Automated Image Analysis for Systematic and Quantitative Comparison of Protein Expression within Cell Populations

by

Louis-François Handfield

A thesis submitted in conformity with the requirements
fer the degree of Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto

Automated Image Analysis for Systematic and Quantitative Comparison of Protein Expression within Cell Populations

Louis-François Handfield Doctor of Philosophy
Graduate Department of Computer Science University of Toronto, 2014

Abstract:

Protein subcellular localization is a major indicator of protein function, and efforts have been made to systematically determine the localization of each protein in budding yeast using fluorescent tags. Based on the fluorescence microscopy images, subcellular localization of many proteins can be classified automatically using supervised machine learning approaches. Budding yeast has a stereotypical reproduction mode, such that cell-stage is related to the presence and size of a growing bud. In this work, I investigate the benefits of a cell recognition method and image features that utilize prior biological knowledge of budding yeast shape and its cell-stage dependent changes.

I show that modeling cell-stage dependency of protein abundance and spatial distribution (expression pattern) within a continuous model for cell growth allows the identification of most previously identified localization patterns in a cluster analysis. Further, I show that similarities between the inferred protein expression patterns explain similarities in protein function better than previous manual categorization of subcellular localization. These results suggest that incorporating prior information about yeast morphology in automated image analysis will yield unprecedented power for pattern discovery in high-resolution, high-throughput microscopy images.

Finally, using these new computational methods, I explore cell-to-cell variability in protein abundance and subcellular localization. I define a mean to quantify deviations in subcellular localizations, and find that the method defined is in agreement with previous measurements of cell-to-cell variability in the case of protein abundance. Hence, I show that cell-to-cell 'spatial variability' is a protein expression property, whose measurement is only possible from microscopy images. This measure allows the systematic detection of many classes of such variability, without the use of any prior knowledge about subcellular localization.

Contents

List of Tables

List of Figures

Introduction

Overview

Advances in technology increased the complexity and resolution of measurements for many experimental assays. Further, automation of some procedures allows the execution of lists of similar independent experiments and produces large collections of measurements (high-throughput experiments). In turn, the analysis of such collections is demanding, so automation of downstream analysis is desirable so to systematically detect outstanding observations or quantify tendencies in large data collections. Machine learning methods have shown to be potent for those two tasks, but the nature of the knowledge uncovered from the application of machine learning methods may be difficult to interpret.

A major goal of research in systems biology is the construction of models that accurately describe the workings of cells, tissues and organisms [174]. The characterization of biological processes often requires an understanding of the sequential changes in biological states, many of which also have spatial dependency. For example, the spatial distribution of compounds and/or proteins in cells and organisms is critical for defining realistic models; specific experiments need to be designed to collect spatial information [132].

In this introduction, I first survey means used to qualify protein localization through the use of automated procedures that can be systematically applied to a large number of parallel experiments (high-throughput), as well as technological advances that motivate the development of automated analyses of data sets produced by microscopy. In the second section, I will discuss methods for automated microscopy, focusing on protein subcellular localization (PSL) computational prediction methods. In the last section, the divisions in this thesis are presented.

1 Biological Introduction

1.1 Microscopy

***Microscope*:** The microscope is the first key tool of cellular biology, which enabled the discovery of the cell and cell division [5]. Only after three centuries of work on improving the microscope did an alternative for the classical imaging setup arise: In 1904, August Kohler, who also enhanced bright field microscopy by designing a lens arrangement that made the light source perfectly defocused while the plane containing the sample is in focus to the observer [143], arranged a light source that is not directly observable, so that only deflected light is recorded. This dark field microscopy setup has better signal to noise and contains less artefacts than bright field microscopy, but requires the use of an intense light source (Figure .1).

***Fluorescence microscopy*:** Cellular structures stained with fluorescent dyes have much better contrast

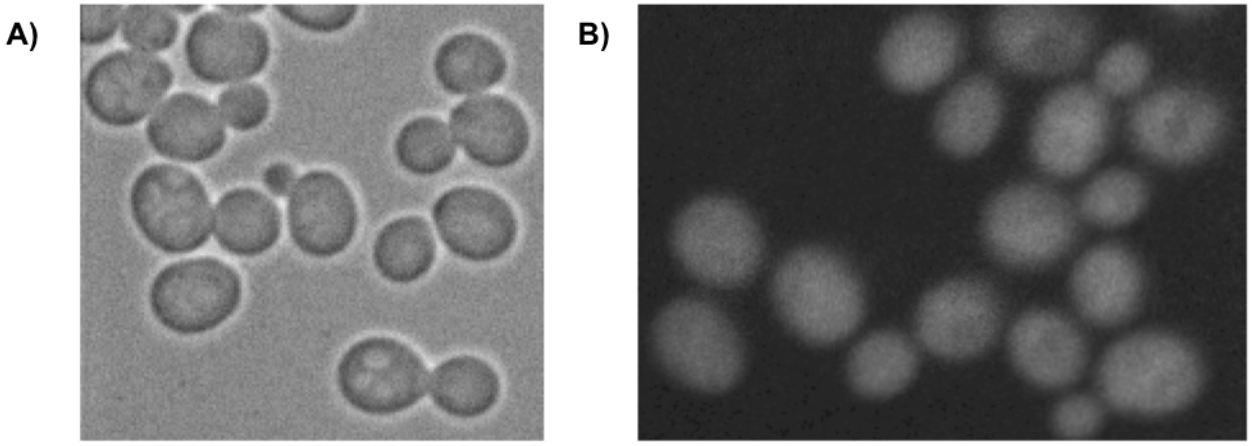


Figure .1: **Microscopy illumination type.** A) Brightfield Microscopy image of budding yeast (image courtesy of Alex Nguyen Ba) B) Darkfield Microscopy, rendered using a red fluorescent protein (RFP). This type of microscopy is exclusively used in this work (image courtesy of Yolanda Chong).

than in brightfield microscopy [84]. Imaging some macromolecule's spatial distribution using a fluorescent antibody (Immuno-fluorescence) is feasible, but this approach may not be applied for live cell imaging, as membranes need to be made permeable so that anti-bodies may enter the cell [112]. The development of charged-coupled devices (CCDs) enabled the numerical capture of images that are 80 fold more sensitive than the human eye [139], and allowed the quantification of intensities that are needed numerical calculations. In the 1980s, the quantification power of digital microscopes was first considered in proposals for means to measure pH and protein concentrations [116].

***Protein Fluorescence*:** In 1962, Shimomura et al. [147] identified the green fluorescent protein (GFP) from jellyfish. Modifications to certain amino acids within the protein change its stability as well as its fluorescent properties [161]. Through these modifications, a large collection of fluorophores was derived, such as RFP (red), YFP (yellow), CFP (cyan), each holding different activation and emission frequencies. By including the gene that encodes for one of these fluorescent proteins in another organism such as yeast or mouse, it is possible to make these organisms express the fluorescent proteins. This can usually be done without altering the cell viability. Huh et al. [76] created a collection of 6,029 yeast strains with a green fluorescent protein (GFP) included in the open reading frame (ORF) of each protein. The fluorescent protein is expected to report the location specifically for each yeast protein, since it is constructed by the yeast cell as a part of the protein of interest, and is therefore referred to as a fluorescent "tag". Its creation, motion and destruction are assumed to be identical to its target. Huh et al. manually classified about 4200 yeast proteins in 22 subcellular protein locations after visual inspection of the fluorescence microscopy images, while allowing

proteins to be assigned to more than one class if needed.

1.2 Advances in Microscopy

In recent years, an overwhelming number of complicated microscope images have been produced, creating challenging computational problems [122]. The vast majority of the protein localization studies were performed using visual interpretation of the images by researchers [112]. But the amount of information collected in each study is growing to the point where years of manual work would be required to analyze certain datasets [122]. New technologies enable that capture of information that is getting harder for an expert to analyze, such as 3D images or time-series of images. The desire to image the cell in finer temporal and spatial scales drove the development of new technologies [166].

***Protein subcellular localization (PSL)*:** Using the fluorescent proteins and microscopy techniques described above, protein spread within cells can be assessed. After visualization, proteins are often assigned subcellular localization labels. Such labels are major determinants of protein function, since one of the major determinants of protein function is sub-cellular localization. For example, protein and lipid biosynthesis occur in the endoplasmic reticulum (ER) [53]. Therefore, a protein of unknown function that localizes to the ER is more likely to be involved in protein or lipid biosynthesis. Similarly, aerobic sugar metabolism occurs in the mitochondria, therefore proteins localized to the mitochondria are likely to be involved in these processes.

***Confocal Microscopy*:** This microscope shines a focused light into a single point on the specimen, and recovers fluorescent signal while filtering out-of-focus components [56]. The sample needs to be scanned in order to recover an image, but the lateral resolution is found enhanced. For that reason, it is typically used for imaging subcellular localizations and organelles. In addition, confocal microscopes can recover 3D maps of protein distributions. Acquisition of 3D images is also possible, but it involves a computational correction for the contribution of out of focal plane signal, inferred from a set of 2D slice images [19].

***Automated Image Acquisition*:** Many microscopes are now programmable, so that the focus [51] and some sample manipulations (such as compound addition [116]) can be automated. Beyond convenience, this helps the standardization of the images. For example, it has been shown that the use of automated microscopy increased the sensitivity and displayed a lower bias than manual microscopy [70]. While generating 3D images by confocal microscopy, the microscope can also selectively image regions of interest [103]; this speeds up acquisition time and reduces unnecessary photobleaching [80].

1.3 Microscopy Datasets

In this work, I analyzed several collections of yeast that were modified to produce fluorescent markers. In all cases, yeast collections are derived from the Huh et al. [76] collection, so that derived collections each contain about ∼4200 yeast strains that have a GFP inserted into the gene encoding one of the ∼6000 yeast proteins. It is important to note that 'one protein' or 'several proteins' refers to protein types produced from a given gene or set of genes, and not to a quantity of molecules. It is possible to add further genetic mutations by crossing the collection with another strain using synthetic genetic array (SGA [158]).

1. In a first collection, a highly expressed RFP (a tdTomato [142] fluorescent protein from the constitutive RPL39 promoter), integrated at the HO locus, was introduced into the GFP collection to mark the whole cell area in order to facilitate automated image analysis. Micrographs were acquired using a confocal microscope (Opera, PerkinElmer). Eight micrographs were imaged (at 1331 x 1017, 12 bit resolution) from each strain, 4 in the red channel and 4 in the green channel, yielding a dataset of 44 Gb of image data. This first collection was used to develop and benchmark cell identification methods, so it is used throughout the thesis. Two additional replicate image collections with identical specifications were also analyzed, but are only used to report on classification accuracy (Section 4) and reproducibility of certain measurements and modeled quantities (cell-to-cell variability; part III). The collection was imaged by Yolanda Chong, and is not currently publicly available.

2. A time series of images was also produced from the above strain collection, where cells were given a mating pheromone (alpha factor). These images show changes in bud morphology that are due to prolonged influence of the mating pheromone. This was used to show classification performance in a collection of strain with altered morphologies (Section 4). This collection was also imaged by Yolanda Chong.

3. One last image collection analyzed was produced in Tkach et al. [156]. It was obtained through Yeast Resource Center [131]. Using the same microscope, they imaged a different strain collection, which had a RFP (mCherry [142]) tagged onto NUP49, instead of a cytoplasmic marker. The image size and resolution are identical to the previous sets, but only 3 images are taken per strain. Three image collections are rendered; one control set and cells population treated with hydroxyurea or methyl methanesulphonate. These three collections are utilized report on reproducibility of modeled quantities (cell-to-cell variability; part III), when the segmentation and normalization method slightly differs.

2 Computational Introduction

Image analysis and image processing are encompassed by the Computer Science field. In general, any microscopy image analysis will follow the same basic steps as used in general image analysis. Methods for image correction, segmentation, object recognition, feature extraction, image classification and object tracking are all used [101, 122]. In this thesis, I will introduce image correction, segmentation and object recognition methodologies in the Background section of Part I, and image classification methods will be introduced in the Background section of Part II. In addition to classical methods from image analysis, I also use methods from computational statistics and machine learning and these are introduced in Part II and Part III. Here I provide some context for bioinformatics/data mining methods for identification of subcellular localization. First, I will cover means for inferring the protein subcellular localization (PSL) that do not require microscopy data.

2.1 Subcellular Localization from Protein Sequence

Proteins that are similar or that can be shown to physically interact are more likely to be in identical subcellular localization. The similarity of proteins is measured in several ways, but the availability of genomes makes amino acids sequences a common information source. Most simply, protein sequences can be aligned and compared using hamming distance or similarity measures based on the chemical properties of the amino acids they share. More sophisticated pattern recognition techniques can be applied to identify interesting subsequences that may predict similar protein function (Pfam [13], InterPro [77]). In addition, some mechanisms are known to transport proteins within cells given that its sequence tail is enriched in certain amino acids [49]. Some other proteins possess short motifs (short linear motifs) that are recognized by the transport machinery [111]. Some methods [49, 111] were proposed to specifically infer localization from these features alone, but the training data for these approaches is limited; hence, the coverage for obtained localization features is small [93, 112].

***Knowledge based*:** Computational methods have been developed to integrate sequence features as well as expert annotations [8] and computationally inferred annotations [55] to predict protein localizations. For example, ISort was the first method to deal with the multiclass classification problem for 22 sub-cellular localization classes [36]. It uses a nearest-neighbour classifier [105] with distance defined as the projection of the feature vector for a protein of interest onto each other protein. Using the manually annotated yeast GFP microscopy images (Huh et al. [76]) as a gold standard this method obtained a 70% success rate in jack-knife cross-validation [63]. Another method, PLPD [91] attempted to solve the multiclass, multi-label and unequal class size using 'Density-induced Support Vector Data Description', a classification algorithm

loosely based on support vector machine (SVM) [151]. They reported an accuracy of 10% higher than Isort. One concern about this approach is that databases of annotations are based on diverse experimental sources, which would now include the work of Huh et al. [76]. There are many other computational methods to predict subcellular localization based on many other machine learning techniques, and these represent an alternative to direct microscopy-based methods.

3 Rationale

Image analysis of microscopy images has been used in several high-throughput studies to identify morphological phenotypes and localization of protein expression within the cell compartments (Nucleus, Mitochondria, Cytoplasm, etc.). Assessment of protein subcellular localization has been performed manually for budding Yeast [76]; then, automated supervised learning techniques were proposed to perform that task on larger or more complex data sources [19]. The classification task has been studied extensively, and subcellular localizations are recovered with high accuracy, except for rare subcellular localizations and subcellular localizations with strong cell-stage dependencies. Proteins often appear in several locations within a cell at the same time, or the same protein may be present in different locations in different cells at different fractions, and this creates difficulties for classification approaches. Therefore, automated approaches rely on manually removing these from both the training set and cross-validation analysis. The definition of biological function for a particular protein is often unclear [55] and protein function may not be fully characterized by a binary assessment into subcellular categories [122]. Therefore, a major goal of my thesis will be to move beyond simple classification of protein expression, by quantifying protein abundance over time and space, without predetermining the possible patterns.

One of the limitations of machine learning techniques that implement supervised learning is that the nature of class discriminative information cannot be extracted and translated into biological knowledge. To that aim, specific experimental methods have been proposed to characterize protein spatial [24, 132] and temporal profiles, and certain expression properties, such as stochasticity [2] and cell-to-cell variation [152]. These studies focused on a single target protein of interest (low-throughput), in order to validate their proposed model of these protein expression features. There are some high-throughput studies whose aim were the characterize cell shape [138] and stochastic levels [113] for all yeast proteins, and succeed in demonstrating that measurements could be linked to biological processes. Therefore, I propose to analyze protein spatial expression in a similar manner, using methods that are both high-throughput and produce biologically interpretable measurements.

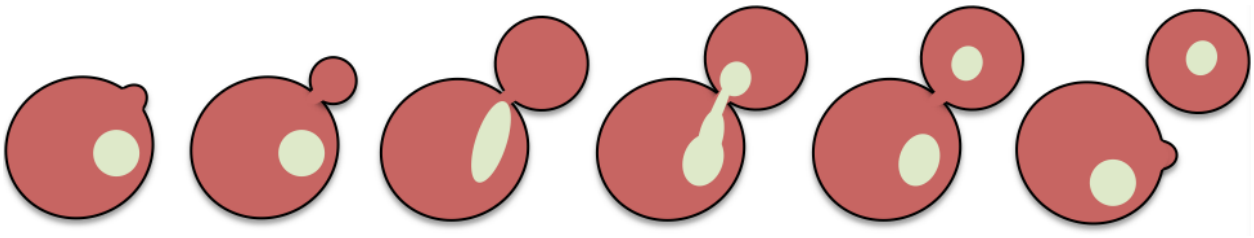


Figure .2: **Cell cycle of budding yeast.** The cell cycle of haploid budding yeast is characterized by the growth of a bud on the periphery of a 'mother' cell. The bud becomes an independent cell after it reaches a certain size. The imaged cells are stained with red and green fluorescent proteins (RFP and GFP respectively), so this represents a possible experimental configuration where the RFP stains the entire cell and GFP is tracking the nucleus that gets included into the growing bud.

3.1 Approach

In order to understand the changes in protein expression over time, I propose to temporally order the cells in still microscope images. One appealing aspect of budding yeast is its asymmetrical reproduction mode that may be used to predict cell-stage from the cell contour [4,69,115] (see figure .2). In principle, this allows the extraction of a quantitative estimate of the cell-stage from cell shape. Using such estimates, I propose to generate time-profiles of image measures made on identified cells. To do so, I need to characterize a model of the cell shape, and extract a robust measure for the cell-stage. For robustness, methods need to handle the varying poses and shapes cells have, but also artefacts in images and misidentified objects. Next, in order to quantify variability of protein expression within and between individual, I propose to define a probabilistic model that would be capable of modeling features of the protein expression, such as differentiating cell-to-cell variability and cell-stage specific cell-to-cell variability.

Hence, this thesis is divided into three specific tasks:

I Develop an image analysis pipeline for high-throughput microscope images of yeast cells including algorithms for (i) segmentation, (ii) cell-finding with confidence estimates, (iii) cell size and stage estimation, which take advantage of prior knowledge about yeast cell shape and growth

II Develop statistical methods to extract and perform time-series analysis of protein expression measurements from temporally ordered cells

III Develop statistical methods to analyze variation in protein expression within cell populations

Part I

Accurate Recognition of Budding Yeast Morphology

Material from: Handfield et al. 2013 [66]

Within Sections: 1.2.5, 2.1, 2.2.*, 2.3

Figures: I.15, I.16, I.17

Overview

Recognition of shapes within digital images or movies has been studied extensively by the computer vision field. One common task many developed methods share is the 'image segmentation', which is to partition the images into a number of regions, so that downstream image analysis methods may be performed on the detected areas, as opposed to the large amount of pixels an image or a movie contains. The detection of cell shape can be performed using image segmentation directly, but the presence of large cell clumps makes individual cell contours harder to resolve.

The morphology of yeast can be characterized either by an ellipse or by a pair of ellipses when the yeast is budding. I utilize this prior knowledge for the recognition of cell contour that accounts for the occurrence of cell clumps in images. I show that this specialized shape recognition can account for misidentified objects and artefacts found in images, which is critical for cell-stage estimation from cell morphology. This chapter describes and compares several methods for their accuracy in the identification of yeast cell morphology. Results motivate the use of prior morphological knowledge to improve cell identification accuracy (Section 1), and to detect inaccuracies of automated cell segmentation methods (Section 2).

0 Background: High-throughput Computational Methods for Cell Morphology Recognition

0.1 Object Identification

All previous studies of automated image analysis have divided the task into subproblems, or presented a complex pipeline of procedures for the particular needs of presented datasets or of proposed hypotheses. Often, the first task is to identify objects of interest from the rest of the acquired data. Identifying objects is desirable if they are scarce in the images (such as neuron dendrites [133]) or if the object instances are to be registered as independent entities (allowing Single-cell measurements). Some classification schemes do not need this done (see 0.1.2 Image Level Features). This image segmentation procedure involves partitioning images into regions belonging to distinct objects, using some knowledge that a priori defines the nature of objects. Hence, the procedure used depends on the nature of the object of interest and the imaging technology used.

***Dendrites and Vascular Systems*:** Excitatory synapses in the cerebral cortex possess a complex morphology, which determines the strength, stability and function of its connections [133]. Hence, characterizing their structures is of profound biological significance. Many algorithms were developed specifically for the task or registering dendrite layouts [3,133,173] and vascular structures [144,160]. Tracing algorithms are fast algorithms that specifically find pairs of anti-parallel edges, which arise for near-linear segments, and use the significant ones as seeds to spawn the linear structures by recursively moving along the edges [144]. Another class of algorithms is skeletonization based, which require segmented images (typically with threshold segmentation). They then find critical points in the inferred gradient of the segmentation [173]. Finally, these points are joined by finding the best set of lines linking them. Skeletons of certain types of cells (HeLa [109]) were also used to generate image features for their classification. While HeLa cells do not possess dendrites, their shape is typically non-circular hence its skeleton topology may be used to characterize its shape.

