Data compression for EMPIAR

Andrii Iudin, EMBL-EBI

Notes

- Sizes are provided as in Mac OS X, that is in decimal GB (Gigabyte) rather than binary GiB (Gibibyte) values. This matches up with how hardware manufacturers measure the capacity of storage devices.
- 7-ZIP by default uses LZMA2 compression algorithm and options for maximum compression.
- By default XZ has standard options but for the multiple image tests (see the last two examples at the end of the document) we have used options for maximum compression.
- Tested on:
  Mac OS X version 10.12.3
  Processor: 2.5 GHz Intel Core i7
  Memory: 16 GB 1600 MHz DDR3
  Storage: 1 TB SSD

Entry 10055

SBF-SEM of late schizont-stage malaria parasite infected red blood cell

Category: micrographs - single frame
Image format: MRC
No. of images or tilt series: 90
Image size: (1600, 1600)
Pixel type: UNSIGNED BYTE
Pixel spacing: (82.2 +, 82.2 +)
Details: SBF-SEM Species: Plasmodium falciparum Microscope: ZEISS MERLIN at 1.2kV Detector: Gatan BSED Z sampling: 700A

File size: 230.4 MB

Compression type	% of compression	Final size (MB)
IMOD mrc2tif "zip"	16	193.5
IMOD mrc2tif "lzw"	-2.3	235.7
IMOD mrc2tif no compression	-0.2	230.8
ZIP	16.6	192.1
7-ZIP	30.3	160.7
BZIP2	27.6	166.7
RAR	23.8	175.5
XZ	30.5	160.2

Entry 10010

Brome Mosaic Virus micrographs - non gain corrected

Category: micrographs - multiframe
Image format: TIFF
No. of images or tilt series: 424
Frames per image: 37
Image size: (4096, 3072)
Pixel type: UNSIGNED BYTE
Pixel spacing: (0.99 +, 0.99 +)

Details: Each frame is an individual TIFF image. The FinalImage.tif included for each micrograph is a sum image of the frames without drift correction and damage compensation. The .box files (found in the directory data/micrographs/non_gain_corrected/box_files)describe particle locations but are rotated 90 degrees relative to the micrograph frames. The dark and gain correction images are found in data/micrographs/non_gain_corrected/gain_correction and are called: dark_frame.tif gain_frame.tif Dataset is related to EMPIAR-10011. This set is a subset of the images in data/micrographs/gain_corrected which were actually used in the final reconstruction.

File size: 25.2 MB

Compression type	% of compression	Final size (MB)
IMOD tif2mrc | IMOD mrc2tif "zip"	56.3	11
IMOD tif2mrc | IMOD mrc2tif "lzw"	54.4	11.5
IMOD tif2mrc | IMOD mrc2tif no compression	0	25.2
ZIP	59.1	10.3
7-ZIP	67.1	8.3
BZIP2	67.5	8.2
RAR	66.7	8.4
XZ	67.1	8.3

Brome Mosaic Virus micrographs - gain corrected

Category: micrographs - multiframe

Image format: MRC

No. of images or tilt series: 647

Frames per image: 37

Image size: (4096, 3072)

Pixel type: 32 BIT FLOAT

Pixel spacing: (0.99 +, 0.99 +)

Details:

These images have been gain corrected and are in MRC format. This is the full set of images collected for the project and includes those that were not finally used for image processing. data/micrographs/non_gain_corrected contains only the subset used in the reconstruction.

File size: 1.86 GB

Compression type	% of compression	Final size (GB)
IMOD mrc2tif "zip"	11.3	1.65
IMOD mrc2tif "lzw"	-12.4	2.09
ZIP	11.3	1.65
7-ZIP	17.7	1.53
BZIP2	10.2	1.67
RAR	13.4	1.61
XZ	14.5	1.59

Entry 10025

Raw movies of T20S Proteasome at 2.8 Å Resolution

Category: micrographs - multiframe

Image format: MRC

No. of images or tilt series: 196

Frames per image: 38

Image size: (7420, 7676)

Pixel spacing: (0.66 +, 0.66 +)

Details: Super-resolution data-acquisition. dark-amibox05-0.mrc and norm-amibox05-0.mrc are used for gain correction. run1_shiny_mp007_data_dotstar.txt is a Relion star file with particle coordinates etc.

File size: 2.16 GB

Compression type	% of compression	Final size (MB)
IMOD mrc2tif "zip"	78.6	462.9
IMOD mrc2tif "lzw"	80.4	423.4
ZIP	78.8	457.4
7-ZIP Ultra in 7zX Mac OS X GUI app	83.1	365
7-ZIP Ultra PPMd 7za command line	82.2	383.3
7-ZIP Ultra LMZA2 7za command line	83.1	364.6
BZIP2	80.3	426.4
KLB	80.2	427
Packbits compression	20.4	1720
Packbits + zip	74.8	543.5
Packbits + 7-zip	80.7	417.9
LZ4	71.8	609.2
BLOSC	50.9	1060
RAR	79.8	435.5
XZ (Mac OS X pre-installed version)	83.0	367.7

Testing on multiple files:

Original size (GB)	7-ZIP Final size (MB)	XZ Final size (MB)
2.16	365	364.5
2.16	362.6	362.3

2.16	364.2	363.8
2.16	363.7	363.4
2.16	363.9	363.6

Entry 10064

VPP_Ribosome, CTEM_Ribosome

Category: tilt series

Image format: MRC

No. of images or tilt series: 11

Image size: (3710, 3710)

Pixel type: 32 BIT FLOAT

Pixel spacing: (2.62 +, 2.62 +)

Details:

CTEM_tomo-->Conventional data set with one defocus value.

File size: 3.36 GB

Compression type	% of compression	Final size (GB)
IMOD mrc2tif "zip"	13.9	2.95
IMOD mrc2tif "lzw"	-9.8	3.69
ZIP	12.5	2.94
7-ZIP Ultra PPMd 7za command line	19.0	2.72
7-ZIP Ultra LMZA2 7za command line	20.5	2.67
BZIP2	13.4	2.91
RAR	13.1	2.92
XZ	19.0	2.72

Testing on multiple files:

Original size (GB)	7-ZIP Final size (GB)	XZ Final size (GB)
3.36	2.67	2.66
3.36	2.68	2.67
3.19	2.59	2.58
3.19	2.58	2.57
3.36	2.68	2.68