Streamlining microscopy datasets by enriching for in-focus frames

We distilled label-free microscopy data by comparing and implementing feature-detection algorithms. Sobel and Laplacian methods outperformed pixel intensity variance in accuracy.

Purpose

Microscopy datasets are notoriously large, making more complex analyses of these data inherently slower or intractable [1][2][3]. Previously, we'd found that filtering large datasets down to smaller, information-dense subsets allowed us to iterate on our experimental troubleshooting in increments of minutes or hours, instead of many hours or days. In that case, we reduced the datasets by 82% by including only frames of cells in focus.

Here, we compare three simple feature-detection algorithms to identify the most reliable metrics for selecting in-focus frames in label-free microscopy data, namely, the variances of the raw image, the Sobel-filtered image, and the Laplacian-filtered image. We found that the variances of the Sobel- and Laplacian-filtered images surpassed the variance of raw pixel intensities in accuracy and that both the Sobel and Laplacian filters provided accurate in-focus frame detection, aligning closely with expert assessments. We hope this comparison will help cell biologists and computational researchers expedite and refine analysis in high-throughput experiments.

- This pub is part of the **platform effort**, "[Microscopy: Visually interrogating the natural world](#)." Visit the platform narrative for more background and context.
- Our **time-lapse data**, **expert annotations**, **Fiji macro**, and **code** in Python are available in [this GitHub repository](#).

The strategy

The problem

The increasing volume of time-lapse microscopy data in experimental biology poses challenges in extracting useful biological information at scale [1][2][3]. Increasing data volume entails longer transfer and processing times, which can slow the pace of biological research. For example, our initial computational workflows to measure the motility, size, and shape of two species of algae took days of processing time, largely due to the total size of the time-lapse data [4]. Also, we found that measuring the morphology of motile cells in bulk within time-lapse data was less accurate than measuring cells in maximal focus [4].

We're performing an interspecies hybridization experiment that involves collecting phenotypic measurements of *Chlamydomonas* cells from time-lapse microscopy data [4]. We want to collect accurate measurements of as many algal cells as possible. We plan to analyze thousands of progeny strains in our experiment.

Our solution

One solution to manage such large and complex image-based datasets is to filter the raw data to enrich information that is useful. We adopted this approach during our preliminary benchmarking of the two interfertile *Chlamydomonas* species that we're hybridizing [4]. In our previous approach, we first selected a subset of frames at random (the first 100 frames of the time-lapse), but quickly realized that this approach can lead to frames without the object in focus. We then applied one technique (variance of Laplacian) for detecting edges in an image, and successfully extracted frames with the cells of interest in focus [4]. Focus-filtering allowed us to speed up our processing time. Also, parsing focal sequences gave

us more accurate morphology measurements of our cells. We measured the cell footprint that had the maximal area in the sequence, as opposed to using measurements of cells that were partially in focus that would inaccurately skew the data [4].

The resource

We evaluated three different focus-filtering methods by comparing them to manual human annotations of in-focus frames. We thought that these methods could be useful for downsampling any dataset where the object of interest moves in and out of the focal plane, or in cases where frames contain superfluous objects. In addition, we wanted to explore two different types of label-free data (DIC and BF) that might be useful for training machine-learning models on microscopy data. We're sharing the results here to serve as a resource to other researchers interested in applying these approaches to various data types.

We chose to assess three distinct focus metrics that can detect features or edges in computer vision workflows [5][6]: variance of pixel intensities, variance of edge sharpness (determined via the Sobel operator), and variance in detailed edge sharpness (determined via the Laplacian operator). Each operator, Sobel or Laplacian, applies an algorithm to convert an image to a filtered image. Then we take the variance of each pixel value in the filtered image to use as a metric. Each approach is summarized in the list below and you can see example visuals in Figure 1.

