Map Challenge: Analysis using Pair Comparison Method

February 6, 2017

Abstract

This document presents the analysis performed over the Challenge dataset using the Pair Comparison Method which is described in an accompanying document (see section entitled "Second method: pair comparison method"). This algorithm is able to sort reconstructions and assign a level of significance to the sorting.

Keywords: structural biology, electron microscopy, 3D reconstruction, high-resolution, benchmarking, challenge

1 Material and Methods

For all datasets (specimens) the following workflow has been followed (small deviations for the general protocol are described in the Results section):

For a given specimen:

- Create a reference volume by randomly rotating the first uploaded 3D map with sampling rate equal to the input data.
- Align each 3D map with respect to the reference volume using Chimera.
  - Load reference and problem volume: `chimera ref.mrc emcd$NUM_$SPECIMEN_unfiltered.mrc`
  - Place origin of coordinates in 3D map center: `viewer -> coordinates -> center`
  - Manually align the different 3D maps with the reference.
  - Refine alignment with command `viewer -> tools -> fit in map` (3D maps with $CC < 0.9$ are dropped)
  - Interpolate aligned map in the reference system of coordinates with the reference sampling rate: `vop resample #1 onGrid #0`
  - Save interpolated 3D map:`viewer -> file -> save_as`
- Mask 3D maps using a sphere with Scipion.
- Apply Pair Comparison method.
  - compute magnitudes called $\overline{FSC_{i,j}}$ and $\overline{FSC_i}$ (which are described in the accompanying material)

Once $\overline{FSC_{i,j}}$ and $\overline{FSC_i}$ have been computed:

- sort the 3D maps based on $\overline{FSC_i}$

- test null hypothesis "two 3D maps can be distinguished" based on $\overline{FSC_{i,j}}$. We will assume that two volumes can be distinguished if the P-value resulting from applying the Wilcoxon test to them is smaller than 0.05.

For some datasets we have also computed the R-factor (and average Fourier shell correlation) as follows:

- In chimera the provided pdb file has been manually fitted with respect to the reference volume (after applying a B-factor correction).
- In chimera the fitted pdb has been further refined by executing a rigid fit (`volume viewer -> Tools -> Fit`)
- Refmac has been executed using the resulting pdb file and the different uploaded 3D maps as input (3D maps have been corrected applying the B-factor).

2 Results

In the following we describe the results obtained for each specimen. In all cases we are going to show two tables. In the first table we just show the 3D maps sorted by the feature $\overline{FSC_i}$ while in the second we show which pairs of 3D maps can be distinguished with a P-value smaller than 0.05. Cells in this second table contain the P-values obtained from comparing the two 3D maps associated to the corresponding row and column. P-values greater or equal to 0.05 will be outlined using a red font. Since 3D maps are also sorted in the second table, in general, pairs of volumes that are close tend to be harder to distinguish than pairs of volumes that are farther away.

To clarify the meaning of the second table, let us discuss in detail one of its rows (see table 1). In this row we see the results of testing the null hypothesis: can the 3D map **emcd104** be distinguished from the map XXX? We see that map **emcd104** cannot be distinguished with P-Value smaller than 0.05 from 3D maps **emcd169** and **emcd120** but it can be distinguished from the rest.

	emcd143	**emcd132**	**emcd165**	**emcd169**	**emcd104**	**emcd120**	**emcd168**	**emcd153**	**emcd158**
emcd104	0.01	0.01	0.01	0.6		0.2	0.02	0.01	0.01

Table 1:

Note1: $\overline{FSC_{i,j}}$ used for sorting and for computing the P-values are available at the end of this document (see appendix A).

2.1 GroEL *in Silico*

$\overline{\text{FSC}_\text{i}}$	3Dmap	R-factor(Rms BondLength)
47.65	emcd143	0.6120(0.0076)
46.1	emcd132	0.6192(0.0076)
45.75	emcd165	0.6092(0.0076)
43.35	emcd169	0.5316(0.0076)
42.6	emcd104	0.5111(0.0089)
42.5	emcd120	0.6003(0.0076)
36.8	emcd168	0.5769(0.0076)
35.65	emcd153	0.5955(0.0076)
35	emcd158	0.5853(0.0076)

Table 2: Sorting based on $\overline{FSC_i}$ for specimen GroEL *in silico*. Third column represents the R-factor and Rms BondLenght computed with refmac. Note the high values of R-factor that indicate a very poor fitting.

	emcd143	emcd132	emcd165	emcd169	emcd104	emcd120	emcd168	emcd153	emcd158
emcd143	0.	0.4	0.01	0.09	0.01	0.01	0.01	0.01	0.01
emcd132	0.4		0.3	0.1	0.01	0.01	0.01	0.01	0.01
emcd165	0.01	0.3		0.6	0.01	0.01	0.01	0.01	0.01
emcd169	0.09	0.1	0.6		0.6	1	0.01	0.01	0.01
emcd104	0.01	0.01	0.01	0.6		0.2	0.02	0.01	0.01
emcd120	0.01	0.01	0.01	1	0.2		0.02	0.01	0.01
emcd168	0.01	0.01	0.01	0.01	0.02	0.02		0.4	0.01
emcd153	0.01	0.01	0.01	0.01	0.01	0.01	0.4		0.6
emcd158	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.6	

Table 3: P-values resulting of comparing all 3Dmap pairs for specimen GroEL *in silico*

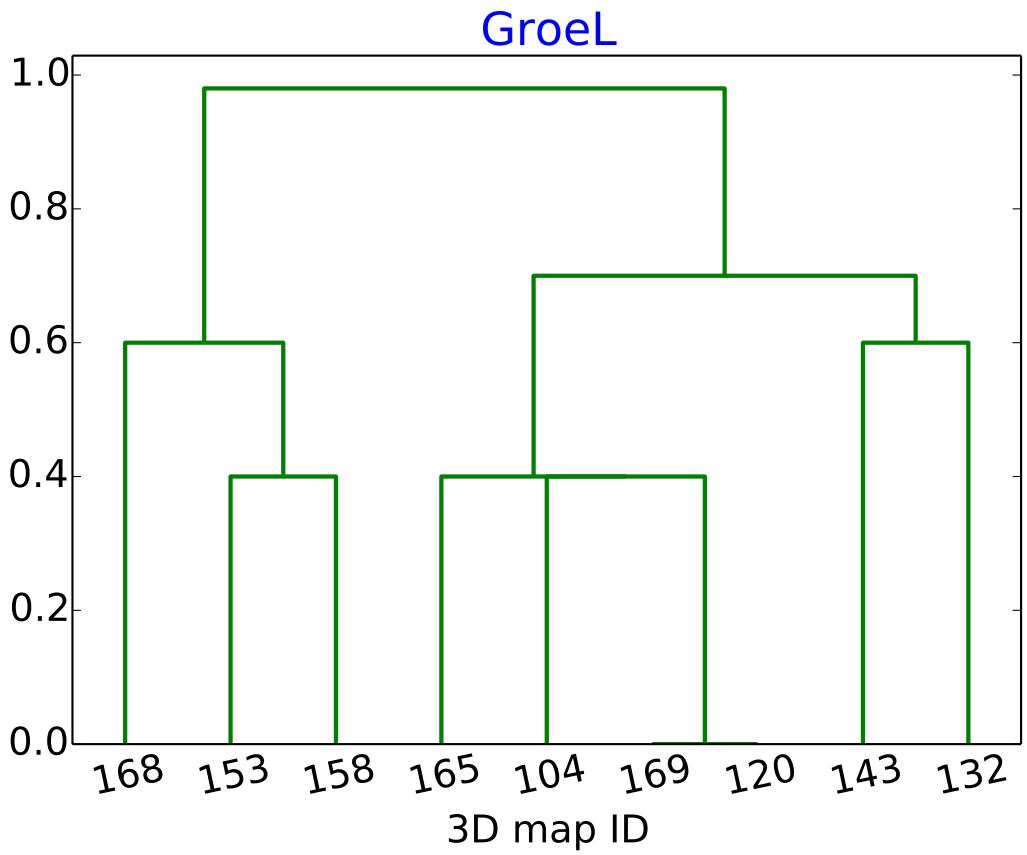

Figure 1: Dendrogram for GroeL hierarchical clustering using as similitude matrix table 3.

