Structure, Volume *27*

Supplemental Information

De Novo Structural Pattern Mining in Cellular Electron Cryotomograms

Min Xu, Jitin Singla, Elitza I. Tocheva, Yi-Wei Chang, Raymond C. Stevens, Grant J. Jensen, and Frank Alber

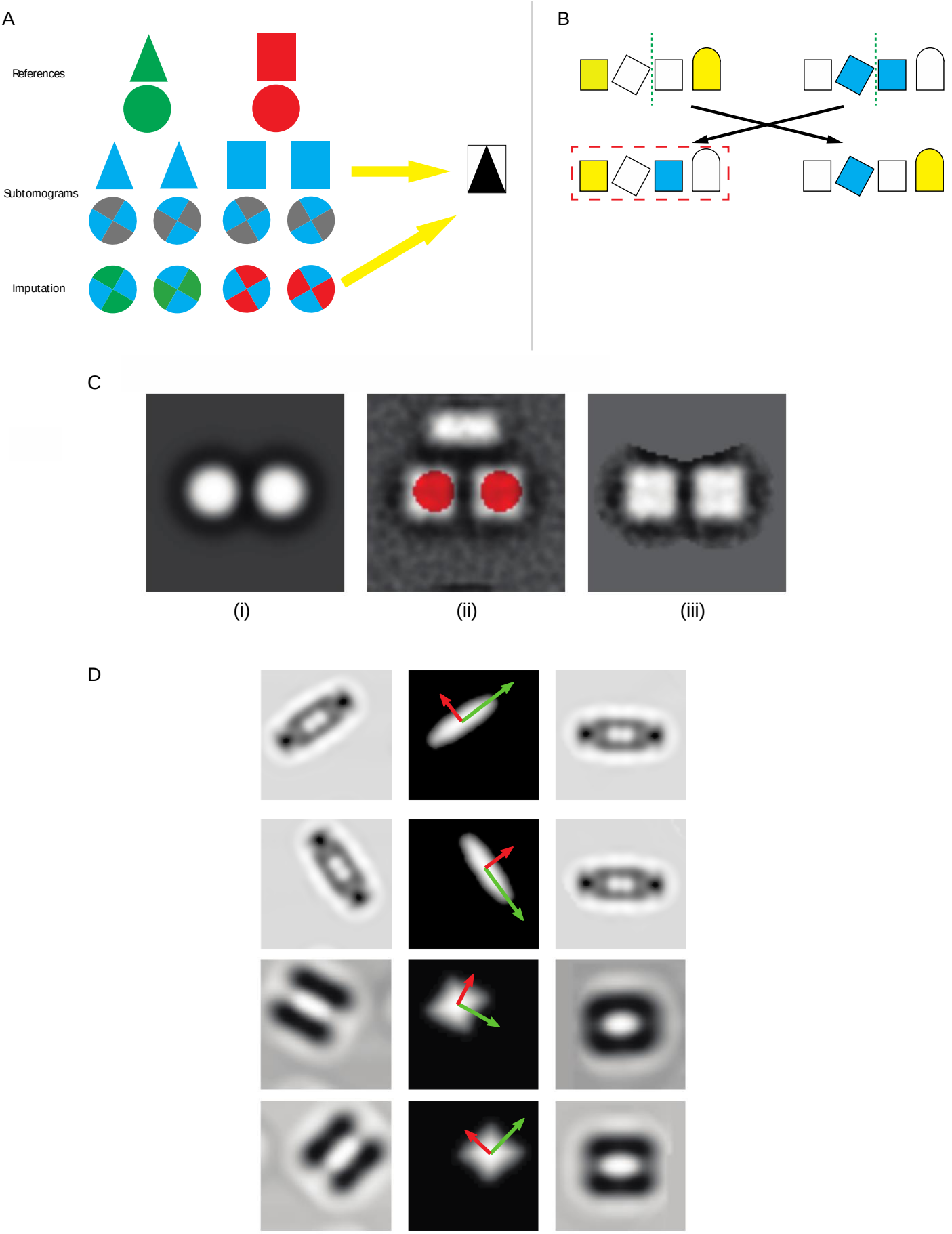
A
References
Subtomograms
Imputation
B
C
(i)
(ii)
(iii)
D

Figure S1: Details of some of the methods used in overall Multi-Pattern Pursuit pipeline. Related to Figure 1 and STAR Methods.

