Exploratory data mining on images of budding yeast with dicentric plasmids

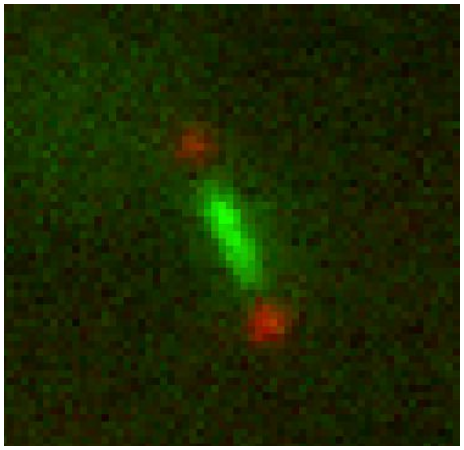
Derek Fulton

Abstract

Computational image analysis is an excellent tool for the investigation of cellular processes. Further, recent advances in imaging technology have enabled the creation of large numbers of image sets. However, such a large amount of complicated data is often difficult to analyze in a timely manner and has become the bottleneck in the research workflow for many groups, because it has not yet become fully automated in a reliable manner. Therefore, it is a good use of resources to exhaust all possible knowledge from the tediously-acquired data in search of any phenotypes.

One fantastic technique for that purpose is data mining. In this project, the data mining process generally involved creatively generating relevant features from the image analysis data and then exploring those features with tools such as classification and data visualization.

In this project, we used wild-type plus three mutants (SIR2Δ, YKU80Δ, and BRN1Δ) and allowed them to transform a dicentric plasmid labelled with a LacO/GFP array on 80% of one side. The spindle pole bodies were labelled with RFP. This allowed the easy visualization of the physical dimensions of the plasmid relative to the spindle and how those changed over time steps.



Picture 1: Image of dicentric plasmid labelled with GFP/LacO array (green) and spindle pole bodies (red) during metaphase. All raw image analysis data (kinetochore microtubule length, centroid position, major axis, spindle length) came from this data.

Analyzing the raw data allowed us to compare mutants by asking physically relevant questions such as “Does the plasmid rock back and forth between the two spindle pole bodies?”, “Are the kinetochore microtubules of similar length?” and “How quickly do the kinetochore microtubules stretch or shrink?” Comparing the data acquired from data mining to the functions of the gene of interest in each mutant allows us to gain knowledge and gain a deeper physical understanding of the morphology of the mitotic spindle in yeast.

Methods

I started with data from an image analysis GUI (Josh). The images were 30-second timesteps with seven z-slices each, and the GUI enabled the selection of the spindle pole body positions as well as the plasmid length. From this data (a double array in MATLAB with each value representing the brightness of a pixel), the image analysis GUI calculated (and therefore I started off with):

- Plane (which z-slice showed the brightest pixels?)
- Major axis length (nanometers)
- Minor axis length (nanometers)
- Centroid position (center of the LacO array in x, y, z)
- 3D Spindle Length (nanometers)
- Spindle Pole Body 1 Position (nanometers)
- Spindle Pole Body 2 Position (nanometers)
- Length of Kinetochore Microtubule 1 (nanometers)
- Length of Kinetochore Microtubule 2 (nanometers)

With this data, I sought to creatively extract features of physical relevance. As each mutant provided me with a couple hundred tiff stacks, I also paid special attention to the dynamic changes in these values over the thirty-second timesteps. With these raw data, I can compare between mutants features such as kinetochore microtubule length, kinetochore microtubule length difference (is one longer than the other?), kinetochore microtubule stretch/shrink rate (how quickly are the kinetochore microtubules stretching/shrinking?), kinetochore microtubule stretch/shrink rate difference (are they stretching evenly?) and so on and so forth.

After parsing all of the data into these features and creating one large dataset, I began using MATLAB for ease of data visualization. It allowed me to plot all of the mutants together, and visually look for relevant differences. I also passed the data into a data exploring software called Weka, which allowed quality data visualization as well as easily applied machine learning algorithms (figure 2).

