Databases and Web services for Electron Microscopy

Kawabata, Takeshi; Suzuki, Hirofumi; Kurisu, Genji

Institute for Protein Research, Osaka University

AsCA 2018, Dec 4, 2018
Data processing for EM Single Particle Analysis

2D images
- 509 DM4 movies (7420 x 7676 x 38; 32 bit)
  - EMPIAR-10013
  - 402.2 GB

EMPIAR

3D reconstruction

3D image
- 340 x 340 x 340 voxels
  - emd_5995.map
  - 150 MB 3.2 Å

EMDB

Atomic modeling

Atomic model
- 32878 atoms
  - pdb3j7h.ent
  - 2.6 MB

wwPDB

Beta-galactosidase
Data processing for EM Single Particle Analysis

2D images
509 DM4 movies (7420 x 7676 x 38; 32 bit)
EMPIAR-10013
402.2GB

3D reconstruction

3D image
340 x 340 x 340 voxels
emd_5995.map
150 MB 3.2Å

Atomic modeling

Atomic model
32878 atoms
pdb3j7h.ent
2.6 MB

Beta-galactosidase

EMPIAR

EMPIAR-PDBj

EMDB

PDBj

wwPDB

EM Navigator

Omokage search

gmfit
EMPIAR: Archive of raw 2D images of EM

Established by Ardan Patwardhan (EBI)

2D images for published 3D maps (most are in EMDB)

Purposes

1) Validation of 3D density maps in EMDB
2) Enhancing developments of new software
3) Training and education

Data size is huge! (110 TB in total, 187 entries)

We are building the mirror site in Japan (EMPIAR-PDBj).
We also plan to open the data-deposition site in Japan.

https://www.ebi.ac.uk/pdbe/emdb/empiar/

Mirror site of EMPIAR in PDBj

Cambridge (UK)

Osaka (Japan)

Time to transfer 1TB
Osaka -> Tokyo : 6-7 hours
UK -> Japan : 15 hours (using Aspera)

EMPIAR-PDBj provides a faster access for Asia and Oceania.

We also plan to open the data-deposition site in Japan.

https://empiar.pdbj.org
File size distribution of EMPIAR

Total file size: **110 TB**

Number of entries in total : 187
Number of entries (< 100 GB) : 105
Number of entries (> 1 TB) : 26
Number of entries (> 10 TB) : 3

Note: EMDB has **7141** entries

Smallest: 220 MB

Largest : 12 TB
Images for Picked Particles (84.4 GB)

169 MRCs (320 x 320 x ~100; 32 bit)

Many small 2D images

Necessary to validate **resolution** of 3D map by FSC.
2.2 Å resolution cryo-EM structure of beta-tubulin with a cell-permeant inhibitor

Bartesaghi A, Merk A, Banerjee S, Matthies D, Wu X, Mline JL, Subramaniam S
Science 348 1147-1151 (2015)
PMID: 25953817
DOI: 10.1126/science.aab1576

Multi-frame images (movies) (12.4 TB)
(7676 x 7420 x 38 x 1539; 32 bit)

1539 MRC movies (7676 x 7420 x 38; 32 bit)

Real raw data (too raw?)
Multi-frame images (movies) (321 GB)
(3710 x 3838 x 1338; 16 bit)

2.6 Å

321.4 GB

1338 TIFF images (3710 x 3838; 16 bit)
with a gain-reference file (K2-gain170629.mrc)

<table>
<thead>
<tr>
<th>Image set</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Unaligned multi-frame micrographs of beta-galactosidase recorded</strong></td>
</tr>
<tr>
<td><strong>Category:</strong> micrographs - multiframe</td>
</tr>
<tr>
<td><strong>Image format:</strong> TIFF</td>
</tr>
<tr>
<td><strong>No. of images or tilt series:</strong> 1338</td>
</tr>
<tr>
<td><strong>Frames per image:</strong> 49</td>
</tr>
<tr>
<td><strong>Image size:</strong> (3710, 3838)</td>
</tr>
<tr>
<td><strong>Pixel type:</strong> UNSIGNED 16 BIT INTEGER</td>
</tr>
<tr>
<td><strong>Details:</strong> Raw, non aligned micrograph moves in compressed TIFF format. The gain reference file is located in the same directory in MRC format. A part of this data was used as a tutorial data for Relion3.0.</td>
</tr>
</tbody>
</table>
EMPIAR-10013