A) The basic idea of imputation-based dimension reduction. Left: Upper row: Green triangle and red rectangle are structures in two reference density maps. The filled circles show missing wedge masks, which are regions of valid Fourier components of the subtomogram image in Fourier space. Middle row: Blue Triangles and rectangles are structures in individual subtomograms. These subtomograms are aligned against its most similar references (top row). The circles are the corresponding missing wedge masks, which indicate regions with valid (colored in blue) and missing (colored in grey) Fourier coefficients. Lower row: Imputation of subtomograms by replacing missing Fourier coefficient regions (previously in grey color) with valid Fourier coefficients from the corresponding references (in green and red colors). Right: The variance of voxel intensities across imputed subtomograms. The region with low variance is represented in black color. The region with high variance is represented in white color. The use of a variance map is only for illustration purpose. **B)** Crossover operation in GA based subtomogram set refinement. Upper row: two parent solutions, where the colored shapes correspond to selected subtomograms, and white shapes correspond to unselected subtomograms. The dashed green line represents the crossover point. Lower row: two children solutions after applying crossover operation. Highlighted in dashed rectangle is a better solution where selected subtomograms contain same shape with same orientation. **C)** Basic idea of reference guided segmentation is illustrated using a toy example. (*i*) A reference density map *a*. (*ii*) A subtomogram *f* that is roughly aligned against *a*. It contains a particle represented by two disjoint cubes, and a rectangular neighboring structure. The red region is a seed corresponding to $R_a^{structure}$ calculated from segmenting *a*. (*iii*) Final masked subtomogram. **D)** Basic idea of level set based pose normalization is illustrated using four toy examples in four rows. Left column: center slice of four simulated subtomograms containing two Proteosome and two GroEL complexes with different orientations and locations. Middle column: the density map is the positive part of the approximation level sets, and the vectors are inferred pose. Right column: pose normalized subtomograms.