1. **Variance of pixel intensities**: This metric gauges the dispersion of pixel intensities from their mean value within each frame. We postulated that frames with in-focus cells would exhibit a heightened pixel intensity variance. The underlying rationale is that such frames, being more feature-rich, would likely have a wider range of pixel intensities.

2. **Variance in edge sharpness**: To understand the spread or variability of edge sharpness in our images, we analyzed the variation in the intensity of edges. We did this by computing the magnitude of the image gradient, which highlights the edges, using the Sobel operator. We hypothesize that images with in-focus cells will show a higher variability in edge intensity. This is because in-focus images tend to have crisper, more defined edges.

3. **Variance in detailed edge sharpness**: Building on the previous idea, we also examined the variation in the sharpness of finer details within the edges themselves. This is achieved by using a second-order differential operator, the Laplacian, to measure the rate of change of the magnitude of the gradient. Our hypothesis was that this measure would further distinguish between in-focus and out-of-focus cells, as sharper detail within edges is more pronounced in the in-focus images.
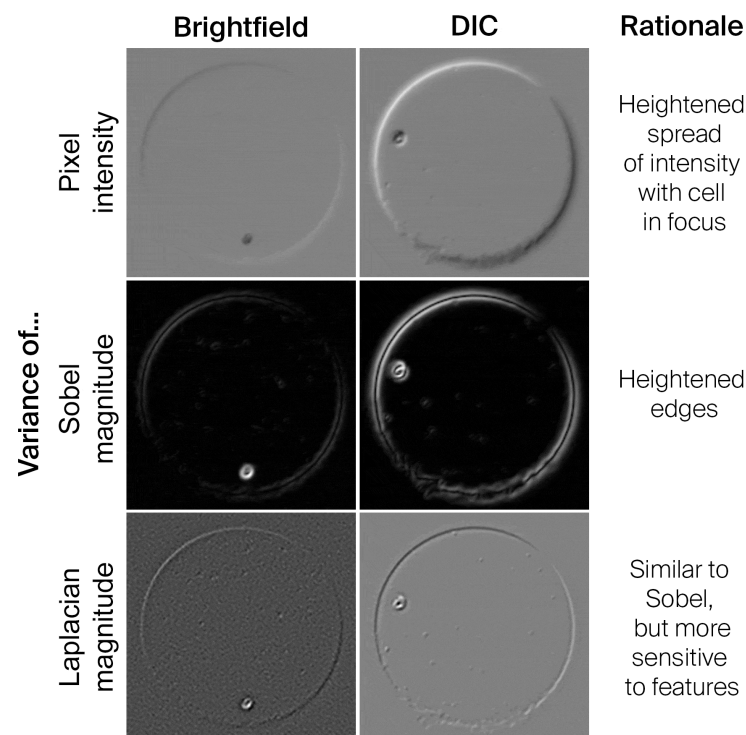

Figure 1. **Application of focus metrics to time-lapse bright-field (BF) and differential interference contrast (DIC) microscopy data**.

Image sequences show a *Chlamydomonas* cell swimming in a well with a diameter of 100 microns. Each filter, Sobel or Laplacian, applies an algorithm to convert an image to a filtered image. Then we take the variance of each pixel value in the filtered image. "Rationale" denotes why we thought each measure might be able to distinguish between frames with cells that are in focus vs. out of focus.

To compare these three algorithmic approaches against a "ground truth," we recruited a few of our scientists who are proficient in cellular microscopy and adept at identifying when cells are in focus. These experts examined time-lapse sequences of motile cells as they swam in and out of the focal plane, which we previously captured using either DIC or BF imaging [4]. Participants classified each frame as either "in focus" or "out of focus." The annotations all agree with one

another for 70% of frames (126/180). At least 3 of the 4 annotations agreed with one another for 92% of frames (166/180).

