BIA4 ICA 1 - List of datasets

Below you will find a list of freely available datasets that you can use for ICA 1.

As a group you need to:

- Discuss which biomedical question(s)/problems your software will help answer (this is key for ICA 2 as well!).
- Choose a dataset that you would like to work on and decide some tasks that you would like to perform on it.
- Write your software to perform these tasks!
- Test your software on the dataset and document the results.

It's important to remember that in some cases you might need to use some other datasets or tools to help you with your task. For example, you might need to use an extra dataset to train or test a machine learning model, or you might need to use another tool to annotate images.

ICA 1 is a group assignment, with groups of 7-8 students, therefore the expectation is that the output will be a group effort from 7-8 students working together, so don't limit yourself to something obvious or trivial. The result might not be perfect but it needs to show group effort.

It's fine to use other datasets if you have a specific question in mind that you want to answer and the dataset proposed here do not fit your needs. The course organizer will be happy to discuss your ideas and might point you to some other more relevant datasets.

1. *Drosophila melanogaster* wings

Description

This dataset contains images of Drosophila wings with various genotypes.

Reference

Sonnenschein et al, 2015 - An Image Database of Drosophila melanogaster Wings for Phenomic and Biometric analysis

Available from

http://gigadb.org/dataset/view/id/100141

Data type and size

Several Gb of TIFF files taken at 20x and 40x with different microscopes.

---

2. Retina scans

Description

This dataset contains images of retina scans from glaucoma positive and negative patients.

Available from

https://www.kaggle.com/sshikamaru/glaucoma-detection

Data type and size

650 JPEG files (~200 Mb).

---

3. Calcium imaging in neurons

Description

These datasets contain calcium imaging data from neurons, as part of the "CodeNeuron" challenge.

Available from

http://neurofinder.codeneuro.org/

---

4. Breast Cancer Cell Biopsies

Description

The BreaKHis dataset contains images from H&E stained biopsies from benign and malignant breast tumors.

Reference

Spanhol et al, 2016 - "A Dataset for Breast Cancer Histopathological Image Classification"

Data type and size

850 Mb of PNG files.

Available from

https://www.kaggle.com/forderation/breakhis-400x

---

5. Nematodes

Description

A dataset of images of different nematodes.

Reference

https://arxiv.org/abs/2103.08335

Data type and size

JPG images ~ 600 Mb.

Available from

https://github.com/xuequanlu/I-Nema

---

6. Cell tracking

Description

The cell tracking challenge aims to improve algorithms in cell tracking in 2D and 3D.

Reference

Ulman et al, 2017 - An objective comparison of cell-tracking algorithms

Data type and size

13 different datasets of several Gb of 2D + time and 3D + time videos of moving cells.

Available from

http://celltrackingchallenge.net/

7. Malaria species

Description

A dataset containing images from cells parasitised by different species of Plasmodium.

Available from

https://www.kaggle.com/saife245/malaria-parasite-image-malaria-species

8. Mouse behaviour

Description

A dataset of videos of different behaviours from mice.

Reference

Jhuang et al, 2010 - Automated Home-Cage Behavioral Phenotyping of Mice

Data type and size

4200 short MPG videos of mice in different behaviours (~1Gb). A full dataset is also available, containing over 10.6 hours of continously labeled video (8 day videos and 4 night videos) for the eight behaviors of

interest: drink, eat, groom, hang, micromovemnt, rear, rest, walk.

Available from

https://cbmm.mit.edu/mouse-dataset

9. Zebrafish movies

Description

A large dataset of videos of zebrafish swimming in a tank, viewed from two different angles.

Reference

Pedersen et al., 2020 - 3D-ZeF: A 3D Zebrafish Tracking Benchmark Dataset

Data type and size

14 Gb of high-resolution videos + ground truth annotations

Available from

https://motchallenge.net/data/3D-ZeF20/

10. The Cancer Imaging Archive

Description

A large collection of medical images from cancer patients in various modalities (MRI, CT, etc).

Available from

https://www.cancerimagingarchive.net/browse-collections/

Data type and size

Depends on the dataset, but can be several Gb of DICOM files. DICOM is a standard format for medical images; you can read DICOM files in Python using the `pydicom` library.

11. The Yeast Resource Center (YRC) dataset

Description

The YRC Public Image Repository (YRC PIR) is a database of fluorescent microscopy images depicting the localization, co-localization and FRET (fluorescence energy transfer) of proteins in cells--particularly in the budding yeast *Saccharomyces cerevisiae*.

Available from

https://images.yeastrc.org/imagerepo/searchImageRepoInit.do

Data type and size

~2.4 Tb of images (~1M images), can be downloaded in smaller batches / subsets.

12. The Cellpose dataset

Description

This is the dataset used to train the Cellpose software, which is a deep learning-based software for cell segmentation. You can use this dataset to test your own cell segmentation software or to perform other tasks.

Reference

Stringer et al, 2021 - Cellpose: a generalist algorithm for cellular segmentation
Stringer and Pachitariu, 2024 - Cellpose3: one-click image restoration for improved cellular segmentation

Available from

https://www.cellpose.org/datasets/

Data type and size

~200 Mb of images.

13. Whole-brain light-sheet imaging data

Description

Data from 18 larval zebrafish, ~100k neurons/fish, during multiple visually-evoked behaviors.

Reference

Chen et al. 2018 - Brain-wide Organization of Neuronal Activity and Convergent Sensorimotor Transformations in Larval Zebrafish.

Available from

https://janelia.figshare.com/articles/dataset/Whole-brain_light-sheet_imaging_data/7272617

Data type and size

~55 Gb of data; data from single fish is ~1-3 Gb.

14. SNEMI3D dataset

Description

A large training dataset of mouse cortex in which the neurites have been manually delineated. An unlabelled test dataset is also available.

Reference

Kasthuri et al. 2015 - Reconstruction of a Volume of Neocortex.

Available from

https://zenodo.org/records/7142003

Data type and size

~400 Mb of data in TIFF format.

15. Dendritic spine dataset

Description

A fully annotated dataset of Two-Photon Laser Scanning Microscopy (2PLSM) images of three types of dendritic spines.

Reference

Ghani et al, 2017 - Shape and appearance features based dendritic spine classification

Available from

https://github.com/mughanibu/Dendritic-Spine-Analysis-Dataset

Data type and size

~8 Mb of PNG files.