A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution MRI: Application to adaptive segmentation of in vivo MRI

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th>Citation</th>
<th>Iglesias, Juan Eugenio; Augustinack, Jean C.; Nguyen, Khoa; Player, Christopher M.; Player, Allison; Wright, Michelle; Roy, Nicole; et al. “A Computational Atlas of the Hippocampal Formation Using Ex Vivo, Ultra-High Resolution MRI: Application to Adaptive Segmentation of in Vivo MRI.” NeuroImage 115 (July 2015): 117–137. © 2015 The Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1016/j.neuroimage.2015.04.042">http://dx.doi.org/10.1016/j.neuroimage.2015.04.042</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>Elsevier</td>
</tr>
<tr>
<td>Version</td>
<td>Final published version</td>
</tr>
<tr>
<td>Accessed</td>
<td>Wed Dec 05 21:29:53 EST 2018</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/108059">http://hdl.handle.net/1721.1/108059</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Creative Commons Attribution-NonCommercial-NoDerivs License</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td><a href="http://creativecommons.org/licenses/by-nc-nd/4.0/">http://creativecommons.org/licenses/by-nc-nd/4.0/</a></td>
</tr>
</tbody>
</table>
A computational atlas of the hippocampal formation using ex vivo, ultra-high resolution MRI: Application to adaptive segmentation of in vivo MRI

Juan Eugenio Iglesias a,b,⁎, Jean C. Augustinack b, Khoa Nguyen b, Christopher M. Player b, Allison Player b, Michelle Wright b, Nicole Roy b, Matthew P. Frosch c, Ann C. McKee d,e,f,g, Lawrence L. Wald b, Bruce Fischl b,h, Koen Van Leemput b,i,j,k, for the Alzheimer’s Disease Neuroimaging Initiative 1

a Basque Center on Cognition, Brain and Language, San Sebastián, Spain
b Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
c C.S. Kubik Laboratory for Neuropathology, Pathology Service, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
d Department of Neurology, Boston University, School of Medicine, Boston, MA, USA
e Department of Pathology, Boston University, School of Medicine, Boston, MA, USA
f United States Department of Veterans Affairs, VA Boston Healthcare System, Boston, MA, USA
g Bedford Veterans Administration Medical Center, Bedford, MA, USA
h Computer Science and AI Lab, Massachusetts Institute of Technology, Cambridge, MA, USA
i Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark
j Department of Information and Computer Science, Aalto University, Finland
k Department of Biomedical Engineering and Computational Science, Aalto University, Finland

A R T I C L E   I N F O

Article history:
Received 5 July 2014
Accepted 20 April 2015
Available online 29 April 2015

A B S T R A C T

Automated analysis of MRI data of the subregions of the hippocampus requires computational atlases built at a higher resolution than those that are typically used in current neuroimaging studies. Here we describe the construction of a statistical atlas of the hippocampal formation at the subregion level using ultra-high resolution, ex vivo MRI. Fifteen autopsy samples were scanned at 0.13 mm isotropic resolution (on average) using custom-designed hardware. The images were manually segmented into 13 different hippocampal substructures using a protocol specifically designed for this study; precise delineations were made possible by the extraordinary resolution of the scans. In addition to the subregions, manual annotations for neighboring structures (e.g., amygdala, cortex) were obtained from a separate dataset of in vivo, T1-weighted MRI scans of the whole brain (1 mm resolution). The manual labels from the in vivo and ex vivo data were combined into a single computational atlas of the hippocampal formation with a novel atlas building algorithm based on Bayesian inference. The resulting atlas can be used to automatically segment the hippocampal subregions in structural MRI images, using an algorithm that can analyze multimodal data and adapt to variations in MRI contrast due to differences in acquisition hardware or pulse sequences. The applicability of the atlas, which we are releasing as part of FreeSurfer (version 6.0), is demonstrated with experiments on three different publicly available datasets with different types of MRI contrast. The results show that the atlas and companion segmentation method: 1) can segment T1 and T2 images, as well as their combination, 2) replicate findings on mild cognitive impairment based on high-resolution T2 data, and 3) can discriminate between Alzheimer’s disease subjects and elderly controls with 88% accuracy in standard resolution (1 mm) T1 data, significantly outperforming the atlas in FreeSurfer version 5.3 (86% accuracy) and classification based on whole hippocampal volume (82% accuracy).

© 2015 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

⁎ Corresponding author at: Basque Center on Cognition, Brain and Language, San Sebastián, Spain.
E-mail address: e.iglesias@bcbl.eu (J.E. Iglesias).

1 Data used in preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.
Introduction

The hippocampal formation is a brain region with a critical role in declarative and episodic memory (Scoville and Milner, 1957; Eldridge et al., 2000), as well as a focus of structural change in normal aging (Petersen et al., 2000; Frisoni et al., 2008) and diseases such as epilepsy (Cendes et al., 1993) and, most notably, Alzheimer’s disease (AD) (Laakso et al., 1998; Du et al., 2001; Apostolova et al., 2006). The hippocampal formation consists of a number of distinct, interacting subregions, which comprise a complex, heterogeneous structure. Despite its internal complexity, limits in MRI resolution have traditionally forced researchers to model the hippocampus as a single, homogeneous structure in neuroimaging studies of aging and AD (Boccardi et al., 2011; Chupin et al., 2009). Even though these studies have shown that whole hippocampal volumes derived from automatically or manually segmented MRI scans are powerful biomarkers for AD (Convit et al., 1997; Jack et al., 1999; Frisoni et al., 1999; De Toledo-Morrell et al., 2000; den Heijer et al., 2006; Wang et al., 2003; Fischl et al., 2002), treating the hippocampus as a single entity disregards potentially useful information about its subregions. In animal studies, these subregions have been shown to have different memory functions (Acsády and Káli, 2007; Hunsaker et al., 2008; Kesner, 2007; Rolls, 2010; Schmidt et al., 2012). In humans, they are also thought to play different roles in memory and learning (Gabrieli et al., 1997; Acsády and Káli, 2007; Kieriem et al., 2006; Kesner, 2013; Kesner, 2007; Reagh et al., 2014; Yassa and Stark, 2011), and to be affected differently by AD and normal aging — as indicated by ex vivo, histological studies (Braak and Braak, 1991; Braak and Braak, 1997; Arnold et al., 1991; Thal et al., 2000; Brady and Mufson, 1991; Simic et al., 1997; Harding et al., 1998). Findings from histological studies on hippocampal samples have sparked interest in studying the hippocampal subregions in vivo with MRI, which has been made possible by recent advances in MRI acquisition. Neuroimaging studies that have characterized the subregions in normal aging and AD with in vivo MRI include (Mueller et al., 2007; Wang et al., 2009; Mueller et al., 2010; Small et al., 2011; Kerchner et al., 2012; Wisse et al., 2012, 2014; Burrgren et al., 2008). Most of these studies rely on manual segmentations made on T2-weighted MRI data of the hippocampal formation. The T2 images are often acquired anisotropically, such that resolution along the direction of the major axis of the hippocampus is reduced in exchange for higher in-plane resolution within each coronal slice. This design choice is motivated by the internal structure of the hippocampus: resembling a Swiss roll, its spiral structure changes less rapidly along its major axis, which is almost parallel to the anterior–posterior direction. In in vivo T2-weighted data, part of this spiral becomes visible as a hypointense band that corresponds to the stratum radiatum, lacunosum moleculare, hippocampal sulcus and molecular layer of the dentate gyrus. These layers separate the hippocampus from the dentate gyrus. Henceforth, for simplicity in writing, we will refer to this band as the “molecular layer”2.

Manual segmentation protocols of high resolution, in vivo MRI data of the hippocampal subfields often rely heavily on this molecular layer, which is the most prominent feature of the internal region of the hippocampus that is visible in MRI. However, manual delineation of the subregions in these high-resolution images is extremely labor-intensive — approximately 50 hours per case. Few laboratories currently possess the resources in neuroanatomical expertise and staffing that are required to carry out such studies. Even within those laboratories, the number of cases that can be used in a study is limited by how time-consuming manually tracing the subregions is, which in turns limits the statistical power of the analysis.

These limitations can be overcome with the use of automated algorithms. Two major methods have been proposed for automated and semi-automated hippocampal subregion segmentation so far. In (Yushkevich et al., 2010) – further validated in (Pluta et al., 2012) – Yushkevich and colleagues combined multi-atlas segmentation, similarity-weighted voting, and a learning-based label bias correction technique to estimate the subregion segmentation in a nearly automated fashion. The user needed to provide an initial partitioning of MRI slices into hippocampal head, body and tail (this requirement was dropped in their recent study, Yushkevich et al., 2015). In (Van Leemput et al., 2009), our group introduced a fully automated method based on a statistical atlas of hippocampal anatomy and a generative model of MRI data.

These two methods approach the segmentation problem from different perspectives — parametric and non-parametric. The algorithm we developed – which follows a generative, parametric approach – focuses on modeling the spatial distribution of the hippocampal subregions and surrounding brain structures (i.e., the underlying segmentation), which is learned from labeled training data. The segmentation, which is a hidden variable in the model, is connected to the observed image data through a generative process of image formation that does not make any assumptions about the MRI acquisition. This is indeed the strongest point of the algorithm, since it makes it adaptive to any MRI pulse sequence and resolution that might have been used to acquire the data — even if multimodal (Puonti et al., 2013). Conversely, the algorithm developed by Yushkevich and co-workers relies on a combination of a registration-based, multi-atlas algorithm (a non-parametric method) and machine learning techniques. Both components of their method effectively exploit prior knowledge about the distribution of image intensities derived from training data — information which our parametric method disregards. While the use of prior knowledge about the image intensities is advantageous when the MRI pulse sequence of the test scan matches that of the training data, segmentation of MRI images with different contrast properties is not possible with such an approach. Nonetheless, both Yushkevich’s method and ours have successfully been used to carry out subregion studies on large populations (Teicher et al., 2012; Das et al., 2012; Iglesias et al., 2013).

Our original subfield segmentation method, which is publicly available as part of the FreeSurfer open-source software package (Fischl FreeSurfer, 2012) (version 5.3), is based on a probabilistic atlas that was built from in vivo MRI data acquired at 0.38 × 0.38 × 0.8 mm resolution (Van Leemput et al., 2009). Henceforth, we refer to this atlas as the “in vivo atlas” (FreeSurfer v5.3). The resolution of this atlas is only sufficient to produce a coarse segmentation of the subregions in standard-resolution MRI (i.e., 1 mm); a more accurate model of anatomy is necessary to analyze newer, higher resolution data where the hippocampal substructures are more clearly visualized. Specifically, the in vivo atlas in FreeSurfer v5.3 suffers from three shortcomings. First, the image resolution of the in vivo training data was insufficient for the human labelers to completely distinguish the subregions, forcing them to heavily rely on geometric criteria to trace boundaries, which affected the accuracy of their annotations. In particular, a problematic consequence is that the molecular layer was not labeled, compromising the ability of the atlas to segment high-resolution in vivo data. A second issue is that the delineation protocol was designed for the hippocampal body and did not translate well to the hippocampal head or tail. Due to the second issue, a third problem is that the volumes of the subregions did not agree well with those from histological studies (Simic et al., 1997; Harding et al., 1998), as pointed out by (Schoene-Bake et al., 2014). In this study, we address these shortcomings by replacing the hippocampal atlas in FreeSurfer v5.3 with a new version (FreeSurfer v6.0) built with a novel atlassing algorithm and ex vivo MRI data from autopsy brains. Since motion effects are eliminated, much longer MRI acquisitions are possible when imaging post-mortem samples. Our ex vivo imaging protocol yields images with extremely high resolution and signal-to-noise ratio, dramatically higher than is possible in vivo, which allows us to accurately identify more subregions with a

2 This band has been referred to as the “dark band” in the literature, but is actually bright when imaged with T1-weighted MRI; therefore, we prefer to use the term “molecular layer”.

