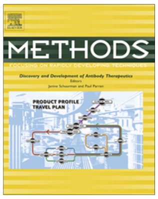
Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth



Visualizing the spatial gene expression organization in the brain through non-linear similarity embeddings

Ahmed Mahfouz [a,b,1], Martijn van de Giessen [a,b,1], Laurens van der Maaten [a,b], Sjoerd Huisman [a,b], Marcel Reinders [b], Michael J. Hawrylycz [c], Boudewijn P.F. Lelieveldt [a,b,*]

[a] *Division of Image Processing, Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands*
[b] *Department of Intelligent Systems, Delft University of Technology, Delft, The Netherlands*
[c] *Allen Institute for Brain Science, Seattle, WA, USA*

ARTICLE INFO

Article history:
Received 7 July 2014
Received in revised form 18 September 2014
Accepted 5 October 2014
Available online xxxx

Keywords:
Allen Brain Atlas
Spatial-mapped gene expression
Brain transcriptome structure
Dimensionality reduction
Stochastic Neighbor Embedding

ABSTRACT

The Allen Brain Atlases enable the study of spatially resolved, genome-wide gene expression patterns across the mammalian brain. Several explorative studies have applied linear dimensionality reduction methods such as Principal Component Analysis (PCA) and classical Multi-Dimensional Scaling (cMDS) to gain insight into the spatial organization of these expression patterns. In this paper, we describe a non-linear embedding technique called Barnes-Hut Stochastic Neighbor Embedding (BH-SNE) that emphasizes the local similarity structure of high-dimensional data points. By applying BH-SNE to the gene expression data from the Allen Brain Atlases, we demonstrate the consistency of the 2D, non-linear embedding of the sagittal and coronal mouse brain atlases, and across 6 human brains. In addition, we quantitatively show that BH-SNE maps are superior in their separation of neuroanatomical regions in comparison to PCA and cMDS. Finally, we assess the effect of higher-order principal components on the global structure of the BH-SNE similarity maps. Based on our observations, we conclude that BH-SNE maps with or without prior dimensionality reduction (based on PCA) provide comprehensive and intuitive insights in both the local and global spatial transcriptome structure of the human and mouse Allen Brain Atlases.

1. Introduction

The mammalian brain is a complex system governing all high-level cognitive tasks. The complexity of this system is reflected in the large number of cell types, organized into hundreds of distinct structures [1]. A major challenge facing the neuroscience community is to collect, integrate and analyze data across different levels and scales to produce new insights about the brain's anatomical and functional organization [2]. At the molecular level, each brain structure has a specific cellular composition with a distinct gene expression signature that dictates its functional role [3]. Therefore, to understand the basic anatomical and functional organization of the brain in relation to gene functions, it is crucial to study the spatial localization of genome-wide gene expressions in the brain.

Given the high cellular diversity in the brain, mapping genes at a sufficient spatial resolution is essential to analyze the transcriptome architecture of the brain. Several studies have previously mapped the expression of genes across the mammalian brain, but they have all been limited either in terms of the number of genes analyzed and/or the number of brain structures assessed [4,5]. The Allen Institute for Brain Sciences provides comprehensive genome-wide maps of gene expression across the mouse and human brain, providing a unique opportunity to study the transcriptome architecture of the mammalian brain. In the Mouse Brain Atlas [6] the expression of ~20,000 genes at a cellular resolution using *in situ* hybridization (ISH) is mapped on an anatomical atlas of the mouse brain. Comparably, the Human Brain Atlas [7] employed micro-arrays to produce a genome-wide map of transcript distribution across the entire human brain. These two resources allow the unprecedented study of how the transcriptome architecture of different brain regions instructs their functional role.

The high diversity of spatially-mapped gene expression patterns in the brain, ranging from globally-expressed genes to

Abbreviations: PCA, Principal Component Analysis; cMDS, classical Multi-Dimensional Scaling; BH-SNE, Barnes-Hut Stochastic Neighbor Embedding; t-SNE, t-Distributed Stochastic Neighbor Embedding; ISH, *in situ* hybridization; API, application programming interface; MR, magnetic resonance; T1W, T1-weighted; T2W, T2-weighted; DT, Diffusion Tensor; AAL, Automated Anatomical Labeling.

[*] Corresponding author at: Leiden University Medical Center, Division of Image Processing (C2-S), Albinusdreef 2, 2333 ZA Leiden, The Netherlands.

E-mail address: b.p.f.lelieveldt@lumc.nl (B.P.F. Lelieveldt).

[1] Shared first authorship.

http://dx.doi.org/10.1016/j.ymeth.2014.10.004
1046-2023/© 2014 The Authors. Published by Elsevier Inc.

highly-specialized regional markers, poses great challenges for computational approaches. Univariate approaches involving the analysis of the expression profiles of few genes of interest using prior knowledge of their site of action in the brain are not suitable to capture the full complexity of the data. In order to capture the complex patterns of expression of thousands of genes across the entire brain (thousands of samples), multivariate approaches should be employed to accommodate the high-dimensionality of the data. However, visualizing high-dimensional data for intuitive interpretation is challenging.