Bacteria and Fungi: Singled-celled bacteria and fungi are often studied using automated microscopy. In designed experimental frameworks, extra information is usually acquired specifically to help the cell segmentation task. For example, a brightfield image of the cells may be used for segmentation. In these images, variations in pixel intensity are explained by light diffraction, typically caused by cell membranes. Cells may be identified by setting a threshold characterizing membrane [19, 90], but Kvarnstrom et al. [41] report that adaptive thresholding is required as the illumination of the cell is not uniform and depends on the cell density. An alternative is the watershed transform [16], which partitions the image independently of the intensity scale. It involves recursively grouping regions of low-gradients. This defines a hierarchical partitioning of the image. Deciding on the final number of partitions is not trivial. Marco et al. [41] proposed a criterion that evaluates the area fitness to an ellipse. This approach allowed the detection of watershed partitions of the background pixels, whose shapes are mainly defined by noise. If a nuclear marker was also present in the images, the problem may be constrained so identified regions have exactly one nucleus [29], because non-dividing cells are expected to have only one nucleus. The watershed transform was also used on darkfield imaging of a tagged protein, which is empirically found to effectively stain the whole cell, but the difference in foreground to background intensity enables to segment the image using adaptive thresholds as well [65].

0.2 Cell Modeling

Cell Cycle determination: The cell cycle stage is known to strongly determine the expression of several proteins that are responsible for critical steps is cell division, whose failure would result in an abnormal

amount of chromatin, for example [27]. Hence, recovering the relationship between protein expression and cell-stage may hint at the protein's function. It is possible to force the cell into a specific cell-stage by inducing an amino acid starvation [4] or by adding a compound that causes cell growth arrest [69]. When the chemical is removed, all the cells begin growing at the same time, and their cell cycles are referred to as "synchronized"'. Using this approach, Niemisto et al. [115] imaged cells stained with a fluorescent nuclear marker. The amount of DNA was measured at several time points following the cell cycle synchronization. This abundance was used to define a cycle-cell phase predictor.

***Shape models*:** Much effort was made to characterize the shape of yeast cells, since these cells use budding as a mean of reproduction, which generates distinctive cell shapes throughout their cell cycle. *Saccharomyces cerevisiae* is a model organism in molecular biology, and databases exist specifically characterizing its genome [35] and morphology [138]. Interestingly, Niemisto et al. [115] also automatically identify the bud neck (the region of a yeast cell connecting the daughter cell to its mother) from cell contours. Their automated method allowed the quantification of the time dependence of the fraction of buds found in cell populations following the cell cycle arrest. Calvert et al. [25] combined the cell-shape-based approach with the nuclear-staining-based approach to infer cell cycle from both the DNA amount (as above) and major axis of budded cell objects. In the most comprehensive analysis of yeast cell shape to date, Ohya et al. [117] analyzed the shape of yeast cell when one of its genes was experimentally deleted. For each deletion mutant, they extracted morphological measures from ellipses best fitting each mother and bud object, as well the position of the expression of actin cytoskeleton and nuclear DNA (available from simultaneous staining) with respect to the ellipse parameters. They report classes of phenotypes, defined as extreme morphological features, were observed in mutants whose deleted protein were more likely to belong to a biological function class. Some of these observations were reported to be specific to a particular cell-stage, or to the presence of absence of a bud. They proposed this framework for function prediction of proteins of unknown function.

1 Cell Contour Segmentation

In this first part, I need to identify the cells in images. For our analysis, it will be important to capture the precise boundaries of each object, as the protein expression will be interpreted as belonging to specific cells (Single-cell measurements), some of which are assumed to be independently occurring objects. As discussed above, there already exists methods for this task, so I here propose an approach designed for budding yeast, and compare the performance to other methods.

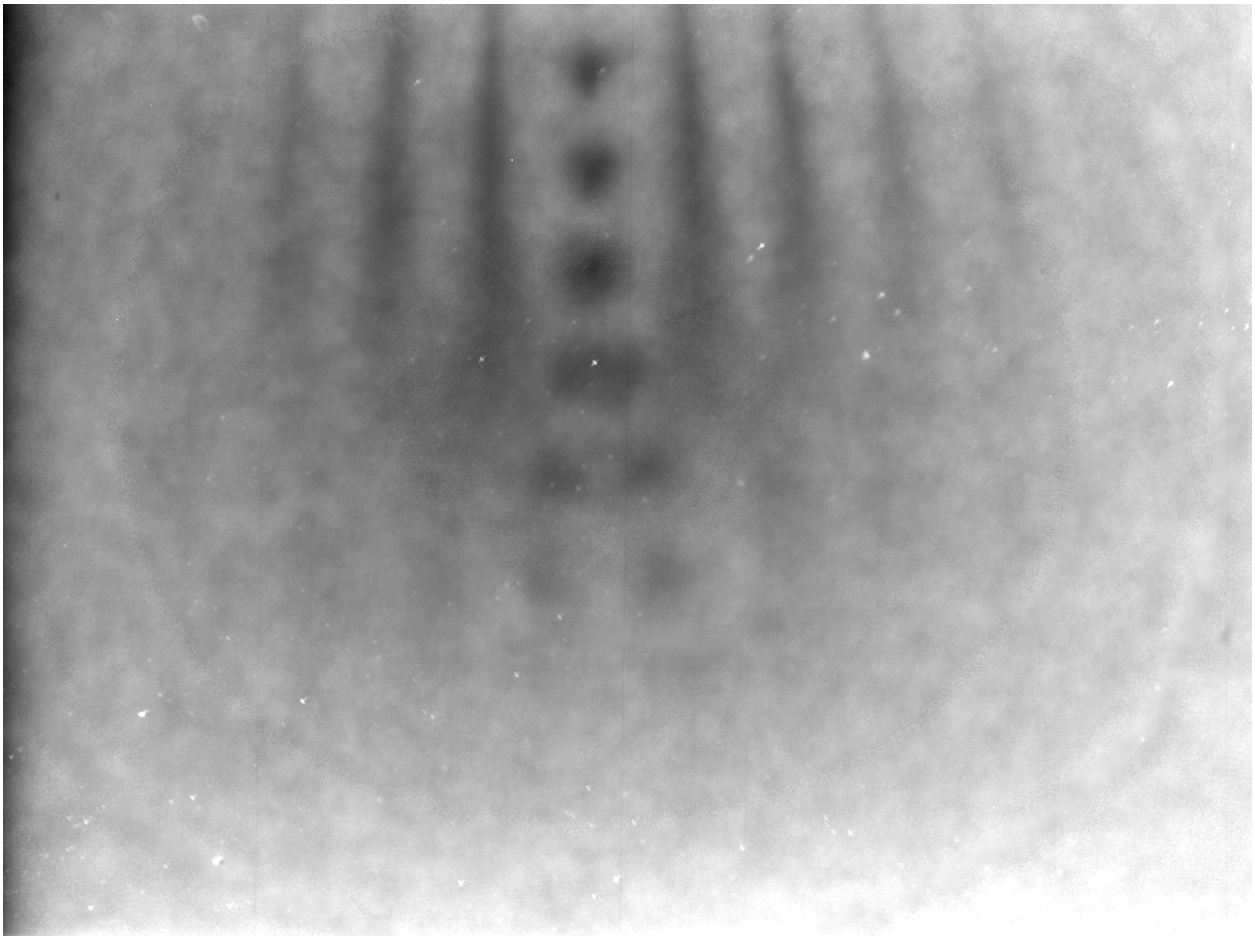


Figure I.1: **Average of the intensity of 4114 images for the RFP channel.** Averaging images shows an uneven background intensity level. This shows that dust, faulty CCD pixels, diffraction patterns, light orientation and the post-processing of the image (vertical lines) by the microscope all contribute in a non-trivial fashion to the background intensity.

1.1 Segmentation

I first need to define the portions of the image that are cells (foreground), so that remains is the background. The experimental paradigm that generated the images dictates what information can be used to separate cells from background. In the image collection that I first analyzed, all the cells are producing a red fluorescent protein (mCherry) that has introduced in yeast strains using synthetic genetic array (SGA [159]). Its expression stains the whole cell, though it appears to be significantly dimmer in yeast vacuoles. Nevertheless, regions of the images without cells display low intensities, so that we may define the boundary of cells from this signal alone.

1.1.1 Image Correction

First, we observe that the luminosity is uneven across the image (Figure I.1). Since I want the segmentation to identify objects independently of their position, such a factor needs to be corrected. A standard approach (available in ImageJ [1] and Cell Profiler [26]) is to fit a 2D polynomial to the image, and subtract it to the image. As we know that the background level is consistent across all images, we instead obtained the average of all the images, which is subtracted to each image. Since the cells all have a typical intensity, and they are distributed randomly in each image, their contribution to the background intensity is negligible compared to other sources.

1.1.2 Adaptive Threshold

One way to divide the foreground objects (cells) from the background is using a pixel intensity threshold that defines the minimal intensity that a pixel from a cell area (cell pixel) can display. Cell Profiler [26] provides several methods that allow the definition of the appropriate threshold. Each of the methods have some assumption for the target image, hence the choice of the method depends on the image type, the fraction of pixel in the foreground/background and the scale of intensities. One of them defines the threshold by fitting a mixture of normal distribution that describes the distribution of pixel intensities that are typical of cell and of background. Provided that the number of cells within an image corresponds to a prior belief, a mixture model allows us to define the intensity range that is indicative of cell boundary pixels. Another threshold method by Otsu [119] does not require such a prior belief, and finds the threshold that best separates the intensity into two classes using a criterion defined on their resulting variance.

1.1.3 Pseudo 2D Hidden Markov Model

One drawback we observe for the threshold methods is that punctate noise in often captured. Such noise can be easily filtered out based on its size, but its occurrence in proximity of cells may affect the contour. We sought to limit its contribution by using a Hidden-Markov Model (HMM, [20]) in order to better define the contour in a more robust manner. HMMs are generative probabilistic models that are used to model simple correlation in sequential observations by positing the existence of an unobserved, so-called 'hidden' sequence of variables that determines the distribution under which the observations were generated. In our case, the pixel intensities are the observed variables and the hidden variables are 'cell', 'background' or 'artefact'. By applying this on rows of an image, punctate noise is corrected whenever bright background pixel are between by dim pixels. The same argument holds for dark vacuole pixels. Another advantage is that we may integrate into the search in the best background and foreground intensity, and update the prior on the fraction of cell pixel in the image from the transition probability, which is a concern for segmentation based on mixture of normal distributions whose performance relies on a provided density parameter for the number cells in images.

$Z_i \in \{``Bg``, ``Fg``, ``Ar``\}$

$$\ell_{Bg}(R_i) = \frac{1}{\sigma_{Bg}\sqrt{2\pi}} e^{\frac{-(R_i-\mu_{Bg})^2}{2\sigma_{Bg}^2}} \qquad \ell_{Fg}(R_i) = \frac{1}{\sigma_{Fg}\sqrt{2\pi}} e^{\frac{-(R_i-\mu_{Fg})^2}{2\sigma_{Fg}^2}} \qquad \ell_{Fg}(R_i) = K_{Ar}$$

In our case, we model the intensity of the red fluorescent marker that is found in each yeast cell staining the whole cell area. We define the probability for each pixel to be from a three components model: (i) cell, (ii) background or (iii) artefact. We model the intensity of foreground and background as normally distributed, and find the probability density value (a constant) that we expect to explain 0.1% of the pixels from a given image. This last probability distribution is similar to a uniform distribution, in that any intensity value has the same likelihood, but the domain of the distribution adapts to all observed intensity values, which allows any density for a finite number of intensities to be defined on a continuous range. It is used to explain outliers in red intensities, which systematically observed in images caused by defective CCD pixels. Maximizing the likelihood under the constraint that only 0.1% of the pixels are expected from artefacts does not yield a close form for K_{Ar} that gives maximum likelihood. Hence a numerical procedure is used to update that parameter within the EM updates (see Appendix 1).

In order to avoid assumptions about artefacts when we model the artefact pixel class, the HMM solely learns the transition probabilities from background to foreground. For efficiency, we use an independence assumption of the hidden state of each pixel so there is no loop in the latent variable dependency map, so that the we expect each hidden variable in linear time of the number of pixel using the forward-backward algorithm paradigm on the unrooted tree structure (see fig I.2). This approximation of the 2D HMM was previously named Pseudo2D HMM [20]. One important variation introduced is that the independence map changes from point to point. This is important as it alters what is the total likelihood that is to be maximized. In this case, we actually maximize the likelihood of every row and columns as independent objects, where every pixel from a row considers the inferred latent variable of its orthogonal column to be an observation. This variation makes each latent variable depend on all other latent variables based on the hamming distance between observed pixel coordinates.

We compare the performance of the Pseudo2D HMM and a simple mixture of Gaussians with prior cell fractions (Figure I.3). While uncommon, we observe that cells that display intensities that are lower than typical cell lose a fraction of their volume to the background. The use of a Pseudo2D HMM fills the dark area and typically links neighbouring cells. Still, mixture of Gaussians may successfully segment such cells by pre-processing the image. Cell Profiler [26] allows users to identify what is the convolution needed to

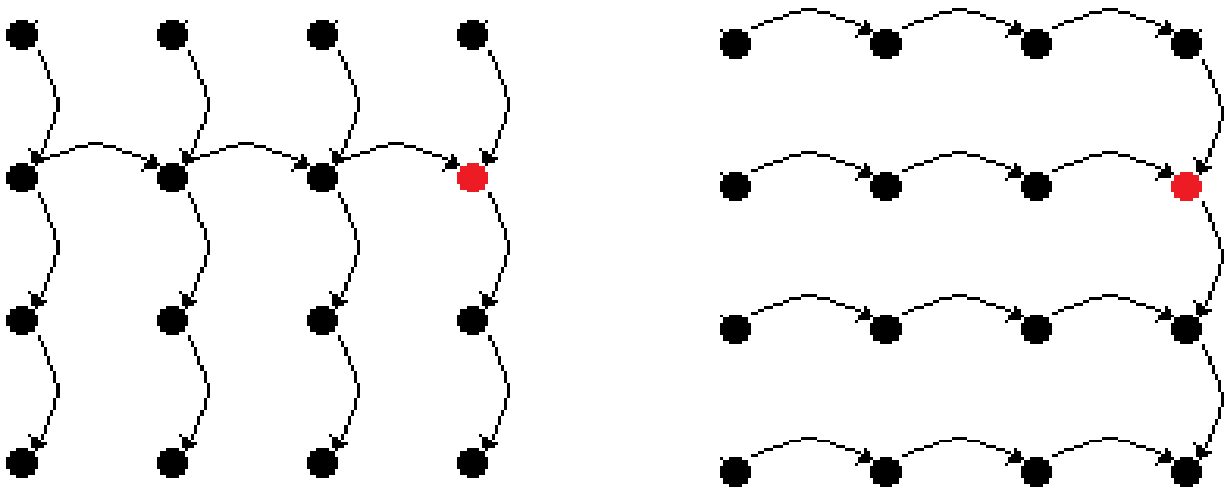


Figure I.2: **Hidden markov network.** The hidden states of each pixel are recovered under the assumption that two equally likely independence maps exist for the hidden state: for the first, all columns are independent, but for the pixels on the same row as the hidden variable of interest (red point), and similarly for the second, rows are independent but for one column of interest.

be applied that is adequately to fill such area, by simultaneously identifying individual cells that are to correspond to provided cell size range. This process effectively fills the space between cells objects and assigns the pixels to the foreground class. Hence, regardless of the segmentation method used, we will have to partition contiguous foreground areas an unknown number of independent cells.

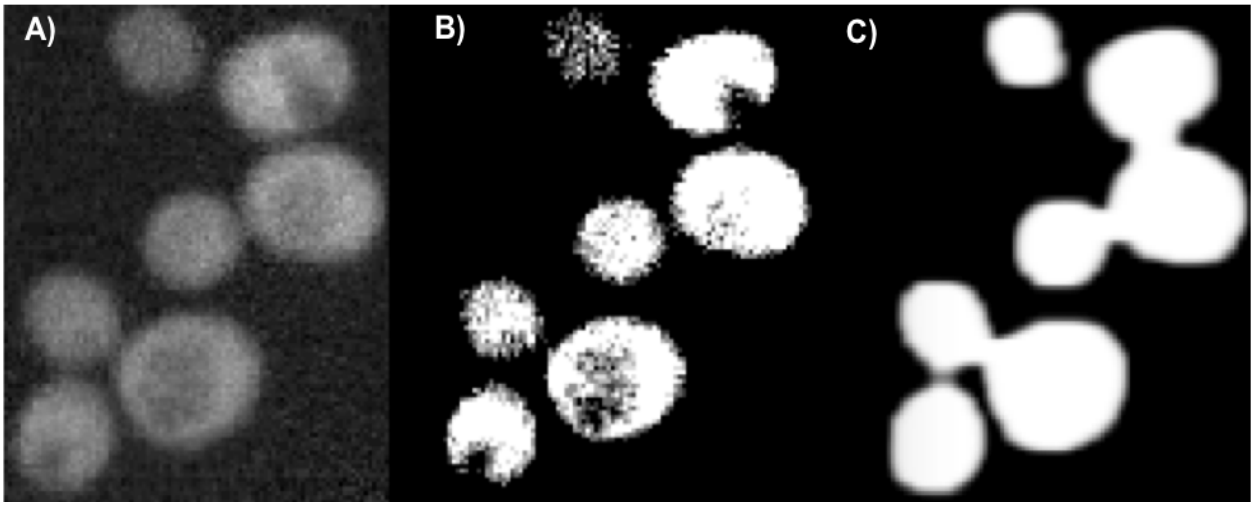


Figure I.3: **Comparison of segmentation methods.** a) Example of a group of cells. b) Threshold segmentation using a mixture of Gaussians. The variance in the mean intensity of RFP between cell assigns certain cells to the background class, and often recognizes vacuoles as background. c) Pseudo 2DHMM segmentation. Darker areas within cells are assigned to the foreground (cell) class instead. On the other hand, the neighbouring cells are typically fused, since transition probabilities learn the typical contiguity of the foreground pixel.

1.2 Shape Identification

Some previously used object recognition methods for cell identification, which use adaptive threshold [41], watershed transform [16], Voronoi segmentation [82] or 'shape' segmentation [165], often prune or merge foreground objects (cell areas) if they are grouped such that a large portion of their boundary contours are touching (cell clumps). It was previously noted that adaptive threshold tend to merge the area associated to neighboring (touching) cells [118]. Similarly, Voronoi segmentation has been noted to crop parts of cell areas, while seeded watershed transform often include areas from the image background into foreground objects [28].

First, I evaluated the ability of several methods to segment the one image collection I plan to analyze. Using Cell Profiler [26], I noted that the watershed transform and 'Shape' segmentation similarly identifies foreground objects whose contour does not always correspond to the contour of cells within image (Figure I.4). These methods tend to accurately identify the separation between foreground objects to the background area, but do not manage to properly partition cell clumps into foreground objects. For that reason, extra fluorescent markers are often used to help the segmentation and decide on the numbers of cell that are in clumps by identifying seeds for foreground objects. The quality of the seeds is critical to produce satisfactory results [29], which usually motivates the use of a nuclear marker. In this image collection, generated seeds are so that foreground objects are often split or merged I.4A).

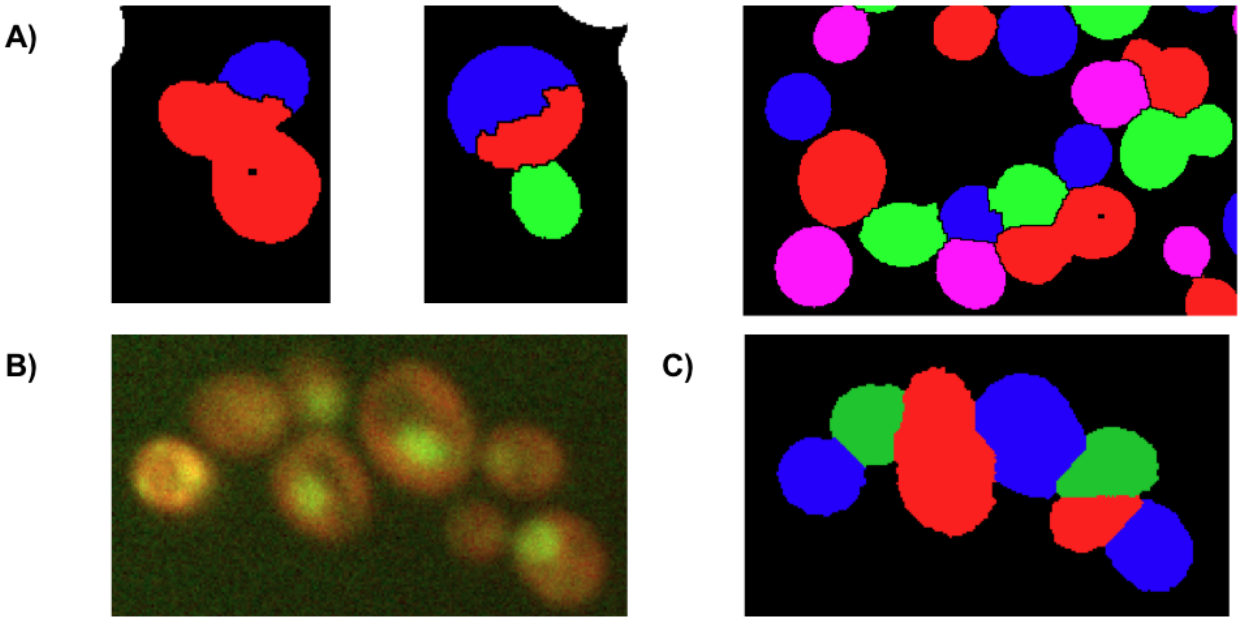


Figure I.4: **Examples of cell segmentation by previously proposed methods.** A) Three images containing cells are segmented by the Cell Profiler implementation of seeded watershed transformation. B) Raw microscope image for a large clump of cells C) Cell Profiler 'Shape' segmentation was applied on the image B). We observe that a number foreground objects are merged by the two segmentation methods.