2.2 T20S Proteasome

$\overline{\mathbf{FSC_i}}$	**3Dmap**	RF 7 2.5	RF 5 2.5	AVE FSC
114.94	emcd108	0.27	0.26	0.76
114.72	emcd103	0.32	0.32	0.68
114.55	emcd107	0.36	0.30	0.63
112.96	emcd141	0.35	0.30	0.68
112.35	emcd145	0.39	0.29	0.46
111.26	emcd144	0.39	0.30	0.45
106.49	emcd162	0.31	0.26	0.62
101.79	emcd130	0.24	0.27	0.47

Table 4: Sorting based on $\overline{FSC_i}$ for specimen T20S Proteasome. 4^{th}, 5^{th} and 6^{th} columns represents the R-factor between frequencies 7-2.5Å, frequencies 5-2.5Åand the average Fourrier shell correlation as reported by refmac. Note that the average Fourier correlation produce a sorting similar to the new method while a sorting produced by R-factor would be quite different.

	emcd108	**emcd103**	**emcd107**	**emcd141**	**emcd145**	**emcd144**	**emcd162**	**emcd130**
emcd108		0.3	0.1	0.3	0.4	0.4	0.02	0.02
emcd103	0.3		0.9	0.02	0.3	0.3	0.02	0.02
emcd107	0.1	0.9		0.2	0.4	0.3	0.02	0.02
emcd141	0.3	0.02	0.2		0.4	0.4	0.02	0.02
emcd145	0.4	0.3	0.4	0.4		0.02	0.7	0.02
emcd144	0.4	0.3	0.3	0.4	0.02		0.7	0.02
emcd162	0.02	0.02	0.02	0.02	0.7	0.7		0.02
emcd130	0.02	0.02	0.02	0.02	0.02	0.02	0.02	

Table 5: P-values resulting of comparing all 3Dmap pairs for specimen T20S Proteasome

Note: 3Dmaps emcd130 and emcd131 are identical (we have ignored emcd131).

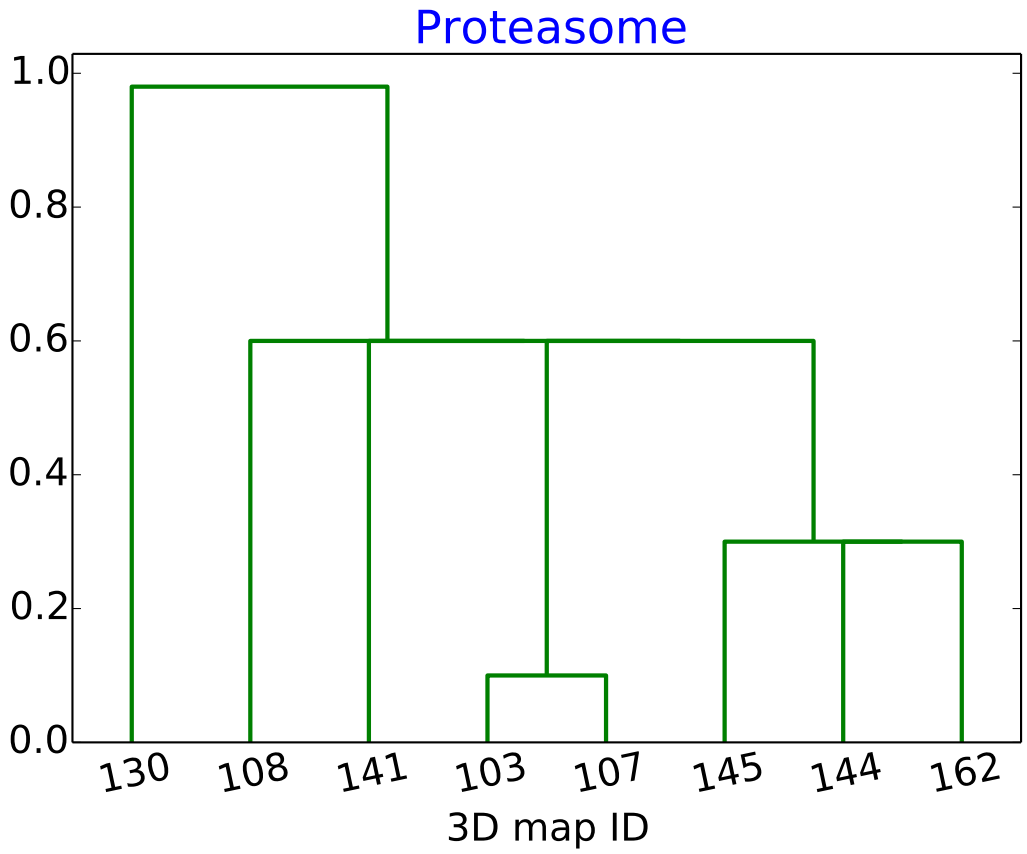

Figure 2: Dendrogram of proteasome hierarchical clustering.

2.3 Apo Ferritin

$\overline{\text{FSC}_\text{i}}$	3Dmap
58.85	emcd166
57.15	emcd118
53.89	emcd121
52.46	emcd112
51.64	emcd124
32.74	emcd155
27.72	emcd147
27.19	emcd122

Table 6: Sorting based on $\overline{FSC_i}$ for specimen Apo-Ferritin.

	emcd166	**emcd118**	**emcd121**	**emcd112**	**emcd124**	**emcd155**	**emcd147**	**emcd122**
emcd166		0.02	0.6	0.6	0.04	0.02	0.02	0.02
emcd118	0.02		0.7	0.7	0.04	0.02	0.02	0.02
emcd121	0.6	0.7		0.3	0.6	0.02	0.02	0.02
emcd112	0.6	0.7	0.3		0.7	0.02	0.02	0.02
emcd124	0.04	0.04	0.6	0.7		0.02	0.02	0.02
emcd155	0.02	0.02	0.02	0.02	0.02		0.04	0.02
emcd147	0.02	0.02	0.02	0.02	0.02	0.04		0.6
emcd122	0.02	0.02	0.02	0.02	0.02	0.02	0.6	

Table 7: P-values resulting of comparing all 3Dmap pairs for specimen Apo-Ferritin

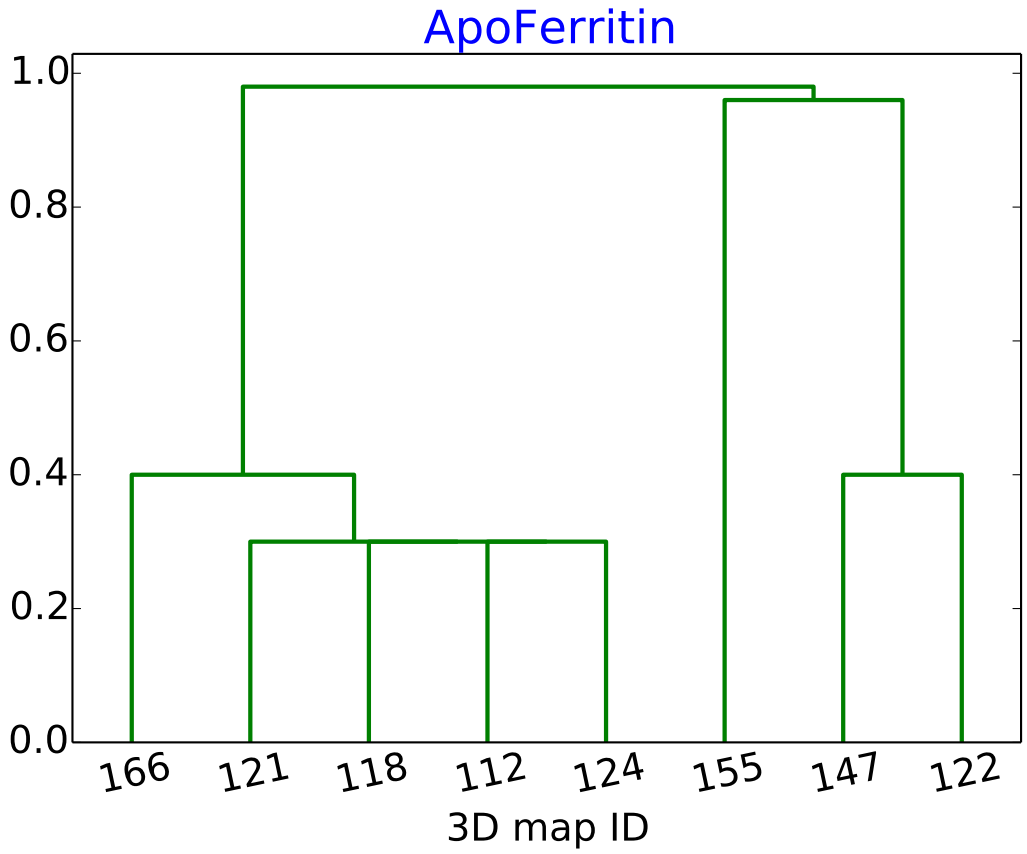

Figure 3: Dendrogram of ApoFerritin hierarchical clustering.

2.4 TRPV1 Channel

$\overline{\text{FSC}_\text{i}}$	3Dmap
47.93	emcd161
47.56	emcd135
46.91	emcd133
46.46	emcd115
44.46	emcd101
36.74	emcd156
36.56	emcd163

Table 8: Sorting based on $\overline{FSC_i}$ for specimen TRPV1 Channel.

	emcd161	emcd135	emcd133	emcd115	emcd101	emcd156	emcd163
emcd161		0.5	0.5	0.1	0.04	0.04	0.04
emcd135	0.5		0.04	0.6	0.5	0.04	0.04
emcd133	0.5	0.04		0.5	0.6	0.04	0.04
emcd115	0.1	0.6	0.5		0.04	0.04	0.04
emcd101	0.04	0.5	0.6	0.04		0.04	0.04
emcd156	0.04	0.04	0.04	0.04	0.04		0.8
emcd163	0.04	0.04	0.04	0.04	0.04	0.8	

Table 9: P-values resulting of comparing all 3Dmap pairs for specimen TRPV1 Channel

Note: 3Dmap emcd146 could not be aligned with cross-correlation coefficient >0.9 and was ignored ion the sorting procedure.

