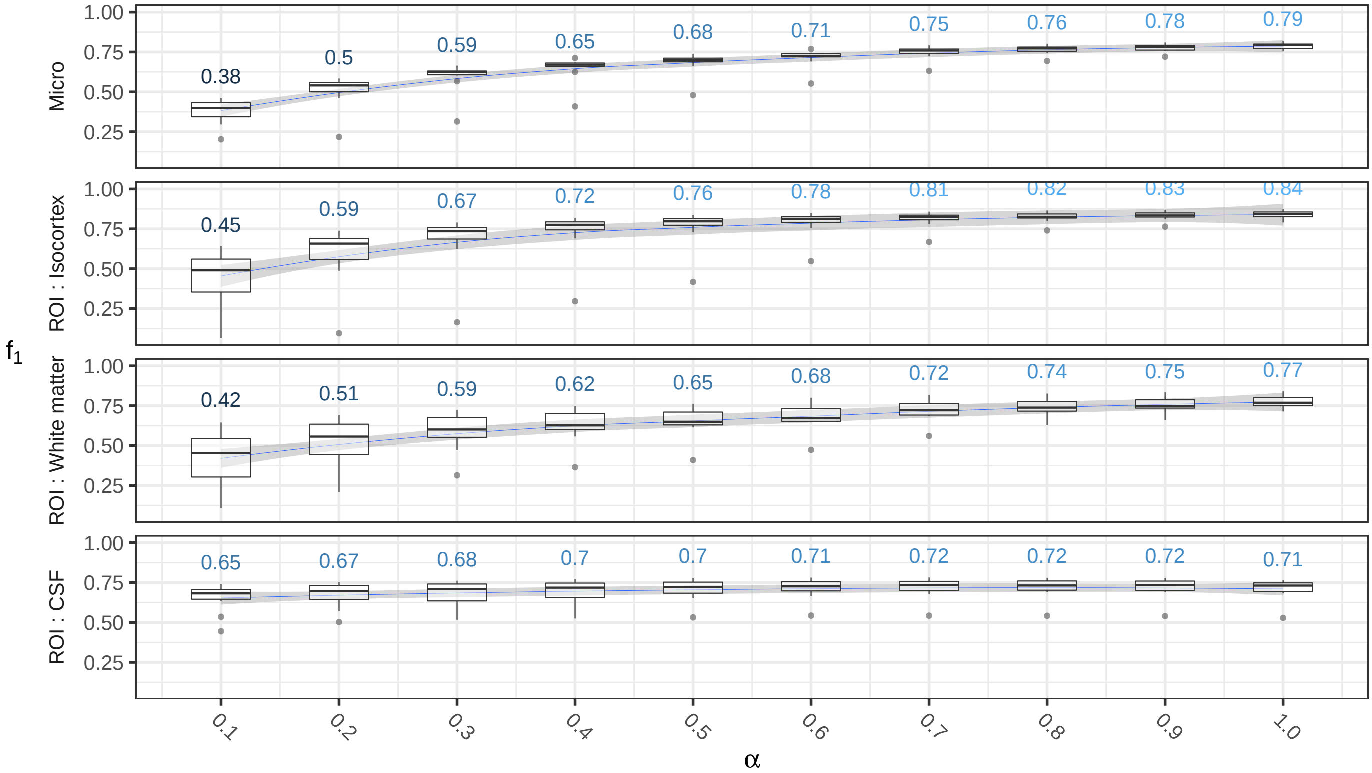
Inline Supplementary Figure 1: Optimization of the β parameter. For each tested value, a Tukey's boxplot represents the different quartiles of the F_1 score for regions CSF, isocortex and white matter as well as those of the micro-F_1 score. The mean score is also indicated in blue. The x axis has a logarithmic scale.



Inline Supplementary Figure 2: Resulting segmentation of an arbitrarily chosen subject with 3 different values of β. The first row shows the segmentation while the second one shows in red voxels which classification differ from the manual segmentation of reference.

Inline Supplementary Table 2: Optimization of the β parameter. Mean F_1 scores for all region of the hierarchy. For each region, the maximum score is shown in red and scores that are equal to the maximum with a precision of 0.01 are bold.