3 We use the term “subfields” to refer to the CA structures (i.e., CA1–4), and subregions to refer to the whole set of hippocampal substructures, including parasubiculum, presubiculum, subiculum, fimbria, molecular layer and hippocampus–amygdala transition area (in addition to the subfields).
seven samples were from the Massachusetts General Hospital Autopsy (Veterans Administration Medical Center, Bedford, VA). The other two different sources. Eight of the samples were from the Framingham Heart Study and Boston University Alzheimer’s Disease Center. The rest of the paper is organized as follows. Section "Atlas construction" describes the MRI data, delineation protocol and mathematical framework that were used to build the statistical atlas, shows the resulting images, and presents results on three datasets with different resolutions and types of MRI contrast. Finally, "Discussion and Conclusion" concludes the article.

Atlas construction

The statistical atlas that we propose is built from a combination of ex vivo and in vivo MRI training data. Here we first describe the acquisition ("Autopsy brain samples and ex vivo MRI acquisition") and manual labeling ("Manual segmentation of ex vivo MRI data: anatomical definitions") of the ex vivo data. Next, we introduce the in vivo training data we used ("In vivo training MRI data"). The algorithm to build the atlas is described in "Algorithm for atlas construction", and the resulting atlas presented in "Statistical atlas: volumes of subregions and sample slices".

Autopsy brain samples and ex vivo MRI acquisition

The ex vivo data consists of fifteen autopsied brain hemispheres from two different sources. Eight of the samples were from the Framingham Heart Study and Boston University Alzheimer’s Disease Center (Veterans Administration Medical Center, Bedford, VA). The other seven samples were from the Massachusetts General Hospital Autopsy Service (Massachusetts General Hospital, Boston, MA). The samples consisted of whole brain hemispheres (left n = 8, right n = 7) from 15 different subjects. Ten of the subjects did not have any neurological conditions, whereas four of them had mild AD and one had mild cognitive impairment (MCI). Eight samples were fixed with periodate–lysine–paraformaldehyde (PLP), and the other seven were fixed with 10% formalin. The demographics of the ex vivo samples were the following: age at death was 78.6 ± 11.9 years, 35.7% were females, 53.3% were left hemispheres and the post-mortem interval was less than 24 h in all cases for which this information was available. The demographics are detailed in Table 1.

A block of tissue including the hippocampus was excised from each ex vivo sample. Depending on its size, the block was placed in either a plastic cylindrical centrifuge tube (60 ml, 3 cm diameter) or, if it did not fit, inside a bag filled with PLP and sealed. In the latter case, air was pumped out using a needle and a vacuum pump in order to minimize the number and size of air bubbles in the samples. Two different pumps were used in the process: a DV-185 N-250 Platinum 10 CFM by JB (Aurora, IL), and a S413801 by Fisher Scientific (Hampton, NH).

The tissue block was subsequently scanned in a 7 T Siemens scanner using a 3D FLASH sequence with TR = 50 msec, TE = 25 msec, α = 20°. Two of the samples were scanned at 0.1 mm isotropic resolution, seven at 0.12 mm, five at 0.15 mm and one at 0.2 mm (see Table 1). Three different coils were used in the acquisition, accommodating variations in sample size: a 4-turn solenoid coil (28.5 mm inner diameter, 44 mm length), a 4-channel phased-array (a linear array of loop coil elements each with 5 cm coil diameter, 1.5 cm overlap between adjacent elements, 16 cm in length) and a small birdcage (24 rings, outer diameter = 19.7 cm, inner diameter = 19.3 cm, length = 12 cm). Despite the fact that different coils were used to scan the different samples, the output images were comparable in quality. The whole procedure received IRB approval before its execution by the Partners Human Research Committee. Fig. 1 displays some sample slices of the data.

Manual segmentation of ex vivo MRI data: anatomical definitions

In this section we describe the protocol for manually labeling the hippocampal subregions in the ex vivo data. The protocol was specifically designed for this study, and is largely based on the histology and morphometry from (Rosen and Van Hoesen, 1987), and partly also on (Lorente de No, 1934; Insauti and Amaral, 2011; Green and Mesulam, 1988). The Duvernoy atlas (Duvernoy, 1988) was also used as an aid in the delineation process. The set of annotated labels, along with the protocol for their annotation, is described in Table 2. The descriptions in the table are based on ex vivo contrast, not histological data. Given the excellent resolution of the ex vivo MRI data (100 μm), most of the

<table>
<thead>
<tr>
<th>Case #</th>
<th>Age</th>
<th>Gender</th>
<th>Laterality</th>
<th>Resolution</th>
<th>Diagnosis</th>
<th>PMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>N/A</td>
<td>Male</td>
<td>Left</td>
<td>100 μm</td>
<td>AD</td>
<td>N/A</td>
</tr>
<tr>
<td>2</td>
<td>N/A</td>
<td>Female</td>
<td>Left</td>
<td>120 μm</td>
<td>AD</td>
<td>21 h</td>
</tr>
<tr>
<td>3</td>
<td>91</td>
<td>Female</td>
<td>Right</td>
<td>120 μm</td>
<td>Control</td>
<td>16 h</td>
</tr>
<tr>
<td>4</td>
<td>83</td>
<td>Female</td>
<td>Left</td>
<td>150 μm</td>
<td>AD</td>
<td>5.5 h</td>
</tr>
<tr>
<td>5</td>
<td>89</td>
<td>Female</td>
<td>Right</td>
<td>120 μm</td>
<td>AD</td>
<td>N/A</td>
</tr>
<tr>
<td>6</td>
<td>82</td>
<td>Male</td>
<td>Right</td>
<td>120 μm</td>
<td>Control</td>
<td>N/A</td>
</tr>
<tr>
<td>7</td>
<td>63</td>
<td>Male</td>
<td>Left</td>
<td>120 μm</td>
<td>Control</td>
<td>N/A</td>
</tr>
<tr>
<td>8</td>
<td>87</td>
<td>Male</td>
<td>Left</td>
<td>120 μm</td>
<td>MCI</td>
<td>21 h</td>
</tr>
<tr>
<td>9</td>
<td>67</td>
<td>Male</td>
<td>Right</td>
<td>150 μm</td>
<td>Control</td>
<td>12 h</td>
</tr>
<tr>
<td>10</td>
<td>N/A</td>
<td>Male</td>
<td>Left</td>
<td>150 μm</td>
<td>Control</td>
<td>6.5 h</td>
</tr>
<tr>
<td>11</td>
<td>N/A</td>
<td>Male</td>
<td>Right</td>
<td>150 μm</td>
<td>Control</td>
<td>N/A</td>
</tr>
<tr>
<td>12</td>
<td>N/A</td>
<td>Male</td>
<td>Left</td>
<td>150 μm</td>
<td>Control</td>
<td>N/A</td>
</tr>
<tr>
<td>13</td>
<td>60</td>
<td>Male</td>
<td>Right</td>
<td>120 μm</td>
<td>Control</td>
<td>&lt;24 h</td>
</tr>
<tr>
<td>14</td>
<td>86</td>
<td>Female</td>
<td>Left</td>
<td>100 μm</td>
<td>Control</td>
<td>12–24 h</td>
</tr>
<tr>
<td>15</td>
<td>N/A</td>
<td>N/A</td>
<td>Right</td>
<td>200 μm</td>
<td>Control</td>
<td>N/A</td>
</tr>
</tbody>
</table>
subregion boundaries were visible in the images, but subtle transitions were still difficult to distinguish. Another source of differences between our ex vivo MRI labels and histology is that many boundaries are oblique (rather than perpendicular) to the imaging planes. Nonetheless, we used previously published anatomical contrast to guide us (Augustinack et al., 2005; Fischl et al., 2009; Augustinack et al., 2010; Augustinack et al., 2013). We also used neuroanatomical knowledge of particular layers to help us identify boundaries.

Distinguishing the boundaries between subiculum, CA1, CA2 and CA3 was more difficult due to the lack of image contrast between those subfields, but we used the pyramidal layer thickness and pyramidal layer intensity for this. We also combined the knowledge of location and pyramidal layer thickness to determine the subregions: the subiculum is widest, CA1 is thinner than subiculum, CA2 is thinner than CA1, and finally CA3 is the thinnest of the subfields. To distinguish the subicular boundaries, we used neighboring neuronal groupings such as entorhinal layer II islands (Augustinack et al., 2005), presubiculum clouds (Green and Mesulam, 1988), and reduced laminations in parasubiculum (compared with presubiculum and entorhinal cortex, Green and Mesulam, 1988). A full description of the histologic architecture is beyond the scope of this work because the atlas described here is based on ex vivo contrast. Nonetheless, our ex vivo MRI delineations represent a significant improvement over the previous FreeSurfer hippocampal segmentation (that only used geometric properties), and are much closer to the underlying subregion boundaries.

Seven manual labelers that were supervised by J.C.A. used the protocol described in Table 2 to annotate the subregions in the 15 ex vivo scans. Annotating each scan took approximately 60 h; a single hippocampus at this resolution contains more voxels than an in vivo scan of an entire brain. The annotations were made using Freeview, a visualization tool that is included in FreeSurfer. The first step in the protocol was to rotate the MRI volume to align the major axis of the hippocampus with the perpendicular direction to the coronal view. Then, assuming a left hippocampus (in right hemispheres, the image was flipped for delineation), the subregions were labeled using the definitions in Table 2. When bubbles were present in the images, the labelers filled them with the label of the structure they believe would be under the bubble. Since this introduces noise in the manual labels, minimizing the number and size of air bubbles in the sample prior to acquisition was crucial. The delineations were made in coronal view, while using information from the other two views (sagittal and axial) to guide the tracing; even though this might lead to slightly jagged boundaries in sagittal and axial view, this roughness is averaged out when the manual segmentations are downsampled and combined into the probabilistic atlas (see Statistical atlas: volumes of subregions and sample slices). In order to ensure the consistency between the manual labelers, J.E.I. and J.C.A. evaluated their delineations and served as quality control for each case, refining the segmentations where necessary. Sample manual tracings, along with the color coding of the subregions (for visualization purposes), are shown in Fig. 2.

In vivo training MRI data

Learning the spatial distribution of labels surrounding the hippocampus from the ex vivo data requires manual delineation of its neighboring structures. Even though it would be possible to trace these structures on ultra-high resolution, ex vivo MRI data, such approach would represent an unnecessary labeling effort for two reasons. First, the neighboring structures only need to provide a course context to assist the segmentation of the hippocampal subregions, and therefore do not require labeling at ultra-high resolution, which is extremely time consuming; delineation at standard resolution (i.e., ~1 mm) is sufficient. And second, there is already a number of publicly available and proprietary in vivo datasets for which such structures have already been manually labeled.

