

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

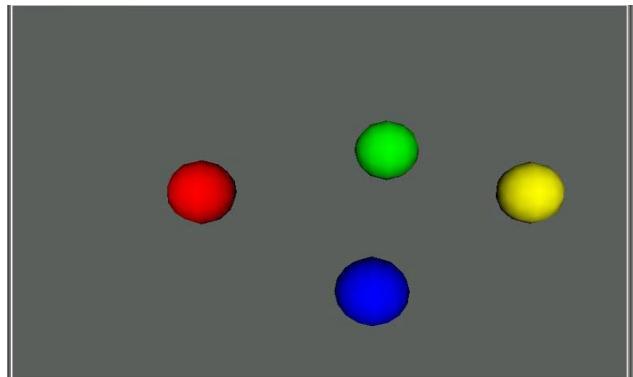
- Live-cell imaging can capture spatiotemporal dynamics of biological phenomena.
- Using image analysis, (x, y, z, t, c) data can be obtained from microscopy images.



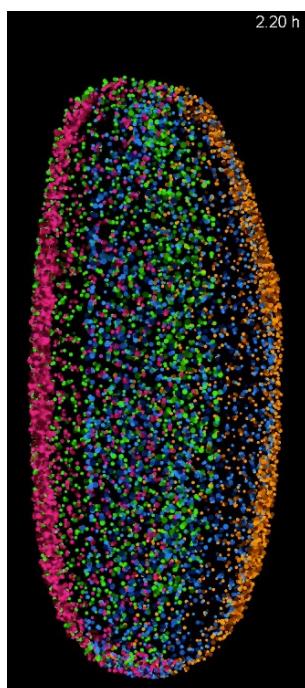
(Keller et al. 2008)

Quantitative biological dynamics data

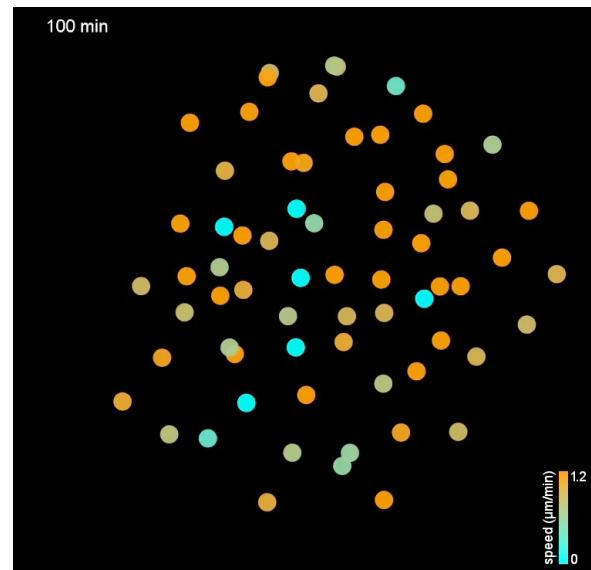
C. elegans (Bao et al., 2006)



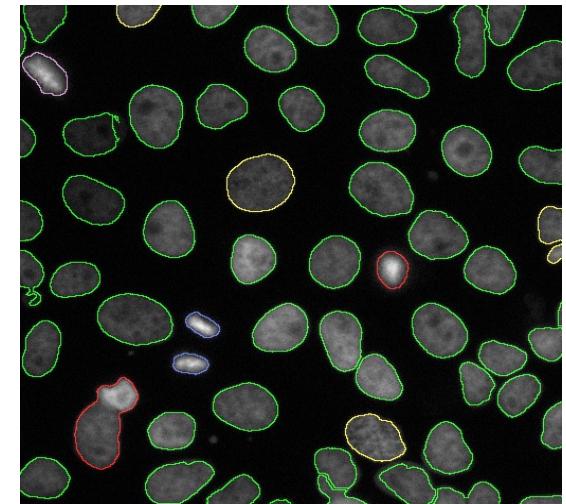
D. melanogaster (Keller et al., 2010)



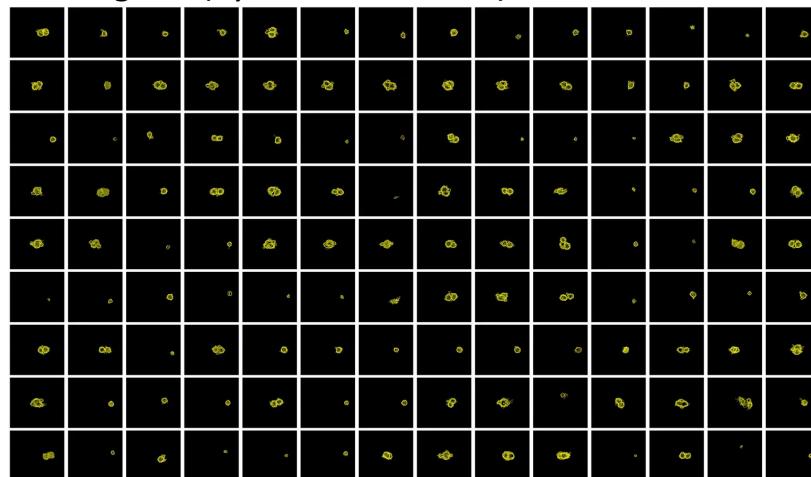
zebrafish (Keller et al., 2008)



HeLa cell (Held et al., 2010)