While the nature of the fluorophore may differ between image collections, the cell morphology (shape) is more consistent. Identifying the number of cells in a clump has been previously performed by fitting to ellipsoidal area to bud and mother cells [41]). Cell identification procedures that identify cells solely from the morphology of their periphery, and use no other marker. Still, I note that the use of the 'Shape' segmentation found in Cell Profiler [26] merges foreground objects (Figure I.4C).

As I proposed to use cell morphology (shape of foreground object areas) to predict cell-stage, I considered that this issue needed to be addressed in order to produce accurate cell-stage estimates. Hence, I implemented many alternative approaches for this problem, each having particular level of success and time complexities. In this section, I describe the approaches I devised for partition cell clumps (grouped foreground objects) into individual cell areas. The evaluation and comparison of the methods are performed in the following section (Section 1.3). I will first I present a statistical circle model that is used to explain the foreground pixels from clump of cells. I also defined a heuristic method that uses the geometrical distances to rapidly find portion of the clump that are locally circular. The last method uses robust regression on ellipse parameters to find

a portion of the clump contour that is locally ellipsoidal.

1.2.1 Probabilistic Cell Model

The characterization of budded yeast cells is a more specific problem than the general segmentation of cells. Yeast cells and their buds appear to be near circular; their shapes are not as variable or complex as other imaged cells, such as HeLa cells or neuron dendrites. The importance of a reliable method that is specialized in the identification of yeast cell is motivated by the use of the bud size as a cell-stage indicator. Moreover, the correspondence of Mother cell to bud needs to be exempt of erroneously subpartitionned cells.

In order to define a constraint based segmentation, one approach was to assume that the image was generated by a probabilistic model, so that the parameters that give the highest likelihood to the image represent the center coordinates and width of the circles characterizing the cells. The characterization of a single circle parameters from coordinates that are sampled from the cell periphery is a well-established problem, and the biases and robustness of alternative methodology are well characterized. The difficulty is that the known results may no longer apply once portion of the periphery are not observed or when local deviations to a circle or ellipse systematically occur. More importantly, if an unknown number of ellipses explains the set of contour pixels that are to be explained, the problem becomes more difficult. For this reason, I first attempted to use the entirety of the foreground pixel, as opposed to estimate of the center coordinate on contour pixels, which are likely to vary significantly depending on the background noise level and proximity or absence of neighboring cells.

I devised two approaches to model circular shapes areas. The first one was to model the probability of a pixel to be within a cell based on prior function of its distance to a cell center of given radius. This model assumes that all the foreground pixels are close to a single circle center, and that the background pixel occurs beyond a distance corresponding to the circle radius. We obtain the likelihood for an image of foreground and background pixel, sampled from the posterior probability that has been characterized from an intensity image in the previous section using the pseudo2D-HMM. Let's define such an image of posterior probability as:

$$i[x_1, x_2] : N^2 \rightarrow [0, 1] \quad \text{where: } 0 \leq x_1 < width, 0 \leq x_2 < height \tag{1}$$

Then the likelihood of a sampled image can be characterized as:

$$P_F(i[x_1,x_2]|\vec{c},rad) = \prod_{\vec{x}}(F(\frac{||\vec{c}-\vec{x}||}{rad})^{i[\vec{x}]}(1-F(\frac{||\vec{c}-\vec{x}||}{rad}))^{(1-i[\vec{x}])}$$
$$\text{where } F(x): R^+ \rightarrow [0,1] \tag{2}$$

Uncovering the maximum likelihood parameters for the circle requires numerical updates. I use gradient ascent of the likelihood surface. The derivatives may be easily computed:

$$\begin{aligned}\frac{dLL(\vec{c},rad)}{drad} &= \frac{-||\vec{c}-\vec{x}||}{rad^2}F'(\frac{||\vec{c}-\vec{x}||}{rad})(\frac{i[\vec{x}]}{F(\frac{||\vec{c}-\vec{x}||}{rad})} - \frac{1-i[\vec{x}]}{1-F(\frac{||\vec{c}-\vec{x}||}{rad})}) \\ \frac{dLL(\vec{c},rad)}{d\vec{c}} &= \frac{d||\vec{c}-\vec{x}||}{d\vec{c}}\frac{1}{rad}F'(\frac{||\vec{c}-\vec{x}||}{rad})(\frac{i[\vec{x}]}{F(\frac{||\vec{c}-\vec{x}||}{rad})} - \frac{1-i[\vec{x}]}{1-F(\frac{||\vec{c}-\vec{x}||}{rad})})\end{aligned} \tag{3}$$

where F' is the derivative of F.

It remains to define the prior function of the probability of a pixel using its distance to a cell center and the circle radius. If we normalize the distance to the circle center by its radius, a perfect circle can be characterized by a step function that returns 1 or 0 if the ratio (fold distance) is smaller than 1 or not (respectively). Since we expect cell shapes to only be approximately circular, we such a discontinuous function is unlikely. Further, a continuous function is required for uncovering the maximum likelihood parameters; I use as prior function the sigmoid function that as centered about 1 and defined with a prior shape parameter $\beta = 10$:

$$F(x) = \frac{1}{1+e^{\beta(x-1)}} \quad \text{so that:} \quad F'(x) = \frac{\beta e^{\beta(x-1)}}{(1+e^{\beta(x-1)})^2} \tag{4}$$

where 'x' is the ratio of distance to circle center to circle radius (fold distance). This approach performs well for cases were the segmentation identified lone unbudded cells. As it stands, there is no gain in evaluating such an approach, the task of fitting single circle is better performed and better understood using alternative methods. As such, I next extend this framework to partition cells that could not be obviously separated by a foreground/background segmentation of the image.

1.2.2 Multi-cell Probabilistic Model

Aggregates of cells are found within images, and we want to identify individual cell objects. First, we partition the image into groups of contiguous foreground pixels, which will treated independently so to reduce the complexity of the partitioning task. The basic idea is to extend the function F to capture an arbitrary number of fold distances to predict the probability that a pixel is in some cell, without explicitly partitioning foreground pixel. If a clump of cells is modeled under the assumption that there are 3 circles explaining the

shape of the foreground pixels, the function F would to be redefined to be:

$$F(x_0, x_1, x_2) = \frac{e^{\beta(x_0-1)}+e^{\beta(x_1-1)}+e^{\beta(x_2-1)}}{(1+e^{\beta(x_0-1)})*(1+e^{\beta(x_1-1)})*(1+e^{\beta(x_2-1)})}$$
$$\text{So that:} \quad \frac{dF(\vec{x})}{dx_j} = \beta e^{\beta(x_j-1)} \cdot \frac{1-\sum_{i\neq j} e^{\beta(x_i-1)}}{(1+e^{\beta(x_j-1)})*\prod_i(1+e^{\beta(x_i-1)})} \tag{5}$$

The above function is a generalization sigmoid function; this function always returns a value between 0 and 1, and if any one fold distance x_i becomes arbitrarily large, in the limit it becomes equivalent to the same function with one less parameter. This prior function defines the probability of each pixel to belong to a specific cell or to a subset of the cells by removing terms from the numerator (the defined probabilities obey bayes rule). Let's define $Z_{\vec{x}} \in \{0, 1..., n\}$ to be the ownership state of a pixel, where the state 0 indicates that a pixel does not belong to a cell. Then:

$$P(Z_{\vec{x}} = j | \vec{c}_{\{1,...n\}}, rad_{\{1,...n\}}) = \frac{e^{\beta\left(\frac{||\vec{c_j}-\vec{x}||}{rad_j}-1\right)}}{\prod_i(1+e^{\beta\left(\frac{||\vec{c_i}-\vec{x}||}{rad_i}-1\right)})} \tag{6}$$

Hence, the above formulation characterizes the probability of each pixel to belong to any cell, so the final image segmentation can either have the form of most probable values for the hidden states 'Z', or the defined posterior probability for the hidden state may be explicitly used for further characterization of the cell shape. Hence, the partitioning of cell clumps is resolved by uncovering the maximum likelihood center and radii, as opposed to previous methods that typically model and detect cell boundaries.

The number of circular objects that are to be found in each cell clump is unknown; this proves to be a challenging aspect for the problem when number of cells increases. Previous methods are capble of uncovering the ellipse coordinates for pairs of adjacent cell accurately [138], but any clump of more than two circular areas are treated as artefacts or "bad data", since the accuracy of the ellipse fit may be impaired. For this problem, I attempted to define a method that could partition cell groups of arbitrary sizes. One major issue is that the parameter space for the maximization of the likelihood increases with the number of cells, so that obtaining the global maximum likelihood parameter from defining a prior distribution of number of cell requires an extensive search. Given the scale of the data to analyze, it is difficult to guarantee maximum likelihood parameters were obtained in a time efficient manner; therefore a greedy search procedure was devised.

In order to find the number of cells in a clump, new cell centers are iteratively added within the search procedure for the maximum likelihood parameters. The initial guess for the first circle center is obtained by the center of mass of the whole shape. Then, a fixed number (20) of updates on the circle parameters

using gradient ascent is performed. Initial guesses for subsequent circles use the coordinates of foreground pixel that are currently poorly explained by the current parameters. I first compute the center of mass of all pixels that are not within any circle. The center is then updated to match the closest pixel to that center that is not within any circle. Finally, the center coordinate is updated to match the mass center of pixels that are not within any circle weighted based on the proximity of the current center guess.

$$\vec{c_0} = \frac{\sum\limits_{\vec{x}\in S} \vec{x}}{|S|} \quad \vec{c_i} = \frac{\sum\limits_{\vec{x}\in S} \vec{x} \frac{10}{10+||c\vec{m}_i - \vec{x}||^2} \prod\limits_{j=0}^{i-1} I(\frac{||\vec{c_j} - \vec{x}||}{rad_j} > 1)}{\sum\limits_{\vec{x}\in S} \frac{10}{10+||c\vec{m}_i - \vec{x}||^2} \prod\limits_{j=0}^{i-1} I(\frac{||\vec{c_j} - \vec{x}||}{rad_j} > 1)} \quad \text{where} \quad c\vec{m}_i = \frac{\sum\limits_{\vec{x}\in S} \vec{x} \prod\limits_{j=0}^{i-1} I(\frac{||\vec{c_j} - \vec{x}||}{rad_j} > 1)}{\sum\limits_{\vec{x}\in S} \prod\limits_{j=0}^{i-1} I(\frac{||\vec{c_j} - \vec{x}||}{rad_j} > 1)} \tag{7}$$

Iteratively, cell centers are added and parameters of all cell centers are updated for a fixed number of steps. The procedure is terminated once either the number of unexplained pixels is zero, or that the parameters of a circle decreased its radius below what can be described as the smallest cell we reasonably expect to detect (5 pixels in width). The maximum likelihood parameters encountered in the search are retained for partitioning the clumps. Since this model allows the parameterization of intersecting circles, I unexpectedly observed that multiple circle parameters often converged to fit the same cell in the clump. In such event, any circle that contained the center of a larger circle was filtered out.

This approach worked well in small clumps of cells, but it did not yield accurate coordinates for elongated cells and in the event that the number of identified cells is incorrect (Figure I.5). Despite such mistakes, we observe that is capable of uncovering the center coordinates for large clumps of cells. On the other hand, having erroneous additional circles treated as mother or bud objects could significantly perturb downstream analyses. For this reason, another probabilistic model was considered.

clump size	1 cell	2 cells	4 cells	8 cells
running time	<1s	2s	25s	240s

Table I.1: **Running time for cell clump partition.** Time required to resolve a single clump of cells based on the clump size.

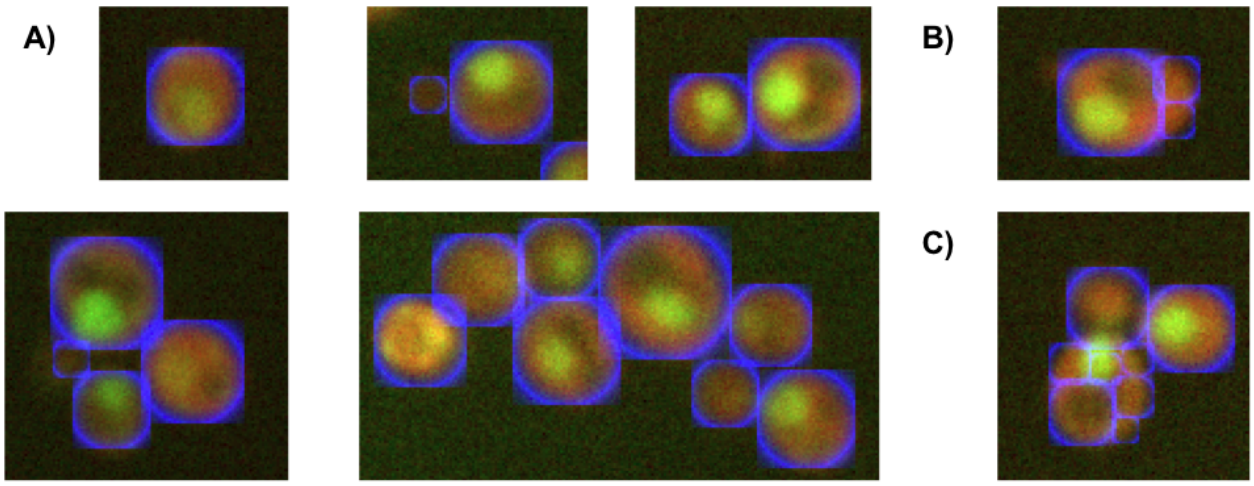


Figure I.5: **Cell segmentation using probabilistic circle model.** A) Example of maximum likelihood circle coordinates. The red and green color represents the abundance of RFP and GFP tagged protein respectively. The blue circles were drawn atop of the microscope image, using the circle coordinates that were fitted by the probabilistic circle model. B) Ellipsoidal cells causes this approach to fail identifying the correct cell number and their circle parameters C) Compact cell groups also are misidentified.

1.2.3 Geometric Probabilistic Model

One correction considered to this model is to modify the distance function to use the distance to an ellipse contour, as opposed currently used to the distance to a circle periphery. Interestingly, such a modification prevents the detection of early buds: small buds and mother cell are fitted by a single ellipse, and I could not devise appropriate initial guesses for the iterative circle addition that would resolve the bud position. One issue with the previous method is that it does not register that the contour of the shape should intersect with the parameterized circles, the circles instead are used to cover the whole clump and gaps between circles have little impact on the likelihood (Figure I.5).

In a segmented image, the knowledge of how close a pixel is away from any background pixel is useful for the characterization for the found objects, as this measure would be available regardless of the nature of the segmented objects. In my case, objects are agglomerates of cells that need to be partitioned into single cells, which are assumed to have a circular shape. Because this distance has an expected value for each point from a theoretical circle, we may use it to uncover portions of the foreground that appear to be locally circular (Figure I.6A).

Given an image for which we know the probability of each pixel to be from the background, we want

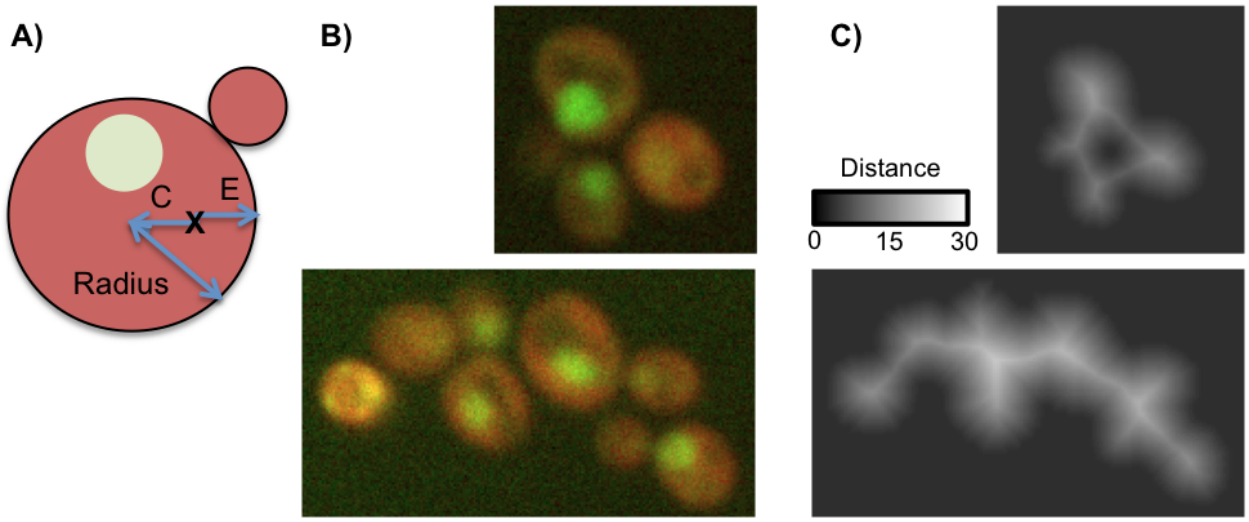


Figure I.6: **Geometrical distance to background.** A) For any point in a circle, the sum of the distance to the circle center'C' and the shortest distance to a point on the edge of the circle 'E' is equal to circle radius. B) Example of raw microscope images of RFP and GFP-tagged proteins. C) Calculation of the 'Distance to Edge' map: The brighter points are so that they are farther from any contour pixel of the cell clumps.

to define a map of geometric distance to background for each foreground pixel. We estimate this quantity using an iterative motion on the image grid (which includes diagonals and knight moves), where transitions from a point deterministically select the neighbour through which the shortest path to background is expected. We then compute the expected path length under the assumption that pixels reached along paths have background/foreground state transitions described by a HMM with the parameters inferred from the segmentation. The transition probabilities for diagonal and knight moves are obtained by exponentiation of the transition matrix by the distance between the two points. Since it is enforced that transitions are only allowed from point of higher expected distance to lower ones, distances can be computed directly by dynamic programming, in linear time of the number of pixels in the image.

To do so, I first characterize 'Distance to Edge' measure, which is a quantity defined for each foreground pixel, that represents the expected minimum physical distance to a background pixel. This quantity is computed by having an agent choose the best path from a foreground pixel to background pixel, where allowed transitions also include diagonal and knight moves on the image grid of pixels. I use an HMM with parameters recovered by the segmentation to determine the expected path length needed to reach the first background pixel. I first assume that this distance value is bounded above for any foreground pixel by the 'Distance to Edge' of the first pixel on a hypothetical infinite linear segment of identical pixels. The purpose of this assumption is upperbound has an initial guess for the whole 'Distance to Edge' map. Then,

the distances and optimal paths may be recovered in linear time by dynamic programming, by iteratively processing the pixel with lowest 'Distance to Edge' and to update its neighbour 'Distance to Edge' values (see algorithm below). Doing so would be equivalent to have an agent will select the best transition as all smaller 'Distance to Edge' are already known on the currently processed pixel.

The time complexity of the generation of the 'Distance to Edge' map is $O(n \cdot log(n))$, where n is the number of pixels in the image. This requires storing the coordinate of pixel and their associated '$DtoE$' value in a heap. Anytime a '$DtoE$' is being updated from having a neighbouring pixel processed, its associated node in the tree has to be moved according to the new '$DtoE$' value, which takes $O(log(n))$ time in the worst case. Since the number of potential updates is bounded by the number of neighbouring pixel, and that pixels are only processed once, the number of insertion, deletion and updates in the heap can be bounded by the number of pixel in the image. Since these operations are all $O(log(n))$ time, then the whole process is $O(n \cdot log(n))$.

