Data and text mining

i2d: an R package for simulating data from images and the implications in biomedical research

Xiaoyu Liang 1,2, Ying Hu3, Chunhua Yan3 and Ke Xu1,2,*

1Department of Psychiatry, Yale School of Medicine, New Haven, CT 06511, USA, 2VA Connecticut Healthcare System, West Haven, CT 06516, USA and 3Center for Biomedical Bioinformatics and Information Technology, National Cancer Institute, Rockville, MD 20852, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on August 13, 2020; revised on October 26, 2020; editorial decision on November 13, 2020; accepted on November 16, 2020

Abstract

Motivation: High-quality imaging analyses have been proposed to drive innovation in biomedical and biological research. However, the application of images remains underexploited because of the limited capacity of human vision and the challenges in extracting quantitative information from images. Computationally extracting quantitative information from images is critical to overcoming this limitation. Here, we present a novel R package, i2d, to simulate data from an image based on digital convolution.

Results: The R package i2d allows users to transform an image into a simulated dataset that can be used to extract and analyze complex information in biomedical and biological research. The package also includes three novel and efficient methods for graph clustering based on simulated data, which can be used to dissect complex gene networks into sub-clusters that have similar biological functions.

Availability and implementation: The code, the documentation, a tutorial and example data are available on an open source at: github.com/XiaoyuLiang/i2d.

Contact: ke.xu@yale.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Imaging is a powerful tool in biomedical and biological research that allows us to improve or extract valuable information on tissue composition, cellular structure, and fundamental molecular processes. Recent advances in biomedical and biological imaging have resulted in an explosion in the quality and quantity of images obtained in a digital format, such as JPEG, TIF, and RAW. However, these images remain underexploited because of the limited visual capacity of humans to detect precise imaging patterns. Computationally extracting quantitative information from an image is critical to overcoming this limitation.

Methods for quantifying information from images have been developed. magick (Ooms, 2018) and EBImage (Pau et al., 2010) are well-established image processing toolboxes that provide functions for reading, writing, processing, and analyzing images. magick can process the basic information of an image such as image filtering, kernel convolution, and layer stacking. EBImage includes a range of fast image processing functions and represents images as multidimensional arrays containing pixel intensity values. However, neither of these methods can directly transform spatial information from images. The interpretation and analysis of images based on pixel intensity are challenging and result in misinformation or incomplete data analysis.

In this article, based on image digitalization and convolution, we aim to develop a tool to simulate data points from an image such that simulated data are closer to real images when the number of simulated points tends to infinity. Here, we present an R package, i2d, depends on two R packages EBImage (Pau et al., 2010) and igraph (Csardi and Nepusz, 2006), and allows users to easily transform an image into a simulated geographic coordinate dataset that can be used to extract and analyze the complex information in biomedical and biological images (Fig. 1). This utility allows users to perform graph clustering analysis using clustering algorithms based on distances among coordinates. More importantly, it includes three methods for graph clustering and trajectory analysis.

2 Materials and methods

2.1 Image digitization

We consider an image as a matrix whose elements are numbers between 0 and 255. The size of this matrix is image height by width by the number of image channels. A grayscale image has 1 channel, while a color image has 3 red-green-blue (RGB) channels. For a color image, the true color image RGB is converted to a grayscale intensity image. A grayscale image that has a one-pixel matrix is used for the subsequent analysis. Here, we summarize the general steps for image digitization:
depending on igraph package (Csardi and Nepus, 2006). Users can manually define the number of clusters. The details of the three methods can be found in the Supplementary Material. We treat each point as a vertex and the Euclidean distance between two points as the weight of the edge. Our goal is to find a partition \( \mathcal{G} \) that partitions \( N \) vertices \( v_1, \ldots, v_N \) into \( K \) disjoint clusters \( G_1, \ldots, G_K \) taking into consideration the edge structure of the graph in such a way that there should be many edges within each cluster and relatively few between the clusters, where \( \mathcal{G} = (G_1, \ldots, G_K) \) with \( G_k = \{v_1, \ldots, v_N\} \) and \( G_i \cap G_j = \emptyset (i \neq j) \). In the meanwhile, the outliers should be detected and removed before the clustering procedure.

The minimum spanning tree (MST)-based graph clustering algorithm (mstGCluster) not only can be used to group the vertices into a specific number of clusters but also can filter the noise in the graph. Prim’s algorithm is used to find an MST of the graph. Then, we utilize the k-means clustering method to detect the noise in the graph. Last, removing \( K-1 \) edges with the largest weights will group the vertices into \( K \) disjoint clusters (Supplementary Figs S1 and S2 in Supplementary Material).

The modularity optimization-based community detection method for graph clustering (codeGCluster) treats all the vertices as a big community. We separate the whole community into several small communities by removing the weak connections. The community detection methods, such as the Louvain algorithm, the fast-greedy modularity optimization algorithm, and the near linear time algorithm are used to find communities in graphs. We use iteration to separate the whole community into a specific number (\( K \)) of communities (Fig. 1c and Supplementary Fig. S3 in Supplementary Material).

The package also includes a function for trajectory analysis, which can be used for gene enrichment analysis and graph clustering. In detail, the longest path through the MST, corresponding to the longest distance of the vertices that were selected from the graph, is referred to as the ‘backbone’ of the tree. We extract the backbone and the long branches on the backbone of MST for trajectory analysis and graph clustering analyses.

### 3 Conclusions

The developed R package i2d is a new tool for digital data simulation from images. The usage of i2d integrated into the R environment provides a flexible and powerful platform for image data handling and analysis. It is user friendly and easy and fast to operate. Furthermore, the simulated dataset can be used for extracting and investigating information in biomedical and biological images.

### Author Contributions

Y.H. and C.Y. were responsible for the R package preparation, data analysis, and manuscript preparation. K.X. contributed to manuscript preparation. X.L. was responsible for the R package preparation and manuscript preparation. All authors read and approved the final manuscript.

### Funding

This work was supported by the National Institute on Drug Abuse [R01DA02691 (Xu), R01DA047063 (Xu and Aouizerat), R01DA047820 (Xu and Aouizerat)].

### Conflict of Interest: none declared.

### References

Csardi,G. and Nepus,T. (2006) The igraph software package for complex network research, Int. J. Complex Syst., 1695, 1–9.

Ooms,J. (2018) Magick: advanced graphics and image-processing in R. CRAN. R package version, 1.

Pau,G. et al. (2010) EBImage—an R package for image processing with applications to cellular phenotypes. Bioinformatics, 26, 979–981.