Structure of β-galactosidase at 3.2-Å resolution obtained by cryo-electron microscopy

Publication: Structure of beta-galactosidase at 3.2-Å resolution obtained by cryo-electron microscopy
Bartesaghi A, Matthies D, Banerjee S, Merk A, Subramaniam S
PMID: 25071206
DOI: 10.1073/pnas.1402809113

Related PDB entry: 3j7h
Related EMDB entry: EMD-5995
Deposited: 2014-07-23
Released: 2014-08-07
Last modified: 2014-08-07
Dataset size: 442.5 GB
Dataset DOI: 10.6019/EMPIAR-10013
Version history: 1 2015-07-01 Directory structure reorganized. Entry has now full set of micrographs including frame alignment shifts, particle locations and defoc.

Contains: 3.2Å micrographs

442.5 GB

Multi-frame images (movie) (442 GB)
(7420x 7676 x 38; 32 bit)

509 MRC images (7420 x 7676 x 38; 32 bit)

Full set of multi-frame micrographs

Category: micrographs - multiframe
Image format: MRC
No. of images or tilt series: 509
Frames per image: 38
Image size: (7420, 7676)
Pixel type: 32 BIT FLOAT
Details: Full set of multiframe micrographs related to the dataset: 10012. Related PDB: 3J7H.
Please note that the number of frames varies between [5, 10, 25 and 38]
Structure of β-galactosidase at 3.2-Å resolution obtained by cryo-electron microscopy (frame-averaged micrographs)

Publication:
Structure of beta-galactosidase at 3.2-Å resolution obtained by cryo-electron microscopy
Bartesaghi A, Matthies D, Banerjee S, Mark A, Subramaniam S
PMID: 25071205
DOI: 10.1073/pnas.1402809111

Related PDB entry: 3j7h
Related EMDB entry: EMD-5995
Deposited: 2014-07-22
Released: 2014-08-07
Last modified: 2014-08-07
Dataset size: 108.0 GB

Dataset DOI: 10.6019/EMPIAR-10012
Version history:
1 2015-10-28 Restructured the entry directory to have "data" and ".private".

Image set
509 MRC images (7420 x 7676 x 1; 32 bit)

Motion-corrected still images

Beta Galactosidase micrographs

Category: micrographs - single frame
Image format: MRC
No. of images or tilt series: 509
Image size: (7420, 7676)
Pixel type: 32 BIT FLOAT
Details: These images were generated by aligning and then averaging multiframe images. All frames (1-38) from the unprocessed data were used. Related PDB: 3j7H.
Various Types of data in EMPIAR

1) **Micrographs** or **Particle images**

2) **Multiple-frame** or **Single-frame**
   - (Movie)
   - (Still image)

3) Gray scale depth: **32 bit** or **16 bit**

Different opinions about what is *raw* data

We accept everything you can provide.
Various Types of data in EMPIAR

1) **Micrographs or Particle images**
2) **Multiple-frame or Single-frame**
   (Movie)    (Still image)
3) **Gray scale depth:** 32 bit or 16 bit

Different opinions about what is *raw* data

We accept everything you can provide.