These are the motivations for using an additional training dataset consisting of in vivo, whole brain MRI scans. The dataset consists of T1-weighted scans from 39 subjects (19 males, 20 females, mean age: 56.3 years, 29 controls, 10 mildly demented) acquired with a MP-RAGE sequence in a 1.5 T scanner with the following parameters: TR = 9.7 ms, TE = 4 ms, TI = 20 ms, flip angle = 10°, 1 mm, isotropic resolution. Thirty-six brain structures, including the whole left and right hippocampi, were manually delineated using the protocol described in (Caviness et al., 1989); see sample slices, as well as a qualitative comparison with the ex vivo delineation protocol, in Fig. 3. Note that these are the same subjects that are used to construct the probabilistic atlas in FreeSurfer.
Bayesian algorithm to optimize the topology of the mesh, which is a model that uses Bayesian inference to learn the parameters of the mesh from manual annotations, assuming that its topology is fixed. Finally, we introduce a Bayesian algorithm to optimize the topology of the mesh, which is a model selection problem. Throughout the rest of this section, we will assume that all the training samples correspond to left hippocampi; samples from right hippocampi are simply flipped before being fed to the algorithm.

**Algorithm for atlas construction**

Here we describe the procedure to build the probabilistic atlas from in vivo and ex vivo data. We will first describe the underlying model, which is based on a tetrahedral mesh representation. Then, we will use Bayesian inference to learn the parameters of the mesh from manual annotations, assuming that its topology is fixed. Finally, we introduce a Bayesian algorithm to optimize the topology of the mesh, which is a model selection problem. Throughout the rest of this section, we will assume that all the training samples correspond to left hippocampi; samples from right hippocampi are simply flipped before being fed to the algorithm.

**Underlying model**

To build the probabilistic atlas of the hippocampal formation from ex vivo and in vivo data, we developed a generalization of our previous method (Van Leemput et al., 2009) that can deal with partial...
Fig. 2. Eight coronal slices from Case 14 and corresponding manual annotations. The slices are ordered from anterior to posterior. Sagittal and axial slices, as well as 3D renderings of the manual segmentation are shown in the supplementary material (Fig. 15, Fig. 16 and Fig. 17).

Fig. 3. In vivo dataset and comparison with ex vivo images. (a) Sagittal slice in vivo. (b) Corresponding manual delineation of brain structures; note that the hippocampus (in yellow) is labeled as a single entity. (c) Coronal slice in vivo. (d) Corresponding manual delineation. (e) Close-up of the hippocampus (in yellow) on a sagittal slice in vivo. (f) An approximately corresponding slice from Case 12 of the ex vivo dataset. (g) Close-up of the hippocampus on a coronal slice in vivo. (h) An approximately corresponding slice from Case 12 (ex vivo).
information. The algorithm aims to produce a compact tetrahedral mesh representation of the atlas, in which each vertex has an associated vector of probabilities for the different hippocampal subregions and surrounding structures. The topology and resolution of the mesh are locally adaptive to the shape of each anatomical region i.e., coarse in uniform regions and fine around convoluted areas. The difference with respect to the original method is that we no longer assume that all the labels of the training dataset are readily available: for the ex vivo data, the labels for the surrounding structures are not given, and for the in vivo data, the hippocampal subregions are not available. Instead, we observe a modification in which some sets of labels have been collapsed into more general labels in a deterministic fashion (Iglesias et al., 2015). The function that collapses the labels is different for each training dataset: for the ex vivo samples, it collapses all the non-hippocampal structures into a single, generic background label. For the in vivo data, the function collapses all the hippocampal subregions into a single label corresponding to the whole hippocampus.

Specifically, let there be M label volumes $C_m$, $m = 1, 2, ..., M$, derived from in vivo or ex vivo data. Each label volume $C_m = \{x_i \mid i = 1, 2, ..., I\}$ has $I$ voxels, where each voxel has a manual label belonging to one of $P$ possible collapsed classes: $c_i \in \{1, 2, ..., P\}$. We model these label images as having been generated by the following process (illustrated in Fig. 4):

a) A tetrahedral mesh covering the image domain is defined. This mesh is described by the position of its $N$ vertices $x^r = \{x_n \mid n = 1, 2, ..., N\}$ and their connectivity $\kappa$. Henceforth, we will refer to $x^r$ as the reference position of the mesh. Each mesh node has an assigned vector of label probabilities $\alpha = (\alpha_1, \alpha_2, ..., \alpha_P)$, where $(1, 2, ..., L)$ is the set of labels before collapsing. The probabilities satisfy $\alpha_0 \geq 0$, $\sum_{i=0}^{L} \alpha_i = 1$.

b) $M$ deformed meshes are obtained by sampling $M$ times from the following probability distribution:

$$p(x^m \mid \beta, \kappa) \propto \exp \left[ - \frac{U(x^m, \kappa)}{\beta} \right] = \exp \left[ - \sum_{r=1}^{T} \frac{U^r_i(x^m, x^r)}{\beta} \right],$$

where $x^m$ is the deformed mesh position, $T$ is the number of tetrahedra in the mesh, $\beta$ is a mesh flexibility parameter, and $U^r_i(x^r)$ is a penalty that goes to infinity if the Jacobian determinant of the $i$th tetrahedron’s deformation becomes zero (Ashburner et al., 2000). This deformation model allows the mesh to describe a broad spectrum of hippocampal shapes, while preventing the Jacobian determinant of

the deformation of each tetrahedron from becoming zero (which is equivalent to collapsing it) or negative (which is equivalent to reversing its orientation). By avoiding collapses and orientation reversals of the tetrahedra, we ensure that the topology of the mesh is preserved. Throughout the rest of this paper, we will assume that $\beta$ is a known constant.

c) From each deformed mesh $x^m$, a latent label $l_i^m \in \{1, 2, ..., L\}$ is generated for each voxel $i$ by sampling from the label probabilities given by the mesh. At non-vertex locations, these probabilities are computed using barycentric interpolation: $p_i(l_i^m \mid \alpha^m, \kappa) = \sum_{n=1}^{N} \alpha_n^m \delta_{l_i^m}(x_i)$, where $p_i$ is the prior probability at voxel $i$, $x_i$ is the spatial location of voxel $i$, and $\delta_{l_i^m}(\cdot)$ is an interpolation basis function attached to node $n$ of mesh $m$ – see details in (Van Leemput, 2009). Assuming conditional independence of the labels between voxels given the deformed mesh position $x^m$, we have that:

$$p(L_m \mid \alpha^m, \kappa) = \prod_{i=1}^{I} p_i(l_i^m \mid \alpha^m, \kappa),$$

where $L_m = \{l_1^m, ..., l_I^m\}$ is the $m$-th latent label volume.

d) Finally, the observed label volumes are given by $c_i = f_\alpha(l_i^m)$ (for in vivo volumes) and $c_i = f_{\text{ex}}(l_i^m)$ (for ex vivo volumes). The function $f_\alpha$ collapses all the hippocampal subregion labels into a single, global hippocampal label, whereas $f_{\text{ex}}$ collapses all the non-hippocampal labels into a single, generic background label. Therefore, we can write:

$$p_i(c_i^m \mid \alpha^m, \kappa) = \sum_{k \in C} p_i(k \mid \alpha, \kappa), \quad (1)$$

where $k \in C$ denotes and its mapping from $L$ to $C$ is bijective (i.e., no labels are collapsed), the generative model is assumed to be as in (Van Leemput, 2009).

Given this probabilistic model, the construction of the atlas is equivalent to solving the following inverse problem: given a set of (collapsed) label volumes (i.e., manual segmentations), we search for the atlas that most likely generated them according to the model. We use Bayesian inference to find the answer, as detailed below.

Optimization of model parameters — mesh deformations and atlas probabilities

Assuming that the mesh connectivity $\kappa$ and reference position $x^r$ are known, the problem to solve is:

$$\{\alpha, \{x^m\}\} = \arg \max_{\alpha, \{x^m\}} \beta(\{x^m\} \mid \kappa, \beta),$$

![Fig. 4. Illustration of the generative model of the manual labels for ex vivo (top) and in vivo (bottom) MNI.](image-url)
where $\alpha$ and $\{x^m\}$ represent the most likely atlas probabilities and atlas deformations, respectively. Using Bayes’ rule, we have:

$$\{\alpha, \{x^m\}\} = \arg\max_{\alpha, \{x^m\}} p(C_m|\alpha, \{x^m\}, K)p(\{x^m\})p(\alpha|\beta, x', K) = \arg\max_{\alpha} \prod_{m=1}^{M} p(C_m|\alpha, x^m, K)p(\{x^m\})p(\alpha|\beta, x', K),$$

where we have assumed a flat prior for $\alpha$, i.e., $p(\alpha) \propto 1$. Now, taking the logarithm of Eq. (2), and expanding the sum over voxels and hidden labels (Eq. (1)), we obtain the objective function for atlas building:

$$L(\alpha, \{x^m\}; \{C_m\}, x', \beta, K) = \sum_{m=1}^{M} \left\{ \log p(\{x^m\}) + \sum_{i=1}^{I} \sum_{k \in c_i^m} p_i(k|x^m_m, K) \right\}.$$

The first term in $L$ in Eq. (3) represents the (negated) cost of warping the mesh according to the deformation model we have borrowed from (Ashburner et al., 2000), whereas the second term represents the data fidelity. We solve Eq. (3) by optimizing $\alpha$ with $\{x^m\}$ fixed and vice versa until convergence. Updating $\alpha$ amounts to re-estimating the label probabilities at each spatial location, whereas the optimization of $\{x^m\}$ with $\alpha$ fixed represents a group-wise, nonrigid registration process. As shown below, the update equations are analogous to those given in analytical form by:

$$\frac{\partial L(\alpha, \{x^m\}; C_m, x', \beta, K)}{\partial x^m} = -\frac{1}{\beta} \sum_{i=1}^{I} \sum_{k \in c_i^m} \frac{\partial p_i(k|x^m_m, K)}{\partial x^m} + \sum_{k \in c_i^m} \frac{\partial p_i(k|x^m_m, K)}{\partial x^m} P_{i},$$

or

$$\frac{\partial L(\alpha, \{x^m\}; C_m, x', \beta, K)}{\partial x^m} = -\frac{1}{\beta} \sum_{i=1}^{I} \sum_{k \in c_i^m} \frac{\partial p_i(k|x^m_m, K)}{\partial x^m} + \sum_{k \in c_i^m} \frac{\partial p_i(k|x^m_m, K)}{\partial x^m} P_{i},$$

where $W_{m,k}$ is given by the E Step:

$$W_{m,k} = \frac{\alpha_k^m \phi_n^m(x_k)(k \in c_i^m)}{\sum_{k \in c_i^m} \alpha_k^m \phi_n^m(x_k)},$$

It can easily be shown that the maximum of $Q(\alpha, \alpha)$ is attained at (M step):

$$\alpha_k^m = \frac{\sum_{m=1}^{M} \sum_{i=1}^{I} \sum_{k \in c_i^m} W_{m,k}}{\sum_{m=1}^{M} \sum_{i=1}^{I} \sum_{k \in c_i^m} 1}.$$

As mentioned previously, the optimization scheme in Van Leemput (2009) is recovered when the mapping from $I$ to $C$ is bijective.

