Toward an EM Time Series: Automated Cell Identification in a Developmental Context

Anthony Santella[3], Irina Kolotuev[1], Zhirong Bao[2]

1 Electron Microscopy Facility Université de Lausanne, Lausanne, Switzerland

2 Developmental Biology Program, Sloan Kettering Cancer Center, New York, New York

3 Molecular Cytology Core, Sloan Kettering Cancer Center, New York, New York

Memorial Sloan Kettering Cancer Center

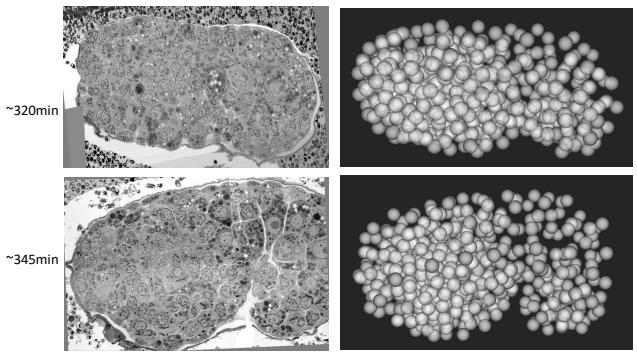
UNIL | Université de Lausanne Electron Microscopy Facility

Abstract:

Though automation makes EM data more accessible, navigation and identification of structures remains a bottleneck for interpretation in large tissues. We present a robust, general method for assigning single cell identities from a template to a sample despite differences in the cells present and their spatial configuration, validating this in the context of the C. *elegans* embryo. We introduce **neighbor graph constraints** to model the invariant spatial structure of a labeled samples and use this to assign a global quality score to a labeling based on its internal consistency with expected cell-cell contacts. This score is used in a novel **gradient descent optimization** of the template sample which removes cells whose presence cause neighbor constraint violations and are therefore hypothesized to be missing in the unlabeled sample. Our final answer is produced by an **instance-based learning** like approach where the sample is independently matched against each example in the **ensemble** of reference data sets and a **consensus** identity is assigned.

We apply this method to identifying all cells in Electron Micrographs of two *C. elegans* embryos at ~320min and 345min p.f.c. Imaging with Focused Ion Beam SEM and a serial array method provides an undistorted image of the worm simplifying the problem of alignment. Time lapse fluorescence microscopy provides the reference atlas data set of cell positions, identities, and division timings. For validation, identities were independently and manually assigned to a subset of cells in the two EM data sets based on position and cell morphology.

This approach has potential in other organisms where establishing single cell alignment between individuals is critical for understanding the extent of single cell consistency. Weaker prior information presents a challenge since individual identities cannot be established by lineaging. The large strokes of our method are universal, though tissue level labels based on anatomy require a different objective function based on counts of expected tissue-tissue contacts. Though we can permute our template absent single cell identities we lose the use of priors for which cells may divide or die, possibly necessitating more advanced optimization methods.



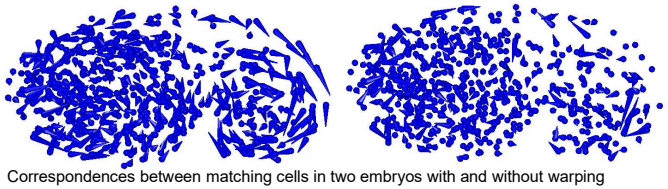
A cross section through each EM data set, canonically reoriented and a corresponding nuclear positions.

EM Data:

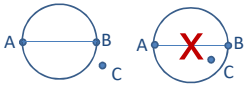
Preparation used High-Pressure Freezing following by freeze substitution using 2% osmium and 0.1% Uranyl Acetate diluted in Acetone. After dehydration, samples were gradually infiltrated in EPON 812 resin and flat embedded. (Kolotuev, 2014, Traffic; Burel et al., 2018, Development). Stacks were generated using FEI HELIOS 650 FIB/(Focused Ion Beam) -SEM microscope, aligned using IMOD software. The early embryo was imaged laterally ~(3600x1800x900px); the late embryo axially ~(5100x4328x1600px).

Algorithm:

An initial linear alignment is nonlinearly refined using a regularized nonlinear alignment method GMM-CPD.