Now that we have a geometrical distance to the background, we want to define a model that utilize the fact that the sum for any pixel in a circle of its distance to the circle center and distance to circle periphery, which is equal to the radius of the circle (Figure I.6a). We use a normal distribution to model the deviation to theoretical expectation for the measured "Distance to Edge" given a set of circle coordinates. The marginal probability for each pixel to belong the i^{th} circle now needs to be explicitly evaluated. Instead of a gradient motion on the likelihood function, we use soft Expectation Maximization (soft-EM) to update the circle parameterization (Eq 8). The maximization step requires iteratively updating the center coordinates, but the radius and variance have closed forms for their maximum likelihood parameterization when the latent variable '$Z_{\vec{x}}$' and circle coordinates are fixed.

$$\begin{aligned} rad_j &= \frac{\sum_{\vec{x} \in S} P(Z_{\vec{x}}=j)(D_{\vec{x}} + ||\vec{x} - \vec{c_j}||)}{\sum_{\vec{x} \in S} P(Z_{\vec{x}}=j)} \\ \sigma_j &= \frac{\sum_{\vec{x} \in S} P(Z_{\vec{x}}=j)(D_{\vec{x}} + ||\vec{x} - \vec{c_j}||)^2}{\sum_{\vec{x} \in S} P(Z_{\vec{x}}=j)} - rad_j^2 \quad \text{where } D_{\vec{x}} \text{ is the 'Distance to Edge'} \end{aligned} \tag{8}$$

In addition, we model RFP abundance in cells using another normal distribution (Eq 8). Cells have relatively uniform RFP intensity over their area, but the mean intensity varies from cell to cell; such information could be also utilized to identify the right number of cell objects.

$$\begin{gathered} L(R_{\vec{x}}, D_{\vec{x}} | Z_{\vec{x}} = j) = \frac{1}{\sigma'_j \sqrt{2\pi}} e^{\frac{-(R_{\vec{x}} - \mu_j)^2}{2\sigma'^2_j}} \cdot \frac{1}{\sigma_j \sqrt{2\pi}} e^{\frac{-(D_{\vec{x}} + ||\vec{x} - \vec{c_j}|| - r_j)^2}{2\sigma_j^2}} \\ \text{where } R_{\vec{x}} \text{ is the RFP abundance} \end{gathered} \tag{9}$$

As before, the parameter space for circle coordinates is large, so that the same greedy search procedure

Input Image

RFP = 1 $P(Z_{0,0} = 1 \mid RFP = 1) = 10^{-8}$	RFP = 11 $P(Z_{1,0} = 1 \mid RFP = 1) = 0.5$
RFP = 5 $P(Z_{0,1} = 1 \mid RFP = 1) = 0.01$	RFP = 20 $P(Z_{1,1} = 1 \mid RFP = 1) = 1 - 10^{-8}$

Initial guess

RFP = 1 DtoE = 1.001 $(= 1 + \frac{p \cdot P(Z=1 \mid RFP=1)}{(1-p) \cdot (1-P(Z=1 \mid RFP=1))})$	RFP = 11 DtoE = 5.76
RFP = 5 DtoE = 1.23 $(= 1 + \frac{p \cdot P(Z=1 \mid RFP=5)}{(1-p) \cdot (1-P(Z=1 \mid RFP=5))})$	RFP = 20 DtoE = $9 \cdot 10^8$

First step

RFP = 1 DtoE = 1.001	RFP = 11 DtoE = $1 + P(Z_{0,0} = 1 \mid Z_{1,0} = 1, RFP_{0,0} = 1) \cdot 1.001$
RFP = 5 DtoE = 1.23	RFP = 20 DtoE = $\sqrt{2} + P(Z_{0,0} = 1 \mid Z_{1,1} = 1, RFP_{0,0} = 1) \cdot 1.001$

Second step

RFP = 1 DtoE = 1.001	RFP = 11 DtoE = $1 + P(Z_{0,0} = 1 \mid Z_{1,0} = 1, RFP_{0,0} = 1) \cdot 1.001$
RFP = 5 DtoE = 1.23	RFP = 20 DtoE = $1 + P(Z_{0,1} = 1 \mid Z_{1,1} = 1, RFP_{0,1} = 5) \cdot 1.23$

Final step

RFP = 1 DtoE = $10^{-8} \cdot 1.001$	RFP = 11 DtoE = $0.5 \cdot (1 + P(Z_{0,0} = 1 \mid Z_{1,0} = 1, RFP_{0,0} = 1) \cdot 1.001)$
RFP = 5 DtoE = $0.01 \cdot 1.23$	RFP = 20 DtoE = $(1 - 10^{-8}) \cdot (1 + P(Z_{0,1} = 1 \mid Z_{1,1} = 1, RFP_{0,1} = 5) \cdot 1.23)$

Table I.2: **Calculation of 'Distance to Edge'.** Example of a possible 'Distance to Edge' calculation run on a 2x2 image. Z is 0 for background pixels, and 1 for foreground pixels. The transition probability from a foreground pixel to another foreground pixel (p) was learned by a HMM, and the transition probability for a diagonal transition (and knight moves) are directly inferred by transforming the transition matrix (Since our HMM is constrained to learn time reversible transition probabilities, we may compute that probability from the learned transition probability matrix trained on row and columns). First, each pixel is initialized to the assumed number of transitions needed to into an identical pixel to reach a background, where the state of the first pixel is assumed to be foreground so that $DtoE \geq 1$. The top left pixel is processed first because it has lowest DtoE. In doing so, two neighbours will have their agent change their optimal path so they go through our processed pixel, so their DtoE may be computed has we are guaranteed that the current DtoE will not be modified in the future. We then process our 2nd pixel, which will update the DtoE value of the bottom right pixel. Before the final step, the two pixels on the right will be process, but no DtoE values would be updated. The final DtoE accounts for the probability of each pixel to be background initially.

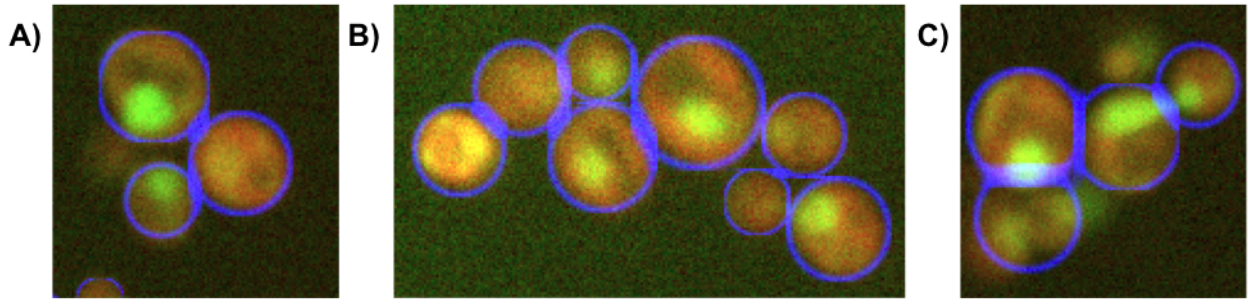


Figure I.7: **Cell segmentation using geometric distance.** Circle coordinates are drawn in blue. A) Example of a missed bud object that was previously found (Figure I.5) B) Properly segmented large clump. C) Objects are often missed, but that have little effect of the parameter of neighbouring identified objects, which was an issue for the previous method.

is used, that is, to incorporate additional circles within the soft-EM procedure. We observe that the circle coordinates properly adapt to the contours for the shape, and that large cells are not partitioned by an arbitrary number of circles (Figure I.7). On the other hand, small objects are often missed.

1.2.4 Heuristic Method

The previous method proved to be slow, and the greedy search may never fit all the objects in large clumps. If an iteratively added circle arises in the wrong location, many objects will be missed because the search cannot go back and correct this mistake. In order to detect small cells (early buds) in a more systematic fashion, one could provide a better initial guess for the number of cell in a clump, and their approximate coordinates. We know that the likelihood surface has many local maxima, so we are required to have reliable estimates for the cell center coordinates in order to robustly identify cells. Hence, I devised a fast heuristic to identify cell centers; it uses dynamic programming and identifies hundreds of cell centers in an image in half a second. This procedure performs independently on how clumped cells are and its running time is $O(n \cdot log(n))$ in terms of the image size.

A prerequisite for the use of the heuristic is a foreground/background segmented image. Further, the transition probabilities inferred by the pseudo 2D HMM are utilized. It would be possible to define the transition probability from a given segmented image, but as other segmentations typically do not assign a foreground posterior probabilities but binary separations, this approach may not be successful. To use this heuristic, we first need to compute the 'Distance to Edge' map, which can be performed in $O(n \cdot log(n))$ (Section 1.2.3). The way the heuristic processes the each pixel is an identical to framework previously defined for the 'Distance to Edge' calculation: a heap is used to store and maintain a certain ordering of the pixels, which are processed one by one in the order defined in the heap. Hence, the time complexity is identical.

The idea for the heuristic is that any pixel belonging to simple ellipsoidal shapes has a distance to the periphery of the shape that can be related to the distance to a potential centroid. The procedure allows any pixel to be centroid candidates and evaluates their fitness to be a circle centers using simple criteria. One important observation is that pixel small 'Distance to Edge' values are unlikely to be centroid candidates if the circle they explain contains pixels with higher 'Distance to Edge'. In such event, a larger circle of foreground pixels is known to intersect with such a candidate circle. The estimates of 'Distance to Edge' are often inaccurate because the background pixels are often missed between adjacent cells (Figure I.6C). For that reason, good centroid candidates are not necessary local maxima in the 'Distance to Edge' map, but the tips of the squeletonization of the shape. This problem is different than the detection of dendrites, as the detection of edges and the pairing process required to define dendrite axes is not designed for circular globular shapes [3].

In order to find candidate centroids, we will have every pixel render random paths toward presumptive cell centers. The transition probabilities are defined from pixel of lower 'Distance to Edge' to pixels of strictly higher 'Distance to Edge'. The probabilities are fully defined from the relative 'Distance to Edge' of neighbours and transition distance (as diagonal and knight moves are allowed). The transition with highest increment of 'Distance to Edge' when normalized by transition distance will have the highest probability. Transition probabilities are hence weighted the agreement of transition and the local gradient in the 'Distance to Edge' map, using the following formula:

$$P(S_{i+1} = \vec{k} | S_i = \vec{j}) = \frac{\left(\frac{D(\vec{k}) - D(\vec{j}))}{||\vec{k} - vecj||}\right)^2 \cdot I(D(\vec{k}) > D(\vec{j}))}{\sum_{\vec{l} \in N(\vec{j}))} \left(\frac{D(\vec{l}) - D(\vec{j}))}{||\vec{l} - veck||}\right)^2 \cdot I(D(\vec{l}) > D(\vec{j}))} \tag{10}$$

where $N(\vec{j})$ is the set of pixel coordinates neighbouring $\vec{j}$

For such a defined a stochastic motion of the coordinates, we expect that paths drawn from background to local maxima 'Distance to Edge' map have lengths that are similar to 'Distance to Edge' values at the local maximum from perfectly circular shapes (Figure I.8). In the event that a local maximum arises due to a failure to identify background pixels between 3 cells, the path to that local maximum will often be reached by first reaching the center of adjacent cells, which increases the path length. In order to capture this phenomenon, we evaluate the expected sum of the path length for a stochastic motion and the 'Distance to Edge' of the starting state, which may correspond to any foreground pixels that are sampled with equal probability. As it is defined, we may propagate the expected distance sum while building paths, by computing the expected sum marginalized on the probability that the current pixel is reached in the stochastic motion. For each pixel, we compute i) the probability that a stochastic path reaches this pixel, ii) the expected sum

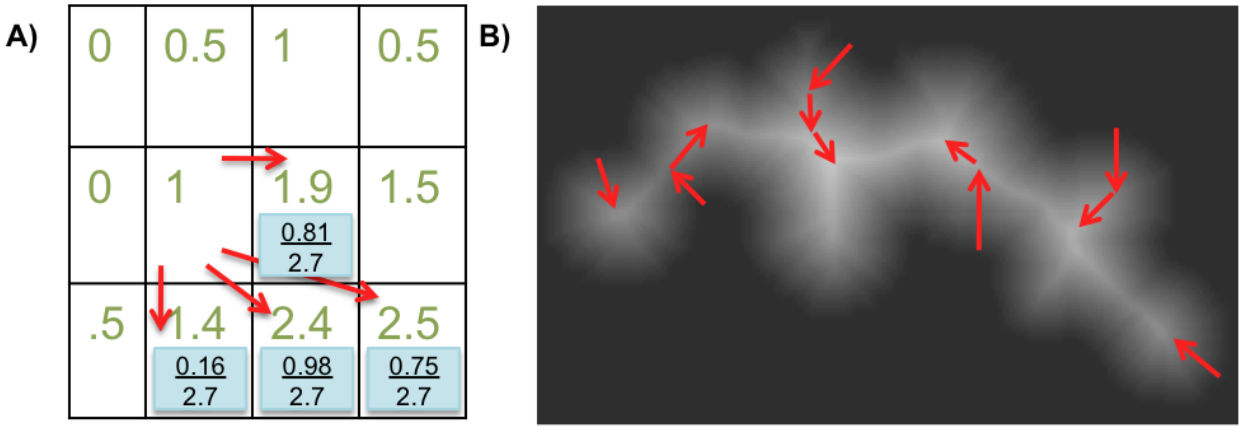


Figure I.8: **Stochastic motion on 'Distance to Edge' gradient.** A) Example of stochastic transition on an image. The 'Distance to Edge' is shown in green, and transition probabilities shown in blue boxes. B) Expected trajectories generated from the motion in the gradient 'Distance to Edge' map. Such paths should have the highest probability under the stochastic motion.

of 'Distance to Edge' and current path length for stochastic path reaching the pixel and iii) the variance in sum of 'Distance to Edge' and current path length. This is done by properly propagating the sufficient statistics along the path, by processing the pixels in increasing order of 'Distance to Edge'.

$$\text{Chebyshev's inequality:} \quad P(|X - \mu| > k) < \frac{Var(X)}{k^2} \tag{11}$$

Having the variance of the total path length allows us to define a bound for the probability that the total path length is larger than a constant, by using the Chebyshev inequality (Eq. 11). As such, we may bound the fraction of the paths reaching a centroid candidate with total path length smaller than 1.25 time the 'Distance to Edge' at the centroid, which is an arbitrary criterion that takes into account that the path length are not optimal due to the probabilistic transitions. Multiplying this fraction by the fraction of paths that go through the centroid candidate produces a value that is proportional the area of the shape that can reach the centroid candidate within the allowed total path length. We use this area criterion to identify the best centroid candidates.

Finally, we rank centroid candidates by the area criterion we evaluated. The best centroid is added to a list of circle coordinates, and its corresponding radius is assigned using the associated 'Distance to Edge' value. Iteratively, circle coordinates from the following centroid candidate are added if they are not intersecting with previously added circles. Candidate centroids with associated area whose size is below a threshold of 16

pixels are ignored, which is an arbitrary threshold that avoid to identify objects that are too small to be buds.

This heuristic method is extremely fast; when provided with images with background and foreground region identified, it produces 1.8Million circle centers and radii within 80 minutes, while using a single thread (375 circles per second). As such, I use these sets of circle coordinates as a starting guess for the parameter search that uses an EM search procedure (Section 1.2.2 & 1.2.3).

Finally, we compute 'Distance to Edge' under the cell-pixel ownership map, so that we may measure the confidence that subpartitions are cell instances, while there might be any background pixel that delimits the boundary between adjacent cells.

1.2.5 Robust Regression

We used robust regression [101] for matching ellipsoidal shapes to the contour of the segmented area. While robust regression is often used to fit shapes in images, this approach was not previously utilized on yeast shape to my knowledge. An ellipse is characterized to be the set of point for which the algebraic error $Err(\vec{x})$ [59] is zero:

$$Err(\vec{x}) = (\vec{x} - \vec{c})^T A(\vec{x} - \vec{c}) - r^2 \tag{12}$$

where $\vec{c}$ is coordinates of the ellipse centre, $\vec{x}$ are the coordinates of the contour points, and r is an additional parameter, proportional to the radius of a circle for a fixed matrix A.

The matrix A makes the set of points with zero algebraic error corresponds to a hyperbole or a line, and a superfluous scale parameter is observed in this parameterization. We therefore constrain the form of the matrix A:

$$A = \begin{bmatrix} \frac{1}{2} + \frac{\cos\theta}{6}\frac{1}{1+e^{\phi}} & \frac{\sin\theta}{6}\frac{1}{1+e^{\phi}} \\ \frac{\sin\theta}{6}\frac{1}{1+e^{\phi}} & \frac{1}{2} - \frac{\cos\theta}{6}\frac{1}{1+e^{\phi}} \end{bmatrix} \tag{13}$$

where θ is the angle corresponding to the orientation of the major axis and ϕ is a parameter that determines the eccentricity of the ellipse. This choice of this matrix to ensure that for any value for the set of 5 parameters (in equation 12) generates an ellipse with major axis length bounded above by twice the minor axis

length, as they are both determined by the eigenvalues of the matrix A and scale with the parameter 'r' [59].

Contour pixels are first identified by finding foreground pixels that are $\leq$5 pixels away from some background pixel (using the Edge Distance Map described above). Initial guesses for ellipses are generated by first fitting a circle to 3 randomly sampled contour points (that circle is unique). Initial guesses are rejected if the circle does not fit within the rectangle clamping the contour points, or if the center is a background pixel. The initial guess ellipse will be set to match width (diameter) and center of an accepted circle. A small eccentricity corresponding to $\phi = 0$ and a random angle θ (drawn uniformly from 0 to π) is used to define its remaining parameters.

If the set of contour pixels matches a single ellipse, we could directly update the ellipse coordinates by minimizing the sum of the algebraic error of all contour pixels. However, if the set of contour pixels is best explained by several ellipses, the sum of algebraic errors is likely to have local minima that are not close to any of the true ellipse parameters. Therefore, we use robust regression [71] and minimize the objective function:

$$\rho(Err) = \frac{\sum\limits_{\vec{x} \in C} Err(\vec{x})^2}{\sigma^2 + \sum\limits_{\vec{x} \in C} Err(\vec{x})^2} \tag{14}$$

where C is the collection of coordinates of contour pixels and σ is the expected error, which is chosen to be 5, matching the thickness of the contour. This effectively weights down the importance of contour points with large deviations to the current ellipse, so that the many local minima can correspond to actual ellipses.

Upon convergence, we discard ellipses that are not bounded by the clamping rectangle, or that have a background pixel at the center. Since a large number of local minima are expected, we generate about 10 fold more set of ellipse parameters than the number of expected ellipses (based on number of contour pixels) and select the ellipse with the best fit. Once we have identified the best ellipse, we remove all contour pixels that have an error smaller than σ, and find the next ellipse using the remaining contour pixels using the same procedure. Since some missed lone pixels may remain, we reject the ellipse and remove the corresponding pixels if the ellipse width is less than 3 pixels or if the number of removed contour pixels accounts for less than 10% of amount expected from the ellipse parameters and known contour width. This process is repeated iteratively, until no more contour pixels can be removed. The running time of the segmentation is linear in the number of pixels in images, and the running time of cell-finding is linear with the number

of randomly sampled circles for the initialization of geometric ellipse fit. On a single 2.83GHz Intel core, 98 seconds were required to analyze a single 1331x1017 image, which on average contained 82 cells and 31 artefacts (see Section 2.2).

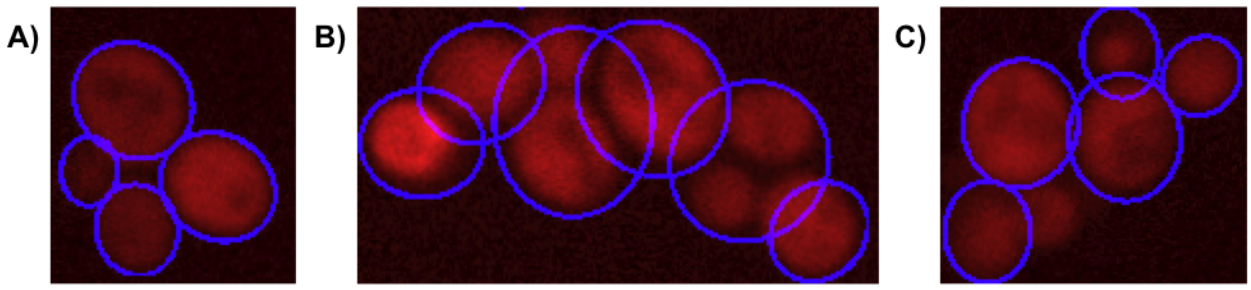


Figure I.9: **Robust ellipse fitting.** A) Ellipse coordinates inferred from robust regression of the contour pixels (drawn in blue). B) The additional freedom of ellipses allows unrelated contour pixels to be fitted by a single ellipse; areas may severely overlap. C) Example of a misidentified cell.

Alternatively, the initial guess may be generated using the set of circle guesses that the heuristic method produces. This only appeared to be marginally beneficial, as a large number of sample circles are still required, except that larger clumps that contain more than 20 cells can be reduced to smaller problems, by only considering contour pixels that are within the reach of a cell found by the heuristic. I will next show how the presented approaches compare to each other, and how to identify cells accurately, even when cells are highly clumped together.

1.3 Identification Performance

In this section, I will show that clumps of yeast cells in images can be partitioned into foreground objects that more accurately match manual identification, which is defined with drawn ellipses, by methods that fit ellipses to foreground object areas, as opposed to two other methods. For that task, a collection of images were manually characterized by Jibril Simmons. Ellipses were drawn atop of microscope images using a color code that was to encode for class labels for identified object types, which includes 'bud', 'mother', 'lone' and 'artefact'. We use this set of 4305 ellipse coordinates drawn on cells to compare the accuracy of different methods that measure object position and size.

1.3.1 Cell Profiler 'Shape' Segmentation

In order to compare the accuracy of the simple cell-finding methods described above with previously proposed methods for cell identification, we compared our results to cell identified using Cell Profiler [26]. This

software allows the user to define pipelines of image modification and annotation, which let the user select the appropriate approach and parameterization that is most effective for the type of microscopy images they wish to analyse. For background correction, we used the polynomial fit to the ensemble of images, and subtracted the resulting amount from each image. We identified the primary objects under Otsu global threshold method, and used the 'Shape' method for defining boundaries between objects and to distinguish the clumped objects. We chose this method because the Cell Profiler documentation suggests that it is proper for clumps of round objects. While many more parameters may alter the intended behaviour, the 'Shape' segmentation was described in by Wahlby et al. [165].

In addition, we need to select a method that is to discover the number of object per clump. In this case, we consider two alternatives: 'Intensity' and 'Shape'. In the first case, the number of objects is detected from intensity peaks; in the second case, it is detected from peaks in distance-transformed image.