**Optimization of mesh topology — model selection**

So far we have assumed that the mesh connectivity and reference position were fixed. Optimizing the mesh topology is important to avoid overfitting to the training data due to the large variability in neuroanatomy across subjects. For instance, at a given spatial location, an atlas built upon a small training dataset could incorrectly assign a zero probability for a given label, if it was not present in any of the training volumes. This problem can be partly overcome by smoothing the atlas (Ashburner and Friston, 2001). As shown in Van Leemput (2009), the mesh topology can be optimized such that an automatically estimated amount of blurring is introduced in the atlas, allowing it to generalize well to previously unseen data. Following Van Leemput’s framework, we compare meshes with different topologies by evaluating their so-called evidence, which expresses how probable the observed training data is for each mesh topology. To compute the model evidence, we follow (Van Leemput, 2009), with the difference that we replace $p_i(l_i^m|\alpha, x^m_m, K)$ by $p_i(l_i^m|\alpha, x^m_m, K) = \sum_{k \in c_i^m} p_i(k|x^m_m, K)$. Recovering our previous algorithm if $f_i^m$ is bijective. The evidence is given by:

$$p((C_m)|\beta, x') \approx \frac{1}{Z} \left\{ \prod_{j=1}^{J} \left[ \sum_{l=1}^{L} \left( N_h + 1 \right) \left( N_b + 1 \right) \right] \right\} \left( \prod_{m=1}^{M} \sum_{k \in c_i^m} p_i(k|x^m_m, K) \right),$$

where $N_h = \sum_{m=1}^{M} \sum_{k \in c_i^m} \sum_{l=1}^{l} W_{m,k}$, $x^m_m$ are the minimizers of Eq. (3), and where the constant $Z$ includes a number of factors that only depend significantly on $\beta$ (which is kept fixed in this work).

In order to compute the most likely connectivity $K$ and reference position $x'$ using Eq. (6), these variables are first initialized with a high-resolution, regular mesh. Then, the mesh parameters $\{\alpha, \{x^m\}\}$ are optimized by solving the problem in Eq. (3). Finally, the mesh is simplified by repeatedly visiting each edge (in random order), comparing the effect on the evidence of either keeping the edge while optimizing $\alpha$ and $x'$, or vice versa.

**Data preprocessing**

To build the atlas, all the training label volumes must be in the same coordinate space. For this purpose, we carried out the following preprocessing steps: 1. manually rotating the FreeSurfer whole brain atlas — described in (Fischl et al., 2002) — so that the major axis of the left hippocampus was aligned with the anterior–posterior axis; 2. extracting from the atlas a binary mask corresponding to the voxels for which the left hippocampus is the most likely label; 3. left–right flipping all right hippocampi in the training data; 4. affine co-registration of the training data (using binary hippocampal masks) to the left hippocampus of the rotated FreeSurfer atlas, using sum of squares as metric; 5. resampling to 0.25 mm resolution (which is above the limit that can currently be achieved in vivo brain MRI scanning); and 6. cropping the training data (using binary hippocampal masks) to the left hippocampus of the rotated FreeSurfer atlas, using sum of squares as metric.
a bounding box around the hippocampi, leaving a 10 mm margin with respect to the boundary as defined by the FreeSurfer atlas — which yielded volumes of dimension $131 \times 241 \times 99$ voxels. The resampling of label volumes was carried out by resampling binary masks for each label separately using cubic interpolation, and picking the label with the maximum value at each voxel. This approach mitigates the block effect caused by nearest neighbor interpolation. This preprocessing pipeline yielded 93 (78 in vivo, 15 ex vivo) volumes with the same size and resolution, in which the hippocampi were affinely aligned, and which were then fed to the algorithm described above. The mesh flexibility parameter was set to $\beta = 0.15$; visual inspection of the results of pilot segmentation experiments using the algorithm proposed in Section "Quantitative results on standard resolution ADNI T1 data" below and 10 T1 scans of the OASIS dataset showed that this value of $\beta$ provided a good compromise between specificity and generalization ability.

**Statistical atlas: volumes of subregions and sample slices**

Coronal slices from the resulting statistical atlas are displayed in Fig. 5. The atlas has a total of 18,417 vertices, so the dimensionality of $x$ — which is equal to the number of degrees of freedom of the nonlinear deformation — is approximately three times this value, i.e., ca. 55,000.
Table 3

Mean hippocampal volumes derived from the proposed in vivo atlas (FreeSurfer v6.0), the in vivo atlas (FreeSurfer v5.3), the UPenn atlas (Yushkevich et al., 2009), and two histological studies of the hippocampus (Simic et al., 1997; Harding et al., 1998). All the volumes are in cubic millimeters. The UPenn atlas is available at http://www.nitrc.org/projects/pennhippoatlas/; we computed its volumes from the most likely labels.

<table>
<thead>
<tr>
<th>Method</th>
<th>Age</th>
<th>PARA.</th>
<th>PRE.</th>
<th>SUB.</th>
<th>CA1</th>
<th>CA2/3</th>
<th>CA4</th>
<th>DG</th>
<th>Tail</th>
<th>ML</th>
<th>HATA</th>
<th>FIM.</th>
<th>ALV.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ex vivo atlas</td>
<td>78.6</td>
<td>51</td>
<td>254</td>
<td>337</td>
<td>520</td>
<td>179</td>
<td>211</td>
<td>244</td>
<td>465</td>
<td>466</td>
<td>50</td>
<td>92</td>
<td>320</td>
</tr>
<tr>
<td>In vivo atlas</td>
<td>22–89</td>
<td>420</td>
<td>521</td>
<td>330</td>
<td>906</td>
<td>496</td>
<td>350</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UPenn</td>
<td>N/A</td>
<td>1754</td>
<td>86</td>
<td>201</td>
<td>167</td>
<td>127</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simica</td>
<td>80.2</td>
<td>404</td>
<td>591</td>
<td>139</td>
<td>197</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Harding</td>
<td>69</td>
<td>321</td>
<td>529</td>
<td>731</td>
<td>138</td>
<td>169</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This study considered CA4 and DG a single structure.

* For this study, we left out the AD cases and averaged the volumes from the elderly controls only.

![Illustration of the generative model of MRI images (monomodal data).](image)

Ex vivo atlas also displays the original in vivo atlas currently distributed with FreeSurfer for comparison. Both atlases show similar levels of blurring in the label probabilities that allow them to avoid overfitting the training data and generalize well to test images. However, the ex vivo atlas follows the internal structures of the hippocampus with much more accuracy than the in vivo version, which relies much more on geometric features—see for example the vertical boundary of CA1 (in red) in Fig. 6. In fact, the in vivo atlas does not describe the molecular layer (dark brown), which is the main feature that will allow the atlas to segment high-resolution T1/T2 data of the hippocampus. The figure also displays the UPenn atlas from (Yushkevich et al., 2009), which has lower resolution than the presented ex vivo atlas, has fewer subregions, and does not model additional surrounding (extrahippocampal) structures.

The improved accuracy of the ex vivo atlas is also reflected on the volumes of the subregions. Table 3 shows the average subregion volumes for the in vivo (FreeSurfer v5.3) and ex vivo atlases (FreeSurfer v6.0), for the UPenn atlas (Yushkevich et al., 2009), and for two different previous histological studies: (Simic et al., 1997; Harding et al., 1998).

Compared with the in vivo atlas, the ex vivo counterpart models a larger number of subregions and also yields volumes for CA1 and (especially) CA2/3 that are much closer to those reported by the referenced histological studies. The UPenn atlas yields accurate volumes for CA4 (for which it agrees well with the ex vivo atlas), but underestimates the volume of CA2/3 and largely overestimates the volume of CA1. Both the ex vivo atlas and the UPenn atlas overestimate the volume of the dentate gyrus, compared with the histological studies.

### Segmentation of in vivo MRI data

In this section, we first introduce an algorithm to segment in vivo MRI data using the proposed statistical atlas (“Algorithm for segmenting in vivo scan”). We subsequently present segmentation results on three different publicly available datasets with different resolutions and MRI contrasts, in order to demonstrate the ability of the algorithm to adapt to monomodal and multimodal data acquired with different MRI protocols and different hardware platforms. In “Qualitative segmentation results on high resolution T1/T2 data from Winterburn et al. (2013),” the algorithm is applied to high resolution (0.6 mm isotropic) T1/T2 data, segmenting the T1 and T2 channels both independently and simultaneously. In “Quantitative results on ADNI T1/T2 data,” we use 1 mm isotropic T1 data and corresponding high-resolution T2 images (0.4 mm in-plane, 2 mm slice thickness) to find group differences in subregion volumes between MCI subjects and elderly controls. Finally, in “Quantitative results on standard resolution ADNI T1 data” we automatically segment 1 mm isotropic T1 scans of AD subjects and elderly controls to compute volumes that are used as feature in classification experiments.

#### Algorithm for segmenting an in vivo scan

Here we describe the algorithm to segment an in vivo MRI scan given the atlas built in “Atlas construction.” We first describe the underlying generative model, which builds on that of “Algorithm for atlas construction,” and then detail an algorithm to estimate the segmentation using Bayesian inference. As in the previous section, we will assume that we are segmenting a left hippocampus. If we wish to segment the right hippocampus, we simply flip the atlas in the left–right direction.

#### Generative model

The built atlas can be used to segment a previously unseen MRI scan acquired with any type of MRI contrast (monomodal or multimodal), using the generative model displayed in Fig. 6. The first layers of the model are the same as in Fig. 4: the atlas, which defines prior probabilities of label occurrences in space, is first deformed according to the model proposed in (Ashburner et al., 2000), and then labels are sampled at each voxel location to obtain a segmentation L. The difference is that now this segmentation is connected to image intensities through a likelihood term, for which we assume that a Gaussian distribution is associated with each label.

In order to reflect the fact that there is very little contrast between different white matter structures in structural MR images of the brain, we assume that the fimbria and the cerebral white matter belong to a global white matter class, described by a single Gaussian distribution. Likewise, the cerebral cortex, amygdala and hippocampal gray matter structures (para-, pre-, and subiculum, CA1–4, GC–DG, HATA) also are assumed to be part of a global gray matter class. The cerebrospinal fluid (CSF) structures (ventricles, hippocampal fissure) share a class as well. The diencephalon, thalamus, pallidum, putamen and choroid plexus each have independent intensity classes. The alveus and molecular layer could in principle be part of the global white matter class; however, due to their thin shape, they are often affected by partial voluming, so we allow them to have their own Gaussian parameters as well.
The observed MRI intensity image \( Y \) is assumed to have been generated by sampling a Gaussian distribution at each voxel \( i \), parameterized by the mean and covariance corresponding to its global class:

\[
p(Y|l) = \prod_i p(y_i|\mu_{G(l)}, \Sigma_{G(l)}) = \prod_i N(y_i; \mu_{G(l)}, \Sigma_{G(l)}),
\]

where \( G(l) \) represents the global class corresponding to label \( l \), \( \mu_c \) and \( \Sigma_c \) are the mean and covariance of the global tissue class \( G \), \( N(\cdot; \mu, \Sigma) \) represents the (possibly multivariate) Gaussian distribution with mean \( \mu \) and covariance \( \Sigma \), and \( y_i \) denotes the intensity in voxel \( i \).

The generative model is completed by a prior distribution on the Gaussian parameters \( \{\mu_c, \Sigma_c\} \). We use a normal-inverse-Wishart distribution for each class – which is the conjugate prior for a multivariate Gaussian distribution with unknown mean and covariance – with covariance-related hyperparameters set to zero (i.e., uninformative prior on the covariance structure):

\[
p(\{\mu_c, \Sigma_c\}) = \prod_c p(\mu_c, \Sigma_c) = \prod_c N(\mu_c; M_c, \nu_c > 0, 0) \propto \prod_c N(\mu_c; M_c, \nu_c^{-1} \Sigma_c) .
\]

Here \( \{M_c\} \) and \( \{\nu_c\} \) are the remaining hyperparameters of the normal-inverse-Wishart distribution. Their interpretation is that prior to observing any image data, we have an initial guess \( M_c \) of the mean of tissue class \( G \), which is assumed to have been obtained as the sample mean of \( \nu_c \) observations (note that \( \nu_c = 0 \) a uniform prior is obtained).