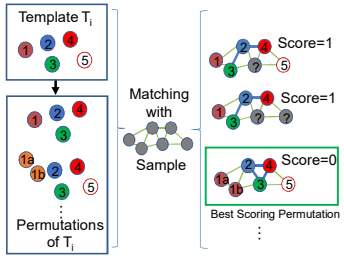


Correspondences between matching cells in two embryos with and without warping



Expected cell-cell contacts are modeled from nuclear positions with a Gabriel graph: A and B are touching if no other cell C appears in the circle defined by them. If A adjoins B in all template data sets it is an expected contact.

Optimization of template data set:
In an optimal match between non-equivalent sets of cells a cell identity can be 'stolen' by a wrong match with a cell that has no match, causing a cascade of errors. This can be detected by a large number of expected contacts missing in the resulting match, and the labeled data set can be modified to minimize violations

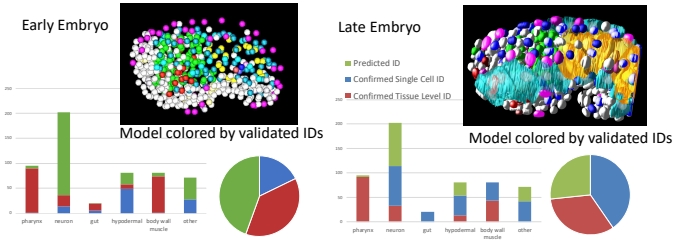


Consensus:

Alignment is repeated for every labeled volume over available embryos within a thirteen minute window with the majority answer being chosen. This entire process can be iterated taking high agreement matches as definitive, then computing in place of the initial nonrigid alignment a nonrigid alignment based on these confident cases

Manual Annotation:

Manual annotation of cell identities was performed for validation purposes. In the late ~350min p.f.c. data set this was done completely independently of matching results, it took ~2 weeks to annotate ~3/4 of the embryo based on cell morphology and relative position. The 300 min p.f.c. early embryo was annotated by checking automated names against morphology and most identities were tissue level, this took approximately 2 days.



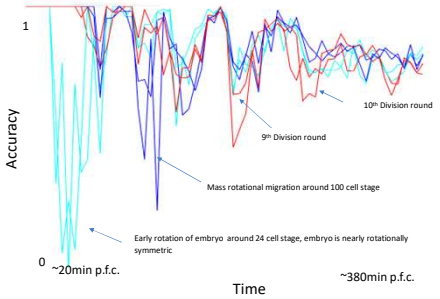
Performance

Performance on EM data using manually identified subset to assess accuracy:

EM Late (222 confirmed IDs, 126 marked as very confident) 74.12% of 222, 91.27% of 126
EM Early (98 confirmed IDs, 80 marked as very confident) 78.35% of 98, 88.75% of 80

Evaluation using 3 lineaged embryos one serving as the unknown and two as templates

Embryo	1	2	3	Consensus	Iterate 1	**Iterate 2**	Iterate 3
1		84.0	77.3	86.79	87.82	**87.31**	86.79
2	82.3		86.5	91.93	91.93	**93.16**	92.63
3	81.9	86.9		89.30	91.05	**92.63**	91.40

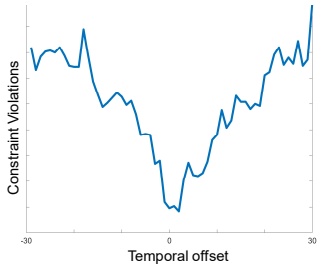


Spatial ambiguity varies over developmental time due to the geometry of the embryo as indicated by the accuracy of matching results at different stages as seen to the left. Additional information such as nuclear size might minimize this ambiguity.

Computational Future Work:

Initial spatial and temporal alignment are currently manual which is sufficient for EM but not for higher throughput imaging modalities where this kind of history-less naming would be useful. A particular area of interest is application of these methods to co-align variant development embryos such as developing zebrafish at the single cell level. Prior information is more limited but the challenges are identical.

A combination of RANSAC like alignment and optimizing temporal alignment to minimize constraint violations could fully automate the process of naming novel data.



Constraint violations at offsets from manual optimal temporal alignment of 320min embryo

Acknowledgments:

Thanks to the entire WormGUIDES team, particularly Ryan Christensen, William Duncan and Brandon Harvey of Hari Schroff's lab for the lineaging of reference embryos and Braden Katzman and Doris Tang for work on the WormGUIDES atlas software.