We evaluated the accuracy of each focus metric as a predictor of the expert-annotated in-focus frames by plotting the receiver operating characteristic (ROC) curves for each metric. The ROC curve is a plot of the true-positive rate (TPR) as a function of the false-positive rate (FPR) for a binary classifier. In this case, we change the value of the threshold systematically to construct the ROC curve.

We calculated and plotted the ROC curves for each focus metric and each of the four experts' sets of manual annotations (Figure 2). We found that the variance of pixel intensity is a poor focus metric for both bright-field and DIC time-lapse data (Figure 2, A). For each of the expert annotations, this metric had roughly equal false- and true-positive rates at most thresholds — ROC curves roughly align with the diagonal of the plot, suggesting it performs as well as a program that selects frames randomly.

Promisingly, we found that the Sobel and Laplacian metrics accurately identified frames of cells in focus. For each expert annotation, these metrics displayed a high true-positive rate and a low false-positive rate at most thresholds. At a false-positive rate of 5%, the median true-positive rate for the Sobel metric was 71% for bright-field images and 79% for DIC images. For the Laplacian metric, it was 57% for bright-field images and 72% for DIC images. Consequently, the ROC curves were aligned along the upper-left portion of the plot. Thus, we see that the Sobel and Laplacian metrics perform well in this particular quantitative assessment.

While the Laplacian metric appears to perform a bit worse on BF images (i.e. 57% for the Laplacian versus 71% for the Sobel), these reportings are sensitive to differences in user annotations, so we don't put a ton of weight on these numbers. That said, this finding is in line with our qualitative assessments. By simple visual inspection, we see that the Laplacian metric performs better on DIC data than on BF data. This fits with our expectation that Laplacian should do better with finer features, as there are more fine features in DIC images. In addition, the Sobel metric identifies frames where the shadow of the cell is visible, whereas the Laplacian metric is less sensitive to the shadow. If we want to

track cells even when the cell is slightly out-of-focus (i.e. a shadow is visible), we might use the Sobel metric. If we only want the frames of the cell truly in focus, we might use the Laplacian metric. In our previous work, using Laplacian focus filtering reduced data volume by 82% while maintaining accurate measurements [4].
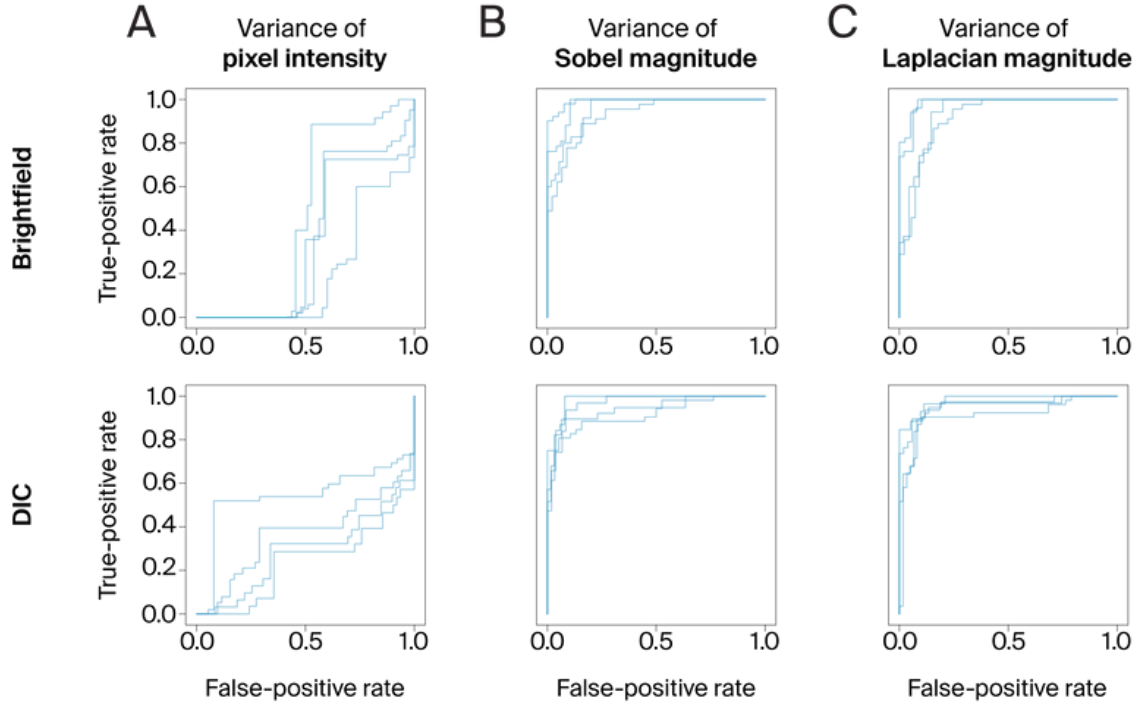