Segmentation as Bayesian inference

Using the described generative model, segmentation is cast as an optimization problem in a Bayesian framework – we search for the most likely labeling given the probabilistic atlas and the observed image intensities. Exact inference would require marginalizing over the model parameters \( x \) (the mesh deformation) and \( \{\mu_c, \Sigma_c\} \), which leads to an intractable integral. Therefore, we make the approximation that the posterior distribution of such parameters in light of the atlas and observed image intensities is heavily peaked. This allows us to first find the maximum-a-posteriori (MAP) estimates of the parameters, and then use these values to derive the segmentation:

\[
L = \arg \max_L p(L|Y, \alpha, x', \beta, K) \\
\approx \arg \max_L p(L|x, \{\mu_c\}, \{\Sigma_c\}, Y, \alpha, x', \beta, K),
\]

where the most likely model parameters are given by:

\[
\{x, \{\mu_c\}, \{\Sigma_c\}\} = \arg \max_X p(x|\{\mu_c\}, \{\Sigma_c\}|Y, \alpha, x', \beta, K).
\] (7)

Using Bayes rule, we rewrite the problem in Eq. (7) and obtain the following objective function for parameter estimation:

\[
\{x, \{\mu_c\}, \{\Sigma_c\}\} = \arg \max_X p(x|\{\mu_c\}, \{\Sigma_c\}|Y, \alpha, x', \beta, K) = \arg \max_X \log p(Y|x, \alpha, x', \beta, K) + \sum_c \log \int \sum_k p(Y|\mu_c, \Sigma_c) p(k|x, \alpha, x', \beta, K) d\mu_c.
\] (8)

where we have introduced the prior for global tissue class \( G \) as:

\[
p_G(x, \alpha, x', \beta, K) = \sum_k p(k|x, \alpha, x', \beta, K). \]

We solve Eq. (8) by alternately optimizing for the mesh deformation \( x \) and the Gaussian parameters \( \{\mu_c\}, \{\Sigma_c\} \). We update the mesh deformation \( x \) by optimizing Eq. (8) directly with a conjugate gradient algorithm, and the Gaussian parameters with an EM algorithm. In the E step, we perform a probabilistic label classification for each voxel:

\[
\Omega^G_i = \frac{p_i(Y|\mu_c, \Sigma_c)p_{G(i|x, \alpha, x', \beta, K)}}{\sum_k p_{G(k|x, \alpha, x', \beta, K)} p_{G(i|x, \alpha, x', \beta, K)}},
\]

and in the M step, the Gaussian parameters are updated as follows:

\[
\mu_c \leftarrow \frac{\nu_c M_c + \sum_i \Omega^G_i y_i}{\nu_c + \sum_i \Omega^G_i},
\]

\[
\Sigma_c \leftarrow \frac{\sum_i \Omega^G_i \{y_i - \mu_c\}(y_i - \mu_c)^T + \nu_c \{\mu_c - M_c\}(\mu_c - M_c)^T}{\sum_i \Omega^G_i}.
\]

Once the optimal parameters \( x, \{\mu_c\}, \{\Sigma_c\}\) have been estimated, the (approximately) optimal segmentation \( \arg \max_X p(L|x, \{\mu_c\}, \{\Sigma_c\}|Y, \alpha, x', \beta, K) \) can be computed voxel by voxel as:

\[
i = \arg \max_k p_i(Y|\mu_{G(k)}, \Sigma_{G(k)}) p_k(k|x, \alpha, x', \beta, K).
\]

Image preprocessing, algorithm initialization and computation of hyperparameters

To initialize the segmentation algorithm and compute the hyperparameters \( \{M_c, \nu_c\} \), we use the output from the standard FreeSurfer pipeline ("recon-all"), which operates on whole brain T1 data at 1 mm resolution. FreeSurfer produces a skull-stripped, bias field corrected volume that we use as T1 component of the input to the hippocampal segmentation algorithm. It also produces a segmentation of this whole brain volume into 36 structures. If additional channels (e.g., high-resolution T2) are available, they are first rigidly registered with the T1 scan with mutual information (FreeSurfer’s “mri_robust_register”), using the brain mask provided by FreeSurfer to eliminate the influence of non-brain tissue on the alignment. The resulting transform is used to map the (skull-stripped, bias field corrected) T1 data and its automated segmentation to the space of the additional scan. The mapped segmentation is used to skull strip the additional channels. The preprocessed data from all the available channels is then resampled to the voxel size at which we desire to compute the segmentation, which is equal to the resolution at which the atlas will be rasterized.

We position the cuboid region that the atlas models by mapping it to the image to be segmented with an affine, sum of squares based registration algorithm (implemented in "mri_robust_register"). The algorithm uses the binary hippocampal segmentation from FreeSurfer as target image, and a soft probability map for the whole hippocampus – computed from the mesh in reference position – as source image for the registration. Henceforth, we refer to the region covered by the mapped atlas cuboid as “atlas region of interest (ROI)”. The voxels outside the atlas ROI are not considered by the segmentation algorithm.

In addition to preprocessing the input data and computing the atlas ROI, the segmentation of the brain into 36 structures generated by FreeSurfer is also used compute the hyperparameters \( \{M_c, \nu_c\} \) as follows: for each global class \( G \), we first find all the voxels segmented
as such structure by FreeSurfer. Next, we set $M_G$ to the modality-wise median intensity of that structure. Then, we set $\nu_G$ to the number of voxels used in the estimation of the median. Using the FreeSurfer segmentation from the whole brain improves the estimate of the Gaussian parameters, especially when the number of voxels for a given class is small within the atlas ROI. For instance, it is not always easy to estimate the Gaussian parameters of the CSF from the atlas ROI, partly due to the presence of the choroid plexus. However, if we look at the whole brain, such parameters can be easily estimated from the full ventricles.

There are however four classes for which the hyperparameters are computed in a different manner: gray matter, white matter, alveus and molecular layer. For the gray and white matter, since there are so many voxels labeled as such in the whole brain, we only use those from the hippocampus and neighboring regions in the FreeSurfer parcellation of each hemisphere. This way, we take advantage of a relatively large number of voxels to inform the model while eliminating the drift in the hypermeans $\{M_c\}$ that could be caused by voxels far away from the hippocampus due to the MRI bias field. For the white matter, we use the superior occipital gyrus, the orbital part of inferior frontal gyrus and the opercular part of the inferior frontal gyrus of the corresponding hemisphere. For the gray matter, we use the parahippocampal, entorhinal and fusiform cortices, as well as the whole hippocampus and amygdala from the corresponding hemisphere.

The hyperparameters of the alveus and molecular layer are computed in a different way because, due to their thin shapes, they are more severely affected by the partial volume effect, such that the global white tissue class does not model their intensity distribution correctly (despite the fact that they are white matter structures). We compute the hyperparameters of these structures by mimicking the partial volume effect as follows. First, we rasterize the mesh at the initial position $x$ at the native resolution (0.25 mm isotropic). Next, we sample a label $l_i$ at each voxel $i$ from $l_i \sim p_i(k|\alpha,\lambda_i)$. to generate a sample segmentation $L$. Then, we assign to each voxel the mean intensity of its corresponding label, while assuming that the alveus and molecular layer belong to the global white matter class. Then, we blur this image with a Gaussian kernel that matches the resolution of this synthetic image to that of the test scan to segment, by setting its FHWM (in voxels) in each direction to the voxel size of the test scan in that direction divided by the native voxel size of the atlas (0.25 mm). Finally, $M_{alv}$ and $M_{ML}$ are set to the median intensity of the fimbria and molecular layer in the synthetic image, whereas we set $\nu_{alv}$ and $\nu_{ML}$ to the volume of these two structures provided by the atlas (see Section “Statistical atlas: volumes of subregions and sample slices”).

Qualitative segmentation results on high resolution T1/T2 data from Winterburn et al. (2013)

In this section, we show qualitative results on a publicly available dataset of high resolution T1/T2 data. The dataset (Winterburn et al., 2013)
(2013) is a public repository of T1 and T2-weighted scans of five subjects (two males, three females, ages 29–57) acquired on a 3 T GE scanner with an 8-channel head coil. Both the T1 and the T2 scans were acquired at 0.6 mm isotropic resolution, and then super-sampled to 0.3 mm isotropic. Manual delineations of five subregions are also available as part of the repository: CA1, CA2/3, dentate gyrus, molecular layer and subiculum. Note, however, that a direct evaluation through comparison of manual and automated segmentations (e.g., with Dice scores) is not possible due to the differences between our subregion labeling protocol (described in Manual segmentation of ex vivo MRI data: anatomical definitions) and theirs. Instead, we present qualitative results: since high resolution images are available for both the T1 and T2 channels (see sample slices in Fig. 7), we can compare the outputs produced by the segmentation algorithm on the T1 data, the T2 data, and both combined. When using a single channel in the segmentation, \( y_i \) is a scalar with the T1 (or T2) intensity, and \( \{\mu_G, \Sigma_G\} \) are also scalars with the means and variances of the tissue types. When we segment both channels simultaneously, \( y_i \) is a 2 × 1 vector with the T1 and T2 intensities at spatial location \( i \), while the means \( \{\mu_G\} \) are 2 × 1 vectors and the covariances \( \Sigma_G \) are 2 × 2 matrices. In this experiment, the work resolution is set to 0.3 mm — equal to the voxel size of the input scans.

Figs. 8 and 9 also show sample segmentations from “subject 3” in the dataset using the T1 scan alone, the T2 scan alone, and both scans simultaneously; segmentations from the other four subjects in the dataset are provided in the supplementary material (Fig. 20 and Fig. 21). Note that we have removed from the final segmentation the neighboring structures of the hippocampus, as well as the alveus; showing very little contrast in in vivo MRI due to its thin shape, its automated segmentation is often unreliable. The method effectively adapts to the MRI contrast in each case, successfully segmenting the hippocampus in all three scenarios. The segmentation based solely on the T1 image accurately captures the global shape of the hippocampus, but often under-segments the molecular layer (marked with blue arrows in the figures), as well as cysts and CSF pockets (red arrows), which are hardly visible in T1. These features are correctly segmented in the T2 image, which, on the other hand, captures the global shape of the hippocampus less accurately than the T1 scan, due to its poorer contrast between gray and white matter (see regions marked with yellow arrows in the figures). The output based on multimodal MRI data takes advantage of the information from both channels to produce a smoother, more accurate segmentation that combines the advantages of the T1 and T2 MRI contrasts.

Figs. 8 and 9 also show the corresponding manual segmentations from the original article (Winterburn et al., 2013). The agreement between the manual and automated segmentations is fair in general, but some differences can be found in the areas where the segmentation is poorly supported by the image contrast (e.g., the medial digitation in Fig. 8) and also in the hippocampal regions where our definitions of the subregions and theirs disagree. First, some of the labels of our protocol do not exist in their labeling scheme: tail, fimbria, GC–DG, HATA, parasubiculum and presubiculum. Second, even though the agreement of the protocols is good in the superior part of the hippocampus (see 3D renderings in Fig. 10), large differences in the definition of the subregions can be generally observed in the inferior part: our subiculum is largely part of their CA1, while our presubiculum and parasubiculum correspond approximately to their subiculum.