C. elegans (Kyoda et al., 2013)

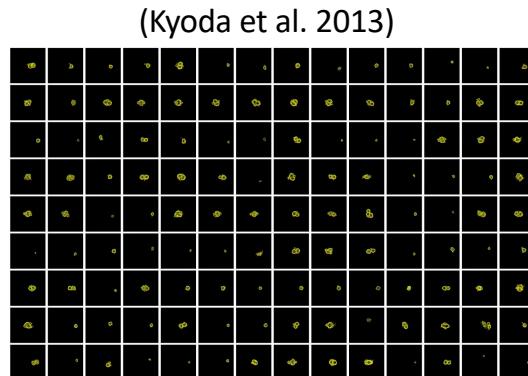


C. elegans (Yemini et al., 2013)



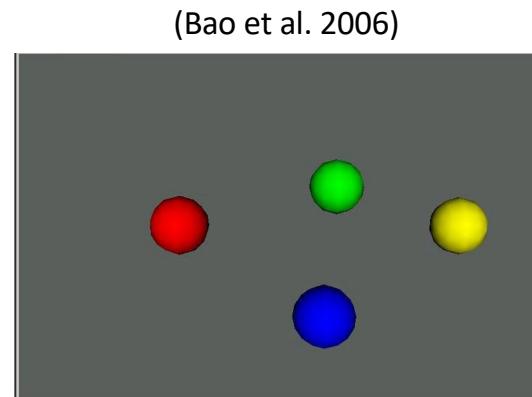
Problem

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



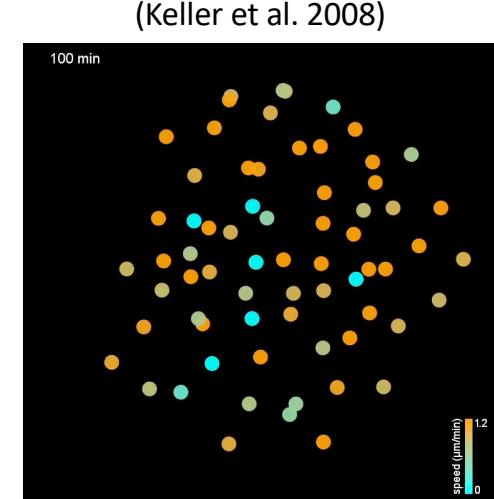
```
0000, 1000, 51.8232, 37.7117, 27.1820, 1, P0, 17  
477, 342, 23, 21111111210111212112.....  
486, 321, 24, 21211222100121222102.....  
.....  
1000, 2000, 33.3146, 32.1893, 25.2702, 2, P0, 17  
320, 257, 23, 211112111121111112.....
```

Text file



```
t001-nuclei → 1, 1, -1, 4, -1, 380, 366, 16.1, 80, EMS, 21281015,  
t002-nuclei 2, 1, -1, 2, -1, 387, 153, 16.6, 86, ABp, 2836266,  
..... 3, 1, -1, 3, -1, 189, 251, 17.2, 88, ABa, 2850348,  
..... 4, 1, -1, 5, -1, 562, 269, 18.1, 80, P2, 2168825
```

Separated text files

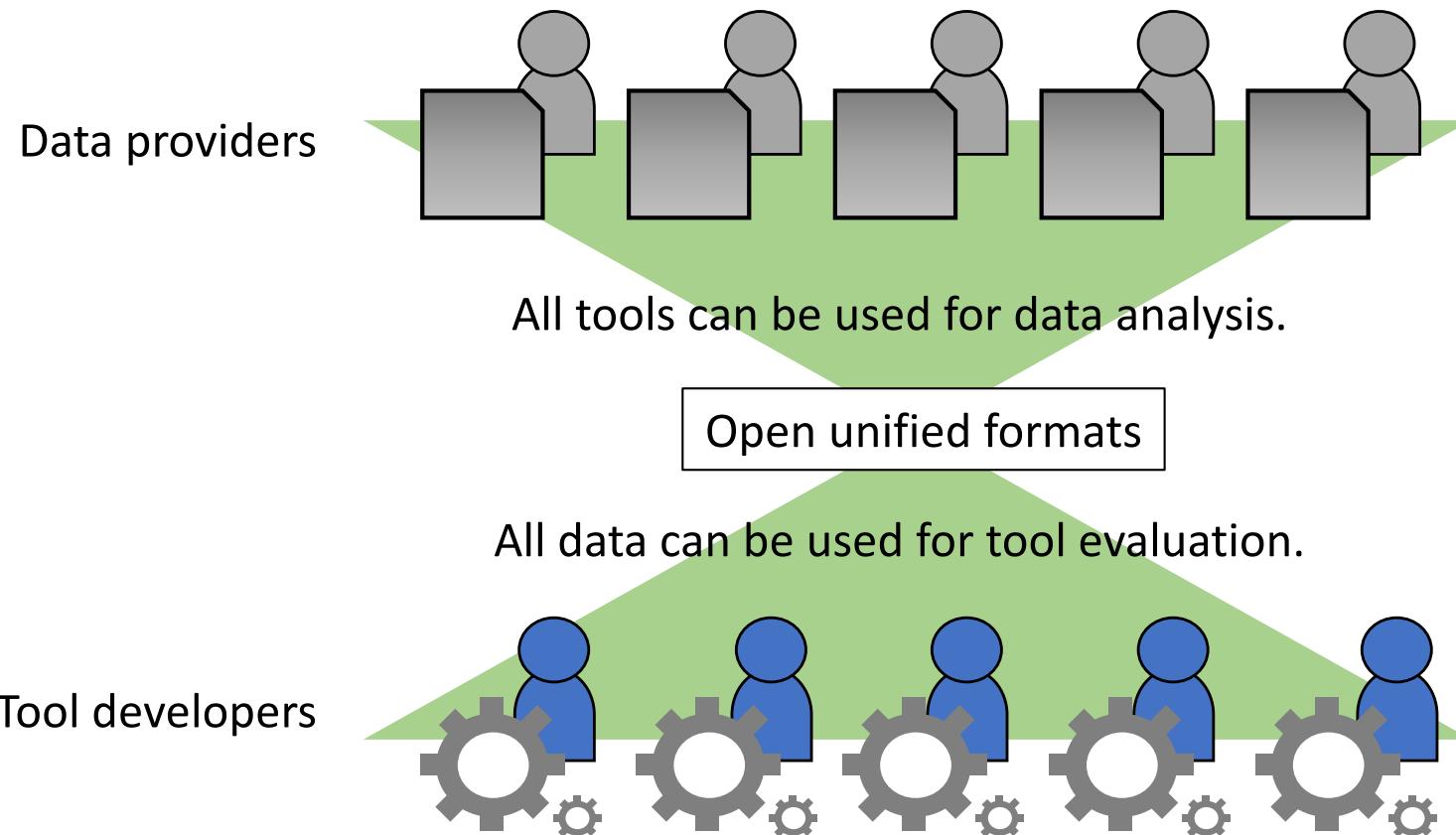


```
<68x17 double> → 611.2904, 563.0863, 80.3444, 13.8991, ....  
<69x17 double> 485.5862, 546.2255, 91.1758, 14.0556, ....  
..... 562.8646, 459.1969, 95.8143, 14.2670, ....  
.....
```