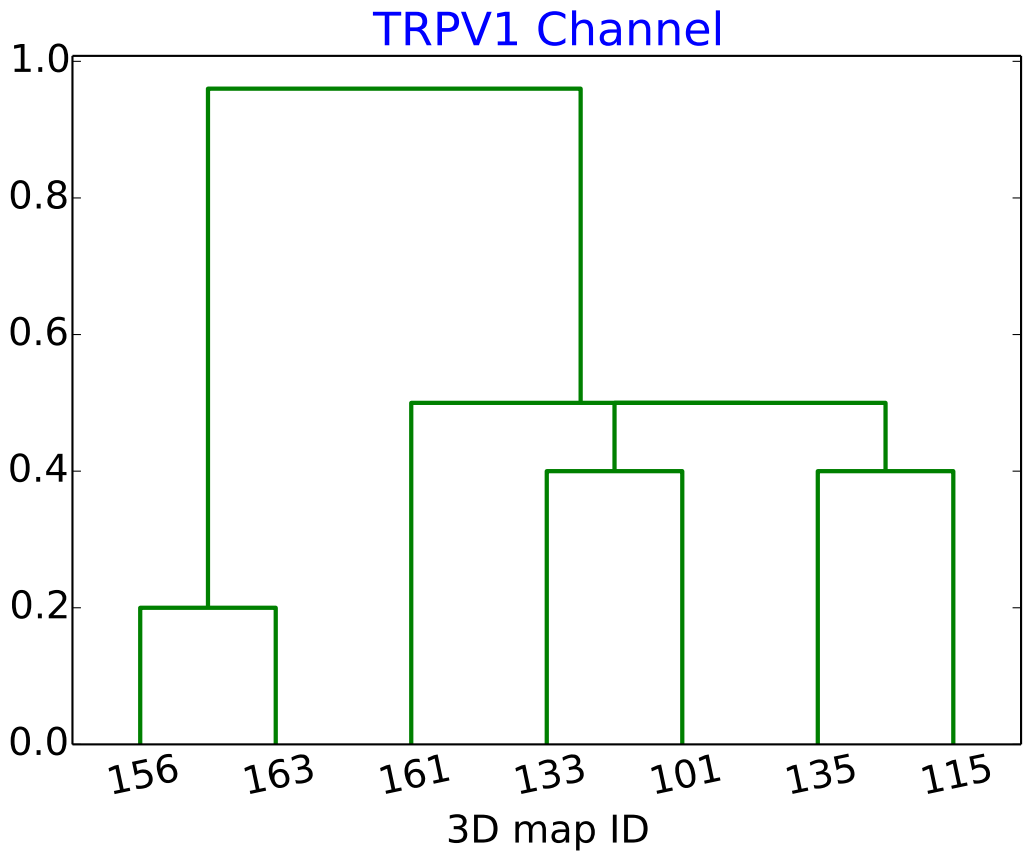

Figure 4: Dendrogram of TRPV1 Channel hierarchical clustering.

2.5 80S Ribosome

$\overline{\mathbf{FSC_i}}$	**3Dmap**	R-factor	ave FSC
218.89	emcd123	0.53(0.005)	0.1959
208.88	emcd151	0.54(0.005)	0.1538
204.22	emcd150	0.54(0.005)	0.1459
203.06	emcd149	0.54(0.005)	0.1364
198.06	emcd126	0.54(0.005)	0.1634
193.87	emcd114	0.53(0.005)	0.2273
190.91	emcd125	0.52(0.005)	0.2704
180.48	emcd127	0.54(0.005)	0.1194
177.90	emcd148	0.55(0.005)	0.0717
176.32	emcd128	0.55(0,005)	0.1044
168.14	emcd119	0.54(0.006)	0.1484
164.88	emcd111	0.54(0.005)	0.1676
141.11	emcd129	0.55(0.006)	0.0382

Table 10: Sorting based on $\overline{FSC_i}$ for specimen 80S Ribosome. 4^{th} and 5^{th} columns represents the R-factor between frequencies 4.5-3Åand the average Fourier shell correlation as reported by refmac. Note that the very low value of the average Fourier correlation and the high value of the R-factor that suggest a poor agriment between pdb file and 3D maps.

	e123	**e151**	**e150**	**e149**	**e126**	**e114**	**e125**	**e127**	**e148**	**e128**	**e119**	**e111**	**e129**
emcd123		0.07	0.03	0.02	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00
emcd151	0.07		0.07	0.09	0.06	0.09	0.01	0.00	0.00	0.00	0.01	0.01	0.00
emcd150	0.03	0.07		0.02	0.4	0.2	0.07	0.00	0.02	0.00	0.03	0.01	0.00
emcd149	0.02	0.09	0.02		0.4	0.2	0.1	0.00	0.02	0.00	0.03	0.01	0.00
emcd126	0.00	0.06	0.4	0.4		0.37	0.1	0.00	0.05	0.00	0.02	0.02	0.00
emcd114	0.01	0.09	0.2	0.2	0.37		0.28	0.09	0.06	0.02	0.03	0.03	0.00
emcd125	0.00	0.01	0.07	0.1	0.1	0.28		0.09	0.06	0.01	0.03	0.01	0.00
emcd127	0.00	0.00	0.00	0.00	0.00	0.09	0.09		0.4	0.00	0.06	0.06	0.00
emcd148	0.01	0.00	0.02	0.02	0.05	0.06	0.06	0.4		0.7	0.1	0.07	0.00
emcd128	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.00	0.7		0.1	0.07	0.00
emcd119	0.00	0.01	0.03	0.03	0.02	0.03	0.03	0.06	0.1	0.1		0.01	0.00
emcd111	0.00	0.01	0.01	0.01	0.02	0.03	0.01	0.06	0.07	0.07	0.01		0.01
emcd129	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	

Table 11: P-values resulting of comparing all 3Dmap pairs for specimen 80S Ribosome. In order to make the page fit in the page the name of the 3Dmaps has been shorten from the canonical form *emcd129* to *e129*

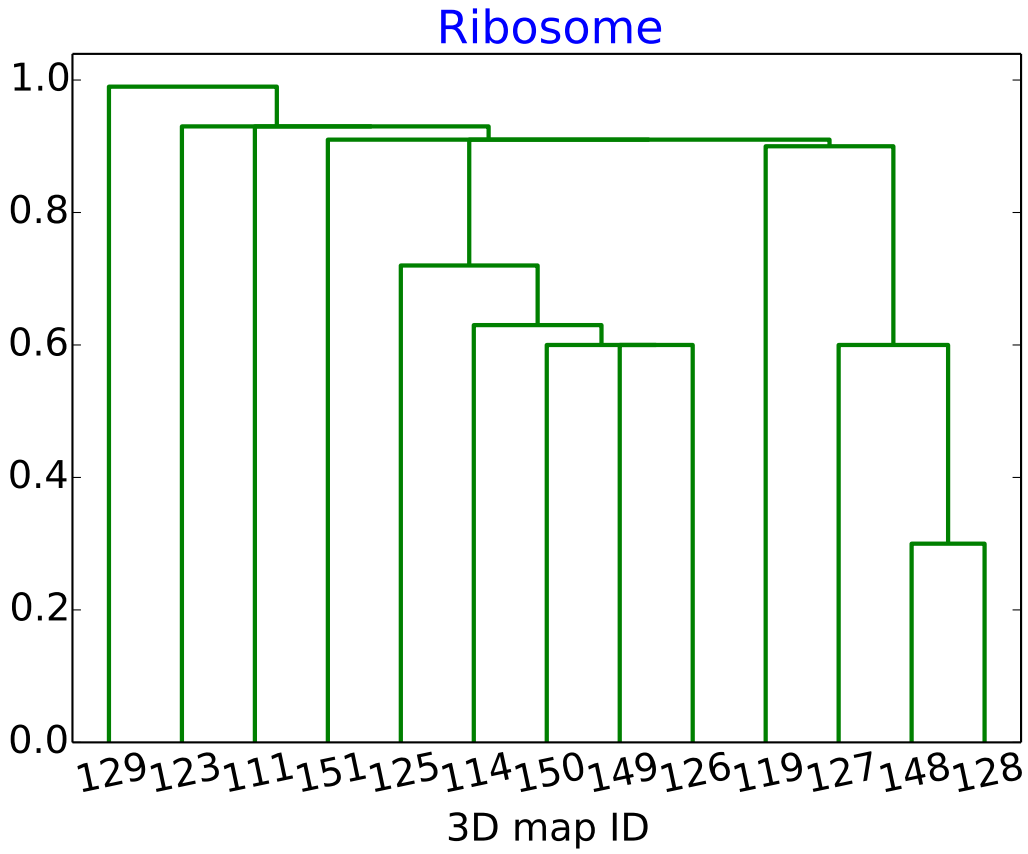

Figure 5: Dendrogram of Ribosome hierarchical clustering.