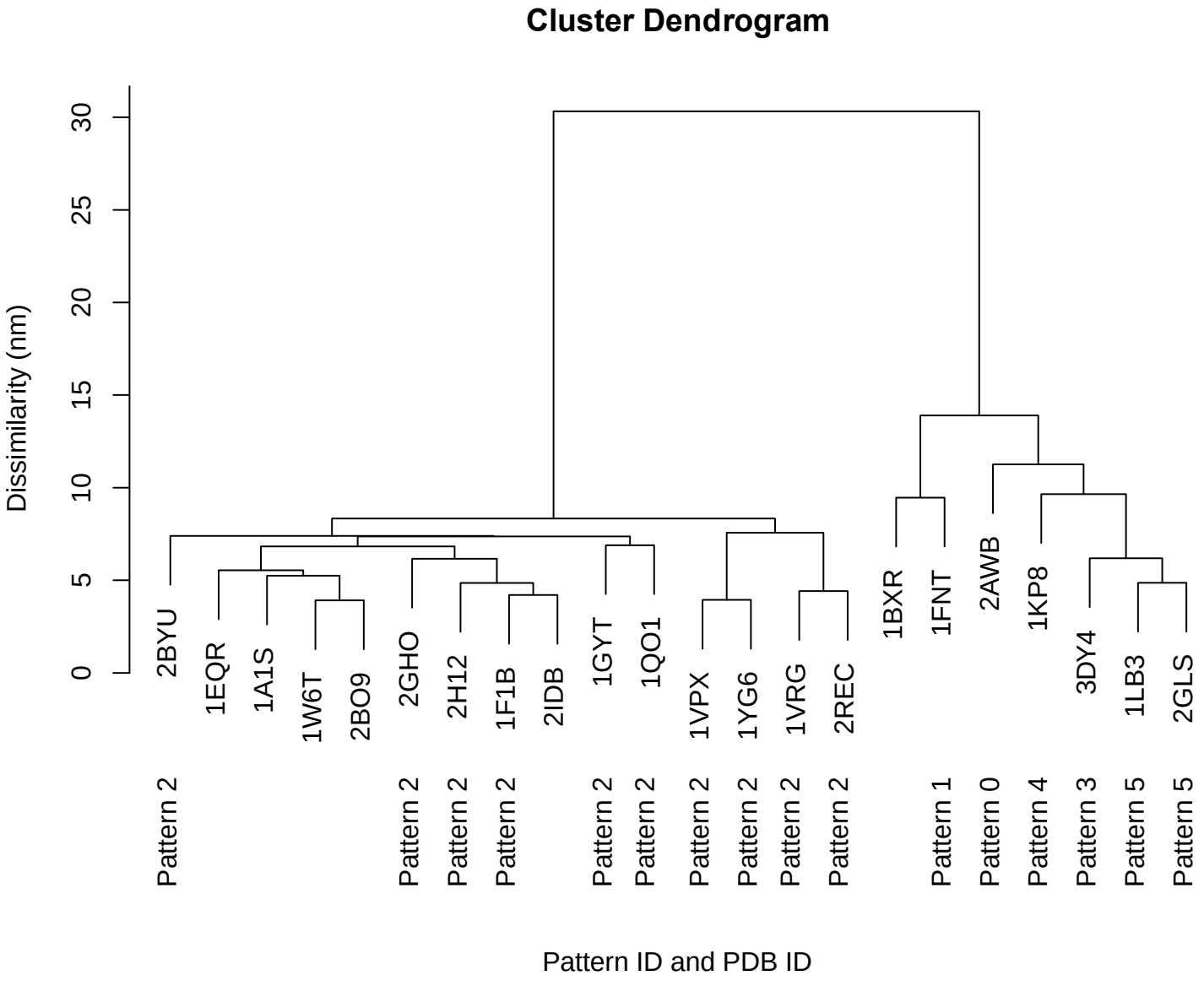

Figure S2: Dendrogram of hierarchical clustering of the templates of macromolecular complexes used for simulation. Related to Figure 3.

Each template is labeled by PDB ID of the complex and the ID of pattern whose subtomograms contain that complex. The hierarchical clustering is based on structural dissimilarity in terms of FSC at 0.5 cutoff.