Matlab file

Open unified data formats

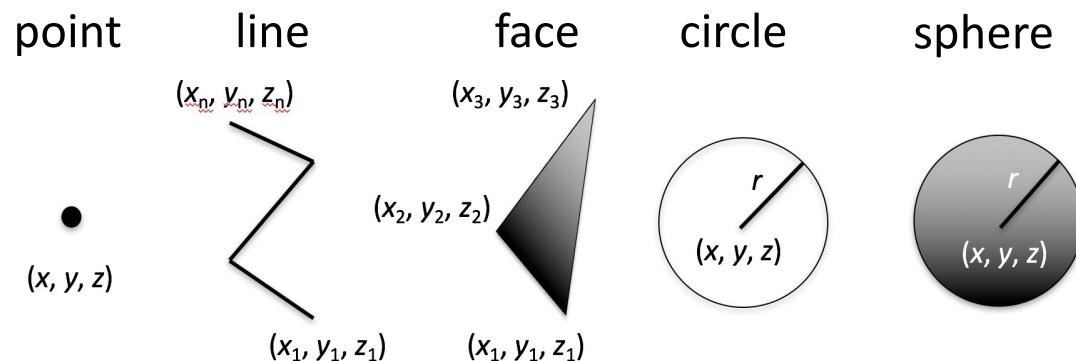
- Allowing
 - Data analysis and comparison
 - Tool development and its evaluation



BDML: Biological Dynamics Markup Language

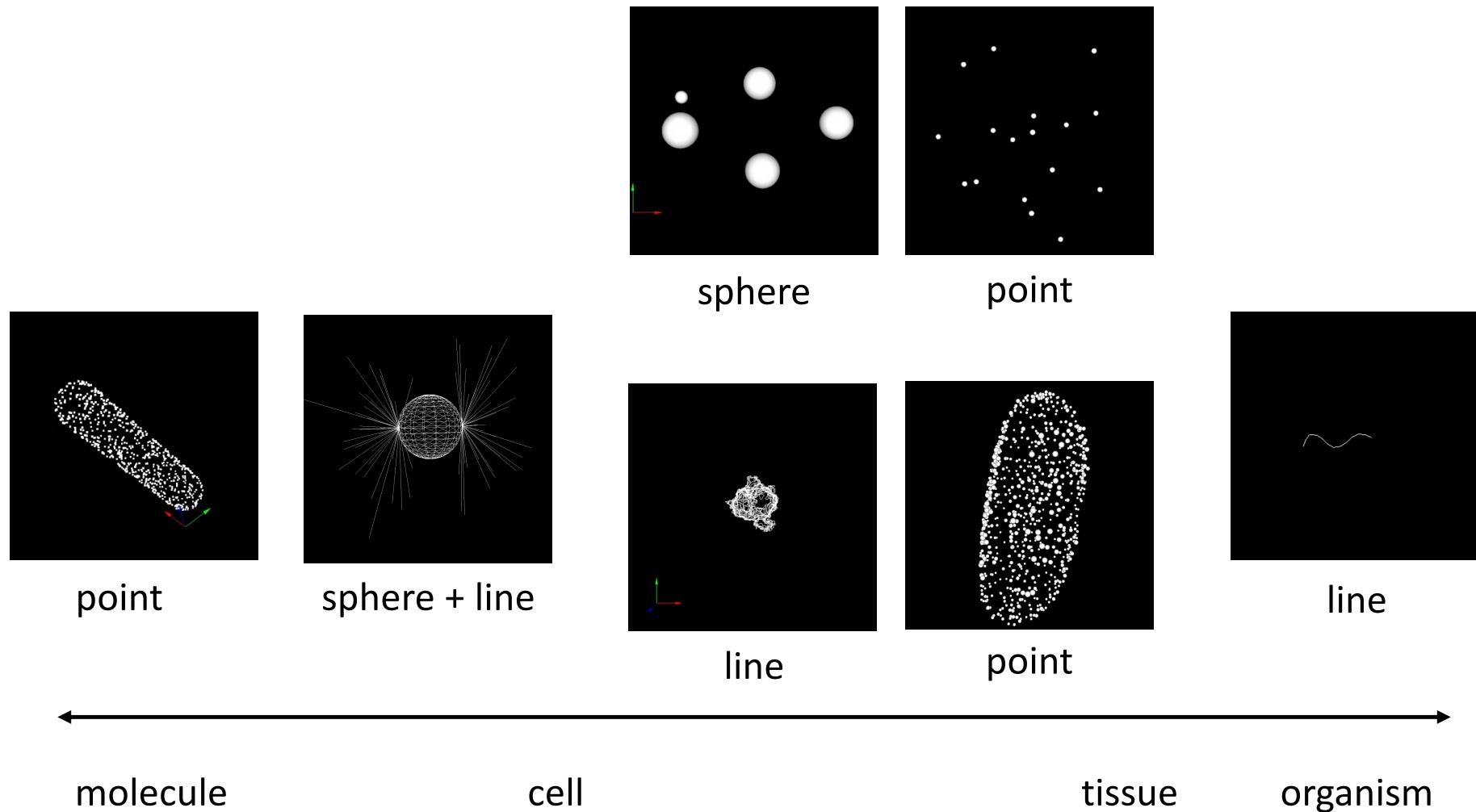
- An open unified format for representing quantitative data of biological dynamics

```
<scaleUnit>
  <tScale>20</tScale>
  <tUnit>second</tUnit>
</scaleUnit>
<component>
  <componentID>100</componentID>
  <time>1</time>
  <measurement>
    <point><xyz><x>10.32</x><y>30.42</y><z>18.32</z></xyz></point>
  </measurement>
</component>
<component>
  <componentID>101</componentID>
  <time>2</time>
  <prevID>100</prevID>
  <measurement>
    <point><xyz><x>9.57</x><y>32.05</y><z>14.91</z></xyz></point>
  </measurement>
</component>
```