2.6 Brome Mosaic Virus

Alignment for this specimen was made based on symmetry. First, the orientation of the symmetry axes was detected and all maps were rotated so that they present i3 orientation (for example, 3D maps with i1 orientation were rotated using the Euler angles (0,63.43494882,0)). We note here that no 3D map presented originally i3 orientation. (Our notation for symmetry orientation is summarized in `http://xmipp.cnb.csic.es/twiki/bin/view/Xmipp/Symmetry`.)

$\overline{\mathbf{FSC_i}}$	**3Dmap**
92.04	emcd142
90.33	emcd137
89.69	emcd140
87.11	emcd102
79.89	emcd136
69.92	emcd110
56.12	emcd152

Table 12: Sorting based on $\overline{FSC_i}$ for specimen Brome Mosaic Virus.

	emcd142	**emcd137**	**emcd140**	**emcd102**	**emcd136**	**emcd110**	**emcd152**
emcd142		0.6	0.6	0.2	0.04	0.04	0.04
emcd137	0.6		0.8	0.6	0.04	0.04	0.04
emcd140	0.6	0.8		0.2	0.1	0.04	0.04
emcd102	0.2	0.6	0.2		0.3	0.04	0.04
emcd136	0.04	0.04	0.1	0.3		0.07	0.04
emcd110	0.04	0.04	0.04	0.04	0.07		0.04
emcd152	0.04	0.04	0.04	0.04	0.04	0.04	

Table 13: P-values resulting of comparing all 3Dmap pairs for specimen Brome Mosaic Virus

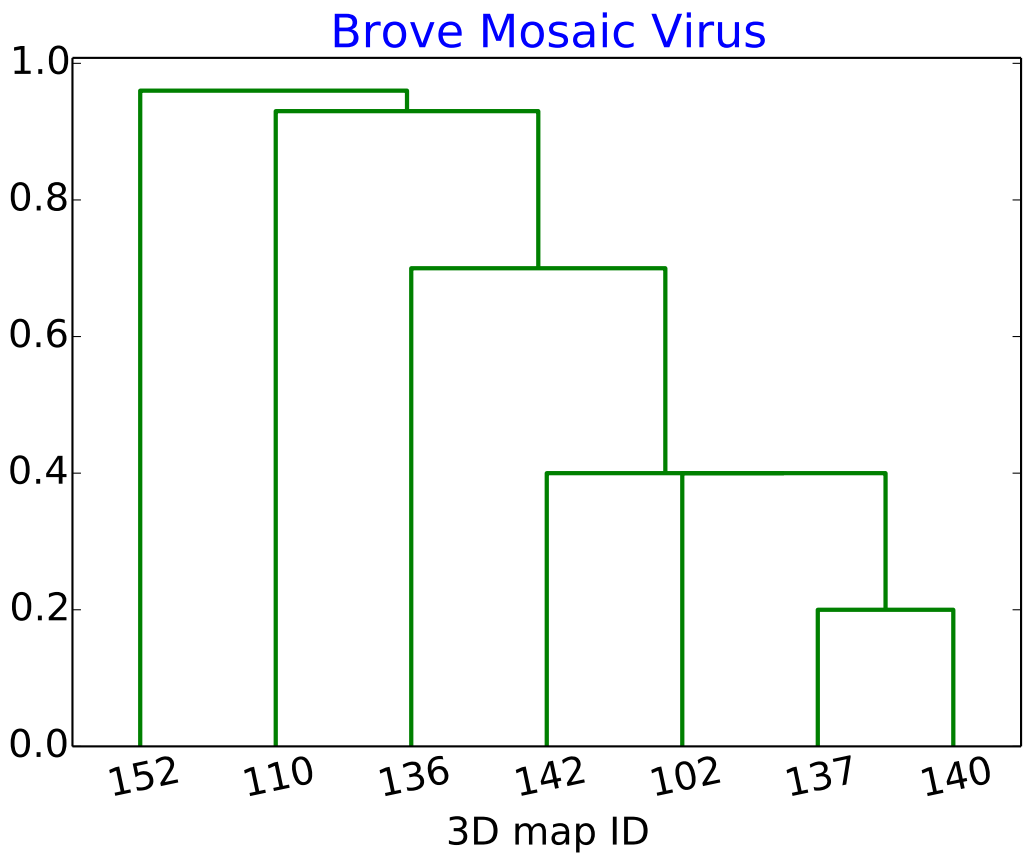

Figure 6: Dendrogram of Brome mosaic virus hierarchical clustering.

2.7 β-Galactosidase

$\overline{\text{FSC}_\text{i}}$	3Dmap
93.44	emcd138
93.17	emcd139
91.57	emcd159
91.5	emcd164
89.76	emcd106
88.2	emcd134
87.63	emcd167
86.57	emcd113
83.18	emcd160
79.77	emcd157
76.41	emcd154

Table 14: Sorting based on $\overline{FSC_i}$ for specimen β-Galactosidase.

	emcd138	emcd139	emcd159	emcd164	emcd106	emcd134	emcd167	emcd113	emcd160	emcd157	emcd154
emcd138		0.3	0.5	0.5	0.5	0.1	0.5	0.06	0.1	0.01	0.007
emcd139	0.3		0.5	0.5	0.5	0.08	0.5	0.06	0.1	0.01	0.007
emcd159	0.5	0.5		0.1	0.5	0.5	0.007	0.5	0.5	0.5	0.5
emcd164	0.5	0.5	0.1		0.5	0.5	0.007	0.5	0.5	0.5	0.5
emcd106	0.5	0.5	0.5	0.5		0.6	0.5	0.05	0.1	0.007	0.007
emcd134	0.1	0.08	0.5	0.5	0.6		0.5	0.4	0.3	0.02	0.01
emcd167	0.5	0.5	0.007	0.007	0.5	0.5		0.5	0.5	0.5	0.5
emcd113	0.06	0.06	0.5	0.5	0.05	0.4	0.5		0.3	0.02	0.007
emcd160	0.1	0.1	0.5	0.5	0.1	0.3	0.5	0.3		0.007	0.01
emcd157	0.01	0.01	0.5	0.5	0.007	0.02	0.5	0.02	0.007		0.08
emcd154	0.007	0.007	0.5	0.5	0.007	0.01	0.5	0.007	0.01	0.08	

Table 15: P-values resulting of comparing all 3Dmap pairs for specimen β-Galactosidase

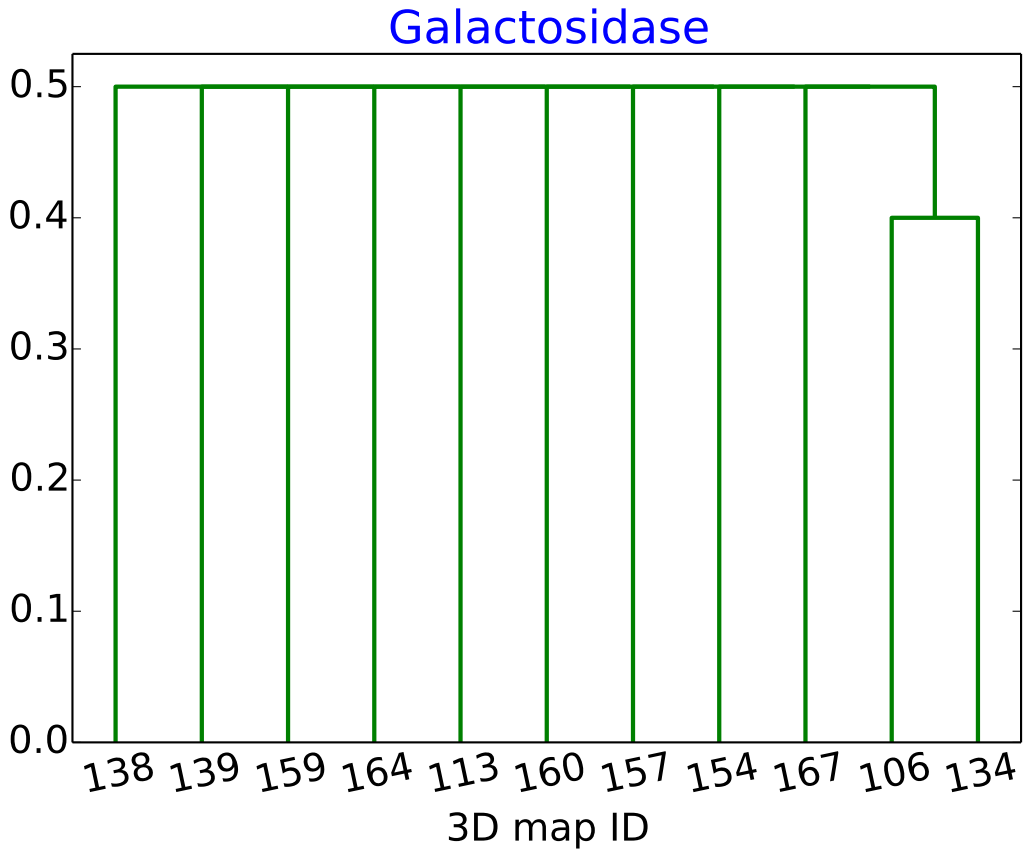

Figure 7: Dendrogram of Galactosidase hierarchical clustering.