While manually drawn shapes are guaranteed to be ellipsoidal, automated segmentation methods do not render ellipsoidal areas only. As such, certain shapes may not be fit with ellipses. Still, we want to quantify the agreement in shape for identified cells circle coordinates. We therefore define ellipse coordinates from computing 6 statistics of the pixel coordinates for the shape (Eq 15). This way, the center and ellipse parameter can be deterministically defined (see Appendix 2).

$$|S|, \overline{x} = \frac{\sum\limits_{\vec{x} \in S} x_0}{|S|}, \overline{y} = \frac{\sum\limits_{\vec{x} \in S} x_1}{|S|}, \overline{x^2} = \frac{\sum\limits_{\vec{x} \in S} x_0^2}{|S|}, \overline{y^2} = \frac{\sum\limits_{\vec{x} \in S} x_1^2}{|S|}, \overline{xy} = \frac{\sum\limits_{\vec{x} \in S} x_0 x_1}{|S|} \tag{15}$$

Equation 15:**Pixel coordinates statistics.** Summary of the size, position and spread of pixel coordinates from a 2 dimensional shape.

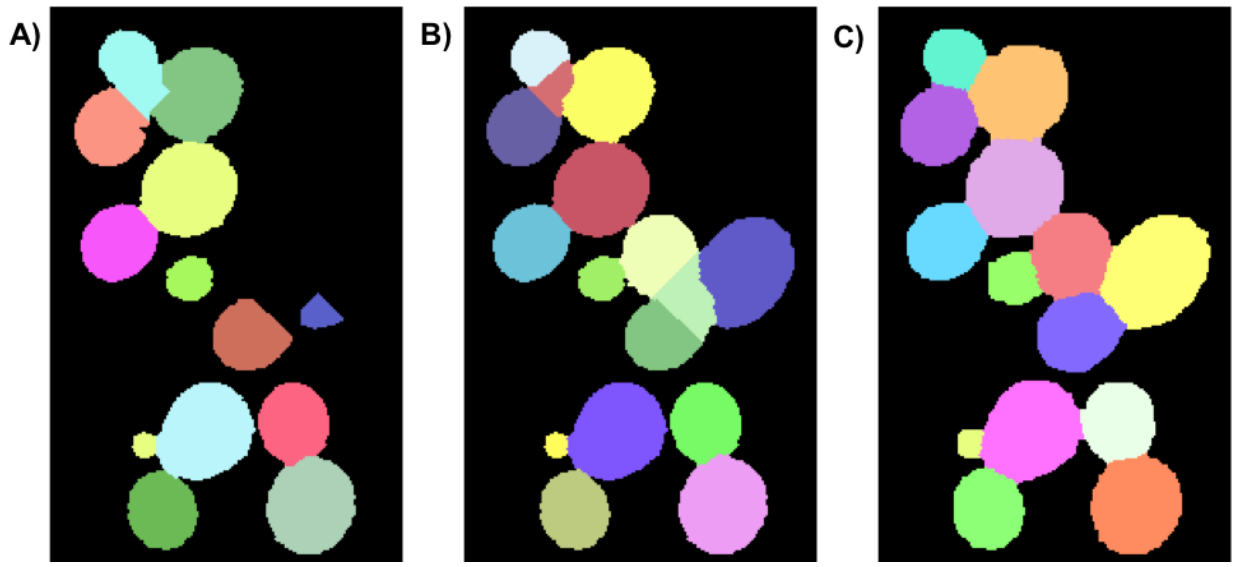


Figure I.10: **Comparison of segmentation methods on an image example.** A) Cell segmentation using Cell Profiler where the number of objects in a clump is found by using local maxima in the RFP intensity, and by then refining the contour by using the 'Shape' based segmentation. B) Cell segmentation using Cell Profiler, which used 'Shape' based segmentation to both define the contour and uncover the number of objects. Triad of cells are frequently mistaken by introducing an erroneous cell at the intersection of the 3 objects. C) Cell segmentation based on ellipse-fit and watershed transformation to define the cell contours.

1.3.2 Method Comparison

In order to evaluate the accuracy of cell identification methods, we obtain a test set of ellipse coordinates. 4305 ellipses have been manually recognized by drawing an ellipse a top of 68 microscope images. A color code for the drawn image was used to additionally define a 'cell type' label for each cell (used in Section 2.1). The ellipse coordinates were recovered from filling the inside of drawn shapes, and using the previously define coordinate statistics (Eq. 15). As such, we now report on the agreement for automated cell segmentation methods and this reference set of ellipse coordinates.

It should be first noted that the time complexity of the considered approaches and combination of approaches differ. For example, the multi-cell fitting approaches are the slowest as they require updating circle parameters in parallel, while methods such as robust regression are capable of detecting cell one by one. It should be noted that the choice of parameter for the searches can modulate by folds the running time of any processes. As such, theoretically faster algorithms were given parameters that increase their running time so they are comparable to slower processes. For the example of robust regression, the number of guess circle coordinates may be increased, so that the global best ellipse fit has a higher probability of being reached.

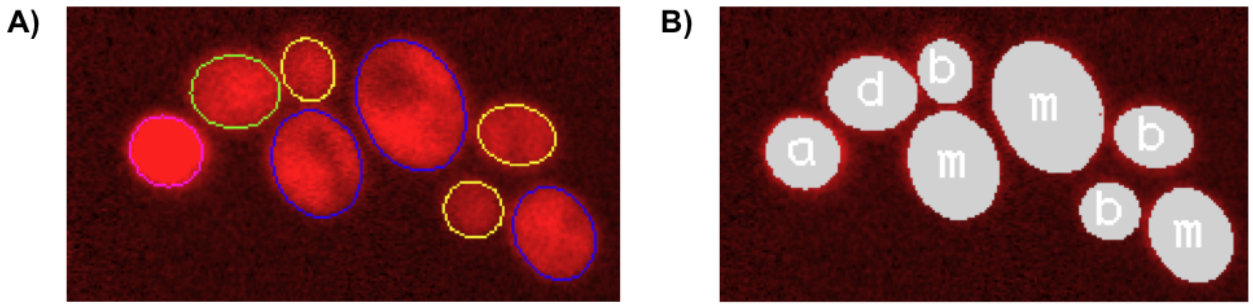


Figure I.11: **Manually identified cells.** A) Raw Microscope image with ellipses drawn using ImageJ [1]. The color code is used to identify the object type. To help the recovery of ellipse coordinates, drawn ellipses were required to never intersect; consequently, the size of the cells are typically underestimated. B) Area of ellipses drawn from coordinates recovered from drawn ellipses.

As such, one have to keep in mind the time cost of methods when comparing method accuracies, as it may be a constraining factor for certain approaches. For this project, the time component had little importance: all approaches managed to analyse the full collection of images within 3 days in the worst case. This was possible to do using a cluster computer that had more than 100 cores.

We want to quantify how effectively segmentation method estimate cell size and cell position; accurate cell identification is critical as bud size is to report for cell-stage estimate. Artefact and misidentified cells may limit the resolution at which we relate protein expression to cell-stage from accounting for high 'noise' levels. While running time requirements differ by folds (Table I.3), we still compare their accuracies, as selected parameterization was chosen to yield higher accuracies.

In order to evaluate the accuracy of circle coordinates, they first need to be paired to drawn ellipses so that deviation may be quantified. It is likely that the total number of ellipse identified differ from the control to the automated identification. Many causes may explain such disagreements. For example, objects that were intersecting with the image contour were not manually identified or certain cells were simply not noticed, sometime because their intensity is lower than expected. Early bud have low intensity, visual inspection indicates that a number of bud were missed in the control. On the other hand, segmentation method may identify more or less cells than needed.

For now, we are to characterize the accuracy of properly identified cells, so that missing and superabundant cells are ignored. The accuracy in artefact detection will be covered in a different section (Section 2.3). We assigned each manually identified ellipse to the automatically identified ellipse with closest center. We assigned manually identified ellipses to automatically identified ones if their centers are occurring within the

average major axis length of the two ellipses. We note that different methods cover different fraction of the manually identified objects; this is due to the number of ellipse automatically detected to differ from the set of manually drawn ellipses, which is detected using this ellipse overlap criterion.

The disagreement in coordinates is measured as distances between ellipse centers and log-ratios of either ellipse areas or major length of ellipses. I know that ellipse area and major axis lengths should be systematically underestimated in the test set, since it was imposed that the drawn ellipses do not intersect, so to simplify the coordinate recovery. For that reason, the standard deviation in log-ratio of ellipse area is most informative of agreement for area measurements. We note that probabilistic methods systematically underestimate cell size, which indicate that circular model tend to fit tips of ellipsoidal shapes. Overall, robust regression has the best agreement with the test set.

Typically, the more distant the ellipse centers are, the less likely the corresponding ellipse has similar areas or major axis length. The choice of threshold criterion for the ellipse pairing significantly contributes for reported average or standard deviations. In order to better compare accuracies of each method, we report deviations for the closest pair of ellipse coordinate only (best 2000 or 3000 ellipse pairs; table I.4). This table allows us to compare estimate accuracies for circle coordinates or cell sizes for a number of cells that does not depends on the identification accuracy, which is defined from the ellipse ovelap criterion previously defined. Still, we note that robust regression yield the best agreement with the test set, deviating from the manually assess area by $\pm 20\%$ std. dev. and $\pm 10\%$ std. dev. for the ellipse width. The probabilistic circle model is shown to be inferior to 'Shape' segmentation of the cell.

Method		Cell Profiler		Probabilistic		Heuristic	Robust
		Intensity	Shape	Shape	Distance		Ellipse fit
Running time		760s	1117s	87420s	54720s	93s	6155s
Recall		64.2%	85%	57.8%	86.2%	75.9%	87.1%
Center distance	mean	11.87	5.67	4.57	4.72	6.96	2.68
	std. dev.	27.92	15.61	5.51	5.37	7.91	4.51
Area log-ratio	mean	-.017	.262	.017	.034	.638	.336
	std. dev.	.822	.390	.568	.624	.525	.343
Major axis length log-ratio	mean	1.168	1.023	-.081	-.069	.231	.148
	std. dev.	.649	.535	.484	.476	.268	.195

Table I.3: **Error distribution for fitted ellipse coordinates.** Comparison of presented methods and two Cell Profiler implementations. The total running time is for processing the 68 images of size 1335x1009. The percentage of drawn ellipses that are paired to identified ellipse is indicated as 'Recall'. For three error measures considered, the mean and standard deviation describes the error over identified ellipses that were each paired to distinct manually drawn ellipses. The first error measure is the distance between the fitted ellipse center to the ellipse center that was manually identified (Center distance). Error for areas and major axis length are reported using log-ratios of corresponding ellipse parameters. Due to a bias, better agreement to manual identification is evaluated by reporting lower standard deviations for log-ratio error measures only (see Figure I.11 a).

Method		Cell Profiler		Probabilistic		Heuristic	Robust
		Intensity (2000)	Shape (3000)	Shape (2000)	Distance (3000)	(2000)	Ellipse fit (3000)
Center distance	mean	1.50	1.75	2.67	2.82	2.59	1.33
	std. dev.	0.88	1.05	1.57	1.76	1.49	0.66
Area log-ratio	mean	.278	.271	.041	.059	.732	.336
	std. dev.	.183	.182	.395	.467	.350	.183
Major axis length log-ratio	mean	1.010	.979	-.051	-.042	.281	.151
	std. dev.	.493	.489	.200	.235	.177	.096

Table I.4: **Method comparison for fitted ellipse coordinates.** Comparison of presented methods and Cell Profiler, which reports the same three error measures as Table I.3. As opposed to the previous table, the agreement in the best 2000 or 3000 closest ellipses pairs is shown (indicated on the table); this allows the objective comparison of errors for methods that each identify a different number of cells per image (different accuracies) and the exclusion of outliers, which may be due to erroneous pairing of identified ellipses to manually drawn ellipses, from the calculation of means and standard deviations for errors measures.

1.3.3 Cell Area Refinement

The methods I introduced were designed to solely identify circle or ellipse coordinates; however, foreground object shapes are not always perfectly ellipsoidal. In this section, I will show that assigning watershed basins to closest ellipses decreases the error of estimates of the center position and size of objects.

Many approaches can uncover foreground object shape. They inherently infer the hidden state $Z_{\vec{x}}$ of each pixel, which results in defining a number of shapes that partition the whole image. Such shapes were already characterized in probabilistic model as to infer the circle coordinates. While the probabilistic inference is possible to partition cell clumps, it inherently is poor to detect cell boundaries, as it mainly relies on the circle coordinates. As such, the area of a theoretical circle agrees better with manual assessment of cell size than the size of the inferred shape (Figure I.12). Using for accurate circle or ellipse coordinates using robust regression so not change this fact. For this reason, I incorporate in the shape recovery process additional information.

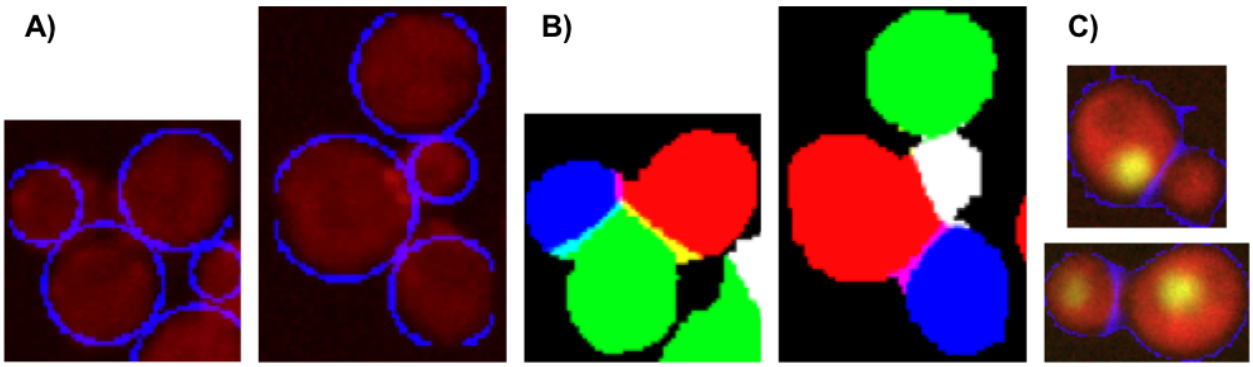


Figure I.12: **Probabilistic shape inference.** A) Example of circle coordinates recovered by Heuristic method (drawn in blue). B) Shape areas inferred from ellipse coordinates. C) Example of bud area with an over-estimated (top) and under-estimated (bottom) area.

The idea is to constrain to space of potential cell boundaries by devising a criterion that forces certain pixel to occur in the same cell. We expect that pixel inside cells are brighter than pixel on the periphery. As such, objects should not be separated in region where the intensity is locally maximal. I use watershed transformation [16] to generate an over segmented image, which will serve as constraint for uncovering cell shapes (Figure I.13). Pixels that share common local maxima from contiguous may then be individually assigned to each cell based on the proximity on the ellipse coordinates. Optionally, the image may be blurred (Figure I.13 B). This reduces the amount of segment and makes the local maxima often occur closer to a potential cell center. Hence, we deterministically assign each segment (watershed catchment basin) based on

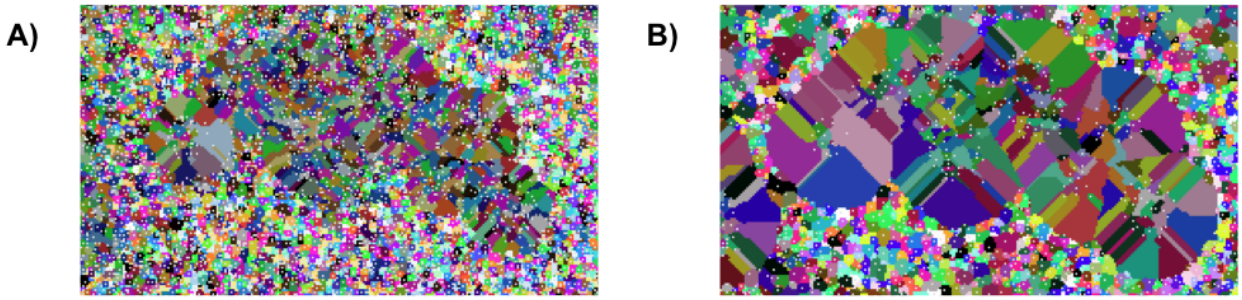


Figure I.13: **Watershed transformation.** A) Watershed segmentation applied directly on pixel intensity (no gradient calculations). White dots represent the local maxima in each segment. B) Watershed segmentation applied on a slightly blurred image using Gaussian kernel ($\sigma = 1$ pixel). The blurring process reduces significantly the number of segment, and local maxima in cell appear closer to the cell center.

the proximity of the local maxima to ellipse coordinates. Uncovering shapes from a 1330x1009 image may be performed within a second; the running time linear in the size of the image and in the number of cell per clump.

For all the ellipse fitting approaches and the test set of ellipse coordinates, we extracted the corresponding shape obtained from assigning watershed segments to ellipses. We next compared the center and area of such shape to the shape that was directly produced by the Cell Profiler 'Shape' based segmentation. As before, the shapes are paired using the proximity of the shape mass center, as longs the deviation does not exceed the average major axis length of the two ellipses corresponding to each shape. We observe that the amount of pair object slightly changed (Table I.5). When comparing the closest 2000 or 3000 shapes to the test set, we observe that robust regression still performs better than Cell Profiler pipelines (Table I.6).

Method		Cell Profiler		Merged Watershed basins			
				Probabilistic		Heuristic	Robust
		Intensity	Shape	Shape	Distance		Ellipse fit
Recall		64.2%	85%	57.6%	86.4%	75.8%	85.8%
Center distance	mean	11.87	5.67	3.39	3.69	5.56	2.78
	std. dev.	27.92	15.61	5.87	5.96	8.20	6.46
Area log-ratio	mean	-.017	.262	.334	.345	.417	.373
	std. dev.	.822	.390	.356	.412	.434	.343

Table I.5: **Error distribution for position and size of cell shapes.** Errors are measured for shapes that defined by assigning watershed basins to closest ellipses. The percentage of drawn ellipses that are paired to identified shapes is indicated as 'Recall'. The first error measure is the distance between the fitted ellipse center to the ellipse center that was manually identified (Center distance). Error for areas is reported using log-ratios of foreground object size to paired drawn ellipse area. Due to a bias, better agreement to manual identification is evaluated by reporting lower standard deviations for log-ratio of areas only (see Figure I.11 A).

Method		Cell Profiler		Merged Watershed basins			
				Probabilistic		Heuristic	Robust
		Intensity (2000)	Shape (3000)	Shape (2000)	Distance (3000)	(2000)	Ellipse fit (3000)
Center distance	mean	1.50	1.75	1.30	1.42	1.41	1.07
	std. dev.	0.88	1.05	0.80	0.95	0.82	0.52
Area log-ratio	mean	.278	.271	.336	.334	.333	.329
	std. dev.	.183	.182	.180	.179	.184	.159

Table I.6: **Method comparison for position and size of identified cell shapes.** Comparison of presented methods and Cell Profiler, which report the same two error measures as Table I.5. As opposed to the previous table, the agreement in the top 2000 or 3000 ellipse pairs is shown (indicated on the table); this allows us to objectively compare methods that each identify a different number of cells per image. Both 'Center Distance' and standard deviation in 'Area log-ratio' are the lower than values reported in table I.4, for all 4 methods that fitted ellipses.

1.4 Discussion

Presented results suggest that clump occurrences in images can be better accounted for using prior knowledge on the cell shape. The pipeline of computational methods has been evaluated on a relatively small set of images; other types of microscopy image of yeast are likely to require different pipeline to achieve similar results. Each of the processes within the pipeline relies on some feature of the image collection analyzed. For instance, the separation of foreground and background pixel assumes that intensities of the two classes have can be separated by a intensity threshold, which is further assumed constant on the whole image area; if this is violated, I observed that clumped cells may either becomes isolated cells or have contours that poorly fits their peripheries. The generation of watershed 'basins' relies on a gradient in pixel intensities within individual cells; the separation of the objects inherently assumes that the intensity is lower between

cells than inside cells, which is not a typical property of bright field microscopy images (Figure .1). Finally, the partition of the clumped cell with the use of ellipse strongly depends on the cell density in an image (total number of cells in image); larger cell clumps have a fewer background/foreground separation pixels per cell. Hence, the higher accuracy and time efficiency of the presented pipeline method is bound to change depending of the image collection analyzed.

While this image collection is the central piece for my published work [66], other image collections will be analyzed by this pipeline, which may be slightly altered to perform better on a given image collection (section II.4.2 and III.1.4.1). While the identification accuracy has been carefully evaluated on a single image collection, the method that identifies foreground objects using ellipse robust regression produces comparable measurements on several different microscopy image collections. The cell identification accuracies cannot be quantified for those image collections since no manual identification was performed. Still, it can be shown that the methodological pipeline is can produce comparable results between image collection, even if the pipeline is adapted to the image collection at hand. This flexibility is allowed by the organization of the processes that were implemented specifically for parallel computing and without any external dependency (C++ source code available, see [66]).

2 Modeling Cell Morphology

Now that the individual cell shapes have been separated, an optional task acknowledges that such shape are instances of time slices from time-dependent biological system. As such, each object has a particular state that describes undergoing biological processes at a given the cell-stage. All the imaged cells are assumed to have the same genetic and environment background, so any observed cell at a given stage may have similar phenotypes. The morphology of budding yeast is informative of the cell-stage, so we expect that the expression of certain proteins is linked to the morphology of identified cell from their cell-stage dependence. Hence, we are to devise a cell-stage estimation that is shown to be robust to the presence of artefacts and segmentation errors that occurred in the cell segmentation.