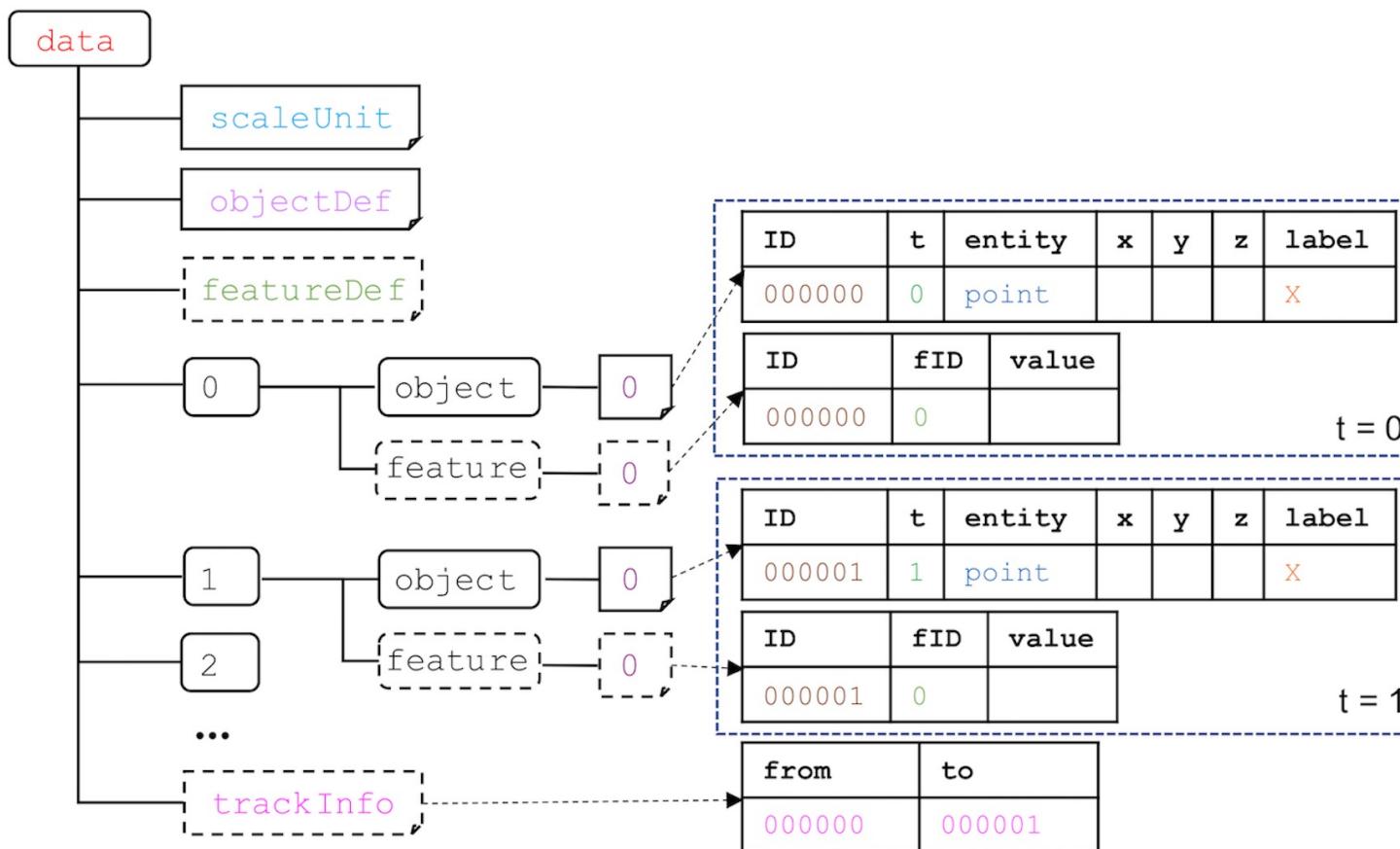
Biological dynamics described in BDML

- Data ranging from molecules to organisms



BD5

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



Example

The figure shows a screenshot of a software interface for managing biological data, likely from a microscopy experiment. The interface consists of several windows and panels.

Left Panel: A file browser showing the contents of the file `081505_L1_bd5.h5`. The tree view includes categories like `data`, `feature`, and `object`, with numerous sub-items numbered 0 through 128.

Top Row Windows:

- objectDef at /data/ [081505_L1_bd5.h5 in /Users/kkyoda/081505_L1_bd5.h5]**: Shows a table with columns `oID` and `name`. One entry is `0 nucleus`.
- 0 at /data/0/object/ [081505_L1_bd5.h5 in /Users/kkyoda]**: Shows a table with columns `ID`, `t`, `entity`, `x`, `y`, `z`, `radius`, and `label`. Data rows include:

ID	t	entity	x	y	z	radius	label
001001	1.0	sphere	163.0	343.0	6.1	1.62	polar1
001002	1.0	sphere	380.0	366.0	16.1	3.6	EMS
001003	1.0	sphere	387.0	153.0	16.6	3.87	ABp
001004	1.0	sphere	189.0	251.0	17.2	3.96	ABA
001005	1.0	sphere	562.0	269.0	18.1	3.6	P2
- featureDef at /data/ [081505_L1_bd5.h5 in /Users/kkyoda/081505_L1_bd5.h5]**: Shows a table with columns `fID`, `name`, and `fUnit`. One entry is `0 totalGFPsig... a.u.`

Middle Row Windows:

- scaleUnit at /data/ [081505_L1_bd5.h5 in /Users/kkyoda/081505_L1_bd5.h5]**: Shows a table with columns `dimension`, `xScale`, `yScale`, `zScale`, and `sUnit`. One entry is `0 3D+T 0.09 0.09 1.0 micrometer 1.0`.
- 0 at /data/0/feature/ [081505_L1_bd5.h5 in /Users/kkyoda]**: Shows a table with columns `ID`, `fID`, and `value`. Data rows include:

ID	fID	value
001001	0	103162.0
001002	0	2181015.0
001003	0	2836266.0
001004	0	2850348.0
001005	0	2168825.0

Bottom Row Windows:

- trackInfo at /data/ [081505_L1_bd5.h5 in /Users/kkyoda/081505_L1_bd5.h5]**: Shows a table with columns `from` and `to`. Data rows include:

from	to
001001	002001
001003	002002
001004	002003
001002	002004
001005	002005
002004	003001
- 3D View:** A 3D coordinate system showing four white spheres representing tracked objects in a black space.

Text at Bottom:

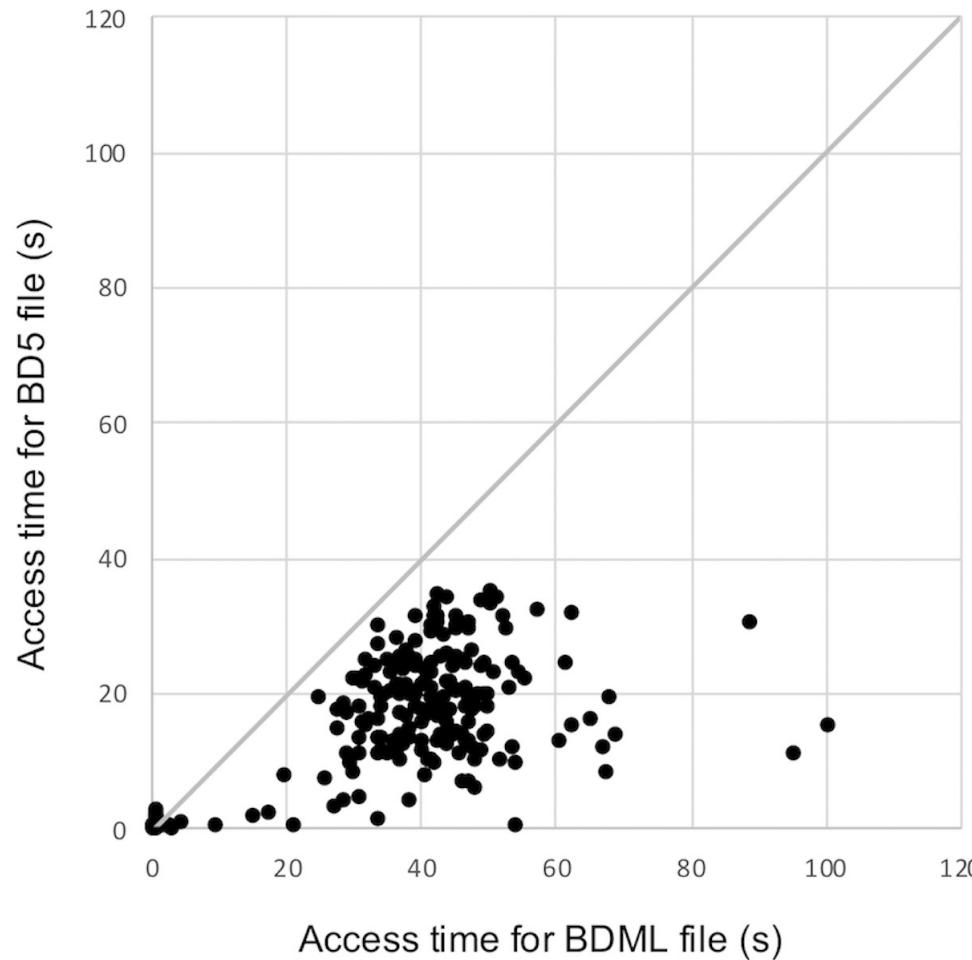
```

featureDef at /data/ [081505_L1_bd5.h5 in /Users/kkyoda] [ dims0, start0, count1, stride1 ]
scaleUnit at /data/ [081505_L1_bd5.h5 in /Users/kkyoda] [ dims0, start0, count1, stride1 ]
trackInfo at /data/ [081505_L1_bd5.h5 in /Users/kkyoda] [ dims0, start0, count23975, stride1 ]
  
```

Citation: (Bao et al., 2006)