3 Discussion

At present, this document cannot comment on the best image processing workflows for the different specimens because the metadata describing the algorithms used to calculate the different 3D reconstructions is not yet available. In the future, and using other statistical tests such as ANOVA we will expand this section.

A Values of the feature $\overline{FSC_{i,j}}$

In this Appendix we show the value of the feature $\overline{FSC_{i,j}}$ for each specimen and pair of reconstructions. These values have been used to compute the sorting and P-values.

A.1 GroEL *in Silico*

	emcd104	emcd120	emcd132	emcd143	emcd153	emcd158	emcd165	emcd168	emcd169	$\overline{FSC_i}$
emcd104	0	5.7	6.2	6.6	4.2	4.3	5.9	4.4	5.3	42.6
emcd120	5.7	0	6.2	6.1	4.5	4.3	5.9	4.5	5.3	42.5
emcd132	6.2	6.2	0	6.7	4.7	4.8	6.3	4.9	6.3	46.1
emcd143	6.6	6.1	6.7	0	5.1	4.6	7.7	5	5.85	47.65
emcd153	4.2	4.5	4.7	5.1	0	3.7	5.05	4	4.4	35.65
emcd158	4.3	4.3	4.8	4.6	3.7	0	4.5	3.7	5.1	35
emcd165	5.9	5.9	6.3	7.7	5.05	4.5	0	4.8	5.6	45.75
emcd168	4.4	4.5	4.9	5	4	3.7	4.8	0	5.5	36.8
emcd169	5.3	5.3	6.3	5.85	4.4	5.1	5.6	5.5	0	43.35

Table 16: $\overline{FSC_{i,j}}$ (columns: 2nd-next to last) and $\overline{FSC_i}$ (last column) for specimen GroEL *in silico*

A.2 T20S Proteasome

	emcd103	emcd107	emcd108	emcd130	emcd131	emcd141	emcd144	emcd145	emcd162	$\overline{FSC_i}$
emcd103	0	16.3	16.16	11.92	11.92	15.52	14.05	14.31	14.54	114.72
emcd107	16.3	0	16.82	11.74	11.74	15.85	13.59	13.82	14.69	114.55
emcd108	16.16	16.82	0	11.63	11.6	16.06	13.77	13.97	14.93	114.94
emcd130	11.92	11.74	11.63	0	20.39	11.66	11.74	11.78	10.93	101.79
emcd131	11.92	11.74	11.63	20.39	0	11.66	11.74	11.78	10.93	101.79
emcd141	15.52	15.85	16.06	11.66	11.66	0	13.92	14.1	14.19	112.96
emcd144	14.05	13.59	13.77	11.74	11.74	13.92	0	19.38	13.07	111.26
emcd145	14.31	13.82	13.97	11.78	11.78	14.1	19.38	0	13.21	112.35
emcd162	14.54	14.69	14.93	10.93	10.93	14.19	13.07	13.21	0	106.49

Table 17: $\overline{FSC_{i,j}}$ (columns: 2nd-next to last) and $\overline{FSC_i}$ (last column) for specimen T20S Proteasome

A.3 Apo-Ferritin

	emcd112	emcd118	emcd121	emcd122	emcd124	emcd147	emcd155	emcd166	$\overline{FSC_i}$
emcd112	0	9.41	11.05	4.5	7.8	4	5.88	9.82	52.46
emcd118	9.41	0	9.69	4.2	11.48	4.28	4.99	13.1	57.15
emcd121	11.05	9.69	0	4.44	8.23	4.27	6	10.21	53.89
emcd122	4.5	4.2	4.44	0	3.65	2.73	3.29	4.38	27.19
emcd124	7.8	11.48	8.23	3.65	0	4.73	4.12	11.63	51.64
emcd147	4	4.28	4.27	2.73	4.73	0	3.23	4.48	27.72
emcd155	5.88	4.99	6	3.29	4.12	3.23	0	5.23	32.74
emcd166	9.82	13.1	10.21	4.38	11.63	4.48	5.23	0	58.85

Table 18: $\overline{FSC_{i,j}}$ (columns: 2nd-next to last) and $\overline{FSC_i}$ (last column) for specimen Apo-Ferritin

A.4 TRPV1 Channel

	emcd101	emcd115	emcd133	emcd135	emcd156	emcd161	emcd163	$\overline{FSC_i}$
emcd101	0	8.58	7.54	7.66	6.27	8.49	5.92	44.46
emcd115	8.58	0	7.89	8.02	6.48	9.21	6.28	46.46
emcd133	7.54	7.89	0	10.74	5.96	8.54	6.24	46.91
emcd135	7.66	8.02	10.74	0	6.06	8.75	6.33	47.56
emcd156	6.27	6.48	5.96	6.06	0	6.56	5.41	36.74
emcd161	8.49	9.21	8.54	8.75	6.56	0	6.38	47.93
emcd163	5.92	6.28	6.24	6.33	5.41	6.38	0	36.56

Table 19: $\overline{FSC_{i,j}}$ (columns: 2nd-next to last) and $\overline{FSC_i}$ (last column) for specimen TRPV1 Channel

A.5 80S Ribosome

	e-111	e-114	e-119	e-123	e-125	e-126	e-127	e-128	e-129	e-148	e-149	e-150	e-151	$\overline{FSC_i}$
emcd111	0	14.12	18.3	15.18	15.12	14.01	12.5	12.27	9.52	12.28	13.65	13.81	14.12	164.88
emcd114	14.12	0	14.2	18.6	17.6	16.83	14.61	14.14	10.46	19.2	16.35	16.62	21.14	193.87
emcd119	18.3	14.2	0	15.73	15.29	14.48	12.86	12.62	9.33	12.39	14.1	14.3	14.54	168.14
emcd123	15.18	18.6	15.73	0	19.57	22.13	18.76	18.19	13.51	16.46	19.99	20.17	20.6	218.89
emcd125	15.12	17.6	15.29	19.57	0	17.52	15.04	14.57	10.86	14.37	16.63	16.87	17.47	190.91
emcd126	14.01	16.83	14.48	22.13	17.52	0	15.01	14.58	11.8	15.57	18.56	18.65	18.92	198.06
emcd127	12.5	14.61	12.86	18.76	15.04	15.01	0	14.05	11.57	15	16.96	16.98	17.14	180.48
emcd128	12.27	14.14	12.62	18.19	14.57	14.58	14.05	0	11.58	14.62	16.51	16.51	16.68	176.32
emcd129	9.52	10.46	9.33	13.51	10.86	11.8	11.57	11.58	0	13.19	13.11	12.96	13.22	141.11
emcd148	12.28	19.2	12.39	16.46	14.37	15.57	15	14.62	13.19	0	15.04	14.99	14.79	177.9
emcd149	13.65	16.35	14.1	19.99	16.63	18.56	16.96	16.51	13.11	15.04	0	22.13	20.03	203.06
emcd150	13.81	16.62	14.3	20.17	16.87	18.65	16.98	16.51	12.96	14.99	22.13	0	20.23	204.22
emcd151	14.12	21.14	14.54	20.6	17.47	18.92	17.14	16.68	13.22	14.79	20.03	20.23	0	208.88

Table 20: $\overline{FSC_{i,j}}$ (columns: 2nd-next to last) and $\overline{FSC_i}$ (last column) for specimen 80S Ribosome. In order to make the page fit in the page the name of the 3Dmaps has been shorten from the canonical form *emcd129* to *e129*

A.6 Brome Mosaic Virus

	emcd102	emcd110	emcd136	emcd137	emcd140	emcd142	emcd152	$\overline{FSC_i}$
emcd102	0	12.87	13.03	15.74	18.05	17.32	10.1	87.11
emcd110	12.87	0	10.75	12.48	12.64	12.62	8.56	69.92
emcd136	13.03	10.75	0	17.71	14.08	15.86	8.46	79.89
emcd137	15.74	12.48	17.71	0	16.59	17.98	9.83	90.33
emcd140	18.05	12.64	14.08	16.59	0	18.71	9.62	89.69
emcd142	17.32	12.62	15.86	17.98	18.71	0	9.55	92.04
emcd152	10.1	8.56	8.46	9.83	9.62	9.55	0	56.12

Table 21: $\overline{FSC_{i,j}}$ (columns: 2nd-next to last) and $\overline{FSC_i}$ (last column) for specimen Brome Mosaic Virus

A.7 β-Galactosidase

	emcd106	emcd113	emcd134	emcd138	emcd139	emcd154	emcd157	emcd159	emcd160	emcd164	emcd167	$\overline{F}$
emcd106	0	10.47	10.54	10.63	10.7	8.77	9.48	6.66	9.49	6.66	6.36	8
emcd113	10.47	0	9.18	10.47	10.24	9.43	8.72	6.5	8.79	6.5	6.27	8
emcd134	10.54	9.18	0	11.61	11.64	7.98	8.37	6.82	8.78	6.82	6.46	
emcd138	10.63	10.47	11.61	0	14.73	8.4	8.77	6.64	9.22	6.64	6.33	9
emcd139	10.7	10.24	11.64	14.73	0	8.33	8.75	6.65	9.16	6.65	6.32	9
emcd154	8.77	9.43	7.98	8.4	8.33	0	7.72	6.04	7.87	6.04	5.83	7
emcd157	9.48	8.72	8.37	8.77	8.75	7.72	0	6.34	9.22	6.34	6.06	7
emcd159	6.66	6.5	6.82	6.64	6.65	6.04	6.34	0	6.99	20.23	18.7	9
emcd160	9.49	8.79	8.78	9.22	9.16	7.87	9.22	6.99	0	6.99	6.67	8
emcd164	6.66	6.5	6.82	6.64	6.65	6.04	6.34	20.23	6.99	0	18.63	
emcd167	6.36	6.27	6.46	6.33	6.32	5.83	6.06	18.7	6.67	18.63	0	8

Table 22: $\overline{FSC_{i,j}}$ (columns: 2nd-next to last) and $\overline{FSC_i}$ (last column) for specimen β-Galactosidase

References

Map Challenge: Data Analysis Proposal

July 29, 2016

Abstract

This work presents two tools for statistically analyzing the results of the Map Challenge. These tools provide an assessment of the performance and reliability of the different software packages for the different data sets.