2.1 Cell Stage Assessment

Since cells that are undergoing the budding process are better characterized by a pair of ellipses [138], we have devised an approach to identify bud and mother cells as separate ellipsoidal objects. Therefore, cell 'types' have to be assigned to each object: either artefact or one of three cell types ('mother', 'bud' or 'lone' cell). These cell type have been manually identified for 4305 foreground objects. In this section, I introduce

a simple heuristic that assigns label types to foreground objects based on adjacency and size of foreground objects; then, I show that its use on foreground objects obtained by ellipse robust regression more accurately labels object types than its use on objects obtained from two Cell Profiler pipeline methods.

Once, artefact are identified (Section 2.2), the remaining objects were assigned types using a simple heuristic based on the cell sizes (Figure I.14 A). Mother-bud pairs were defined as reciprocally smallest and largest adjacent cells, and in addition buds were not allowed to have any smaller neighbouring object. Any other cell is considered unbudded or 'lone'. With this definition, a mother-bud pair may be independent cells in G1 phase that are found to be adjacent: we still consider them as a pair since it is likely that such a connection existed in the very recent past if one of the two cells still small. The adjacency of cells is determined from the Pseudo-2D HMM segmentation, which appears to systematically clump adjacent cell, due to a higher intensities for background pixels that are between cells. By applying this simple rule, we characterized 405359 mother-bud pairs, and 494680 remaining lone cells, so that a total of 1.3 million cells were identified. The size of the bud is here considered as cell-stage indicator. Alternative quantities have been previously shown to be explaining cell-stage, such a major to minor axis of the shape of the joined bud and mother cell. The major axis is expected to be aligned in the bud to mother orientation, and increase in length as a bud grows or elongate (different yeast strain). Since the dataset we does not have cell-stage indicator, we cannot evaluate the accuracy of the cell-stage estimation directly: it will be evaluated in the next chapter.

We first check that bud size can be related to cell-stage at all. The probability density of object radii are different and depend of the cell type that is assigned to objects (Figure I.14B). Interestingly, the number of bud increases in density on the left of the mode for the probability distribution. Since microscope images are time slices of yeast colonies that are assumed in an exponential growth phase, this is incoherent with the notion that bud grow in radii linearly. Assuming that the steady state for the probability density of bud sizes is reached, bud of any positive radii should have equal likelihood (or density). Further assuming that buds stop growing once they becomes a lone or mother cell forces the probability density of bud radii to be strictly decreasing. As such, the observed density function cannot be explained by a radii growth that is linear in time. Similarly, bud areas do not appear to growth linearly. Cell volume, which can be estimated from the cell area, appears to have a density that is relatively constant for small buds (Figure I.14C). This indicates an agreement for the linear growth in volume of small buds. While the density increases eventually, this may be coherent with the notion that the bud growth eventually slow down. Some that reason, we consider '$area^{\frac{3}{2}}$' of the bud as cell-stage to be a proper cell-stage indicator, which is coherent with the observe distribution of bud size.

The association of bud and mother cell is critical to properly utilize the bud size as cell-stage indicator for its mother. I compare the previously defined method in their capacity to predict the cell type from using this heuristic type assignment. As before, identified objects are paired with coordinates of manually identified cells. For the fraction that have close enough ellipse centers (same criterion as before), we report the accuracy and false-positive rate of cell type assessment (Table I.7). It is clear that any of the proposed approaches out-perform Cell Profiler 'Shape' segmentation. This method does not strongly assume any shape model, if not for some notion of convexity; methods that specifically tackle ellipse shape recovery are more suited. Robust regression is the method that has the best classification power.

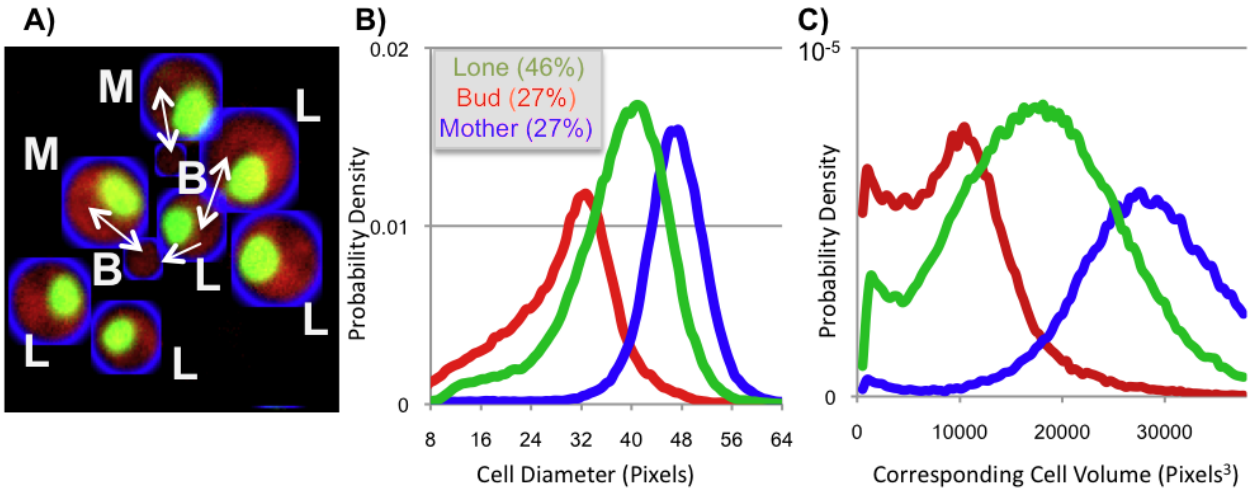


Figure I.14: **Prediction of 'cell type' using a simple heuristic.** A) Example of the application of mother-bud assignment heuristic. Pairs of circular objects that reciprocally have largest and smallest sizes among neighbouring areas are said to be 'mother' cells (indicated by M) and 'bud' cells (indicated by B, mother-bud pairs indicated by bidirectional arrows), unless the potential 'bud' cell has a smaller neighbour than itself (indicated by a unidirectional arrow). Any other cells are labelled as 'lone' cells (L). B) Diameters of the identified cell, dependent of the type assessment. C) Estimation of the cell volume, from transforming the cell area.

	Cell Profiler		Probabilistic		Heuristic	Robust
Precision	Intensity	Shape	Shape	Distance		Ellipse fit
Mother	40.4%	34.6%	51.7%	56.2%	49.7%	55.7%
Bud	25.3%	22.9%	58.0%	59.5%	42.5%	62.4%
Lone	56.4%	53.8%	80.9%	79.8%	70.1%	83.7%
Recall						
Mother	17.9%	19.3%	78.4%	79.9%	58.8%	84.9%
Bud	12.7%	13.9%	12.6%	79.7%	78.6%	84.7%
Lone	79.9%	72.9%	52.6%	55.3%	48.4%	53.3%
Accuracy	51.7%	48.3%	49.5%	66.0%	57.3%	67.1%

Table I.7: **Classification of 'cell types' .** Classification performance for cell types with respect to manual assessment. Since three object labels are considered ('Mother', 'Bud' and 'Lone'), we can define precision and recall for each class (one-against-the-rest binary classification). We note that the Cell Profiler based cell type assessments are biased to rarely label cells 'lone', so that the recall for identifying 'Mother' and 'Bud' is worse than random cell type assessment (a recall of 33.3% is expected). We note that among all the methods robust regression displays the highest precision and accuracy.

2.2 Cell Confidence

Because automated identification of clumped cells in images with artefacts is a challenging computational task, we expect a fair fraction of the identified objects to be misidentified objects and/or non-trivial artefacts. Indeed, close examination of example images revealed a significant number of artefact classes: Noise in image corners, ruptured cells, cells that lost RFP, defective CCD pixels, contamination, and out of focus objects were sometimes erroneously identified by our pipeline.

Instead of trying to characterize each artefact class, we defined 3 quality measures based on object shape and contour, which have known distributions for circular or ellipsoidal objects (see 'Cell confidence' in Methods). We also use the mean RFP signal within the object as an additional quality measure. We model variation in each quality measure using a Normal distribution whose parameters are a function of object size and infer parameters using a set of cell contours obtained from the set of manually fit ellipses (see Methods). A uniform distribution is used to model the quality measures from 'non-cell' objects, allowing us to compute the posterior probability that an object is a cell under the model that the objects in our images are drawn from a two-component mixture of cells and non-cells:

$$P(Cell|\vec{q}, size) = \frac{P(\vec{q}|size, Cell)P(Cell)}{P(\vec{q}|size, Cell)P(Cell) + P(\vec{q}|size, non-Cell)(1 - P(Cell))} \tag{16}$$

where $\vec{q}$ is the vector of quality measures and RFP intensity, and $P(Cell)$ is a mixing parameter that can be

thought of as the prior probability for an object to be a properly identified cell. We use EM to re-estimate that mixing parameter, while the cell class parameters are inferred from our set of manually identified cells and are not updated. We refer to this posterior probability as the 'cell probability' for each individual cell. The majority of cells in the images show high-confidence ($\geq 95\%$) (Figure I.17). We define the probability of a mother or bud as the product of the two cell probabilities. We also allow these cells to be partially assigned to the lone cell class based on the cell probability of the putative related mother or bud.

2.2.1 Quality Measures

In addition to the mean RFP intensity in the object, we define three shape measurements based on geometrical properties of ellipses and circles. First, we compute the best fit of an ellipse to an arbitrary shape 'S' by evaluating 6 statistics (eq. 15) on the coordinates of pixels in the shape (Appendix 2). As described in the Appendix, the ellipse fit is based to matching the coordinate statistic a theoretical ellipse produces.

$$F_{\vec{c},A,r,D}(\vec{x}) = \begin{cases} D & (\vec{x}-\vec{c})^T A(\vec{x}-\vec{c}) \leq r^2 \\ 0 & (\vec{x}-\vec{c})^T A(\vec{x}-\vec{c}) > r^2 \end{cases} \tag{17}$$

Since the coordinates are drawn from a bitmap, we observe that the recovered density 'D' is the ratio of number of pixel to fitted ellipse area. The measured densities typically bounded above by 1, but for smallest objects. Any shape whose density is above or equal to 1 is assigned to the artefact class, and we use $q_1 = log(1-D)$ as a first quality measure for each ellipse.

The second quality measure is based on the relationship between the perimeter and the area of an ellipse. We compute the perimeter of the shape by counting the number of pixels that have 3 or more background pixels among their 8 neighbouring pixels. The theoretical relationship between the perimeter length of an ellipse and the parameters has no simple form, but may be approximated using the Ramanujan first approximation [12]:

$$L = \pi(3(a+b) - \sqrt{10ab + 3(a^2+b^2)}) \tag{18}$$

where 'a' and 'b' are the minor and major axis length of the ellipse. The log-ratio for the Ramanujan approximation to perimeter of the fitted ellipse and number of contour pixel is our second confidence measure q_2.

A third quality measure captures the deviation of the shape to a circle, by reporting the log. coefficient

Quality Measures	Artefacts		Cells	
	mean	std.dev.	mean	std.dev.
Ellipse Density*	0.9360	0.0637	0.9804	0.0373
Contour Pixel Deviation*	1.0123	0.1413	0.9879	0.0760
Edge Distance Deviation*	1.0445	0.0349	1.0081	0.0177
RPF Deviation to Exp.	-0.0196	0.3173	0.105	0.2318

Table I.8: **Distribution of 'Quality Measures' .** Distribution of quality measures for objects that were manually identified as artefacts or cells. Artefacts display significantly higher level of variance. *quality measures are not log. transformed.

of variation of the sum of the distance to the ellipse center and the distance to the edge for each pixel in the area(eq 19). In a theoretical circle, there should not be any variance, since the two quantities are to sum up to be exactly the radius of the circle.

$$q_3 = \frac{1}{2} log \left(\frac{\frac{1}{|S|} \sum_{x \in S} (D_{edge}(\vec{x}) + ||\vec{x} - \vec{c}||)^2}{(\frac{1}{|S|} \sum_{x \in S} D_{edge}(\vec{x}) + ||\vec{x} - \vec{c}||)^2} - 1 \right) \tag{19}$$

The last quality measure is the mean RFP intensity $q_4 = \frac{T_R}{|S|}$. We model each of the quality measures using Normal distributions. We observe that the quality measure spread displays a non-trivial dependency on cell size. For example, the mean RFP intensity increases with bud size, but then decreases for mother cells beyond a certain size (due to the larger dark vacuoles). For this reason, we define 7 Normal distributions, for each of the 4 quality measures that correspond to the distribution of quality measure for 7 bins of cell sizes. The quality measure vector is then modeled by the linear interpolation of a pair characterized random variables X_i and X_{i+1} (eq. 20).

$$\begin{aligned} &\vec{q} = (i + 1 - \tfrac{|S|}{500}) X_i + (\tfrac{|S|}{500} - i) X_{i+1} && where\ |S| \in [500 \cdot i, 500 \cdot (i+1)] \\ &X_i \sim N(\mu_i, \Sigma_i) && where\ i \in \{0, 1, 2, 3, 4, 5, 6\} \end{aligned} \tag{20}$$

where $\vec{q} = \{q_1, q_2, q_3, q_4\}$ and Σ_i are diagonal covariance matrices.

Further, we can show that the distribution of the quality measure for manually identified artefacts and cell differ significantly (Table I.8).

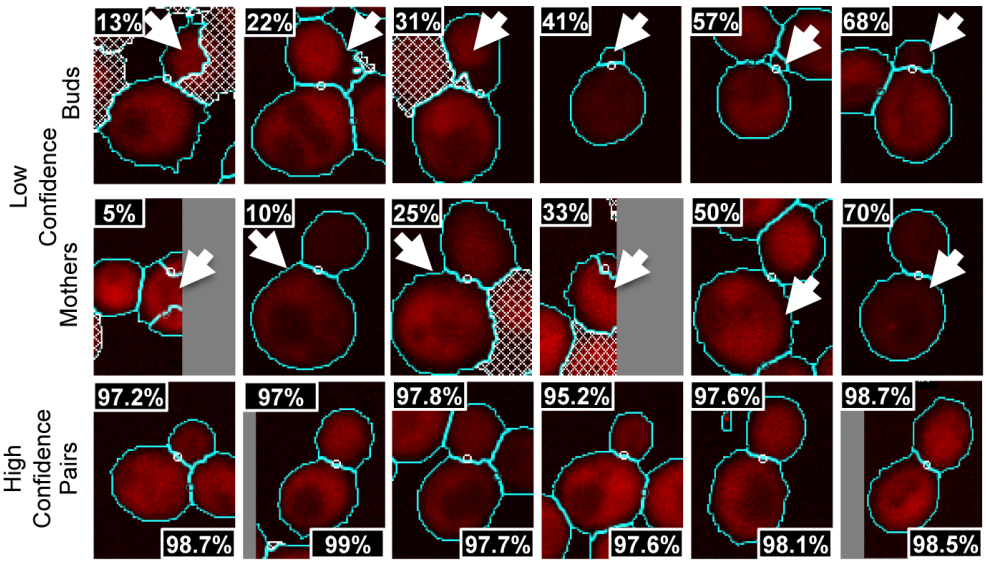


Figure I.15: **Example of low and high confidence objects.** The cyan lines in each image represent the cell contours produced, and the white dots indicate the predicted bud neck position. The dashed objects represent obvious artefacts that were filtered using thresholds (See text for details). Objects on the edge of images were not automatically filtered out, but are expected to have low confidence.

2.3 Agreement with Manual Identification of Artefacts

In this section, I show that modeling the distributions of image features in foreground objects (4 quality measures of cell shape and RFP intensity) allows the detection of artefacts in images. 139 artefacts were manually identified. We used this set to compute the false positive rate by pairing automatically identified cell areas to the manually identified cells and artefacts (Figure I.16). We also computed the false-positive rate as a function of cell probability threshold. For example, filtering all cells that have a cell probability below 0.8 reduces the false positive rate. This is in agreement with previously reported results using post processing [28]. While filtering objects based on cell probability preferentially excludes artefacts (Figure I.16), we also found that small buds typically have lower cell probability (Figure I.17B), so defining cell confidence thresholds also preferentially filters small buds. Hence, we use these cell probabilities to weight individual cells when computing averages over cell populations.

Note that the nature of manually identified artefacts is confined to abnormality in microscopy images; failure to 'properly' identify object and can affect downstream analyses. In the next chapter, we will perform such downstream analysis, which will allows so to note that probability threshold reduces the robustness of the time-profiles of protein expression (Figure II.7B), and that the quality of unsupervised analysis is higher when the cell confidence weighting approach is used(Table II.4& II.5).

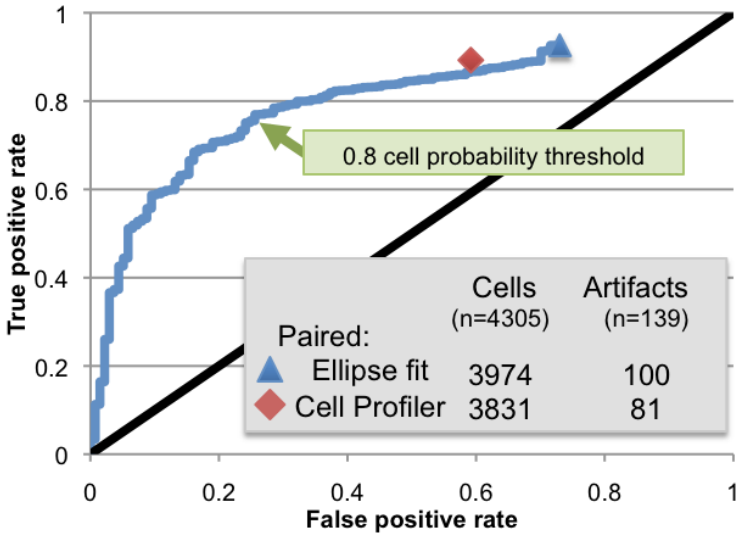


Figure I.16: **ROC curve for cell identification with confidence scores.** A test set of 4305 cells and 139 artefacts were identified by manually drawing ellipses around objects in images. Automatically identified cell areas were paired to manually drawn ellipses if their centers were found within 10 pixels. Other manually identified cells were considered false negatives. The false-positive rate (number of artefacts/number of predictions) and true positive rate (or recall, which is the number correctly identified cells/number of manually identified cells) are plotted as a function of cell confidence. As a reference, we also display the performance using a Cell Profiler pipeline (red diamond) and the baseline accuracy of our method (blue triangle) without a cell probability cut-off. The expected performance of random guessing corresponds to $y = x$ in this plot (thick black trace).

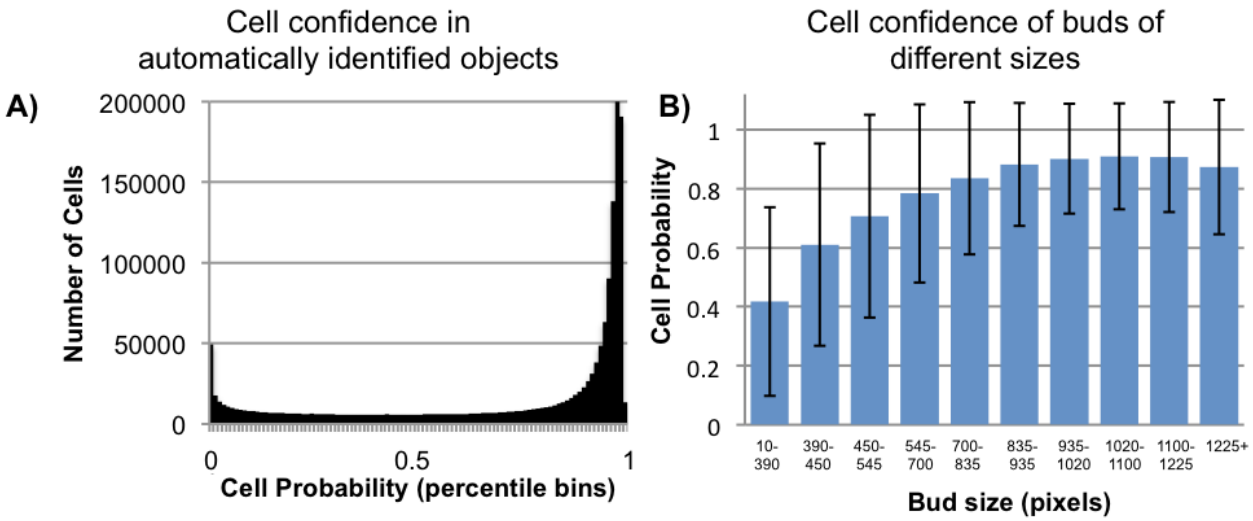


Figure I.17: **Confidence estimates for automatically identified cells.** a) Histogram of Cell probability for automatically identified objects. Cell probability is calculated for each of the 1.3 million identified cells as defined in the text. The assigned cell probabilities are displayed using 100 bins. The majority of the identified shapes have a probability to belong to the cell class that is above 95%. b) Dependence on bud size for cell confidence on bud cells. The set of 405359 identified buds was partitioned into 10 groups based on bud size, such that each group has the same number of cells. The mean and standard deviation in the measured cell probabilities is shown (grey bars). Smaller buds tend to have lower cell probabilities.

Part II

Modeling Protein Expression

Material from: Handfield et al. 2013 [66]

Within Sections: 1.1, 1.1.1, 1.2, 1.2.1, 2.2, 2.3, 2.4.*, 3, 3.1.2, 3.2.*

Figures: II.3, II.7, II.8, II.9, II.10, II.11, II.12, II.13, II.14

Overview

The collection of fluorescent yeast mutants allows the analysis of the subcellular localization of ∼4000 yeast proteins. Many protein localizations are hard to describe using words (annotations), which is the prevalent mean to describe protein subcellular distributions: many proteins are present in multiple subcellular localizations simultaneously with varying fractions, and several proteins change in abundance or subcellular localization at specific cell-stages. Machine learning provides automated means to predict subcellular localization from images (pattern recognition). Since such methods relies on the accuracy of annotation assigned to images by experts (training set), they have low accuracies for proteins that are not well described by the frequent subcellular localization classes.