ROI β	0.025	0.05	0.075	0.1	0.25	0.5	0.75	1.0	2.5	5.0	7.5	10.0
${F_1}^{\text{micro}}$	0.62	0.73	0.76	0.77	**0.79**	0.78	0.77	0.77	0.75	0.74	0.74	0.74
Intracranial	0.96	0.96	**0.97**	**0.97**	**0.97**	**0.97**	0.96	0.96	0.96	0.96	0.96	0.96
CSF	0.71	0.71	**0.72**	**0.72**	0.71	0.67	0.61	0.58	0.52	0.51	0.49	0.47
Ventricles	0.39	0.65	0.70	**0.71**	**0.71**	**0.71**	0.70	0.70	0.68	0.65	0.62	0.59
Brain	0.94	0.94	0.94	0.94	**0.95**	**0.95**	0.94	0.94	0.93	0.93	0.93	0.93
Forebrain	0.93	0.94	0.94	0.94	**0.95**	**0.94**	**0.94**	0.94	0.93	0.93	0.93	0.92
Midbrain	0.65	0.75	0.77	0.78	0.79	0.79	0.79	**0.80**	**0.80**	0.79	0.79	0.79
Hindbrain	0.87	0.91	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**	0.91	0.90	0.90	0.89
Isocortex	0.75	0.78	0.80	0.82	**0.84**	**0.84**	**0.84**	0.83	0.82	0.81	0.81	0.81
Pallium	0.62	0.69	**0.70**	**0.70**	**0.70**	**0.70**	**0.70**	**0.70**	0.69	0.69	0.69	0.69
Dorsal pallium	0.68	**0.76**	0.75	0.74	**0.76**	0.74	0.74	0.73	0.71	0.69	0.68	0.68
Thalamus	0.72	0.78	0.79	0.80	0.81	**0.82**	**0.82**	**0.82**	0.81	0.80	0.80	0.80
Hypothalamus	0.07	0.20	0.49	0.52	**0.54**	**0.54**	**0.54**	0.53	0.53	0.52	0.51	0.51
White matter	0.55	0.70	0.74	0.75	0.77	**0.78**	**0.78**	0.77	0.76	0.76	0.76	0.76
Corpus callosum	0.45	0.47	0.49	0.50	0.61	**0.66**	**0.66**	**0.66**	0.64	0.62	0.60	0.57
Subpallium	0.44	0.69	0.77	0.78	**0.80**	**0.80**	**0.80**	0.79	0.79	0.78	0.78	0.78
Pallidum	0.27	0.51	0.60	0.64	**0.70**	**0.70**	**0.70**	**0.70**	0.69	0.69	0.69	0.69
Striatum	0.46	0.73	0.79	**0.80**	**0.80**	**0.80**	**0.80**	**0.80**	0.79	0.79	0.79	0.78
Caudate nucleus	0.70	**0.74**	**0.74**	**0.74**	**0.74**	**0.74**	**0.74**	**0.74**	0.73	0.72	0.72	0.71
Putamen	0.31	0.70	0.81	0.83	**0.84**	**0.84**	**0.84**	**0.84**	0.83	0.83	0.83	0.83
Pons	0.69	0.74	0.75	0.75	**0.76**	**0.76**	**0.76**	**0.76**	0.74	0.73	0.73	0.72
Medulla	0.59	0.75	0.75	0.75	**0.76**	**0.76**	**0.76**	**0.76**	0.74	0.73	0.73	0.73
Cerebellum	0.77	0.91	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**	0.90	0.89	0.89	0.89



Inline Supplementary Figure 3: Optimization of the α parameter. For each tested value, a Tukey's boxplot represents the different quartiles of the F_1 score for regions CSF, isocortex and WM as well as those of the micro-F_1 score. The mean score is also indicated in blue.

Inline Supplementary Table 3: Optimization of the α parameter. Mean F_1 scores for all regions of the hierarchy. For each region, the maximum score is shown in red and scores that are equal to the maximum with a precision of 0.01 are bold.