Keywords: structural biology, electron microscopy, 3D reconstruction, high-resolution, benchmarking, challenge

1 Introduction

Two of the goals of the 2015/2016 Model Challenge are to *Compare and contrast the various modeling and analysis approaches...* and to *Evolve criteria for evaluation and validation of 3DEM map-derived models.* From the point of view of analyzing the Challenge results the input data is the collection of reconstructions submitted by the challenge participants. These reconstructions can be grouped following different factor, for example, specimen (seven different data sets were made available) or image processing workflow (IPW). The goal is therefore to correlate these factors with the the reconstructions "quality". In this way, we will be able to help prospective users to select the better processing protocols for a given specimen.

Given a quality criteria, in this document we present two methods able to: (1) sort reconstructions, (2) assign a level of significance to this sorting and (3) detect the variables more relevant in the classification.

2 Theoretical Background

Let us image that we have a measurement (as can be resolution, R-factor, etc) computed for each reconstruction uploaded to the challenge website that gives us an estimation of the reconstruction quality. How can we use the whole sets of measurements simultaneously to decide which image processing workflows are more convenient for a given specimen? In the following we describe two methods that answer this question. The first method is based on factorial ANOVA while the second one has been developed specifically for this challenge.

2.1 First method: factorial ANOVA

One of the more popular tests for comparing sets of measurements are the t and z tests. A basic limitations of these statistical tests is that they cannot be applied for more than two groups (that is, they can only test a single null

hypothesis at the same time). This limitation is a major problem in our case since very basic criteria (as can by specimen) produce seven different groups since there are seven different data sets.

ANOVA is a statistical method that stands for analysis of variance. It was developed by Ronald Fisher in 1918 and is able to test several null hypothesis at the same time. For example, we may want to compare the achieved resolution based on two (or more) independent variables as could be software/workflow and specimen. Two-way ANOVA can be used to see the effect of one of the factors after controlling for the other, or it can be used to see the interaction between both factors.

The logic behind this procedure has to do with how much variance there is in the population. ANOVA compares the differences in the samples to see if they are the same or statistically different while still accounting for sampling error.

In order to clarify the methodology let us describe an example. Let us imagine that we have been able to evaluate the resolution of all uploaded reconstructions by comparing them with a PDB reference structure. This value may be used as estimation of the reconstruction quality. Let us assume, that we believe that resolution depends mainly on two factors: image-processing-workflow (IPW) and specimen (the number of factors may be increased if needed). These two factors may be used to classify the reconstruction in groups.

In this particular example, by applying ANOVA we will answer the following questions:

- Does the resolution depend on the specimen? That is, Is the resolution for some specimens higher than the resolution for other specimens regardless of the used IPW?
- Does the resolution depend on the IPW? That is, is the resolution for some IPWs higher than the resolution for other IPWs regardless of the used specimen?
- Does the resolution depend on the software and specimen at the same time? That is, no pattern can be found changing only the specimen or only the IPW. Therefore, both factors are needed for establishing a pattern.

ANOVA tests tell us whether we have an overall difference between groups, but it does not tell us which specific groups differ. That is, the test helps us to detect if there is a significant effect of IPWs on resolution,

but it will not tell us which of these IPWs is better and if there is a significant difference between two particular IPWs. These information can be obtained using the so called $post - hoc$ tests. If the data meet the assumption of homogeneity of variances (groups have similar variance), Tukey's honestly significant difference test (HSD) is a good candidate.

For the more mathematically oriented readers we include in Appendix. A a brief summary of the mathematical background behind ANOVA.

2.2 Second method: pair comparison method

ANOVA is a well known technique for analyzing how data depend on different variables. As we plan to use it, ANOVA requires measurements related with the reconstruction quality and it will tell us if some of the IPWs provide reconstructions with better quality. Nevertheless, it is not the best method if the variables used to group the measurements are not independent, then, we can not use the *post-hoc* methods described above and therefore, we cannot sort the reconstructions by quality.

For this eventuality we present a second comparison method that has the extra advantage of not requiring a reference PDB file and therefore could be used in future challenges in which this information is not available.

For a given specimen the method requires:

- Compute all possible pairs of reconstructions. That is, if 4 reconstructions were uploaded (r_1, r_2, r_3 and r_4-)6×2 pairs will be created ((r_1, r_2), (r_1, r_3), (r_1, r_4), (r_2, r_3), (r_2, r_4) and (r_3, r_4) plus six extra pairs in which the first and second member are interchanged.)

- For each pair formed by the ith and jth reconstructions, compute the Fourier Shell Correlation ($FSC_{i,j}$) between the first member of the pair and the second member of the pair. (Obviously ($FSC_{i,j} = FSC_{j,i}$).

- Then compute a weighted Fourier shell correlation average defined as $\overline{FSC_{i,j}} = \int \nu FSC_{i,j}(\nu)$ where ν is the frequency.

- For each reconstruction r_i compute $\overline{FSC_i} = \sum_{j=1, i \neq j}^{j=J} \overline{FSC_{i,j}}$ where J is the number of reconstructions. (We apologize by the slight abuse of notation introduced here since the "bar" symbol is used with two different meanings. In the first case -$\overline{FSC_{i,j}}$- is a weighted average while in the second case -$\overline{FSC_i}$- is a plain average)

- $\overline{FSC_i}$ will be used for sorting the 3D maps.

In the absence of systematic bias, the higher the resolution of the reconstruction r_i, the higher will be the value of $\overline{FSC_i}$ and therefore this magnitude may be used to sort the reconstructions. Unfortunately, even if we can rank our 3D maps, we do not know if two consecutively ranked reconstructions r_α and r_β are statistically different. This is an important questions because if they are statically different we could claim that the IPW used to produce r_α is superior to the method used to produce r_β (for the particular specimen under study). On the other hand, if r_α and r_β are not statistically different we cannot reject the null hypothesis that both IPWs perform equally well.

In order to answer the question if two reconstructions r_α and r_β are statiscaly different we follow this approach

- Let $\overline{FSC_{\alpha,k}}$ and $\overline{FSC_{\beta,k}}$ be the set of weighted Fourier correlation averages related with r_α and r_β respectively.
- For a given k, $\overline{FSC_{\alpha,k}}$ and $\overline{FSC_{\beta,k}}$ are correlated and therefore we may used a paired test to compare the two population means.
- The best known paired test is *paired t-test* but it assumes that the sample is normally distributed which very likely will not be the case. A second option is the paired Wilcoxon test.

2.3 Test on the performance of the new sorting method

ANOVA is a well known test that has been validated by an extensive use. This is not the case for the new proposed method ("pair comparison method") so a collection of experiments has been performed in order to judge how reliable and robust this method would be. Two different phantoms have been used. The first is totally asymmetric (a ribosome) while the second one presents high symmetry (an icosahedral virus). Since results are very similar for both cases, in the following we present in detail the experiments performed with the second phantom. This phantom is based on the quasi-atomic model of bacteriophage T7 procapsid shell described in Agirrezabala et al. (2007) and deposited in the PDB with accession number 3IZG. A surface rendering can be seen in Fig. 1.

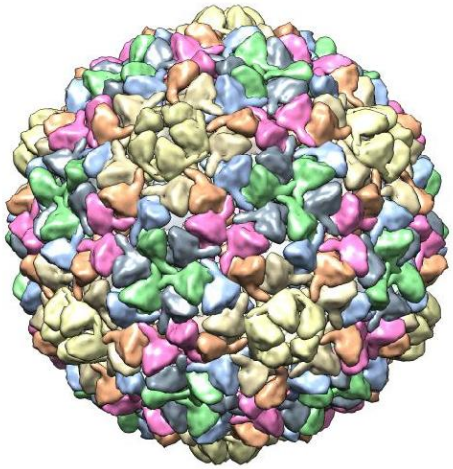

Figure 1: Surface rendering of bacteriophage T7 procapsid shell (PDB Id = 3IZG).

In a nutshell the design of the experiments is as follows, a large set of projections is created. These projections are divided in subsets and reconstructed. The proposed method is applied to the reconstructions in order to sort them by "quality". Since we are working with phantoms, we can compare this sorting with a control one and check if the new algorithm is working properly. Finally, for those reconstructions that are in different positions in the sorting produced by the new algorithm and by the control, we test if this disagreement is statistically significant or not.