### Quantitative results on ADNI T1/T2 data

In this section, we present segmentation results on a dataset of 30 T2 MRI scans from the Alzheimer’s Disease Neuroimaging Initiative.

![Sample coronal slices of “subject 3” from (Winterburn et al., 2013), from anterior (left) to posterior (right). Top row: T1 image. Second row: T2 image. Third row: segmentation computed with the T1 scan. Fourth row: segmentation computed with T2 scan. Fifth row: segmentation computed with T1 and T2 scans simultaneously, overlaid on the T2 images. Bottom row: manual segmentation from the original study. The red arrow marks a CSF pocket, the blue arrows mark the molecular layer, and the yellow arrow marks the medial digitation.](image-url)
The ADNI (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The main goal of ADNI is to test whether MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to analyze the progression of MCI and early AD. Markers of early AD progression can aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as decrease the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California – San Francisco.

ADNI is a joint effort by co-investigators from industry and academia. Subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. These three protocols have recruited over 1500 adults (ages 55–90) to participate in the study, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the corresponding protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see http://www.adni-info.org.

The choice of the subset of ADNI scans for this analysis was motivated by the fact that these are the exact same images that were used in (Mueller et al., 2013), which includes MCI classification results derived from segmentations produced by a number of automated and semi-automated methods, as well as a manual delineation protocol that

Fig. 9. Sample sagittal slices of “subject 3” from Winterburn et al. (2013), from medial (left) to lateral (right). See caption of Fig. 8 for an explanation of the different rows. The red arrow marks a CSF pocket, the blue arrows mark the molecular layer, and the yellow arrows mark segmentation errors in the whole hippocampal shape.

Fig. 10. 3D renderings of segmentations of the high-resolution T1/T2 data. (a) Manual segmentation from Winterburn et al. (2013), anterior view. (b) Automated segmentation using T1 and T2 volume simultaneously, anterior view. (c–d) Inferior view of (a–b).
only considers five coronal slices in the body of the hippocampus (Mueller et al., 2010). Using this dataset enables direct comparison of our results with those from (Mueller et al., 2013). The 30 scans correspond to an acquisition protocol that has recently been added to the ADNI study, with the following parameters: TR = 8020 ms, TE = 50 ms, resolution 0.4 × 0.4 × 2.0 mm (coronal), 24–30 slices, acquisition time 8 min. Of the 30 scans, 16 correspond to subjects with early MCI (ages 74.3 ± 7.6), and the other 14 to age-matched healthy controls (ages 70.8 ± 7.2). Since these scans are part of the ADNI, the corresponding T1-weighted images (sagittal 3D MPRAGE scans at 1 mm resolution) are also available. A sample coronal slice of a T2 scan from ADNI is shown in Fig. 11a, and a sagittal slice (overlaid on the corresponding T1-weighted scan) in Fig. 11b. The sagittal slice shows how narrow the field of view of these scans sometimes is, failing to cover the whole hippocampus.

To segment these ADNI data, we take advantage of not only the high-resolution T2 images, but also the 1 mm isotropic T1 scans, which provide complementary information. On the one hand, the T2 data have good contrast between subregions, but large slice separation and a narrow field of view. On the other hand, the T1 scans provide isotropic data throughout the whole brain — though with little information on the sub-regions. Therefore, we segment both channels simultaneously, i.e., \( \{y_i\} \)

---

**Fig. 11.** (a) Coronal slice from T2 scan from ADNI, and close-up of the hippocampi. (b) Sagittal slice from a T2 scan from ADNI, overlaid on the corresponding T1 volume. This view illustrates the limited field of view of the T2 scans in ADNI. The in-plane resolution of the T2 scans is 0.4 mm, and the slice separation is 2 mm. The T1 scans are 1 mm isotropic.

**Fig. 12.** Inputs (T1, T2) and segmentations for five representative cases of the ADNI T1/T2 dataset. The resolution of the T1 scans is 1 mm isotropic, whereas the T2 scans have 0.4 mm in-plane resolution (coronal) and 2 mm slice separation. (a) A cyst being segmented as hippocampal fissure. (b) A case with good contrast, well-segmented. (c) A case where the partial volume effect has misguided the segmentation, such that part of the lateral ventricle (marked by the arrow) is labeled as CA2/3. (d) A case with poor contrast; the internal segmentation of the hippocampus is largely determined by the prior. (e) A case in which the field of view of the T2 scan does not cover the whole hippocampus; the segmentation of the tail relies solely on the T1 data. All slices are coronal except for (e), which is sagittal. More slices are displayed in the supplementary material (Fig. 22).
and \( \mu_c \) are a \( 2 \times 1 \) vectors and the covariances \( \Sigma_c \) are a \( 2 \times 2 \) matrices. The information of the T1 scan is particularly important when the T2 intensities are missing for some hippocampal voxels due to the limited field of view of the T2 scan (as in Fig. 11b, where the hippocampal tail is only visible in the T1 scan). In that case, the equations for Gaussian parameter optimization in Algorithm for segmenting an in vivo scan needs to be modified in order to account for this missing information — see details in Appendix A. In this experiment, the work resolution is 0.4 mm isotropic — equal to the in-plane resolution of the T2 images.

Fig. 12 shows slices from automated segmentations of some representative cases in the ADNI T1/T2 dataset. When the quality of the scan is good and the molecular layer is visible (as in Fig. 12b), the model generally produces a good segmentation. However, mistakes occur sometimes due to image artifacts. In Fig. 12c, CSF and white matter mix in the same voxel, making it resemble gray matter (partial volume effect) and thus misleading the segmentation algorithm. In Fig. 12d, motion artifacts render the molecular layer invisible. In such cases, the internal boundaries of the hippocampus are mostly determined by the statistical atlas (rather than image intensities), and to a lesser extent by other features such as cysts or the hippocampal fissure. Finally, Fig. 12e shows an example of when the lower-resolution T1 is most useful, which is when the field of view of the T2 scan does not cover the whole hippocampus. In that case, the algorithm can still produce a seamless, smooth segmentation of the whole hippocampal formation.

Direct evaluation of the produced segmentations would require manual delineations for the T2 data made with the ex vivo protocol from Section “Manual segmentation of ex vivo MRI data: anatomical definitions”, which are not available (and would be extremely difficult to make due to the lack of resolution). Therefore, we use indirect evaluation methods instead, based on assessing the ability of the automatically estimated subregion volumes to discriminate two populations (MCI and elderly controls) in a group analysis framework. First, we compute the subregion volumes from the soft segmentations with Eq. (9). Then, we correct these estimates for age and intracranial volume (ICV) by regressing them out with a general linear model. This step is important because the subregion volumes are strongly correlated with these two variables, which can easily confound the analysis — subjects with large ICV and/or of younger age are expected to have larger hippocampi; see for instance (Mueller et al., 2010). Moreover, such correction was used in (Mueller et al., 2013), so we used this correction as well in order to directly compare the results.

Once the corrected volumes have been computed, we compare the volumes of the two groups for each subregion independently with unpaired, two-sample t-tests. Since we have a strong hypothesis that the volume of the subregions does not increase in MCI or AD (except for the fissure, which tends to increase in AD), we can conduct one-tailed tests — for the fissure, we just invert the sign of the test. In addition, we also conduct a power analysis in which we compute the sample size required by the test to have a power of 0.50 — which was the value used in (Mueller et al., 2013) — as well as the power provided by the actual sample size of the dataset.

The group analysis of the subregion volumes is summarized in Table 4. The table also includes two sets of results reported in (Mueller et al., 2013): a group analysis for subregion volumes derived from the manual segmentation protocol, and another group analysis for the volumes given by the semi-automated algorithm from (Yushkevich et al., 2010). Our algorithm finds differences in the left and right CA1, molecular layer, dentate gyrus, CA4 and whole hippocampus, as well as the left fimbria and CA2/3 region. The manual annotations (for which the volumes were left-right averaged) show differences in CA1 and dentate gyrus, whereas Yushkevich’s semi-automated method yields significant differences in CA1–3 and CA4–DG — also with left–right averaged volumes. Our results are quite consistent with theirs, given that their methods do not consider the smaller subregions contained in our protocol. In addition, our results show strong consistency with prior work using

and \( \mu_c \) are a \( 2 \times 1 \) vectors and the covariances \( \Sigma_c \) are a \( 2 \times 2 \) matrices. The information of the T1 scan is particularly important when the T2 intensities are missing for some hippocampal voxels due to the limited field of view of the T2 scan (as in Fig. 11b, where the hippocampal tail is only visible in the T1 scan). In that case, the equations for Gaussian parameter optimization in Algorithm for segmenting an in vivo scan needs to be modified in order to account for this missing information — see details in Appendix A. In this experiment, the work resolution is 0.4 mm isotropic — equal to the in-plane resolution of the T2 images.

Fig. 12 shows slices from automated segmentations of some representative cases in the ADNI T1/T2 dataset. When the quality of the scan is good and the molecular layer is visible (as in Fig. 12b), the model generally produces a good segmentation. However, mistakes occur sometimes due to image artifacts. In Fig. 12c, CSF and white matter mix in the same voxel, making it resemble gray matter (partial volume effect) and thus misleading the segmentation algorithm. In Fig. 12d, motion artifacts render the molecular layer invisible. In such cases, the internal boundaries of the hippocampus are mostly determined by the statistical atlas (rather than image intensities), and to a lesser extent by other features such as cysts or the hippocampal fissure. Finally, Fig. 12e shows an example of when the lower-resolution T1 is most useful, which is when the field of view of the T2 scan does not cover the whole hippocampus. In that case, the algorithm can still produce a seamless, smooth segmentation of the whole hippocampal formation.

Direct evaluation of the produced segmentations would require manual delineations for the T2 data made with the ex vivo protocol from Section “Manual segmentation of ex vivo MRI data: anatomical definitions”, which are not available (and would be extremely difficult to make due to the lack of resolution). Therefore, we use indirect evaluation methods instead, based on assessing the ability of the automatically estimated subregion volumes to discriminate two populations (MCI and elderly controls) in a group analysis framework. First, we compute the subregion volumes from the soft segmentations with Eq. (9). Then, we correct these estimates for age and intracranial volume (ICV) by regressing them out with a general linear model. This step is important because the subregion volumes are strongly correlated with these two variables, which can easily confound the analysis — subjects with large ICV and/or of younger age are expected to have larger hippocampi; see for instance (Mueller et al., 2010). Moreover, such correction was used in (Mueller et al., 2013), so we used this correction as well in order to directly compare the results.

Once the corrected volumes have been computed, we compare the volumes of the two groups for each subregion independently with unpaired, two-sample t-tests. Since we have a strong hypothesis that the volume of the subregions does not increase in MCI or AD (except for the fissure, which tends to increase in AD), we can conduct one-tailed tests — for the fissure, we just invert the sign of the test. In addition, we also conduct a power analysis in which we compute the sample size required by the test to have a power$^6$ of 0.50 — which was the value used in (Mueller et al., 2013) — as well as the power provided by the actual sample size of the dataset.

