BD5: an open data format for representing quantitative biological dynamics data

Koji Kyoda¹, Kenneth H.L. Ho², Hiroya Itoga¹, Yukako Tohsato³, Shuichi Onami¹
¹RIKEN BDR, ²Francis Crick Institute, ³Ritsumeikan University
Bioimage informatics

- Live-cell imaging can capture spatiotemporal dynamics of biological phenomena.
- Using image analysis, (x, y, z, t, c) data can be obtained from microscopy images.

(Keller et al. 2008)
Quantitative biological dynamics data

*Caenorhabditis elegans* (Bao et al., 2006)

*Caenorhabditis elegans* (Kyoda et al., 2013)

*Drosophila melanogaster* (Keller et al., 2010)

Zebrafish (Keller et al., 2008)

HeLa cell (Held et al., 2010)

*Caenorhabditis elegans* (Yemini et al., 2013)
Problem

- Research groups used different data formats.
- It is often difficult to reuse their data because of:
  - intricate data structure
  - the lack of detailed explanations

(Kyoda et al. 2013)

(Bao et al. 2006)

(Keller et al. 2008)

Text file

Separated text files

Matlab file
Open unified data formats

• Allowing
  • Data analysis and comparison
  • Tool development and its evaluation

Data providers

All tools can be used for data analysis.

Open unified formats

All data can be used for tool evaluation.

Tool developers
BDML: Biological Dynamics Markup Language

• An open unified format for representing quantitative data of biological dynamics

Kyoda et al. (2015) Bioinformatics
Biological dynamics described in BDML

- Data ranging from molecules to organisms
BD5

- HDF5-based data format for representing quantitative biological dynamics data

Kyoda et al. (2020) PLoS One
Example

(Bao et al., 2006)
Fast data access

- Compared with BDML, BD5 enables fast access to quantitative data owing to random access to the HDF5-based file.
File size reduction

- BD5 enables fast transfer of large quantitative data because the file size is dramatically reduced.
Bioimaging data format

• Next generation file format for bioimaging data
  • ome-zarr is a zarr-based format for storing bioimaging data.

(Moore et al., Nat. Methods, 2021)
**BD-zarr**

- Dynamics data is stored in AnnData (https://anndata.readthedocs.io)
  - Store coordinates information of biological objects in $X$ array
  - Store features information as separate $obs$ array
  - Store tracking information as separate $obsp$ array
Example

- Early worm embryogenesis data (Kyoda et al., 2020)

```
wt-N2-081015-01
|--- 0  image data |
|    |--- t |
|    |--- c |
|    |--- z |
|    |--- y |
|--- labels |
|    |--- 0  Pixel-based ROI data |
|    |    |--- t |
|    |--- ... |
|--- dyn  Dynamics data |
|    |--- X |
|    |--- obs |
|--- obsp  tracking data |

X position data

<table>
<thead>
<tr>
<th>t</th>
<th>z</th>
<th>y</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.0</td>
<td>39.922649</td>
<td>109.316254</td>
</tr>
<tr>
<td>1</td>
<td>2.0</td>
<td>39.498207</td>
<td>113.885712</td>
</tr>
<tr>
<td>2</td>
<td>3.0</td>
<td>39.999203</td>
<td>111.549751</td>
</tr>
<tr>
<td>3</td>
<td>4.0</td>
<td>40.121613</td>
<td>112.833496</td>
</tr>
<tr>
<td>4</td>
<td>5.0</td>
<td>42.206738</td>
<td>115.653198</td>
</tr>
</tbody>
</table>

obs feature data

<table>
<thead>
<tr>
<th>id</th>
<th>entity</th>
<th>name</th>
<th>sphericity</th>
<th>volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1000</td>
<td>P0</td>
<td>0.837884</td>
<td>76238.3</td>
</tr>
<tr>
<td>1</td>
<td>2000</td>
<td>P0</td>
<td>0.811657</td>
<td>88108.6</td>
</tr>
<tr>
<td>2</td>
<td>3000</td>
<td>P0</td>
<td>0.854491</td>
<td>110026.0</td>
</tr>
<tr>
<td>3</td>
<td>4000</td>
<td>P0</td>
<td>0.771704</td>
<td>107802.0</td>
</tr>
<tr>
<td>4</td>
<td>5000</td>
<td>P0</td>
<td>0.815484</td>
<td>138333.0</td>
</tr>
</tbody>
</table>

obsp tracking data

array([[0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       ...
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0]], dtype=int8)
```
Visualization of BD-zarr data

- with napari image viewer (https://napari.org/)
**Data sharing of bioimaging data**

- SSBD:database (https://ssbd.riken.jp) stores and shares quantitative data and image data of biological dynamics with rich meta data.
Summary

• We have developed BDML/BD5 based on XML/HDF5 for representing quantitative data of biological dynamics.
• Compared with BDML, the BD5 format has two advantages:
  • faster access and retrieval of quantitative data
  • Smaller file size, faster transfer of files in large datasets
• Following the current development in the bioimaging community, we are working on developing a Zarr-based format that are functionally compatible with BD5, HDF5-based format.
Acknowledgement

• RIKEN OLSP (Funding)

• Norio KOBAYASHI (RIKEN R-IH)

• Hideyuki Jitsumoto (RIKEN R-IH)

• Information Systems Division

• Bioimaging Community
  • Josh Moore (University of Dundee)
  • Kevin Yamauchi (ETH Zürich)

• Special acknowledgement to EMBL-EBI’s Embassy Cloud and the BioImage Archive for providing valuable S3 storage and support for this project.