20,000 noisy unaligned projections were created with a sampling rate of

1.5 Å/px. From this data set, 13 independent subsets of projections were generated with: 700, 1020, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000, 5500, 6000 and 10000 projections respectively. These subsets were reconstructed using RELION (Scheres, 2012) and sorted applying the described algorithm. (In the rest of the article we use the symbol r_{xxxx} to denote a reconstruction obtained from $xxxx$ projections.)

The first step before analyzing the results is to establish which is the correct reconstruction order. *A priori* we expect that there should be a strong correlation between the reconstruction position and the number of projections. Nevertheless, it is possible that reconstructions obtained from fewer projections may have higher "quality". Using the phantom as reference, we define as correct order the one given by sorting the value $\overline{FSC_{i,pham}}$. This value is the average weighted Fourier shell correlation of the ith reconstruction and the noiseless phantom. Table 1 (fourth column) shows the results of sorting by this control magnitude and we see that in a few cases, for example for r_{4000} and r_{5000}, a reconstruction obtained from a higher number of projections present a lower "quality" than a reconstruction obtained from less projections.

Table 1 (second column) shows the reconstructions sorted by the pair comparison method. We see that the sorting provided by this method and the control one, although similar, is not identical (see for example r_{6000} and r_{5500}). The question that arises now is if r_{6000} and r_{5500} (and r_{4500}, r_{4000} and r_{5000}) are indistinguishable, that is, if we repeat the experiment will the order remain or will it change. To answer this question, we apply the Wilcoxon test. In Table 2 we show the p-value obtained from comparing the set of values $\overline{FSC_{\alpha,j}}$ and $\overline{FSC_{\beta,j}}$ related with the reconstruction r_α and r_β. We define that two reconstructions are statistically distinguishable if the p-value between them is higher than 0.05 (that is, if we repeat the experiment we should obtain the same result 95 out of 100 times). In Table 2, where p-values higher than 0.05 are marked in red, we see that the pairs (r_{4000}, r_{4500}), (r_{4000}, r_{5000}), (r_{4500}, r_{5000}) and (r_{5500}, r_{6000}) are indistinguishable. In Table 1 indistinguishable reconstructions have the same background color. From these data, we conclude that the order provided by the pair comparison method and the reference one are different but statistically equivalent.

2.4 Using alternative experimental setups

As a way to further validate the proposed pair comparison method, we decided to modified it incorporating two variants. In the first case we used instead of $\overline{FSC_i}$ the magnitude known as R-factor (see Eq. 1b) for com-

$\overline{\mathbf{FSC_{i,j}}}$	**sort using pair comparison method**	$\overline{\mathbf{FSC_{i,pham}}}$	**sort using Phantom**
182.41	r_{10000}	19.67	r_{10000}
176.63	r_{6000}	18.34	r_{5500}
176.55	r_{5500}	18.28	r_{6000}
172.59	r_{4500}	17.78	r_{4000}
172.22	r_{4000}	17.71	r_{4500}
171.33	r_{5000}	17.46	r_{5000}
167.81	r_{3500}	16.88	r_{3500}
162.94	r_{3000}	16.22	r_{3000}
156.23	r_{2500}	15.15	r_{2500}
153.56	r_{2000}	14.86	r_{2000}
142.44	r_{1500}	13.53	r_{1500}
134.85	r_{1020}	12.60	r_{1020}
125.04	r_{700}	11.37	r_{700}

Table 1: Comparison of the sorting provided by the pair comparison method (first and second columns) vs the control (third and fourth columns). Cell with the same color contain subsets of reconstructions that are statistically indistinguishable.

	700	**1020**	**1500**	**2000**	**2500**	**3000**	**3500**	**4000**	**4500**	**5000**	**5500**	**6000**	**10000**
700	NA	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
1020	0.02	NA	0.02	0.03	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02
1500	0.02	0.02	NA	0.05	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.02
2000	0.02	0.03	0.05	NA	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
2500	0.02	0.03	0.03	0.02	NA	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
3000	0.02	0.03	0.03	0.02	0.02	NA	0.02	0.02	0.02	0.02	0.02	0.02	0.02
3500	0.02	0.02	0.03	0.02	0.02	0.02	NA	0.02	0.02	0.02	0.02	0.02	0.02
4000	0.02	0.02	0.03	0.02	0.02	0.02	0.02	NA	0.9	0.2	0.02	0.02	0.02
4500	0.02	0.02	0.03	0.02	0.02	0.02	0.02	0.9	NA	0.4	0.02	0.02	0.02
5000	0.02	0.02	0.03	0.02	0.02	0.02	0.02	0.2	0.4	NA	0.02	0.02	0.02
5500	0.02	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02	NA	0.7	0.02
6000	0.02	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.7	NA	0.02
10000	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	NA

Table 2: Wilcoxon test. Red colored cells mark pairs of reconstructions that are indistinguishable, that is with p-values greater than 0.05

puting the control sorting. In the second variant we check the influence of applying a tight mask to the reconstructed 3D map.

2.4.1 R-factor

In this experiment R-factors were calculated applying the macromolecular refinement program REFMAC (Murshudov et al., 1997) in the frequency range 20-3.5 Å. Using these values a new control sorting was generated. The results, which are presented in Table 3 (second column), are partially in disagreement with the sorting computed using the pair comparison method (Table 3, fourth column). This discrepancy cannot be justified even if we take into account the information provided by the Wilcoxon test that points out which reconstructions are indistinguishable. On the other hand, if the algorithm was executed using as meassurement of quality another magnitude produced by REFMAC called *average Fourier shell correlation*; the results are quite close (Table 3, sixth column) to the ones produced by the original pair comparison method. The only significant difference is that the reconstruction in the ninth and tenth positions are swapped. Finally, we were able to reconcile the sorting produced by all magnitudes if the R-factor was computed only for high frequencies in the range 5-3.5 Å (Table 3, eight col-

umn). In this case, both variants of the algorithm, the one based on average weighted Fourier shell correlations and the one based on R-Factor produce equivalent results except for the reconstructions in the last two positions which are swapped.

The divergence between both measures of quality (R-factor and $\overline{FSC_i}$) is due to the fact that R-factor depends more heavily on the low frequency values than the FSC. FSC is computed by rings (see Eq. 1a), and the value of each ring does not depend on the absolute magnitude of the Fourier components at that ring but on the similarity between the compared 3D maps at that frequency. On the other hand R-factor (see Eq. 1b) is a summation over the whole Fourier space and even after applying a β-factor to the reconstruction, it is more sensible to similarities at low frequency than FSC.

$$FSC(r) = \frac{\sum_{r_i \in r} F_1(r_i) \cdot F_2(r_i)^*}{\sqrt{\sum_{r_i \in r} |F_1(r_i)|^2 \cdot \sum_{r_i \in r} |F_2(r_i)|^2}} \tag{1a}$$

$$Rfactor = \frac{\sum_{r_{min}}^{r_{max}} ||F_{\text{obs}}| - |F_{\text{calc}}||}{\sum_{r_{min}}^{r_{max}} |F_{\text{obs}}|} \tag{1b}$$

where, F_1 is the Fourier transform of the first 3D map, F_2^* is the complex conjugate of the Fourier transform of the second 3D map 2, and r_i is the individual voxel element at radius r. F_{obs} and F_{cal} are the Fourier transforms of the reconstructed and the reference 3D map from the PDB file respectively, the sum extends over all the space between a range of frequencies r_{min} and r_{max}.

2.4.2 Applying tight masks

In our last experiment we want to check the behavior of the proposed method when a tight mask was applied to some of the reconstructions. In this way, before performing the sorting, we applied to half of the reconstructions a mask obtained using the *post-process* algorithm provided by RELION. *post-process* was executed with the default parameters except for the *binarization threshold*, *mask pixel extension* and *add soft edge* that were set to 0.02, 3px and 3px respectively. For each reconstruction its corresponding mask were calculated and applied. That is, the masks applied to the different volumes are similar but not identical. The control sorting computed for this data set

R-factor (20 to 3.5)	sort using R-factor (20 to 3.5)	$\overline{\text{FSC}_{i,j}}$	sort using pair comparison method	averag. FSC	sort using averg. FSC	R-factor (5.0 to 3.5)	sort using R-factor (5.0 to 3.5)
0.1885	r_{10000}	182.41	r_{10000}	0.861	r_{10000}	0.321	r_{10000}
0.1991	r_{5000}	176.63	r_{6000}	0.8231	r_{5500}	0.3422	r_{5500}
0.2024	r_{4500}	176.55	r_{5500}	0.8182	r_{6000}	0.3465	r_{6000}
0.2034	r_{6000}	172.59	r_{4500}	0.8089	r_{4500}	0.3515	r_{4000}
0.2054	r_{5500}	172.22	r_{4000}	0.8017	r_{4000}	0.3567	r_{4500}
0.2068	r_{4000}	171.33	r_{5000}	0.7995	r_{5000}	0.3615	r_{5000}
0.2131	r_{3000}	167.81	r_{3500}	0.7805	r_{3500}	0.3772	r_{3500}
0.2163	r_{1020}	162.94	r_{3000}	0.7609	r_{3000}	0.3907	r_{3000}
0.2184	r_{3500}	156.23	r_{2500}	0.709	r_{2000}	0.3909	r_{2500}
0.2193	r_{2500}	153.56	r_{2000}	0.7087	r_{2500}	0.4008	r_{2000}
0.2326	r_{2000}	142.44	r_{1500}	0.6799	r_{1500}	0.4339	r_{1500}
0.2345	r_{1500}	134.85	r_{1020}	0.6402	r_{1020}	0.4507	r_{700}
0.2389	r_{700}	125.04	r_{700}	0.6204	r_{700}	0.4526	r_{1020}