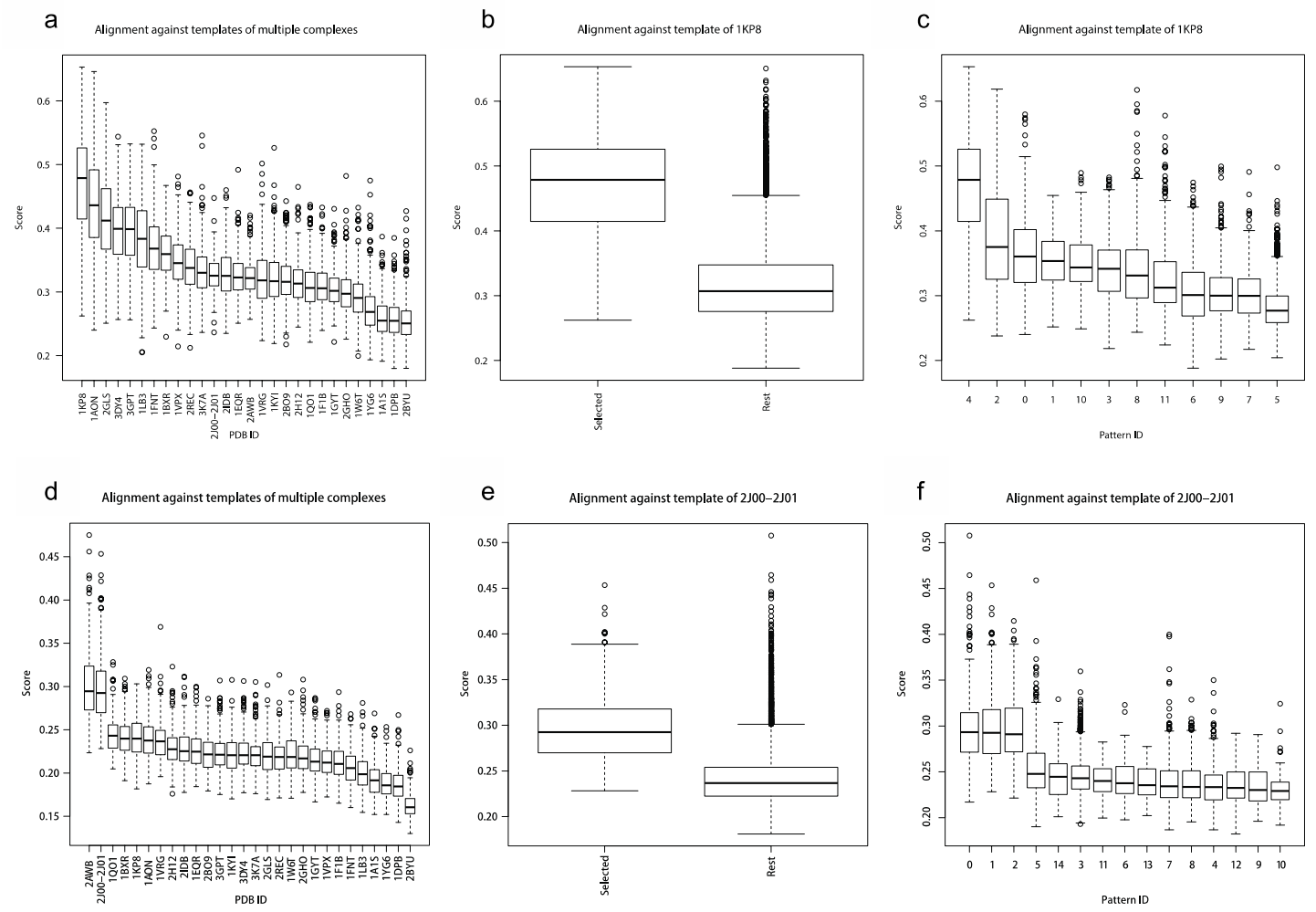

Figure S3: Analysis of patterns from tomogram of *A. longum* and *H. gracilis.* Related to Figure 5.
Analysis of patterns from tomogram of *A. longum.* **a)** Box plot of the distribution of alignment scores of the subtomograms of pattern 4 against all different template complexes (denoted by PDB ID). The complexes are ordered according to median score in descending order. **(b)** (left) Box plot of the alignment score distribution of subtomograms in pattern 4 against the GroEL template complex (PDB ID: 1KP8) and (right) box plot of the alignment score distribution of all other extracted subtomograms against the GroEL template. **(c)** Box plot of alignment score distributions of the subtomograms in all patterns against the GroEL template (PDB ID: 1KP8). The patterns are ordered according to median score in descending order. Analysis of patterns from tomogram of *H. gracilis*: **(d)** Box plot of the distribution of alignment scores of the subtomograms of pattern 1 against all different templates complexes (denoted by PDB ID). The complexes are ordered according to median score in descending order. **(e)** (left) Box plot of the alignment score distribution of subtomograms in pattern 1 against the ribosome template complex (PDB ID: 2J00-2J01) and (right) box plot of the alignment score distribution of all other extracted subtomograms against the ribosome template, **(f)** Box plot of alignment score distributions of the subtomograms in all patterns against the ribosome complex template (PDB ID: 22J00-2J01). The patterns are ordered according to median score in descending order.