Figure 2. **Variance of pixel intensity is a poor predictor of in-focus frames, while variances of Sobel and Laplacian magnitudes perform well on bright-field and DIC data**.

We've plotted ROC curves for the variances of pixel intensity, Sobel magnitude, and Laplacian magnitude to see how they predict in-focus frames on bright-field and DIC time-lapse data. Each curve represents data from a single user annotation. There were four user annotations.

In conclusion, our comparative analysis of focus metrics within label-free microscopy data has revealed useful distinctions in performance between feature- and edge-detection algorithms. Our study demonstrated that the Sobel and Laplacian filtering methods align closely with manual assessments of DIC data by expert microscopists, with high true-positive rates. Sobel performed better than Laplacian filtering for BF data. The variance of pixel intensity proved to be a poor focus metric for both types of data. Complementary to our work, independent researchers found that the Laplacian metric is effective at segmenting regions of objects in focus in micrographs [7]. They also compared

many other metrics and developed a deep learning pipeline for segmenting objects in focus.

We previously used the Laplacian filtering method to identify in-focus frames of algal cells [4]. Given its strong performance here with DIC data, we'll stick with this approach when we move onto the higher-throughput phase of that project because we plan to image using DIC. For future work that involves tracking cells and cell shadows in time-lapse data, we'll consider Sobel filtering.

We hope these findings serve as a resource to guide the use of specific focus metrics for cell biologists and computational imaging specialists, particularly those working with label-free microscopy data (DIC and BF). This work also validates the use of focus filtering as a method to enrich datasets with information that is useful, enabling us to phenotype more strains of cells in our future experimental workflows. Specifically, applying the filtering methods we explored here should improve the speed and reliability of phenotypic measurements in our interspecies hybridization experiments.

Methods

We based our assessment on a time-lapse dataset containing 180 frames of *Chlamydomonas* cells swimming in and out of focus. We collected these images with either DIC (90 frames) or bright-field (90 frames) microscopy for a prior pub [4]. We manually curated the experimental data so that we had cells that were clearly going in and out of the focal plane. We aimed to have an equal representation of "in focus" and "out of focus" frames. In a pilot experiment, we showed frames to the experts in a random order. However, the annotations were highly variable, because the experts had a hard time assessing whether a cell was in focus without seeing the transition between in and out of focus. In our subsequent attempt presented here, we showed all frames in sequence.

We opened the dataset in Fiji [8], and ran a macro that asked the user for an annotation of either "in focus" or "out of focus." Four experts annotated each frame. We analyzed the results of the experiment by receiver operating

characteristic curve (ROC) analysis. The code to perform the analysis and plot ROC curves can be found in our GitHub repository.

We used ChatGPT to write, clean up, and comment code. We also used ChatGPT to write text that we edited, suggest wording ideas, streamline and clarify text, and rearrange text to fit the structure of our "Resource" pub template. Last, we asked ChatGPT for a list of suggested focus metrics that we could use in this study, and we selected three from the list.