ROI α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$F_1{}^{\text{micro}}$	0.38	0.50	0.59	0.65	0.68	0.71	0.75	0.76	0.78	**0.79**
Intracranial	0.96	0.96	0.96	0.96	0.96	**0.97**	**0.97**	**0.97**	**0.97**	**0.97**
CSF	0.65	0.67	0.68	0.70	0.70	0.71	**0.72**	**0.72**	**0.72**	0.71
Ventricles	0.29	0.36	0.43	0.48	0.53	0.56	0.62	0.66	0.68	**0.71**
Brain	0.89	0.91	0.91	0.92	0.93	0.93	0.94	0.94	**0.95**	**0.95**
Forebrain	0.80	0.85	0.88	0.91	0.92	0.93	0.94	0.94	0.94	**0.95**
Midbrain	0.13	0.27	0.47	0.61	0.68	0.74	0.76	0.78	**0.79**	**0.79**
Hindbrain	0.21	0.48	0.68	0.80	0.84	0.88	0.90	0.91	**0.92**	**0.92**
Isocortex	0.45	0.59	0.67	0.72	0.76	0.78	0.81	0.82	0.83	**0.84**
Pallium	0.12	0.42	0.57	0.65	0.68	0.69	**0.70**	**0.70**	**0.70**	**0.70**
Dorsal pallium	0.34	0.46	0.56	0.62	0.67	0.71	0.73	0.75	0.75	**0.76**
Thalamus	0.20	0.40	0.52	0.63	0.70	0.75	0.78	0.80	**0.81**	**0.81**
Hypothalamus	0.05	0.26	0.45	0.48	0.50	0.51	0.52	0.53	**0.54**	**0.54**
White matter	0.42	0.51	0.59	0.62	0.65	0.68	0.72	0.74	0.75	**0.77**
Corpus callosum	0.08	0.12	0.16	0.23	0.29	0.39	0.45	0.49	0.53	**0.61**
Subpallium	0.15	0.25	0.32	0.38	0.42	0.53	0.69	0.75	0.77	**0.80**
Pallidum	0.12	0.18	0.25	0.31	0.35	0.40	0.45	0.57	0.60	**0.70**
Striatum	0.09	0.22	0.27	0.32	0.35	0.49	0.70	0.77	0.79	**0.80**
Caudate nucleus	0.29	0.57	0.60	0.67	0.69	0.71	0.73	**0.74**	**0.74**	**0.74**
Putamen	0.01	0.02	0.03	0.04	0.03	0.31	0.68	0.78	0.82	**0.84**
Pons	0.15	0.30	0.42	0.56	0.63	0.69	0.72	0.74	0.75	**0.76**
Medulla	0.27	0.41	0.60	0.69	0.71	0.74	**0.76**	**0.76**	**0.76**	**0.76**
Cerebellum	0.19	0.48	0.71	0.81	0.86	0.89	0.91	**0.92**	**0.92**	**0.92**

Inline Supplementary Table 4: Optimization of modules combination. Mean F_1 scores for all regions of the hierarchy. For each region, the maximum score is shown in red and scores that are equal to the maximum with a precision of 0.01 are bold.

Denoising	no	no	no	no	yes	yes	yes	yes
MRF	no	no	yes	yes	no	no	yes	yes
ROI Bias	no	yes	no	yes	no	yes	no	yes
F_1^{micro}	0.75	0.76	0.79	**0.80**	0.75	0.76	0.78	0.79
Intracranial	**0.97**	**0.97**	**0.97**	**0.97**	**0.97**	**0.97**	**0.97**	**0.97**
CSF	0.57	0.60	0.67	0.70	0.60	0.62	0.69	**0.71**
Ventricles	0.70	0.71	0.72	**0.73**	0.70	0.71	0.71	0.71
Brain	0.92	0.93	0.94	**0.95**	0.93	0.93	0.94	**0.95**
Forebrain	0.92	0.93	0.94	0.94	0.92	0.93	0.94	**0.95**
Midbrain	**0.81**	**0.81**	**0.81**	**0.81**	**0.81**	**0.81**	0.80	0.79
Hindbrain	0.91	0.91	**0.93**	**0.93**	0.91	0.91	0.92	0.92
Isocortex	0.80	0.81	0.84	**0.86**	0.79	0.80	0.83	0.84
Pallium	0.68	0.69	0.70	**0.72**	0.67	0.68	0.67	0.70
Dorsal pallium	0.74	0.74	0.76	**0.77**	0.74	0.74	0.74	0.76
Thalamus	**0.82**	**0.82**	**0.82**	**0.82**	**0.82**	**0.82**	0.80	0.81
Hypothalamus	0.51	0.52	0.54	0.54	0.52	0.52	**0.55**	0.54
White matter	0.77	0.78	0.79	**0.80**	0.76	0.77	0.76	0.77
Corpus callosum	**0.68**	0.67	**0.68**	**0.68**	**0.68**	**0.68**	0.62	0.61
Subpallium	**0.81**	**0.81**	**0.81**	**0.81**	**0.81**	0.80	0.80	0.80
Pallidum	0.70	0.70	0.70	**0.71**	0.69	0.69	0.69	0.70
Striatum	**0.82**	**0.82**	**0.82**	**0.82**	**0.82**	0.81	0.81	0.80
Caudate nucleus	**0.78**	0.77	**0.78**	0.77	0.77	0.75	0.77	0.74
Putamen	0.84	**0.85**	**0.85**	**0.85**	0.84	**0.85**	0.84	0.84
Pons	0.74	0.74	**0.78**	**0.78**	0.74	0.74	0.76	0.76
Medulla	0.77	0.76	**0.79**	0.78	0.76	0.76	0.76	0.76
Cerebellum	0.91	0.91	**0.93**	0.92	0.91	0.91	**0.93**	0.92