Depositing biological segmentation datasets FAIRly

Dr Elaine Ho[1], Dr Dimitris Ladakis[1], Dr Michele C. Darrow[1]
[1]Artificial Intelligence and Informatics, The Rosalind Franklin Institute, Harwell, United Kingdom

Background incl. aims
Segmentation of biological images identifies regions of an image which correspond to specific features of interest, which can be analysed quantitatively to answer biological questions. This task has long been a barrier to conducting large-scale biological imaging studies as it is time- and labour-intensive. Modern artificial intelligence segmentation tools can automate this process, but require high quality segmentation data for training, which is challenging to acquire. Biological segmentation data has been produced for many years, but this data is not often re-used to develop new tools as it is hard to find, access, and use. Recent disparate efforts [1-4] have been made to facilitate deposition and re-use of these valuable datasets, but more work is needed to increase re-usability. In this work, we review the current state of publicly available annotation and segmentation datasets and make specific recommendations to increase re-usability following FAIR (findable, accessible, interoperable, re-usable) principles [5] for the future.

Methods
A collection of publicly available segmentation and annotation datasets for 3-D volumetric electron microscopy and associated metadata was assembled from searches in the EMDB, EMPIAR, and Open Organelle databases, and a literature search with Pubmed from 2012-2023. Characteristics about these datasets were collected and trends over time were investigated, e.g., the purpose of the segmentations, data formats, biological feature type and size scale, imaging modality, and others.

Results
Whilst there were many examples of publicly available segmentation data that could be easily reused, these datasets were few and far between. We identified barriers to reusing published segmentation data according to FAIR principles. Many publications we reviewed were not findable or accessible as they were deposited at defunct URLs, were only available on request, or required researchers to search separately through citations or online to download the data. We found that data was deposited across at least 12 different formats, and 9 different online repositories, limiting the interoperability of these datasets as significant effort would be required to parse these formats into a single unified database for training segmentation tools. We found that there were considerable differences in the definitions of certain terms such as “segmentation”, “reconstruction”, and “ground truth”. A consensus is required across the breadth of the bioimaging community to ensure that these datasets can be re-used appropriately without misinterpretation.

Enhanced metadata capture and search capabilities would help developers find suitable datasets for their use case and determine how much (if any) manual curation would be required to bring segmentation data to the required standard for their study. This metadata could include the number and size of the feature being segmented relative to the image, the quality of the segmentation or the intended use of the dataset (qualitative visualisation, quantitative morphology, etc).

Conclusion
Re-use of biological segmentation data is important for enhancing development of segmentation software tools, particularly those using artificial intelligence techniques requiring large amounts of training data. Enhanced metadata capture, quality evaluation,

community consensus in ontology and standardisation in data formats is required for encouraging re-use of these precious segmentation datasets.

Keywords:

Segmentation; annotation; volume electron microscopy

Reference:

1. Iudin A, et al. (2023). “EMPIAR: the Electron Microscopy Public Image Archive.” Nucleic Acids Res., 51, D1503-D1511. https://doi.org/10.1093/nar/gkac1062
2. C. Shan Xu, et al. (2021) “An open-access volume electron microscopy atlas of whole cells and tissues.” Nature. https://doi.org/10.1038/s41586-021-03992-4.
3. J. T. Vogelstein, et al. (2018) “A Community-Developed Open-Source Computational Ecosystem for Big Neuro Data.” Nature Methods, (11)15:846–847. https://doi.org/10.1038/s41592-018-0181-1
4. Chan Zuckerberg Imaging Institute 2023, CryoET Data Portal, accessed 3rd Apr 2024. https://cryoetdataportal.czscience.com/browse-data/datasets
5. Wilkinson, M., et al. (2016) “The FAIR Guiding Principles for scientific data management and stewardship.” Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18