Table 3: Comparison of the sorting provided by R-factor (refmac, first and second columns). Average Fourier Shell Correlation (refmac, third and fourth columns) and averaged weighted FSC (new method, fifth and sixth columns). Cell with the same color contain subsets of reconstructions that are statistically indistinguishable.

using $\overline{FSC_i}$ is shown in Table 4 (second column). In this table we differentiate the reconstructions with and without masks by adding the character m to the reconstruction name, in this way rm_{10000} is a reconstruction from 10,000 projections that has been masked while r_{6000} is a reconstruction from 6000 projections that has not been masked. The table clearly shows that all masked reconstructions are in the first positions. Therefore, we may conclude that, as it is well known, applying a tight mask has a major impact comparing reconstructions. If we form two subgroups containing the masked and unmasked reconstructions we see that within each group the higher is the number of projections the better is the reconstruction. One of the obvious conclusions is that $\overline{FSC_i}$ is not a robust magnitude for sorting data sets in which mask and unmasked reconstruction are mixed together.

So far we have used R-factors only for computing the control sorting. At this point we decided to check if R-factor was more robust than $\overline{FSC_i}$ for mixed data sets. In this way, inside the pair comparison method the magnitude $\overline{FSC_i}$ was replaced by the R-factor. Our preliminary results are inconclusive. On one hand, when used for creating a control sorting, the

results seem to be quite independent from the presence/absence of mask (see Table 4). On the other hand when R-factors are being computed between each pair of reconstructions the results are not conclusive since the Wilcoxon p-values show that most of the reconstructions are indistinguishable (data non shown). Undergoing experiments will shed light on this subject.

$\overline{\mathbf{FSC_{i,j}}}$	**sort using pair comparison method**	**R-factor (5.0 to 3.5)**	**sort using R-factor (5.0 to 3.5)**
228.33	rm_{10000}	0.3118	rm_{10000}
227.03	rm_{5500}	0.3265	rm_{5500}
224.86	rm_{4500}	0.3404	rm_{4500}
223.27	rm_{3500}	0.3465	r_{6000}
219.09	rm_{2500}	0.3515	r_{4000}
209.77	rm_{1500}	0.3571	rm_{3500}
195.98	rm_{700}	0.3615	r_{5000}
190.26	r_{6000}	0.3694	rm_{2500}
186.02	r_{4000}	0.3907	r_{3000}
183.93	r_{5000}	0.4008	r_{2000}
173.73	r_{3000}	0.4107	rm_{1500}
162.25	r_{2000}	0.412	rm_{700}
139.98	r_{1020}	0.4526	r_{1020}

Table 4: Comparison of the sorting provided by average Fourier shell correlation (first and second columns). R-factor (refmac, third and fourth columns).

A Mathematical Explanations behind ANOVA

[1] ANOVA is the classical method to compare means of multiple (≥ 2) groups. Suppose $\boldsymbol{N}$ observations were sampled from k groups and define $n = \boldsymbol{N}/k$. Let x_{ij} be the jth observation from the ith group. Here we assume a balanced design i.e the number of samples from each group remain the same. Denote $\bar{x}$ the grand sample mean and $\bar{x_i}$ the sample mean of group i. Observations can be re-written as

$$x_{ij} = \bar{x} + (\bar{x_i} - \bar{x}) + (x_{ij} - \bar{x_i}) \tag{2}$$

This leads to the following model

$$x_{ij} = \mu + \alpha_i + \epsilon_{ij} \tag{3}$$

where μ and α_i are grand mean and ith group mean respectively. The error term ϵ_{ij} is assumed to be independent and identically distributed from a normal distribution.

$$\epsilon_{ij} \sim N(0, \sigma^2) \tag{4}$$

The null hypothesis in ANOVA is that all group means are the same i.e

$$\alpha_1 = \alpha_2 = \ldots = \alpha_k \tag{5}$$

If this is true, the error term for group differences $\bar{x}_i - \mu \sim N(0, \sigma^2/n = \bar{\sigma}^2)$. However, you cannot test this by using one-sample t-test (discard x_{ij} and only use $\bar{x_i}$). Suppose you have $\sigma^2 = 5$ and $\bar{\sigma^2} = 1000$ e.g between group difference is much larger than within group difference. In this case, data from individual groups are similar but groups are quit different, so we should reject the null hypothesis although the one sample t-test may fail to reject. It is really the relative magnitude of within and between group differences what matters. You cannot say much by only looking at one of them.

Now consider the sum of squares for between group difference

$$SSD_B = \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{x}_i - \bar{x})^2 = n\sum_{i=1}^{k}(\bar{x}_i - \bar{x})^2 \tag{6}$$

and for within group difference

[1] Text from: `http://stats.stackexchange.com/questions/78920/mathematical-explanations-behind-anova`

$$SSD_W = \sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2 \tag{7}$$

where SSD_B has a degree of freedom of $k-1$ and SSD_W has a degree of freedom of $N-k$. If there is no systematic difference between the groups, we would expect the mean squares

$$MS_B = SSD_B/(k-1) MS_W = SSD_W/(N-k) \tag{8}$$

would be similar. The test statistic in ANOVA is defined as the ratio of the above two quantities:

$$F = MS_B/MS_W \tag{9}$$

which follows a F-distribution with $k-1$ and $N-k$ degrees of freedom. If the null hypothesis is true, F would likely be close to 1. Otherwise, the between group mean square MS_B is likely to be large, which results in a large F value. Basically, ANOVA examines the two sources of the total variance and sees which part contributes more. This is why it is called analysis of variance although the intention is to compare group means.

B Tables with weighted averaged Fourier shell correlation

	700	**1020**	**1500**	**2000**	**2500**	**3000**	**3500**	**4000**	**4500**	**5000**	**5500**	**6000**	**10000**	$\overline{\mathbf{FSC_i}}$
700	0	10.03	10.87	9.87	10.01	10.13	10.36	10.55	10.5	10.53	10.65	10.64	10.9	125.04
1020	10.03	0	11.96	10.61	10.79	10.98	11.19	11.46	11.4	11.35	11.62	11.56	11.9	134.85
1500	10.87	11.96	0	11.2	11.35	11.6	11.86	12.16	12.06	12.08	12.31	12.31	12.68	142.44
2000	9.87	10.61	11.2	0	12.5	12.8	13.48	13.6	13.7	13.45	14.04	13.96	14.35	153.56
2500	10.01	10.79	11.35	12.5	0	13.29	13.47	13.87	13.77	14.02	14.12	14.14	14.9	156.23
3000	10.13	10.98	11.6	12.8	13.29	0	14.3	14.81	14.55	14.71	14.96	14.98	15.83	162.94
3500	10.36	11.19	11.86	13.48	13.47	14.3	0	15.08	15.3	15.03	15.8	15.78	16.16	167.81
4000	10.55	11.46	12.16	13.6	13.87	14.81	15.08	0	15.76	15.65	16.15	16.23	16.9	172.22
4500	10.5	11.4	12.06	13.7	13.77	14.55	15.3	15.76	0	15.5	16.55	16.72	16.78	172.59
5000	10.53	11.35	12.08	13.45	14.02	14.71	15.03	15.65	15.5	0	15.91	15.89	17.21	171.33
5500	10.65	11.62	12.31	14.04	14.12	14.96	15.8	16.15	16.55	15.91	0	17.03	17.41	176.55
6000	10.64	11.56	12.31	13.96	14.14	14.98	15.78	16.23	16.72	15.89	17.03	0	17.39	176.63
10000	10.9	11.9	12.68	14.35	14.9	15.83	16.16	16.9	16.78	17.21	17.41	17.39	0	182.41

Table 5: $\overline{FSC_{i,j}}$ and $\overline{FSC_i}$ (last column) values for the virus phantom.

References

Agirrezabala, X., Velazquez-Muriel, J. A., Gomez-Puertas, P., Scheres, S. H., Carazo, J. M., Carrascosa, J. L., Apr 2007. Quasi-atomic model of bacteriophage t7 procapsid shell: insights into the structure and evolution of a basic fold. Structure 15 (4), 461–472.

Murshudov, G. N., Vagin, A. A., Dodson, E. J., May 1997. Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr. D Biol. Crystallogr. 53 (Pt 3), 240–255.

Scheres, S. H., Dec 2012. RELION: implementation of a Bayesian approach to cryo-EM structure determination. J. Struct. Biol. 180 (3), 519–530.