*Let’s submit your 2D images to EMPIAR!!*
EM Navigator: lovely EMDB server, user-friendly interface

The latest released data are shown in the top page.
Omokage Search

Fast shape similarity search against EMDB and PDB

EMDB ID / PDB ID can be used for the query shape.
Within a minute, similar-shape maps and atomic models are shown.
Map vs Map

Fitted by **gmfit**

Visualized by **MolMil**

*WebGL-based viewer*

If you click icons, superimposed structure is shown.
Gmfit: rigid-body fitting using GMM

GMM: Gaussian mixture model

Subunit1: CyclinA (1finB), Subunit2: Cyclin-dependent kinase (1finA)

Kawabata (2008), Biophys. J.
Resolution Revolution of EMDB

EMDB: 2018/11/28 u 1:3
EMDB: 2017/10/25 u 1:3
EMDB: 2015/09/30 u 1:3

Nentry vs Resolution (Angstrom)

EMD-4116 2.2Å
EMD-8751 4.4Å
EMD-2548 13Å
EMD-1209 25Å

Dates:
- 2018/11/28
- 2017/10/25
- 2015/09/30
Resolution revolution of GMM

3D density map
200³ voxels
Resolution: 4.1Å
EMD-8623

GMM
Ngauss=10
EM algorithm
Kawabata (2008), Biophys. J.
Resolution revolution of GMM

3D density map
- 200³ voxels
- Resolution: 4.1 Å
- EMD-8623

GMM
- Ngauss = 10
- EM algorithm
- Kawabata (2008), Biophys. J.

Resolution revolution !

GMM
- Ngauss = 546
- Down-sampling Gaussians
Types of rigid fitting problem

**single subunit**

- **global**
  - Atomic model
- **local**
  - Subunit

**multiple subunits**

- **global**
  - Subunit 1
  - Subunit 2
  - Subunit 3
- **local**
  - Subunit 1
  - Subunit 2
Types of rigid fitting problem

single subunit

multiple subunits

Old gmf

global

WEB service

Density map

Atomic model

local

Subunit

Density map

Subunit 1

Subunit 2

Subunit 3
Types of rigid fitting problem

- **single subunit**
  - **global**
    - Old gmfit
    - WEB service
    - Density map
    - Atomic model
  - New gmfit
    - WEB service
    - Density map

- **multiple subunits**
  - **local**
    - WEB service
    - Subunit
    - Subunit 1
    - Subunit 2
    - Subunit 3
Masked Segmentation & Fitting (MaskSegFit)

Case for “partial” fitting.

Introduce “mask” regions.

segmentation

fitting

segmentation

fitting
Example of rigid fittings

Global-Single

Local-Single

Global-Multiple

3sn6_ABGNR

3sn6_ABGN

4l6r_A

4l6r_A + 3sn6_ABGN
Gmfit WEB : 1-to-1 global fitting

Pairwise gmfit
Superimpose two density maps or atomic models by gmfit

<table>
<thead>
<tr>
<th>DENSITY MAP</th>
<th>ATOMIC MODEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input ID</td>
<td>Upload Map File</td>
</tr>
<tr>
<td>emdb id</td>
<td>level</td>
</tr>
<tr>
<td>TARGET</td>
<td>(be superimposed)</td>
</tr>
<tr>
<td>REFERENCE</td>
<td>(fixed)</td>
</tr>
</tbody>
</table>
Gmfit WEB : bad for 1-to-1 local fitting...

Fit one subunit on the map of homo tetramer

Ngauss=20

Failed ....
New Gmfit WEB : 1-to-1 local fitting

Local/Partial fitting : masked segfit gmf

Superimpose an atomic model on a density map by gmf
using masked segmentation fitting algorithm.
Local/partial fitting can be performed one given atomic model versus a given density map.

This new Web service will be available soon!

<table>
<thead>
<tr>
<th>input form</th>
<th>map or atomic model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density map</td>
<td>EMDB-ID: 5995 or upload map file:</td>
</tr>
<tr>
<td>(fixed)</td>
<td>cutoff level: 0.0224</td>
</tr>
<tr>
<td>Atomic model</td>
<td>PDB_ID: 1f4h or upload PDB file:</td>
</tr>
<tr>
<td>(to be superimposed)</td>
<td>chainID: A  assemblyID:</td>
</tr>
</tbody>
</table>

Fit one subunit on the map of homo tetramer

Fit
on
Ngauss=20

Success!
Ngauss=100