Several studies have used Principal Component Analysis (PCA) or classical Multidimensional Scaling (cMDS) to reduce the dimensionality of the voxel level genome-wide gene expression data of the mouse brain [7–9]. These low-dimensional maps are then used either to enable visual exploration of the gene expression patterns or as an input to a clustering algorithm where the resulting clusters are compared to the classical neuroanatomy. Classical methods such as PCA and cMDS focus on appropriately modeling large pairwise distances between gene expression profiles [10]. The focus on modeling large pairwise distances comes at the price of substantial errors in modeling small pairwise distances. However, it is exactly this local similarity structure that is essential in clustering and visual exploration: the goal of clustering is to find groups of nearby data points and, similarly, the goal of visual exploration is generally to determine which parts of the data are similar to a reference data point [11]. Therefore, we advocate to employ embedding techniques that focus on preserving local similarity structure, as is done by techniques such as t-Distributed Stochastic Neighbor Embedding (t-SNE) [12]. Since its introduction in 2008, t-SNE has been proven to outperform linear dimensionality reduction methods, but also non-linear embedding methods such as ISOMAP [13], in several research fields including machine-learning benchmark datasets and hyper-spectral remote sensing data [14].

Recently, t-SNE has been employed to analyze high dimensional proteomic and genomic data. Shekhar et al. [15] used t-SNE to differentiate between cellular phenotypes of the immune system based on mass cytometry data. Ji [16] used t-SNE to analyze the relationship between gene expressions and neuroanatomy in the developing mouse brain showing that t-SNE is able to capture the local similarities in the high-dimensional space. Fonville et al. [17] have shown that t-SNE outperforms PCA and self-organizing maps when used for modeling of mass spectrometry imaging data, where each pixel represents a molecular mass spectrum. All the previously mentioned applications demonstrate the high potential of t-SNE in the visual analysis of high-dimensional molecular data.

The goal of this work is to explore the effectiveness and limitations of t-SNE for spatial mapping of gene expression patterns in both the mouse and the human Allen Brain Atlases. By applying Barnes-Hut-SNE (BH-SNE) [22], a recently developed optimization algorithm for t-SNE, we show the consistency of the low dimensional embedding across the 6 human brains as well as between the sagittal and coronal experiments of the mouse brain. In addition, we quantitatively show the superiority of BH-SNE over PCA and cMDS in separating neuroanatomical regions in the low-dimensional 2D embeddings. Finally, we assess the effect of higher-order principal components on the local and global structure of the spatial transcriptome similarity maps.

2. Materials and methods

2.1. Mouse brain gene expression

The Allen Mouse Brain Atlas [6,18] provides genome-wide cellular-resolution *in situ* hybridization (ISH) gene expression data for approximately 20,000 genes of the 8-week old adult C57BL/6 J male mouse brain. For each gene, sagittal ISH sections were sampled at 25 μm intervals across the entire brain and the high-resolution 2D image series from each experiment were reconstructed in 3D and registered to the Nissl stain-based reference atlas (Allen Reference Atlas). The data were then aggregated into isotropic voxels defined by a uniform 200 μm grid in the reference space by averaging the expression levels and densities of all pixels (in the high-resolution ISH sections) within each voxel. The ontology of the reference atlas is used to label individual voxels with their anatomical nomenclature. In addition, coronal sections are available for a set of approximately 4000 genes that showed marked regional expression patterns in the sagittal plane [3]. More information about the ISH sections alignment and registration to the Allen Reference Atlas can be found in [19].

We retrieved all expression energy volumes from [18] using the Allen Brain Atlas application programming interface (API). Expression energy is a measurement combining the expression level (the integrated amount of signal within each voxel) and the expression density (the amount of expressing cells within each voxel) [20].

We focused our analysis on a subset of high confidence genes for which coronal and sagittal experiments are available, as in [8]. For each gene, we computed the Spearman's rank correlation between the corresponding coronal and sagittal experiments and selected genes in the top-three quartiles of correlation (3241 genes). The coronal and sagittal experiments corresponding to those 3241 genes were retained for further analysis (Supplementary Table 1). For genes with more than one sagittal experiment, the maximum correlation value was used. A mask was applied to exclude all non-brain voxels, resulting in a 61,164 × 3241 (voxels × genes) matrix for the coronal experiments and a 27,365 × 3241 matrix for the sagittal experiments.

2.2. Human brain gene expression

The Allen Human Brain Atlas [7,21] includes RNA microarray data collected from the postmortem brains of six donors, with no known neuropsychiatric or neuropathological history; see Table 1 for detailed information about the donors. Magnetic resonance (MR) T1-weighted (T1W), T2-weighted (T2W) and Diffusion Tensor (DT) images were collected in-cranio, prior to dissection for anatomic visualization of each brain.

Approximately 1000 samples were dissected using manual macrodissection for large regions and laser captured microdissection for smaller regions from two donor brains (H0351.2001 and H0351.2002), representing all structures across the whole brain. For the other four donor brains, approximately 500 samples were taken from one hemisphere only. Each sample is associated with a 3D (*x*, *y*, *z*) coordinate on its corresponding donor's MRI volume. Moreover, the MNI coordinates of each sample is reported (registration to the MNI reference space was done using FreeSurfer software). The dataset contained expression profiles of 29,191 genes represented by 58,692 probes, with 93% of known genes represented by at least 2 probes. The data was already normalized across samples and across different brains using the procedure

Table 1
Human donors information.

Donor ID	Number of samples	Sex	Age (years)	Race/ethnicity
H0351.2001	946	Male	24	African American
H0351.2002	893	Male	39	African American
H0351.1009	363	Male	57	Caucasian
H0351.1012	529	Male	31	Caucasian
H0351.1015	470	Female	49	Hispanic
H0351.1016	501	Male	55	Caucasian