The group analysis of the subregion volumes is summarized in Table 4. The table also includes two sets of results reported in (Mueller et al., 2013): a group analysis for subregion volumes derived from the manual segmentation protocol, and another group analysis for the volumes given by the semi-automated algorithm from (Yushkevich et al., 2010). Our algorithm finds differences in the left and right CA1, molecular layer, dentate gyrus, CA4 and whole hippocampus, as well as the left fimbria and CA2/3 region. The manual annotations (for which the volumes were left-right averaged) show differences in CA1 and dentate gyrus, whereas Yushkevich’s semi-automated method yields significant differences in CA1–3 and CA4–DG — also with left–right averaged volumes. Our results are quite consistent with theirs, given that their methods do not consider the smaller subregions contained in our protocol. In addition, our results show strong consistency with prior work using

### Table 4

<table>
<thead>
<tr>
<th>Structures</th>
<th>Left (this study)</th>
<th>Right (this study)</th>
<th>L/R average (Mueller et al.)</th>
<th>L/R average (Yushkevich et al.)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p value</td>
<td>Sample size</td>
<td>Power</td>
<td>Sample size</td>
</tr>
<tr>
<td>Parasubiculum</td>
<td>0.15</td>
<td>73</td>
<td>0.27</td>
<td>0.22</td>
</tr>
<tr>
<td>Presubiculum</td>
<td>0.21</td>
<td>121</td>
<td>0.20</td>
<td>0.08</td>
</tr>
<tr>
<td>Subiculum</td>
<td>0.14</td>
<td>66</td>
<td>0.29</td>
<td>0.18</td>
</tr>
<tr>
<td>CA1</td>
<td>0.03</td>
<td>24</td>
<td>0.58</td>
<td>0.02</td>
</tr>
<tr>
<td>CA2/3</td>
<td>0.02</td>
<td>18</td>
<td>0.71</td>
<td>0.10</td>
</tr>
<tr>
<td>CA4</td>
<td>0.03</td>
<td>23</td>
<td>0.59</td>
<td>0.04</td>
</tr>
<tr>
<td>GC–ML–DG</td>
<td>0.02</td>
<td>20</td>
<td>0.63</td>
<td>0.01</td>
</tr>
<tr>
<td>Molec. layer</td>
<td>0.02</td>
<td>20</td>
<td>0.67</td>
<td>0.05</td>
</tr>
<tr>
<td>Fimbria</td>
<td>0.12</td>
<td>10</td>
<td>0.68</td>
<td>0.04</td>
</tr>
<tr>
<td>Hippo. fissure</td>
<td>0.45</td>
<td>&gt;1000</td>
<td>0.10</td>
<td>0.26</td>
</tr>
<tr>
<td>HATA</td>
<td>0.20</td>
<td>111</td>
<td>0.21</td>
<td>0.07</td>
</tr>
<tr>
<td>Hippo. tail</td>
<td>0.45</td>
<td>&gt;1000</td>
<td>0.10</td>
<td>0.49</td>
</tr>
<tr>
<td>Whole hippo.</td>
<td>0.04</td>
<td>27</td>
<td>0.54</td>
<td>0.04</td>
</tr>
<tr>
<td>Entorhinal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perirhinal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$^6$ Defined as the probability of correctly rejecting the null hypothesis when it is false.
manual and semi-automated segmentation procedures on the same type of images: even if direct comparison across studies is not possible due to differences in labeling protocols, the differences in CA1, CA4–DG and whole hippocampus have been previously described in (Pluta et al., 2012), which is based on the method from (Yushkevich et al., 2010). In addition, our algorithm also finds statistically significant differences in the molecular layer and fimbria; a decline of the former, which shows great discrimination power in both the left and right hippocampus, has been previously described in the literature (Kerchner et al., 2010, 2012).

In order to quantify the impact of the high-resolution T2 scan in the segmentation, we also compare the ability of the subregion volumes to discriminate the two groups when they are measured with the combined T1/T2 data and when they are derived from the T1 images alone. Table 5 displays the p-values of the corresponding t-tests, which show that the measurements using both modalities better capture the differences in the subregions between the two groups. When only the T1 scans are used, the resolution is insufficient to distinguish the molecular layer. In this case, the volumes of the subregions depend largely on the volume and shape of the whole hippocampus, which reduces the ability of the individual subregions to separate the two groups. Only the fimbria, which is visible at 1 mm resolution, seems to provide comparable discrimination power after the high-resolution T2 scan is removed from the analysis.

Quantitative results on standard resolution ADNI T1 data

Here we present results on a dataset of standard resolution scans from the ADNI. The dataset consists of 400 baseline T1 scans from the study, for which high-resolution T2 data are not available. These scans were acquired with MPRAGE sequences at 1 mm isotropic resolution. The MRI data were processed through the standard FreeSurfer pipeline, including the current hippocampal subfield module, which is useful to compare the segmentations yielded by the in vivo atlas with those produced by the ex vivo atlas we are introducing in this study. This hippocampal subfield processing did not complete successfully for 17 scans (due to software crashes), which were removed from the analysis. The demographics of the remaining 383 subjects are as follows: 56.2% elderly controls (age 76.1 ± 5.6 years), 43.8% AD patients (age 75.5 ± 7.6); 53.6% males (ages 76.1 ± 5.6), 46.4% females (ages 75.9 ± 6.8). The resolution at which we rasterized the atlas and computed the segmentation in this experiment was 1/3 mm.

As in Section “Quantitative results on ADNI T1/T2 data”, we use the performance in a group analysis as a surrogate for segmentation quality. In this case, we validate the segmentation by comparing the subregion volumes of AD patients and elderly controls. As in the previous section, the resolution of the ADNI T1 scans is insufficient to distinguish the molecular layer, making the segmentation much less reliable. Therefore, the volumes of the subregions are largely determined by the whole hippocampal volume, such that most subregions show large discriminative power, but little differentiation between the subregions is observed. This is illustrated by the results in Table 6, which displays effect sizes for each subregion – for both the in vivo and ex vivo atlases – measured with Cohen’s d, i.e., the difference between two means divided by the standard deviation of the data. We used effect sizes rather than p-values in this experiment because, due to the large sample size and strong effect, all p-values were very small and the differences between the effects on the subregions were harder to appreciate. Large or very large effect sizes are observed for all subregions, except for the hippocampal fissure. The effect sizes given by the segmentations based on the in vivo and ex vivo atlases are quite similar, even though they are – on average – slightly larger for the ex vivo atlas. The main difference between the results of the two atlases is the effect for CA1, which is smaller in the in vivo version.

For the reason mentioned above (i.e., the lack of internal contrast of the hippocampus at 1 mm resolution), the results in Table 6 must be interpreted with caution. Therefore, we use another method to separate the two populations, based on a statistical classifier that discriminates the groups using all the subregion volumes simultaneously. The process is the following. First, we average the subregion volumes from the left and right hippocampi, as this boosts the power of the analysis without compromising the generalization ability of the classifier by increasing the dimensionality of the data. Subsequently, we perform a correction for age and ICV, for each subregion independently, in the same way as described in Section “Quantitative results on ADNI T1/T2 data”. Then, we concatenate all the subregion volumes of each subject into a vector, and use it as input to a linear discriminant analysis (LDA) classifier (Fisher, 1936). The use of a simple, linear classifier such as LDA ensures that the classification accuracy is mainly determined by the quality of the input data (i.e., the subregion volumes) rather than stochastic variations in the classifier.

We compared the performance of the segmentations given by the new ex vivo atlas with those produced by the in vivo atlas in FreeSurfer v5.3 (we used the “off-the-shelf” implementation of the segmentation algorithm). We used two metrics in the comparison: the area under the receiver operating characteristic (AUROC) of the classifiers and their maximum classification accuracy. The latter is given by the threshold corresponding to

<table>
<thead>
<tr>
<th>Structures</th>
<th>Left</th>
<th>Right</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ex vivo</td>
<td>In vivo</td>
</tr>
<tr>
<td>Parasubiculum</td>
<td>0.97</td>
<td>0.94</td>
</tr>
<tr>
<td>Presubiculum</td>
<td>1.34</td>
<td>1.27</td>
</tr>
<tr>
<td>Subiculum</td>
<td>1.79</td>
<td>1.76</td>
</tr>
<tr>
<td>CA1</td>
<td>1.28</td>
<td>1.22</td>
</tr>
<tr>
<td>CA2/3</td>
<td>1.37</td>
<td>1.32</td>
</tr>
<tr>
<td>CA4</td>
<td>1.45</td>
<td>1.39</td>
</tr>
<tr>
<td>GC–ML–DG</td>
<td>1.57</td>
<td>1.49</td>
</tr>
<tr>
<td>Molecular layer</td>
<td>1.65</td>
<td>1.50</td>
</tr>
<tr>
<td>Fimbria</td>
<td>0.67</td>
<td>0.58</td>
</tr>
<tr>
<td>Hippocampal fissure</td>
<td>0.78</td>
<td>0.72</td>
</tr>
<tr>
<td>HATA</td>
<td>1.45</td>
<td>1.40</td>
</tr>
<tr>
<td>Hippocampal tail</td>
<td>1.71</td>
<td>1.65</td>
</tr>
<tr>
<td>Whole hippocampus</td>
<td>0.12</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Table 6 Effect sizes (Cohen’s d) of the group analysis for the hippocampal subregion volumes in the AD discrimination task. Larger effect sizes correspond to larger differences between the two groups.

Fig. 13. ROC curves for the AD discrimination task using a LDA classifier on the hippocampal subregion volumes estimated by the ex vivo atlas (FreeSurfer v6.0) and the in vivo atlas (FreeSurfer v5.3 and earlier), as well as for discrimination based on whole hippocampal volume as estimated by FreeSurfer (“aseg”) and by the ex vivo atlas (adding up the volumes of the subregions).
the “elbow” of the receiver operating characteristic (ROC) curve, i.e., the point closest to FPR = 0, TPR = 1. We used leave-one-out cross-validation to compute the ROC. Significance in the difference in AUROC was assessed with a non-parametric test (DeLong et al., 1988). In addition, we also compared a classifier based solely on whole hippocampal volume, which allows us to quantify the benefit of using the subregion volumes with respect to the whole hippocampus. We used two different estimates of the volume: the sum of the subregion volumes given by the ex vivo atlas (except for the fissure) and the whole hippocampus estimate provided by FreeSurfer’s automated segmentation (“aseg”).

The ROC curves for the AD vs. elderly controls discrimination task are shown in Fig. 13, whereas the areas under the curves and accuracies are displayed in Table 7. The ex vivo atlas outperforms the in vivo counterpart, especially around the elbow of the ROC, which is the region where a classifier typically operates. The increment in the AUROC is moderate (1.4%), but statistically significant (p < 0.02). This indicates that the segmentations based on the ex vivo atlas provides more informative estimates of the volumes than those based on the in vivo version. Both the in vivo and the ex vivo atlas outperform the whole hippocampal segmentations, which yields 82% (“aseg”) and 84% (sum of subregions) classification accuracy. The difference in AUROC between the ex vivo atlas and the whole hippocampal segmentations is noteworthy (5.9% and 4.0%, respectively) and statistically significant (p < 0.01 in both cases). These results indicate that the subregion volumes carry useful information, even when the images display limited contrast on the internal subregion boundaries. Automated segmentations of a test scan are shown in Fig. 14.