In this chapter, I first extract several image features (Section 1) to infer the cell cycle dependencies of measurements from the collection of identified cell population (Section 2). I then show that the high resolution images contain more biological information than what can be conveyed by subcellular localization annotations (Section 3). For example, proteins that are bound to each other (such as complex subunits) can be shown to share specific cell-stage dependencies subcellular distribution. Such a characterization allows the extraction of fingerprints of protein expression that not only allow subcellular localization recognition, but are often sufficient to uniquely recognize a protein among the ∼4000 protein expressions across replicate experiments (Section 4).

0 Background: Characterizing Protein Spatial Expression

0.1 Image Features

Computational identification of proteins with particular subcellular localization amounts to a classification problem. Machine learning approaches do not typically classify the phenotypes directly from raw image data: they all suffer from the Curse of dimensionality [14], since the search space for classifiers that model each individual pixel is too large for practical purposes. For the tractability of these algorithms, we must define a far smaller set of functions of the raw input image, whose outputs alone are considered for classification. This is referred to as feature extraction. An alternative is to apply dimensionality reduction algorithms, and choose the most informative dimensions. This is referred to as feature selection. Typically, feature selection is used in tandem with feature extraction [75]. Feature extraction is preferred as the input pixel covariance has structure [78], and certain functions of the input are a priori known to not hold discriminative

power. Sometimes customized features are used, either based on some numerical procedure (such as skeletonation) or using extra information provided by the experimental setup (intensity overlap with an additional marker) [174]. Murphy et al. described and maintained a set of subcellular localization features (SLFs) that contains several classes of image features. They have been using this base set for characterizing protein spatial distributions for nearly all their classification needs [19, 32, 74, 74, 75, 79, 108, 109, 124, 145, 174].

***Texture Features*:** Haralick texture measures [67] form a collection of 14 classes of features measures. It was designed to cover a broad spectrum of possible texture properties, such as fineness, coarseness, smoothness, granulation, randomness, lineation, or being mottled, irregular, or hummocky. In that collection, one feature measures the correlation of pairs of pixel found with a set relative offset position within the image. Another feature is to normalize image by the variance of pixels in the image, so that identical images with different time exposures or transparencies would be indistinguishable under this feature. Some features ignore the gray scale interpretation of the image, and are defined as the transition probabilities of neighbouring pixels conditioned on their integer value. Also, there are some edge features, which are generated by convolving the image with a small matrix that enhances lateral contrast on some orientation. These presented features are operations on images: their output by definition is an image (a two dimensional map of values). In order to extract a single feature value per feature class, Murphy et al. counted pixels above a threshold, for the edge feature [32]. Similar operations can be performed on the whole image or within an identified area.

***Rotation and Scale Invariance*:** There exist several classes of image features that are defined using a scale and an orientation: Gabor filters, which are convolution filters that model the eye perceptrons [78], Zernike moments, which provides a set of orthogonal features defined on a scalable unit circle. Assuming that a procedure is capable to perfectly recovering the orientation of an object, these can be made rotation and scale invariant by measuring the intensities in the objects corresponding coordinate frame. Murphy et al. [75] used the major axis of the segmented areas (spatial variance), and extract these two feature types in the coordinate frame defined by the major axis of variance and its magnitude (scale). The second major axis used to decide if a reflection is necessary. Two identical objects, with different poses and sizes, can generate the same set of features; hence the measured feature values possess invariance properties.

***Temporal Features*:** In time-lapse images, we may recover intensity fluctuation within objects. Hu et al. [74] produced 3D movies of fused GFP strain of the NIH 3T3 cell lines (mouse fibroblast cells). They considered Haralick texture measures [67] in the temporal dimension (instead of spatial dimensions). They evaluated 26 such temporal features at 5 different time points (130 total), for 5 types of images. They report a mean classification accuracy increased from 75.32% to 85.06% by the inclusion of these new features (to a

basic set of 94 non-temporal features). These features are acquired while ignoring the cell movement, which is known to be a concern in other methods [136].

Hence, Hu et al. [74] proposed and defined a similarity metric tracking scheme, which uses cell position, size and total intensity in tandem. Objects are now identified using contiguous areas of intensity above a threshold. Objects in N subsequent frames are linked using Nearest Neighbour under their custom metric. In addition to the feature introduced above, they extract 'Normal flow features', which estimate local motion from intensity changes. This estimation to the displacement field assumes smoothness, that is, absence of discontinuities. Discontinuities systematically arise in recovering depth from stereo images [6, 106], so that detecting these discontinuities would be suggested in that case. Since there is no occlusion in objects of interest, there is a priori no reasons to observe these. Instead of including static feature their previous approach, they defined features to be the weights obtained from linear regression in the time dimension. Non-temporal feature values at each time point were inferred by regression using the k previous frames (autoregression). For 11 classes, they report this time that the overall accuracy increased from 66% to 78% by including (31, 260 features respectively).

0.1.1 Feature Reduction

An infinitely large set of features can be derived from the above classes. Feature space needs to be reduced either by selecting the most informative ones or by recombining them by defining a function mapping from the feature space into a space of lower dimension. Feature selection involves searching for a subset of low ordinality that retains most descriptive power. For that task, heuristic search procedures are used, as this task is NP-Hard [32]. The alternative involves studying the covariance structure of the features in order to find the axes containing the most class discriminative information content. Linear Discriminant Analysis (LDA) is defined to recover the major axes of covariance between two classes of data points [23]. In this case, the reduction would be equivalent to a projection into a smaller space. The LDA is a linear method so that it fails to model the information for features with nonlinear dependencies. For these problems, if might be advantageous to extend the feature space in a higher dimension using kernels, and then recover the most discriminative recombined sets of features using LDA in that space.
One noteworthy aspect of these approaches is that this last reduction scheme is defined for binary classification. Many authors proposed the scheme "one against the rest", in order to allow this methodology to be used for multi-class multi-label classification [23].

0.1.2 Image Level Features

Most of the classification schemes require identifying the cellular objects in images, in order to relate the seen intensities in protein subcellular localizations. However, it is not an easy task to identify individual objects. It is possible to extract features out of the whole image, while ignoring the actual position and orientation of the cells in images. Murphy et al. [75] have applied Gabor filters on local windows throughout the image and realigned them individually using their most varying component. This enables patterns of different orientations to generate identical profiles, enabling the classification of patterns with no knowledge of the number of cells and their orientation. They report that this method outperformed their previous method [29] on the Huh et al. annotated proteins. [76]

0.2 Localization Classification

The machine learning field provides a large collection of methods that predict class labels using labeled set of instances (training set) provided by an expert (Supervised Learning Approach). Among them, the ones that have been considered for classification of cell phenotypes are: Neural networks [40, 108, 145], Support vector machine (SVM) [65, 75]. Several aspects of the problem make the classification task hard. First, the number of possible localizations is large (>20), while the classifiers are typically defined under a binary classification framework. For example, the SVM approach is to find the best hyper-plane separation between two sets of vectors in an augmented space spawned using kernel functions [172]. Extending SVM for multiple classifications is a subject of ongoing research [7]. Secondly, the problem is imbalanced, as few combined localizations (Nucleus, Cytoplasm, Mitochondria) holds 80% of the manually classified proteins, while many localization would account for less than 1% of the training set [76]. This feature induces a classification bias, it has been proposed to compensate by considering penalty functions, different weights for misclassification of positive and negatives samples [169]. Indeed, the classification accuracy is reported to drop with low numbers of training images available [29].

0.2.1 Validation

Once a classification method has been applied to microscope images, it is important to evaluate the performance in a systematic and quantitative manner. The most basic method to validate a classifier is to produce the confusion table [141]. A confusion table (Figure II.6) for a N-classifier is a N by N table that reports, for each class, the fraction of the classified instances that comes from each training set (N training sets for each classes). In particular, the diagonal entries of the table are the class specific recall.

***Cross-Validation*:** These reported values are typically measured with cross-validation. Given a sufficient number of parameters, many machine learning have some guaranty of 100% recall if the tested instances are all part of the training set (overfit) [155]. On the other hand, if the number of parameters is small, it is possible to overtrain, which may lower the predictive capacity [155]. High recall values from overfitting are not informative as the aim is to measure the true discriminative power for unseen examples. K fold cross validation involves separating the labeled instances in K groups, and evaluate the accuracy and recall of instance in a group by classifiers trained on the labeled instances from the K-1 other groups. If K is equal to the number of labeled instance, then K-fold cross-validation becomes leave-one-out validation [87]. This last approach benefits from being unbiased, but requires training K models. Training so many models may be time inefficient, unless the class of machine learning considered possess a property that derives efficiently those K models from a single trained model, which is trained once on the whole dataset, in a rapid manner, such as Gaussian process based classification [130, 130].

***Benchmarking*:** The reported accuracies are solely used for benchmarking; for example, labeling an image has been shown to be harder than labeling groups of 10 images [145]; and labeling individual cell objects in an image is harder than labeling the whole image [29]. Certain human faculties cannot be emulated or outperformed using our current computer technologies. Text, face and semantic recognition are often said to be AI-Hard problems, and are used in security protocols to block malicious automated computer programs [99]. Murphy et Al. [109] report that a biologist with no prior knowledge of fluorescence microscopy was trained to classify subcellular localizations, and achieved a final overall accuracy of 83%. The accuracy of automated classification was 86%, which was reported to be on par with the manual classification. However, it is important to note that the reported accuracies cannot be related to the biological significance of classifiers [141].

0.3 Challenges for Protein Localization Studies

Unsupervised Analysis: On the other hand, unsupervised learning has also been extensively applied for image recognition. For example, distinguishing images of text from scenery images can be done is a probabilistic framework, where the pixel dependencies in 8x8 pixels patches in images were learned under the assumption that two models explain the dependencies, but we are missing each pixel's hidden class labels [92]. One disadvantage of unsupervised pattern recognition is the difficulty of incorporating prior knowledge (especially relational knowledge) and complex structural knowledge [104]. Classifying subcellular localizations for this approach is then hard, since: Some classes of localization are so similar (Ex: endosomal and lysosomal) that only a few pixels hold discriminative power [174]. One possible solution is a semi-supervised approach, where only a fraction of the dataset labels are provided [170]. They report that many types of

supervised classifiers may not achieve the accuracy of their proposed classifier, which is semi-unsupervised as it has only access to 30% of the training labels. Their approach relies on the co-training paradigm, where a pair of classifiers individually ranks the unlabeled instances by their confidence, and iteratively adds their most confident predicted labels to the training set of the other classifier.

***Generative Shape Model*:** Murphy [61] mentions that subcellular localization label may only account for limited information. A generative probabilistic model of subcellular localization patterns was developed for a three fluorescence channel dataset of 447 images of HELA cells. Nuclear shapes were modeled using B-splines [72], and protein localization patterns were modeled using mixtures of 2D Gaussian distributions and estimated parameters using EM. The stated aim for developing generative models is to characterize the information that is to be integrated into predictive cell models. This model has since been extended to model 3D microscopy images, but to our knowledge it has not yet been used for classification [125]. They conclude stating that further work and refinements on generative model is required so that their utility can ultimately be evaluated by suitable for simulations of a whole cell model. Indeed, the task of characterizing a model capable of emulating cell behaviour is of immense difficulty. Being capable of identifying finer and simpler structures may be key in breaking this problem.

0.3.1 Mixed Localization:

In certain systems, such a mouse liver, it is known that about 39% of the proteins are present in multiple subcellular localizations [53]. The classification of these proteins has been reported to be harder, for knowledge based approaches [93] and for image classification approaches [124]. Typically, this aspect of the problem is artificially ignored [36]. Murphy and Al. [29] report that, in their study, they selected 2713 yeast strains (out of the 4156 which showed GFP expression) which were not assigned to 'ambiguous' and 'punctate composite' in the UCSF Database (Huh et al. [76]). Selecting manually the images for training sets includes a considerable observer bias. [141].

***Unmixing*:** Images of proteins of mixed localizations would be arbitrarily assigned to one pure class, which was considered at the moment of the generation of the training set. Increasing the dimensionality of the label is an issue, as experts would possibly disagree on complicated patterns, or even individually not reproduce their own labels [174]. Murphy et al. propose, as an alternative, to infer localization fractions from the feature space. They apply their approach to a human cell line, HeLa cells. A two steps algorithm would first use K-mean clustering [68] in a manually selected feature space to create what would become a basis to explain the feature space in the second step. Using Nearest Neighbour to the centroid of each

cluster (in the feature space), they noted that indeed there is no guaranty that this unsupervised recovered one pure subcellular localization. Each subcellular localization training set would have its protein assigned to different clusters in different factions. The second step then is to use theses fractions and centroids to build an inference map of localization fractions. They finally used synthetic data to evaluate the certain classes of inference maps.

In a follow-up work, Murphy et al. [124] apply a similar methodology to valuate this approach performance on non-synthetic data. Organelle fluorescent markers for Lysosome (Lysotracker green) and mitochondria (Mitotracker green) are used to generate an image set with known proportions of the markers. They design an experiment that, in essence, aims at measuring the capability of classifiers trained on pure localization to infer the proportion in generated images. They first clustered the generated images (using K-means [68]), and then recovered the proportion of the image of pure Lysosome and pure Mitochondria that were assigned to each cluster using Nearest Neighbour. Finally, they compared more inference maps classes by reporting correlation from known to inferred fractions for each class: linear unmixing (0.73), multinomial unmixing (0.77), Fluorescence fraction unmixing (0.83). The only inconvenient aspect of this validation experiment, compared to their previous work, is that this only displays the capacity of distinguishing between two classes, where no data imbalance is found since each marker's signature is equally present in images.

0.4 Model-based Analysis of the Cell

Beyond classification, microscopy gives the promise of providing the sufficient information for mathematical modeling of molecular pathways [73]. While databases of biological pathways exist [85], and that several mathematical models for certain processes were proposed [27], the biological variability in the shape of cells offers a very serious obstacle to the application of this type of theory [22]. Ignoring spatial inhomogeneities can be compensated by the introduction of time delays in models, which preserve their basic equation form [43], but this is insufficient to account for multiple compartments. Hence, much work in progress involves modeling the cell membranes and compartments.

0.5 Temporal Models of Protein Abundance

We observe that the cell shape itself regularizes protein expressions, and it is also critical in controlling the cell mitosis (yeast *Schizosaccharomyces pombe* [126]) or budding *ustilago maydis* [163]. Understanding the morphological causes of protein spatial distributions is a hard problem, which may not be performed in a high-throughput manner given the complexity of possible modes of regulation. For example, Robbins et al. [132] wanted to explain an observed gradient in the spatial distribution of IcsA on cell periphery of

Shigella fexneri. This cell has an elongated shape, and the IcsA expression is mainly observed on one tip of the cell, but a gradient on the whole cell is observed. They proposed two mathematical models that would explain the presence of this gradient: The protein is sent to the cell periphery randomly, then some translocation mechanism brings the protein to the tip; or alternatively, the protein is specifically introduced in the cell membrane at the tip, and from there the protein freely diffuses (on the 2D cell membrane). They managed to rule out the first explanation, by observing the changes in the gradient when the cell is treated with chlorpromazine, which changes the protein diffusion rate on the outer membrane.

1 Single-cell Protein Expression Measurements

As previously mentioned, the images we analyzed contain pairs of bud and mother cell that have their cell-stage estimated from the bud size. In order to utilize the cell-stage measurement, any other observation for protein expression is used to be define to cell-stage dependencies. Among the many image feature (Section 0.1) that allows the recognize protein expression pattern, cell morphology dependencies are expected, but have no biological interpretation. In order to best understand the protein expression, I consider cell measurements whose value and cell-stage dependency can be directly interpreted and related to known biological phenomena.

The first type of measurements I considered relates to the intensity distribution of the pixel from each cell and the second type relates to the spread of proteins within the cell area or the mean distance of proteins to a point of interest. One of the issues that will be discussed with each measure is their normalization, as some have an explicit dependency on object size, cell shape and sometimes with RFP intensity.

1.1 Pixel Intensity Distribution

We next sought to characterize the protein expression phenotype using a small number of measurements that are biologically interpretable. The intensity of GFP signal in each cell relates to the level of protein expression [11, 113]. Therefore, as a first measurement, we use the ratio of total GFP intensity to RFP intensity within in each cell area.

$$f_{\text{Intensity}} = \mu_R(|S|) \cdot \frac{\sum_{\vec{x} \in S} G(\vec{x})}{\sum_{\vec{x} \in S} R(\vec{x})} \tag{21}$$

where $G(\vec{x})$ and $R(\vec{x})$ are the GFP and RFP intensities in the image at coordinate $\vec{x}$. $\mu_R(|S|)$ is the expected

RFP intensity as a function of the cell area and ensures that the intensity ratios are comparable for cells of different sizes. This was necessary to correct for a systematic dependence of RFP intensity on cell size, which was characterized using the entire collection of identified cells (see 'Protein expression measurements' in Methods).

1.1.1 Normalization

We characterize the protein expression phenotype within each cell object using the absolute intensity of the GFP, as well as geometrical distances between proteins to identified points of interest. In both cases, we use the RFP signal to normalize the observations made for the GFP signal. The RFP intensity was found to be dependent on the object size, so we characterized the expected RFP, $\mu_R(|S|)$, and used to normalize the GFP signal by the fold difference to the expectation of the mean RFP intensity (eq. 21). We defined $\mu_R(|S|)$ using three linear function segments that fits the mean level of RFP in the 1.4M automatically identified cells:

$$\mu_R(|S|) = \begin{cases} 15 + \frac{12 \cdot |S|}{1000} & |S| < 1000 \\ 27 + \frac{|S| - 1000}{260} & 1000 \leq |S| < 1650 \\ 29.5 + \frac{|S| - 1650}{1175} & |S| \geq 1650 \end{cases} \tag{22}$$

Moreover, it appears that the cell autofluoresces significantly and systematically contributes to the recorded GFP intensities; the protein that has minimum mean GFP intensity for all identified cells correspond to 45% of the median of the mean GFP intensities in the strain collection. While this occludes protein of low abundance, the contribution is systematic to all strain, so that comparison of protein expression should be minimally affected. For now, this contribution will be ignored, but will be critical for the interpretation of certain quantities, such as stochasticity (Chapter III).

1.1.2 Higher Moments of Intensity Distribution

I will briefly introduce the measurements make on protein expression, as they are captured in the context of identified objects. The first type of measurements relates to the intensity distribution of the pixel from each cell (Figure II.1). The interest of higher moment in the intensity distribution is to capture uneven level of intensities captured within a cell. These include the variance, skewness and kurtosis of the intensity distribution. If the intensity distribution were a mixture of two normal distributions, high kurtosis is an indication the two classes of intensities significantly differ, and skewness determines if the majority of the

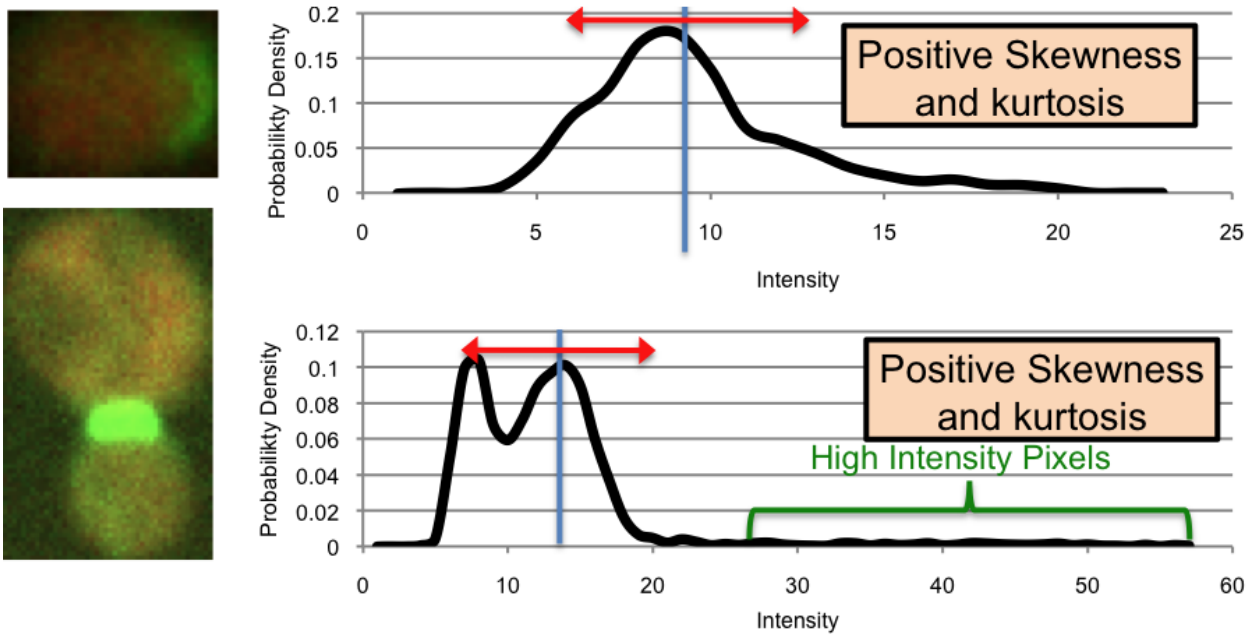


Figure II.1: **Pixel intensity distributions.** Example of intensity distribution captured from the strain DNF2 and LRG1. In both case, the protein expression appear to display a region of high protein abundance, but the distribution of high intensity pixel are different and arguably would not be best modeled by a normal distribution.

pixel are bright or dark. In practice, the intensity distribution does not match such a mixture of two Gaussian distributions due to intensity gradients over the cell, yet the relative abundance and intensity of bright pixel will be captured nonetheless; one attractive feature of higher moments is they can be computed regardless of the true underlying pixel intensity distribution.