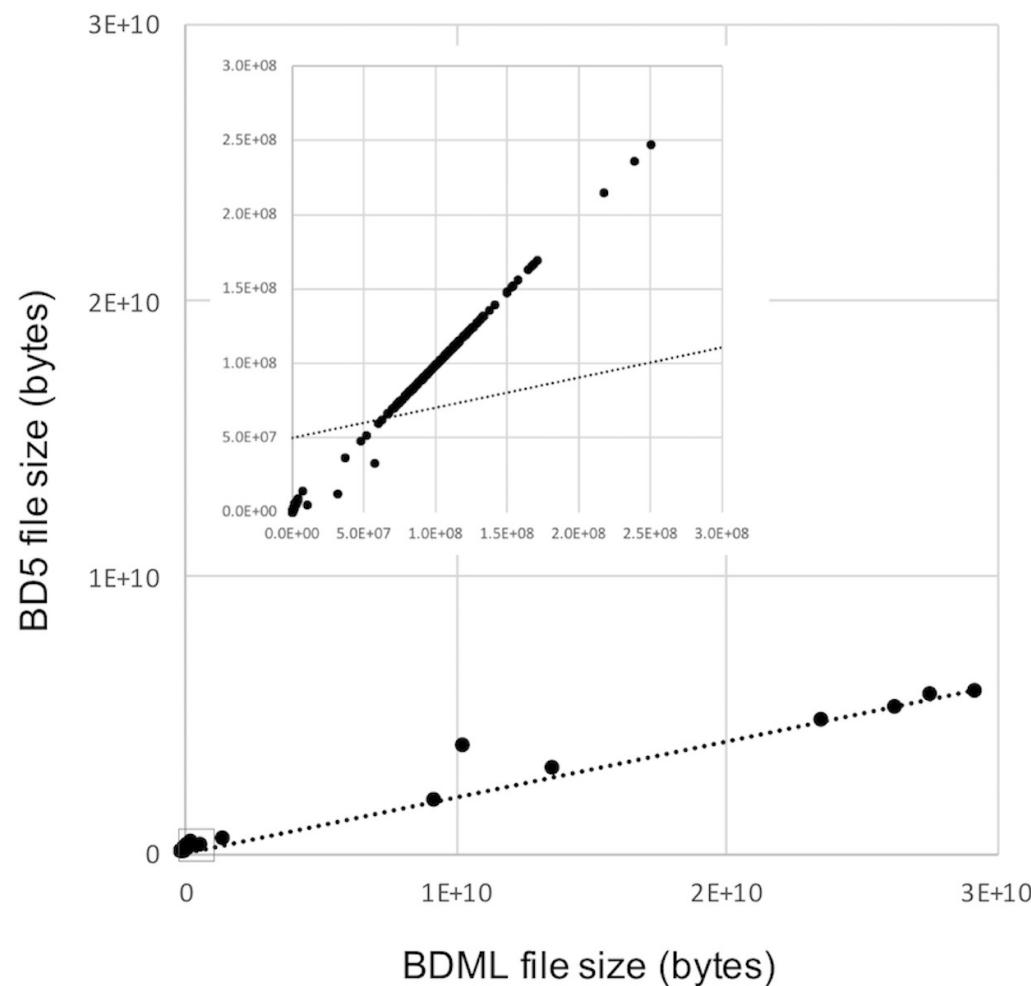
Fast data access

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



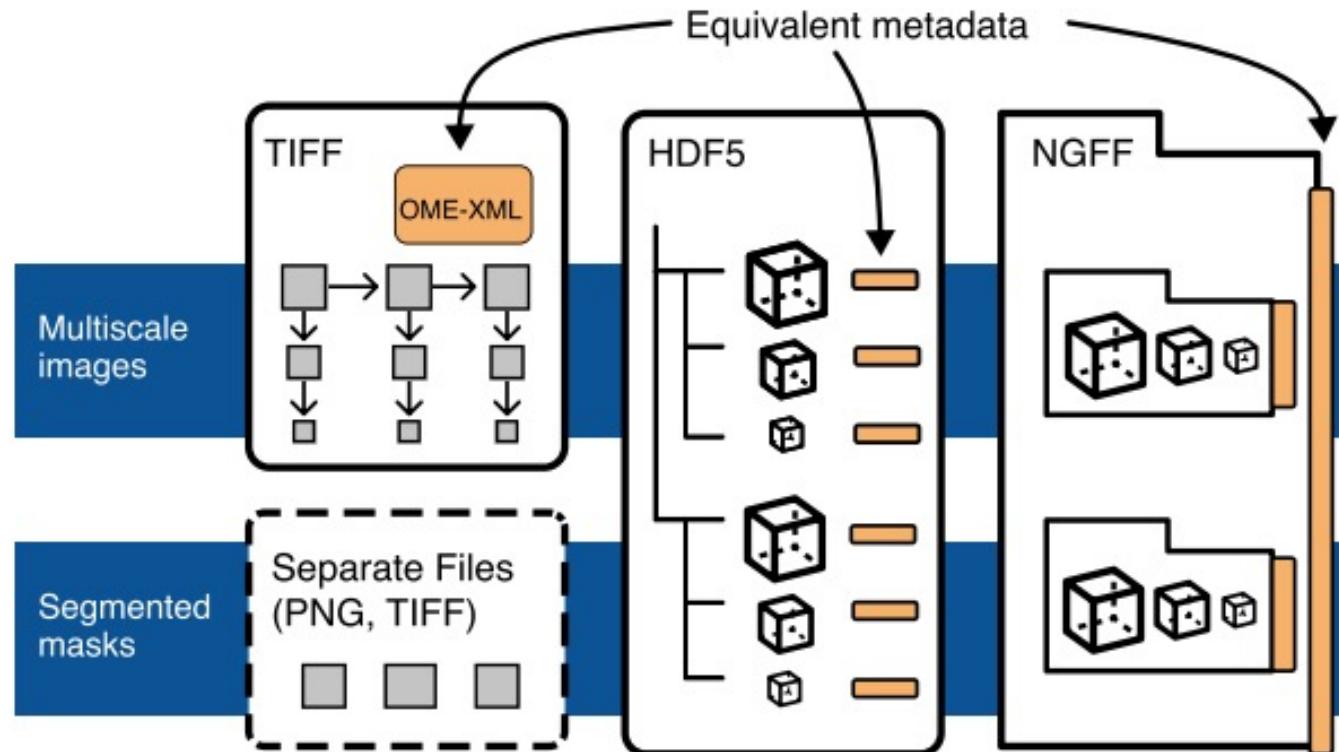
File size reduction

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



Bioimaging data format

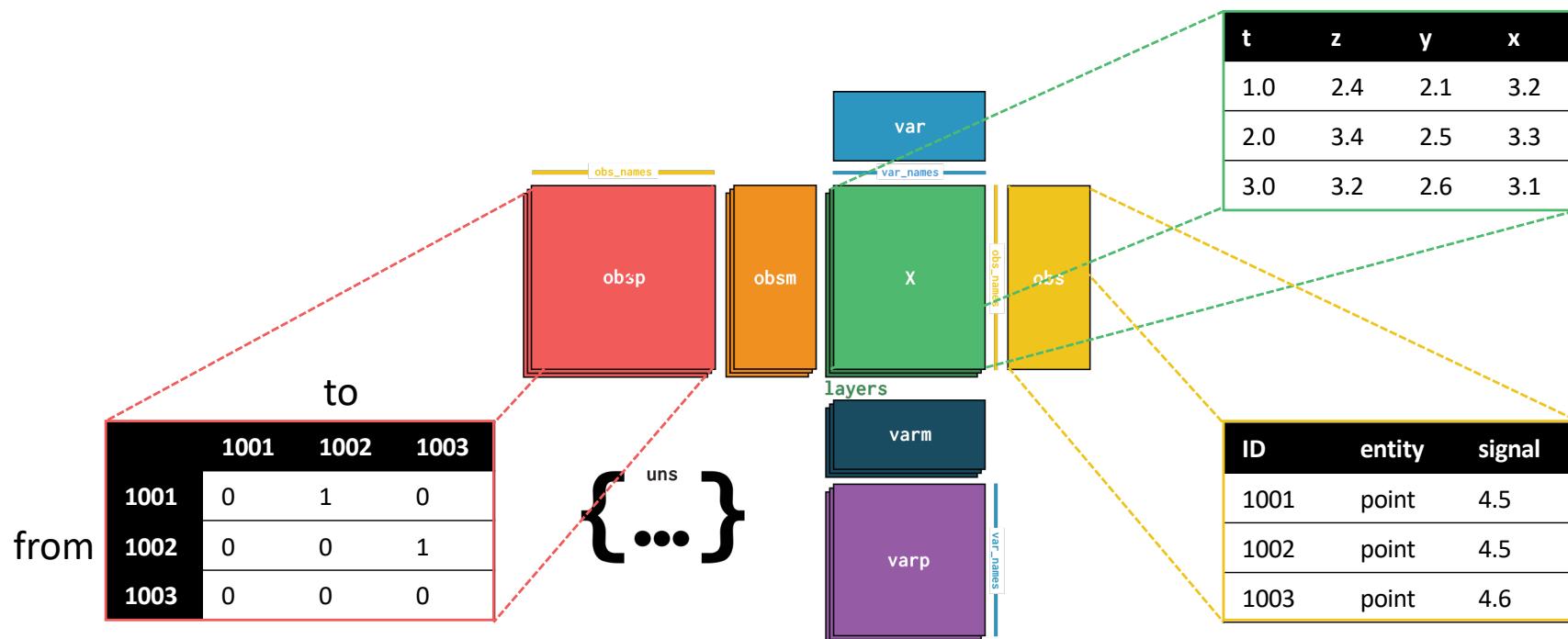
- Next generation file format for bioimaging data
 - ome-zarr is a zarr-based format for storing bioimaging data.



(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
|   |   |   |--x
```



```
|--- labels
```

```
|   |
|   |--0  Pixel-based ROI data
|   |   |--t