Discussion and conclusion

In this paper we have presented the construction of a statistical atlas of the hippocampus at the substructure level using a combination of ex vivo and in vivo MRI data. Manual annotations of the hippocampal subregions (on ex vivo images) and of the neighboring structures (on in vivo data) were combined into a single atlas using a novel algorithm. Using Bayesian inference, the constructed atlas can be used to automatically segment the hippocampal subregions in in vivo MRI scans. Given the generative nature of the framework, the segmentation method is adaptive to MRI contrast and can naturally handle multi-contrast inputs. The segmentation algorithm was validated on three publicly available datasets with varying MRI contrast and resolution (Winterburn, ADNI T1/T2, and ADNI T1). We plan to release the atlas as part of the next release of FreeSurfer, replacing the in vivo atlas of the hippocampal subfield module in the current version of the package (v5.3).

The presented atlas improves previous high-resolution atlases of the hippocampus in several directions. Compared with the in vivo version, the ex vivo atlas was built upon data of much higher resolution, which allowed us to accurately trace the molecular layer with very little dependence on geometric criteria. As a consequence, the atlas yields subregion volumes that better matched values to previously reported histological studies (Simic et al., 1997; Harding et al., 1998). Compared to the ex vivo UPenn atlas presented in (Yushkevich et al., 2009), we have extended their work in four directions. First, we have scanned the samples at .13 mm isotropic resolution on average, which yields voxels four times smaller than those in their atlas. Second, we have modeled a larger number of hippocampal structures (13 vs. 5). Third, we have used a greater number of cases (15 vs. 5). And fourth, our ex vivo atlas models not only the hippocampal formation but also the surrounding structures. This is a critical difference, since it enables us to use the atlas in a Bayesian framework to directly segment in vivo MRI data of arbitrary contrast properties.

We have tested the ability of the volumes derived from automated segmentations given by the atlas to find differences between controls and subjects with MCI and AD. Using high-resolution T2 data, we can reliably fit the atlas to the internal structure to the hippocampus; this was not possible with the previous atlas in FreeSurfer (v5.3 or earlier), which did not model the molecular layer. In a group experiment with MCI subjects and controls, our new method reproduced the results of previous manual and semi-automated algorithms. Moreover, we found differences in the molecular layer and the fimbria, which the aforementioned methods do not segment. Since the molecular layer is a thin structure, it is possible that the lower volume estimates are due to motion artifacts, to which the MCI group is more susceptible. Regarding the fimbria, visual inspection of the images reveals that the appearance of this subregion tends to shift towards that of gray matter in aging and AD. It is important to note that we cannot conclude from our volumetric analysis whether this is a true biological process or the result of motion artifacts.

We also used the atlas to segment standard resolution (1 mm) T1 data. In this case, the molecular layer is not visible, and the fitting of the internal structure of the atlas mostly relies on the prior information encoded in it. Therefore, volumetric results from individual subregions must be interpreted with caution. Still, we hypothesize that the segmentations will be very useful as seed and target regions in functional and

![Fig. 14. Automated segmentation of the hippocampal subregions of a sample case from the ADNI T1 dataset (T1-weighted, 1 mm isotropic) using FreeSurfer automated segmentation (“aseg”), as well as the in vivo and ex vivo atlases. a) Slices of the segmentation. b) 3D renderings of their shape. The color map is the same as in Figs. 2 and 5.](image-url)
diffusion MRI studies, in which the large voxel sizes make the analysis less sensitive to small segmentation errors. Moreover, we have shown that the subregion segmentation carries, despite the lack of internal contrast of the hippocampus, useful information that is not conveyed by its whole segmentation: the ex vivo atlas significantly outperforms the in vivo atlas in the AD classification task, which in turns significantly outperforms the segmentation of the whole hippocampus.

The segmentation algorithm runs in approximately 20 min on a desktop computer — approximately twice as long when the input consists of two MRI modalities. This is in contrast with multi-atlas methods which are currently used in hippocampal subregion segmentation — such as (Yushkevich et al., 2010) — which are intrinsically slow (typically 10–20 h) due to the need to nonlinearly register a number of atlases to the test scan. On the other hand, our algorithm requires at this point that the data have been processed with the standard FreeSurfer pipeline, which takes approximately 10 h on a single core.

A limitation of the atlas presented in this study is that, even with ultra-high resolution MRI, there are boundaries that cannot be seen in the training data, e.g., the interfaces between the CA fields along the pyramidal layer of the hippocampus or the CA4/GC–DG interface. This remains an open problem in the hippocampal subregion MRI literature, in which the discrepancy and variability in subregion definitions remains rather large (Yushkevich et al., in press). This effect can be immediately noticed by comparing the heterogeneous manual annotations used in works such as (Yushkevich et al., 2009, 2010; Winterburn et al., 2013; Mueller et al., 2007, 2010; Van Leemput et al., 2009; Wisse et al., 2012). Another potential limitation of the proposed atlas is that it was built from manual delineations in elderly subjects only. Therefore, the atlas might include slight hippocampal atrophy that could decrease its applicability to studies of younger centers. Future work will follow four directions. First, we will evaluate the usefulness of the segmentations on 1 mm data as seed and target regions on diffusion and functional MRI studies. Second, we plan to extend the atlas to include the hippocampal tail — which will require a careful histologic analysis of the autopsy samples — and also other subcortical structures of interest, such as the thalamic and amygdaloid nuclei. Due to the generative nature of the segmentation framework, we are not constrained to derive the manual delineations from MRI data: histology or optical coherence tomography can be used, too (Augustinack et al., 2014; Magnain et al., 2014). Third, we will use Bayesian techniques to automatically infer the most likely value of the mesh flexibility parameter $\beta$, which controls the tradeoff between accuracy and robustness. And fourth, we would like to include explicit models of the partial volume effect, which makes it very challenging to accurately describe thin white matter structures such as fimbria, and the molecular layer. In this study, we tackled this problem by guiding the image intensity distributions of these structures with hyperparameters derived from subject-specific simulations of partial voluming. However, we believe that explicitly incorporating the partial volume effect in the model, for instance as in (Van Leemput et al., 2003), will further increase the accuracy of our segmentations.

Acknowledgments

Support for this research was provided in part by the National Center for Research Resources (P41-EB015986, and the NCCR BIRN Morphometric Project BIRN002, U24 RR021382), the National Institute for Biomedical Imaging and Bioengineering (RO1EB013565, RO1EB006758), the National Institute on Aging (AG022381, 5R01AG008122-22, KO1AG028521, BU Alzheimer’s Disease Center P30AG13846, BU Framingham Heart Study R01AG1649), the National Center for Alternative Medicine (RC1 AT005728-01), the National Institute for Neurological Disorders and Stroke (RO1 NS052585-01, 1R21NS072652-01, R01NS070963, R01NS083534), and was made possible by the resources provided by Shared Instrumentation Grants 1S10RR023401, 1S10RR019307, and 1S10RR023043. Additional support was provided by The Autism & Dyslexia Project funded by the Ellison Medical Foundation, by the NIH Blueprint for Neuroscience Research (5U01-MH093765), which is part of the multi-institutional Human Connectome Project, and the Finnish Funding Agency for Technology and Innovation (ComBrain). JEI acknowledges financial support from the Gipuzkoako Foru Aldundia (Fellows Gipuzkoa Program). This research was also supported by NIH grants P30-AG010129 and K01-AG030514, as well as the ADNI 2 add-on project “Hippocampal Subfield Volumetry” (ADNI 2-12-233036). The authors would also like to acknowledge Rohit Koppula and Lianne Cagnazzi for their help with the manual segmentation of the ex vivo MRI data; and Susanne Mueller and Paul Yushkevich for their help with the comparison with their methods (Table 4).

The collection and sharing of the MRI data used in the evaluation was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Mars Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Appendix A. Estimation of Gaussian parameters with missing data

In the segmentation of a multimodal test scan, it can happen that the MRI data for some channels are missing at a given voxel (as in Figs. 11b or 12c). In that case, the M step of the Gaussian parameter estimation (Section “Algorithm for segmenting an in vivo scan”) becomes complicated. Instead, we use a generalization of the EM algorithm called “Expectation Conditional Maximization” (ECM) (Meng and Rubin, 1993), in which the M step is replaced by two iterative conditional maximization (CM) steps that update the means and covariance matrices, respectively. Let $\mathbf{y}^{t}_{i}$ and $\mathbf{y}^{m}_{i}$ represent the observed and missing parts of vector $\mathbf{y}_{i}$, respectively; let $\mu_{i}^{c}$ and $\mu_{i}^{m}$ represent the corresponding parts of the mean $\mathbf{\mu}_{i}$; let $\Sigma_{c}^{m}$ and $\Sigma_{m}^{c}$ represent the corresponding parts of the covariance matrix $\Sigma_{c}$; and let $\Sigma_{i}^{m}$ be the part of the covariance matrix describing the cross-correlation between the observed and missing components of $\mathbf{y}_{i}$. The update of the mean is then:

$$
\mu_{i}^{c} \leftarrow \frac{\nu_{c} \mathbf{m}_{c}^{m} + \sum_{t=1}^{T} \Omega_{t}^{c} \mathbf{y}^{t}_{i}}{\nu_{c} + \sum_{t=1}^{T} \Omega_{t}^{c}}
$$

where $\mathbf{y}^{t}_{i} = \mathbf{y}^{c}_{i} - \mathbf{y}^{m}_{i}$, $\mathbf{y}^{m}_{i} = \mu_{i}^{c} + \nu_{c} (\mathbf{\mu}_{c} - \mathbf{m}_{c})$, and $\Sigma_{c}^{m} = \sum_{t=1}^{T} \Omega_{t}^{c}$. The update of the covariance is:

$$
\Sigma_{c}^{m} \leftarrow \sum_{t=1}^{T} \Omega_{t}^{c} \left[ \mathbf{y}^{t}_{i} - \mathbf{\mu}_{i} - \mathbf{\mu}_{c} + \nu_{c} (\mathbf{\mu}_{c} - \mathbf{m}_{c}) \right]^{\top} + \nu_{c} (\mathbf{\mu}_{c} - \mathbf{m}_{c})^{\top} (\mathbf{\mu}_{c} - \mathbf{m}_{c})^{c}.$$

$\nu_{c} = 1$ if channel $c$ is observed and $0$ otherwise.
where \( \hat{\Psi} \) is defined as the same way as above, and the matrix \( \hat{\Psi} \) has observed and missing parts \( \hat{\Psi} = \begin{pmatrix} \Psi_{(o)}^{(0)} & \Psi_{(m)}^{(0)} \\ \Psi_{(m)}^{(1)} & \Psi_{(o)}^{(1)} \end{pmatrix} \) (with the part describing their cross correlation equal to zero). The first step of the ECM algorithm is the same as in the EM counterpart, with the difference that the likelihood term is now evaluated as:

\[
p(y_j | \mu_j, \Sigma_j) = \sum_i N(\mu_{(i)}^{(j)}, \Sigma_{(i)}^{(j)}, 1)\]

In the specific case that the test has two channels and one of them is always observed — which is the case of the experiments of the ADNI T1/T2 data in Section "Quantitative results on ADNI T1/T2 data", it can be shown (Provost, 1990) that the M step is closed form, which leads to faster convergence of the algorithm. In our implementation of the presented Bayesian segmentation algorithm, we use these update equations whenever it is possible — including the experiments with the ADNI T1/T2 data in this paper.

Appendix B. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2015.04.042.

References


Ashburner, J., Andersson, J., Friston, K., 2000. Image registration using a symmetric\( \hat{\Sigma} \).\( \hat{\Psi} \).