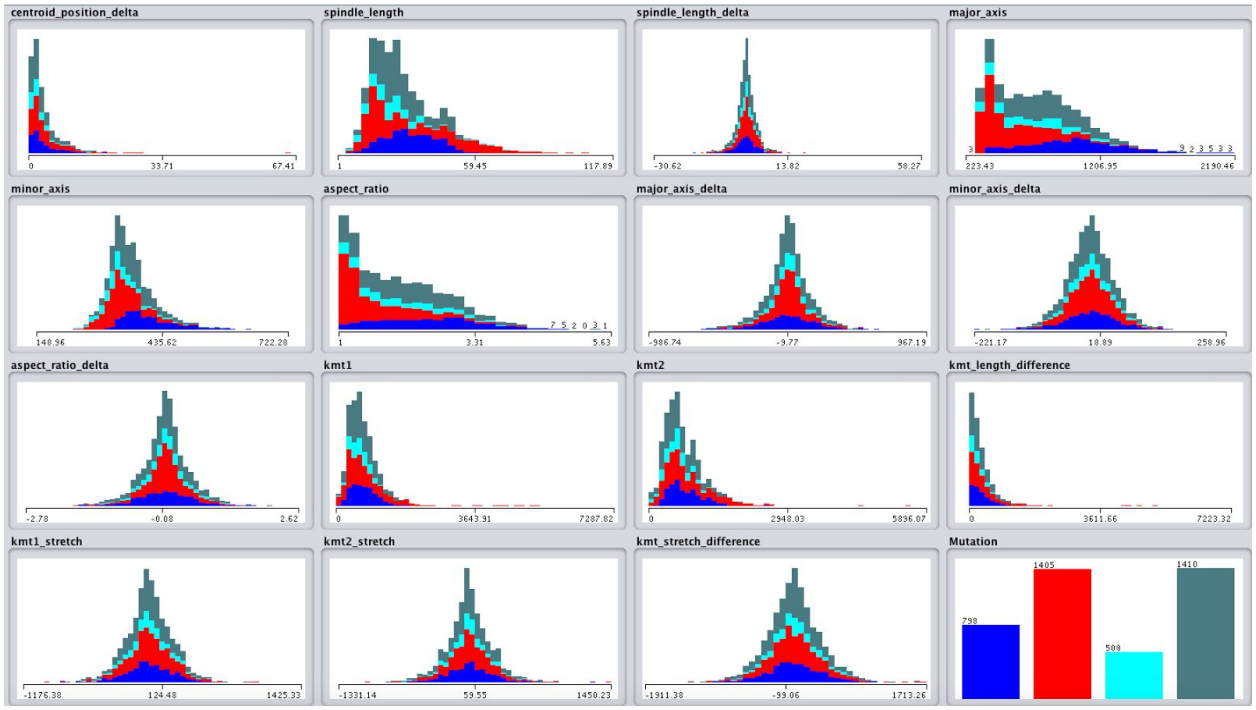
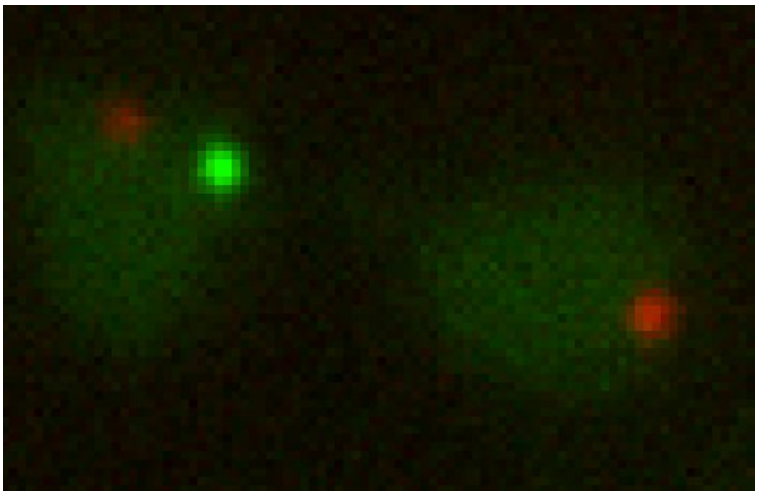


Figure 1. Exploratory histograms of multiple features of the dicentric plasmids for BRN1Δ(dark blue), SIR2Δ(red), wild-type(cyan) and YKU80Δ(gray). Note roughly normal distributions in all of the features except for aspect ratio, major axis and spindle length.

Results

After exploring the data with classification algorithms and data visualization tools, the most promising possible phenotype was the spindle length of the SIR2Δ. Not only did the SIR2Δ mutant exhibit spindle lengths significantly greater than any of the other mutants, it appeared that the plasmid was “rocking” back and forth more rapidly in the SIR2Δ mutant than in other mutants, even for spindles of similar size (figure 2).



Unfortunately, after further analysis, it was discovered that an idiosyncrasy of the original Image Analysis GUI was responsible for the dramatic kinetochore microtubule length difference in the SIR2Δ mutants. Though it appeared that SIR2Δ mutants did display uneven kinetochore microtubule lengths at first, which is a discernible phenotype worth more investigation, it turns out these results were caused due to detachment of the plasmid from one of the kinetochores (picture 2). As the kinetochore microtubule length measurement is only the euclidean distance between the user-selected spindle pole body and the edge of the plasmid signal, this "detachment event" would give highly uneven kinetochore microtubule lengths and suggest a phenotype which does not exist.

Regarding the machine learning component of this data mining project, no well-known classification algorithms were able to correctly classify the phenotypes more than 60% of the time. Practically speaking, this is even less relevant because of the many SIR2Δ outliers (long spindle and high kinetochore microtubule length difference) which are trivially easy for a sophisticated classification algorithm to classify.

Discussion

Overall this was a fruitful process. I do believe that with more time, more mutants and more data, a solid phenotype could be discovered and we could learn more about the physical mechanics of the spindle. When I resume work on this project, I would like to spend some time acquiring more data of wild-type cells, as well as removing any of these "detachment" events or similar kinks in the process.

One thing that should be done is data mining on the raw image data of each plasmid to investigate its shape and take into account one variable previously ignored: intensity. Intensity is difficult to work with for a number of reasons relating to microscopy, but with sufficient data and robust algorithms I believe this can be overcome. We will be able to ask questions about the plasmid shape such revolving around the intensity, which will provide a proxy measurement for concentration of LacO/GFP array binding, which can then be compared with the spindle measurements from this project to hopefully discover new facts about the spindle.

One example is a graph of distance from brightest pixel on the x-axis and intensity on the y-axis. One would expect a reasonably good fit, as well as a negative slope.