|   |   |   |--...
|   |
|--- dyn  Dynamics data
```

X	position data			
	t	z	y	x
0	1.0	39.922649	109.316254	307.414764
1	2.0	39.498207	113.885712	309.511322
2	3.0	39.999203	111.549751	315.685699
3	4.0	40.121613	112.833496	321.046295
4	5.0	42.206738	115.653198	332.908386

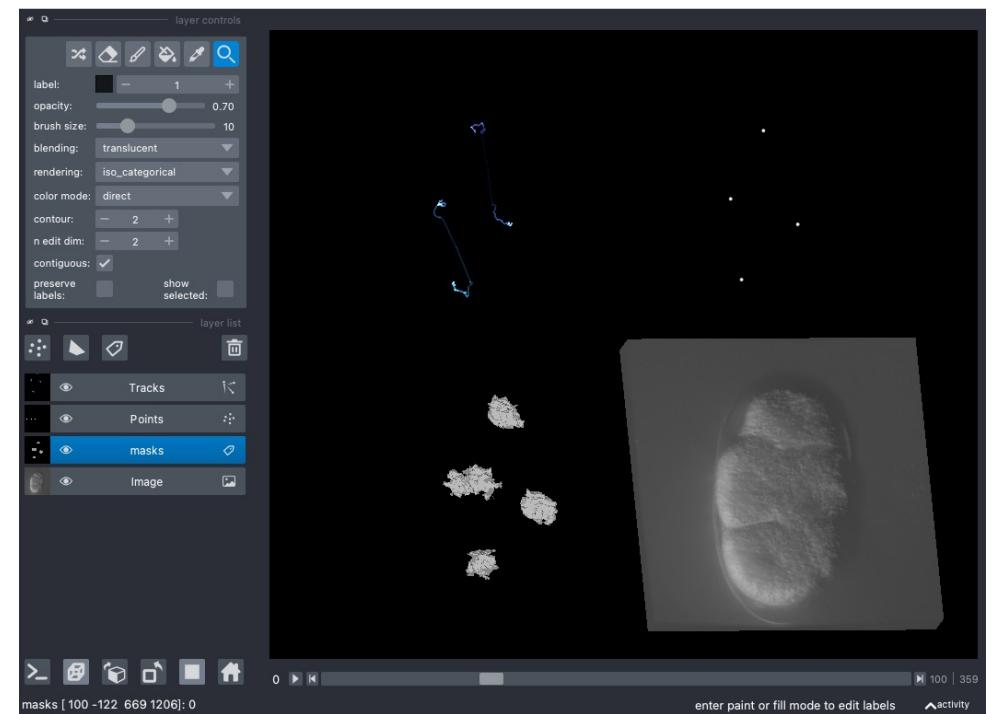
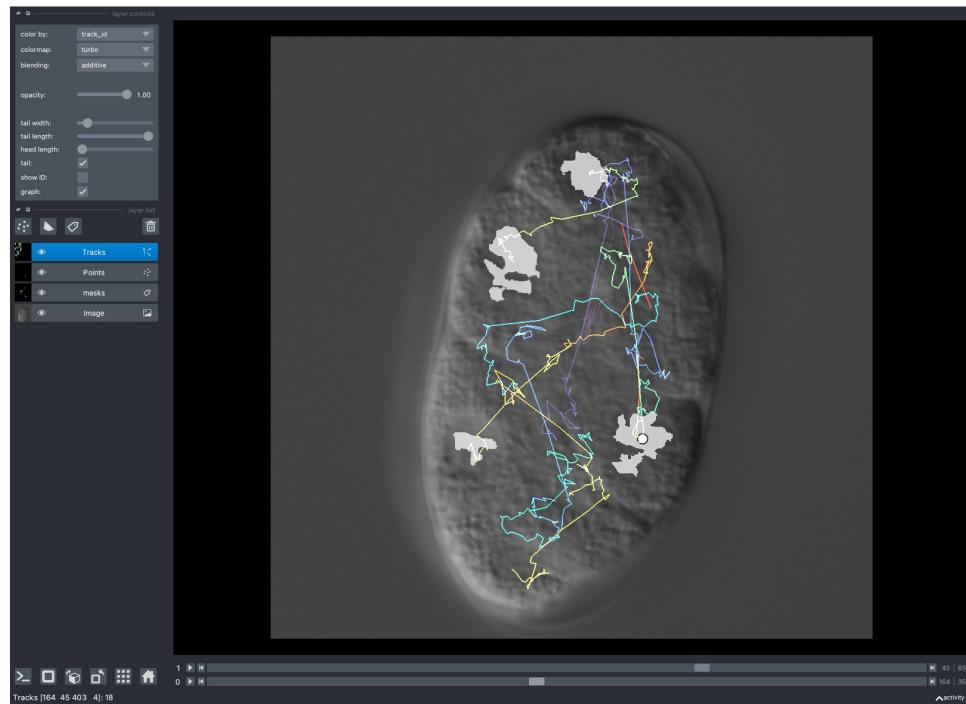
obs	feature data				
	id	entity	name	sphericity	volume
0	1000	line	P0	0.837884	76238.3
1	2000	line	P0	0.811657	88108.6
2	3000	line	P0	0.854491	110026.0
3	4000	line	P0	0.771704	107802.0
4	5000	line	P0	0.815484	138333.0