Key takeaways

1. Sobel and Laplacian filtering methods can accurately identify frames of cells in focus that agree with human microscopy user assessments.
2. The variance of the raw pixel intensities in the image fails to accurately identify frames of cells in focus.
3. We will continue to use the Laplacian metric for identifying in-focus cells in DIC data.
4. In future work, we will consider using the Sobel metric for identifying cells and cell shadows.

Next steps

In the next phase of our research, we'll apply Laplacian focus filtering of DIC time-lapse data. The goal is to perform high-throughput phenotyping of progeny from our _Chlamydomonas_ species hybridization. We'd appreciate any feedback on this pub, especially questions that would help you replicate the work and insights from anyone who may have compared these approaches in analyzing other types of microscopy data.

Contributors (A–Z)

- **Feridun Mert Celebi**: Supervision, Validation

- **Keith Cheveralls**: Formal Analysis, Software, Validation, Visualization

- **Seemay Chou**: Supervision

- **Tara Essock-Burns**: Conceptualization, Supervision

- **Megan L. Hochstrasser**: Editing, Visualization

- **Galo Garcia III**: Conceptualization, Formal Analysis, Investigation, Software, Visualization, Writing

References

1. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. https://doi.org/10.1038/sdata.2016.18

2. Moore J, Basurto-Lozada D, Besson S, Bogovic J, Bragantini J, Brown EM, Burel J-M, Casas Moreno X, de Medeiros G, Diel EE, Gault D, Ghosh SS, Gold I, Halchenko YO, Hartley M, Horsfall D, Keller MS, Kittisopikul M, Kovacs G, Küpcü Yoldaş A, Kyoda K, le Tournoulx de la Villegeorges A, Li T, Liberali P, Lindner D, Linkert M, Lüthi J, Maitin-Shepard J, Manz T, Marconato L, McCormick M, Lange M, Mohamed K, Moore W, Norlin N, Ouyang W, Özdemir B, Palla G, Pape C, Pelkmans L, Pietzsch T, Preibisch S, Prete M, Rzepka N, Samee S, Schaub N, Sidky H, Solak AC, Stirling DR, Striebel J, Tischer C, Toloudis D, Virshup I, Walczysko P, Watson AM, Weisbart E, Wong F, Yamauchi KA, Bayraktar O, Cimini BA, Gehlenborg N, Haniffa M, Hotaling N, Onami S, Royer LA, Saalfeld S, Stegle O, Theis FJ, Swedlow JR. (2023). OME-Zarr: a cloud-optimized bioimaging file format with international community support. https://doi.org/10.1007/s00418-023-02209-1

3. Poger D, Yen L, Braet F. (2023). Big data in contemporary electron microscopy: challenges and opportunities in data transfer, compute and management. https://doi.org/10.1007/s00418-023-02191-8

4. Avasthi P, Braverman B, Essock-Burns T, Garcia III G, MacQuarrie CD, Matus DQ, Mets DG, York R. (2023). Phenotypic differences between interfertile Chlamydomonas species. https://doi.org/10.57844/arcadia-35f0-3e16

5. Mir H, Xu P, van Beek P. (2014). An extensive empirical evaluation of focus measures for digital photography. https://doi.org/10.1117/12.2042350

6. Pech-Pacheco J, Cristobal G, Chamorro-Martinez J, Fernandez-Valdivia J. (None). Diatom autofocusing in brightfield microscopy: a comparative study. https://doi.org/10.1109/icpr.2000.903548

7. Li R, Kudryashev M, Yakimovich A. (2023). A weak-labelling and deep learning approach for in-focus object segmentation in 3D widefield microscopy. https://doi.org/10.1038/s41598-023-38490-2

8. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B, Tinevez J-Y, White DJ, Hartenstein V, Eliceiri K, Tomancak P, Cardona A. (2012). Fiji: an open-source platform for biological-image analysis. https://doi.org/10.1038/nmeth.2019