```
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.

SSBD:database

Organism: ex. C. elegans | Search | Clear

Introduction

Systems Science of Biological Dynamics database (SSBD:database) is an added-value database for biological dynamics. It provides a rich set of open resources for analyzing quantitative data and microscopy images of biological objects, such as single-molecule, cell, tissue, individual, etc., and software tools for analysis. Quantitative biological data and microscopy images are collected from a variety of species, sources, and methods. These include data obtained from both experiments and computational simulations.

Find the dataset from the search box above, or see the dataset list on the [Resources](#).

See [Citation Policies](#) (PDF) for citation instructions.

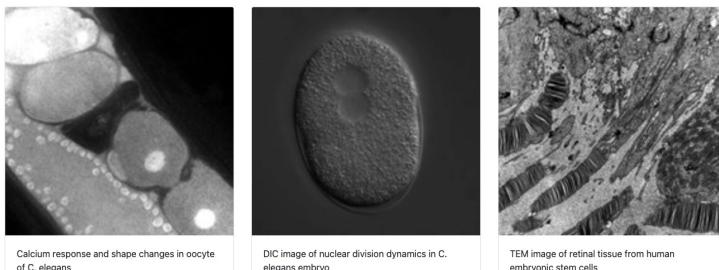
SSBD:database started operation in 2013, under the life science database integration promotion project of the Japan Science and Technology (JST), National Bioscience Database Center (NBDC). Currently, it is funded by RIKEN, JST, and Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science, and Technology of Japan (MEXT).

For the overview of the SSBD:database/repository, please refer to the paper, Tohsato, Y., Ho, K. H. L., Kyoda, K., and Onami S. (2016) "SSBD: a dataset of quantitative data of spatiotemporal dynamics of biological phenomena." *Bioinformatics*, 32(22): 3471-3479. <https://doi.org/10.1093/bioinformatics/btw417>

Japanese / 日本語

Samples

Microscopy Images

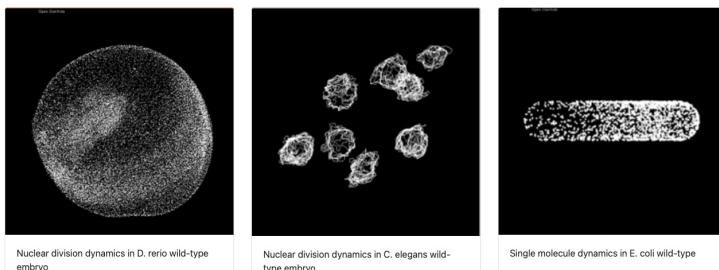


Calcium response and shape changes in oocyte of *C. elegans*

DIC image of nuclear division dynamics in *C. elegans* embryo

TEM image of retinal tissue from human embryonic stem cells

Quantitative Data



Nuclear division dynamics in *D. rerio* wild-type embryo

Nuclear division dynamics in *C. elegans* wild-type embryo

Single molecule dynamics in *E. coli* wild-type

Funding



OLSP
KAKENHI
CREST
NBDC
RIKEN

Tweets

Nov 16, 2021

SSBD:database / SSBD:repository Retweeted @onamix

Onami Lab @BDR_RIKEN is seeking a Scientist who will work for R&D of the SSBD @ssbd_en, an open added-value DB and repository for bioimage and biodynamics data. Those who are interested in data visualization, API, analysis framework are welcome to apply riken.jp/en/careers/res...



Seeking a Research Scientist, Postdoctoral Researcher or Technical Scientist (D02102)
Nov 16, 2021

SSBD:database / SSBD:repository Retweeted @onamix

Introduction movie for NeuroGT is up! NeuroGT is an image database of neurogenic tagging driver mouse lines. Watch how you can explore neurogenic-tagged neurons in the brain for 4 driver lines tagged on each single day during the neurodevelopmental period. ssbd.riken.jp/neurogt/misc/N...

Nov 13, 2021

SSBD:database / SSBD:repository Retweeted @ssbd_en

"Redundant roles of EGFR ligands in the ERK activation waves during collective cell migration" doi.org/10.26508/lsa.2 ... has published on Life Sci Alliance @LSAjournal Congrats! SSBD:repository shares the original images of the paper. See doi.org/10.24631/ssbd....

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.