1.2 Morphological Distances in Protein Expression

We define an additional set of 5 distance measures that characterize the spread of the protein within the cell. Assuming GFP intensities are proportional to protein amount, we can define the probability that a randomly chosen protein is located at a certain pixel coordinate as the fraction of protein found in that pixel. We compute this at each coordinate $\vec{x}$ as the ratio of pixel intensity, $G(\vec{x})$, to the sum of the pixel intensities for that particular cell $T_G = \sum_{\vec{x} \in S} G(\vec{x})$, where '$S$' is the set of pixel coordinates that are within the area of a cell. Using this probability distribution over coordinates $\vec{x}$, we derive the expected value for geometrical distances with respect to the position of a randomly selected protein. For example, for a pixel at coordinate $\vec{x}$, the distance to the cell center is given by $||\vec{x} - \vec{c}||$. Therefore, we can define the expected distance of protein to the cell center:

$$E(Dist_{\text{Proteins to Cell Center}}) = \sum_{\vec{x} \in S} ||\vec{x} - \vec{c}|| \frac{G(\vec{x})}{T_G} \tag{23}$$

Similarly, we define the average distance between proteins, to the protein mass center, to the cell center, to the cell periphery, and to the bud neck (see 'Protein expression measurements' in Methods). We refer to these measurements as expected distances, but it is important to note that they are actually estimates of protein proximities in 2-dimensional images and do not necessarily reflect the true 3-dimensional proximities. Nevertheless, these distances are easily interpretable summaries of protein expression patterns. In order to compare these expected distances between objects of different sizes, we also compute the distances for the RFP signal in each object and use these to normalize the distances obtained for the GFP signal (eq. 24). We report the log-ratio of the expected distances, so that a negative value implies that distances are smaller for the GFP-tagged protein than the approximately uniformly expressed RFP and a positive value indicates that distances are greater for the RFP than for the GFP-tagged protein. While distance log-ratios are dimensionless quantities, we refer to these 5 ratios as 'morphological distances' to emphasize that they measure the spatial spread in GFP intensity within each cell. For example, the 'morphological distance' to the bud neck, $f_{\text{bud neck}}$ is defined as:

$$f_{\text{bud neck}} = log_e \left(\frac{\sum_{\vec{x} \in S} ||\vec{x} - \vec{bn}|| \frac{G(\vec{x})}{T_G}}{\sum_{\vec{x} \in S} ||\vec{x} - \vec{bn}|| \frac{R(\vec{x})}{T_R}} \right) \tag{24}$$

where $\vec{bn}$ is the coordinates of the bud neck, $T_R = \sum_{\vec{x} \in S} R(\vec{x})$ and $log()$ is the natural logarithm.

Some of the morphological distances require us to identify the coordinates of points of interest; the cell center, protein mass-center and bud-neck position are obtained by averaging the coordinates of the cell pixels, GFP intensity and Mother-bud separation contour pixels, respectively. Assuming GFP intensities are proportional to protein amount, we derive the expected value for geometrical distances with respect to the position of a randomly selected protein. The position of cell center, protein mass center and bud neck are given by:

$$\vec{c} = \frac{1}{|S|} \sum_{\vec{x} \in S} \vec{x} \quad , \quad \vec{mc} = \sum_{\vec{x} \in S} \vec{x} \frac{G(\vec{x})}{T_G} \quad , \quad \vec{bn} = \frac{1}{|Sep|} \sum_{\vec{x} \in Sep} \vec{x} \tag{25}$$

where $T_G = \sum_{\vec{x} \in S} G(\vec{x})$ is the sum of GFP intensities and 'Sep' is the set of contour pixels that separates the bud from the mother cell objects. The other two distances have a slightly different form: first, the distance to the perimeter for any coordinate has been computed using Edge Map distance, so that:

$$E(Dist_{\text{Proteins to Periphery}}) = \sum_{\vec{x} \in S} D_{edge}(\vec{x}) \frac{G(\vec{x})}{T_G} \tag{26}$$

Second, we derive the equation for the expected distance between proteins:

$$E(Dist_{\text{between Proteins}}) = \sum_{\vec{x} \in S} \sum_{\vec{y} \in S} ||\vec{x} - \vec{y}|| \frac{G(\vec{x})}{T_G} \frac{G(\vec{y})}{T_G} \tag{27}$$

1.2.1 Normalization

One appeal of such morphological measurements is that they are inherently rotation invariant. One other desirable property is scale invariance, which enables us to recognize that two cells of different size exhibit the same pattern. To that aim, I again use the RFP marker to normalize these expected distances; the log-ratio of expected distances for the GFP and RFP signal is computed (see figure II.2). In the case of distance between proteins, the distance is normalized by the expected distance between a protein and a RFP marker. For that case, the reported log-ratio representing a morphological distance would be:

$$f_{\text{between Protein}} = log \left(\frac{\sum_{\vec{x} \in S} \sum_{\vec{y} \in S} ||\vec{x} - \vec{y}|| \frac{G(\vec{x})}{T_G} \frac{G(\vec{y})}{T_G}}{\sum_{\vec{x} \in S} \sum_{\vec{y} \in S} ||\vec{x} - \vec{y}|| \frac{G(\vec{x})}{T_G} \frac{R(\vec{y})}{T_R}} \right) \tag{28}$$

To analyze and display the morphological distances extracted for each cell for each GFP-tagged strain, we averaged the log-ratios over the cells of each type (weighting cells by their cell probabilities) and display these averages as a heat map (e.g., Figure II.3). In these heatmaps, red indicates positive values (i.e., on average greater values for the GFP-tagged protein than for the RFP) and green indicates negative values

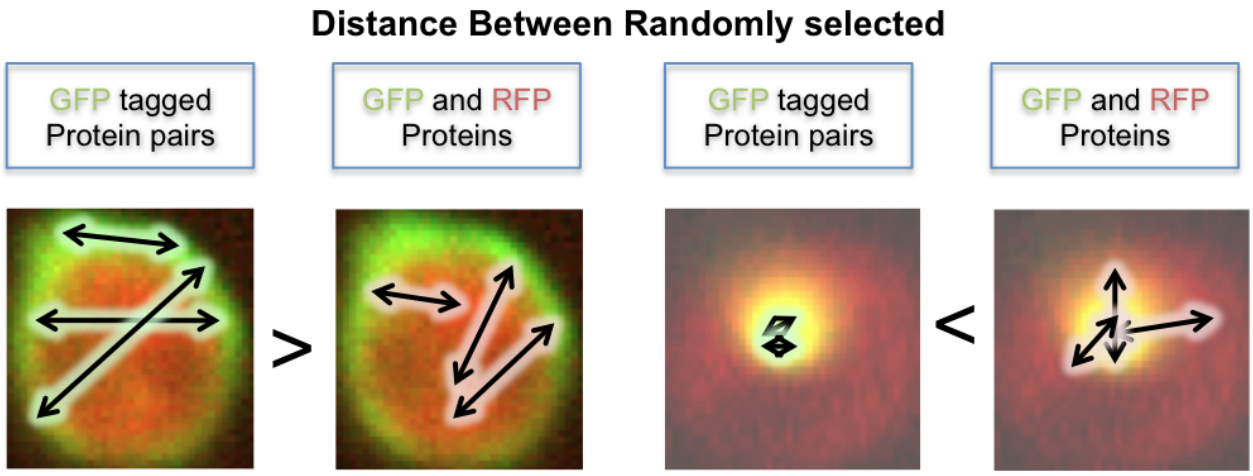


Figure II.2: **Definition of 'Morphological distances' .** 2D Distance Measured between randomly selected GFP tagged proteins (green). Alternatively, the distance between a randomly selected GFP tagged protein and a RFP protein. Two different motifs can be characterized by either larger expected distances to the RFP channel (left panels), and smaller expected distances (right panels).

(i.e., on average smaller values for the GFP-tagged protein than for the RFP).

To illustrate the use of our morphological distances, we clustered the GFP-tagged strains using averages of the 4 distances (see 'Protein expression measurements' in Methods) for each of the 3 types of cells. As expected, clusters of proteins that show the smallest relative distance (i.e., closest) to the cell center were previously reported to be localized to the nucleolus and, on the other hand, the proteins displaying a large relative distance to cell center were previously reported to localize to the cell periphery (Figure II.3). In contrast, if we consider the distance to the cell periphery, we see the opposite pattern, where nucleolar proteins show maximum distances, and cell-periphery proteins show minimum distances. This illustrates that the values we obtain for these expected distance features are related in a relatively simple way to spatial expression pattern of the protein. We note that this result does not imply that the morphological distances are superior to previously defined image features [65, 109] with respect to classifying subcellular localizations; in fact, simple classifiers based on the morphological distances are less accurate (see Section 4).

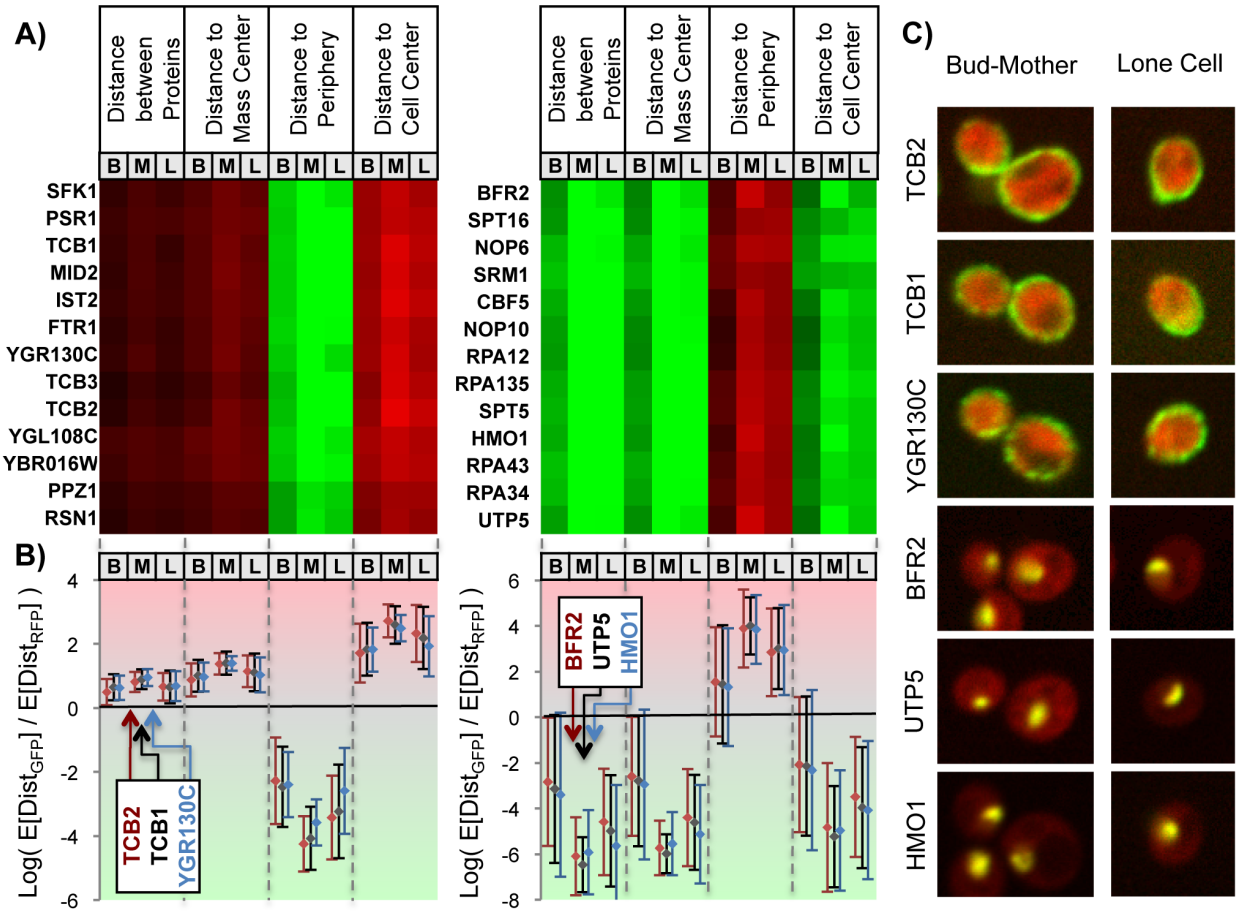


Figure II.3: **Morphological distance.** s a) Heatmap of the mean morphological distance features for each of the 3 cell classes automatically labelled: 'bud', 'mother' and 'lone' (columns indicated by 'B', 'M' and 'L' respectively). The proteins at the two extremes are enriched in cell periphery and nucleolus proteins. b) Three examples of the morphological distances extracted from the heatmap. Although the heatmap only shows the mean, we also compute the standard deviation (error bars). c) Examples of cells from the strains indicated in b). The spread of GFP fluorescence is greater than the RFP for the first three proteins, and less than RFP for the last three.

2 Inference of Time-Profiles of Protein Expression

Now that measurements of interest are available for each cell, we cumulate the shared information into a coherent model, which will capture the cell-stage dependencies of measurements. The purpose of defining cell cycle expression models is to enable pairwise comparisons of protein expression patterns that are not biased by the number of identified objects or their particular distribution in terms of cell-stage estimates.

I first present a simple approach that involves defining bud size bins, so that the averages of cell measurements that have been assigned to each bin characterize the cell-stage dependency of protein expression. Due to issues related to sampling variance, I next present an alternative approach that uses local regression [95] (LOESS) to define time-profiles.

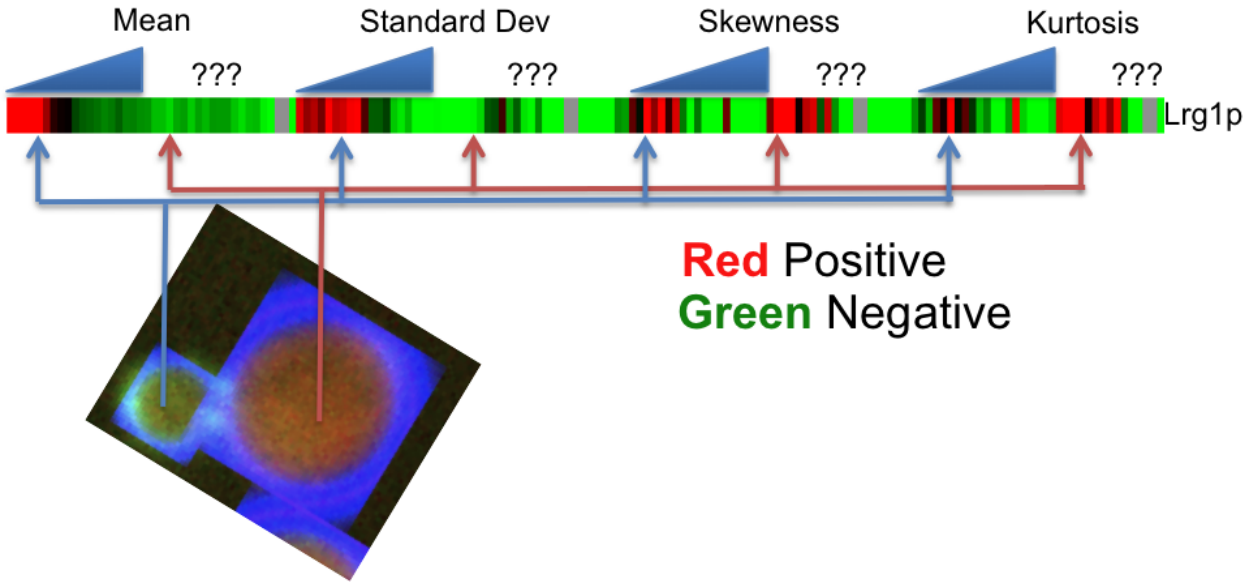


Figure II.4: **Protein 'Time-Profile'.** Time series of protein expression are inferred from a population of identified cell using their associated cell type, estimated cell-stage and cell confidence. We concatenate time series of different cell type and feature measurement in a single vector (4 'Bud' time series and 4 'Mother' time series interleaved). We refer to such vector as a 'Time-Profile', which characterizes the protein expression and its temporal and spatial component, as they captured from our choice of cell cycle model and image feature selection. Triangles represent the size of object associated with bins, which is unknown in the 'Mother' time series.

2.1 Binning

A simple approach involves assigning each cell measurements to different bins, depending on the cell type ('Bud' vs. 'Mother') and/or on cell-stage, here characterized by defining ranges of bud sizes for bins. One advantage or this approach is that the effect of sampling variance (variation in the number of sample in each bin) can be well characterized, which is desirable if one wants to test hypotheses based on a set of observations. One disadvantage is that the resolution of the bins is limited by the abundance of samples, and that poor choice of ranges of bud sizes may cause a fraction bin in profiles to have little or no samples assigned to them.

As a first attempt, I defined a 'time series' of 20 bins of protein abundance in bud objects, and a 'time series' of 15 bins for their corresponding 'Mother' cells. In this analysis, I allowed 'lone' cells to contribute in the 'bud' time series, as the separation of mother and bud cell makes 'bud' become small 'lone' cells under the heuristic cell type assessment. If a cell is assigned in the i^{th} bin of the 'bud' time series, its 'mother' cell, if it exists, is assigned to the i^{th} 'mother' time series. The bin sizes are unequal; the intervals are selected so that the previous assumption for bud growth rate (Section 2.1) makes them of equal size in theoretical time. The number of bins differs in the two time series, because buds were infrequently detected with sizes beyond 1100 pixels. Concatenated 'time series' of feature measurements will be referred to as 'time-profiles' (see Figure II.4).

In a preliminary analysis, I could show that the intensity moment features (Section 1.1) allowed the recovery of major subcellular localization classes from an unsupervised analysis (Figure II.6). I had to select a clustering strategy that can be performed with occurrence of missing values [162]. In this case, I used the 'C Clustering Library ' [42], which systematically ignores missing entries in the metric calculation, so that time-profiles with a high number of missing entries are closer to each other under a correlation or Euclidean metric, as the distances are computed in the subspace where each corresponding bins have some data. Hence, a typical behaviour is that proteins with many missing data points will cluster together. Still, significant subcellular localization enrichments were detected for protein within selected clusters (Table II.1).

Location	Bud	Nucleolus	Mitochondia	Cytoplasm	Vacuole	Vac. Membrane
Fold Enrichment	7.8	3.2	5.1	1.6	16.8	5.6

Table II.1: **Enrichment of subcellular localization.** Fold enrichment in subcellular localization within selected clusters within the hierarchical clustering.

While a large number of mother-bud pairs have been identified, for 50 out of the 4004 strains, less than 10 pairs were identified (Figure II.5). Defining a large number of bins would make means and variances that are

subject to significant sampling noise. Since many strains had a large number of mother-bud pair identified, it would be preferable that the resolution of cell-stage dependence is not dictated by the strains that have little data. In the next section, I used local regression to tackle the heterogeneous sampling variance contribution on the collection of time-profiles.

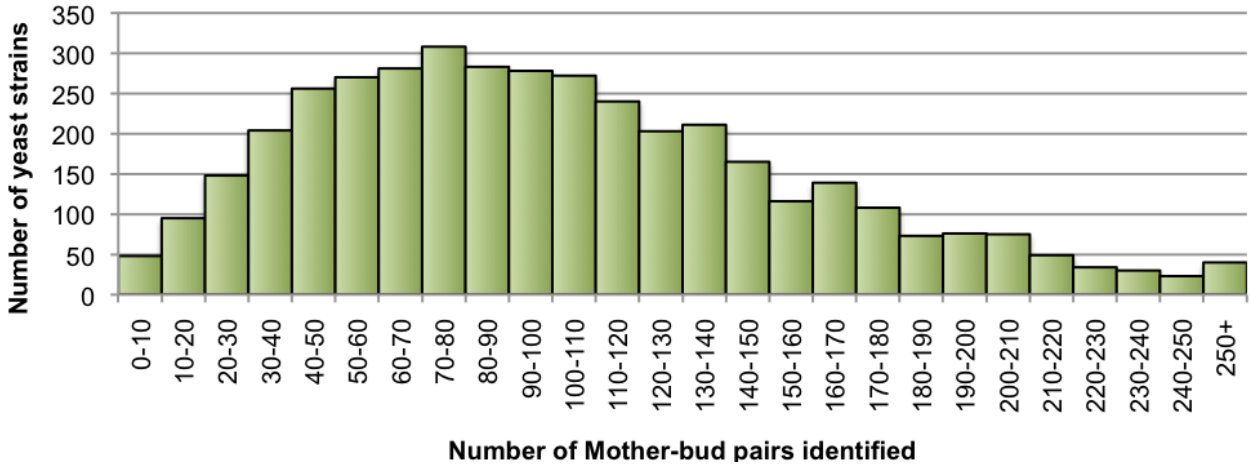


Figure II.5: **Number of mother-bud pairs identified per yeast strain.** Histograph of for the number of identified mother-bud pairs ($\mu = 102$).

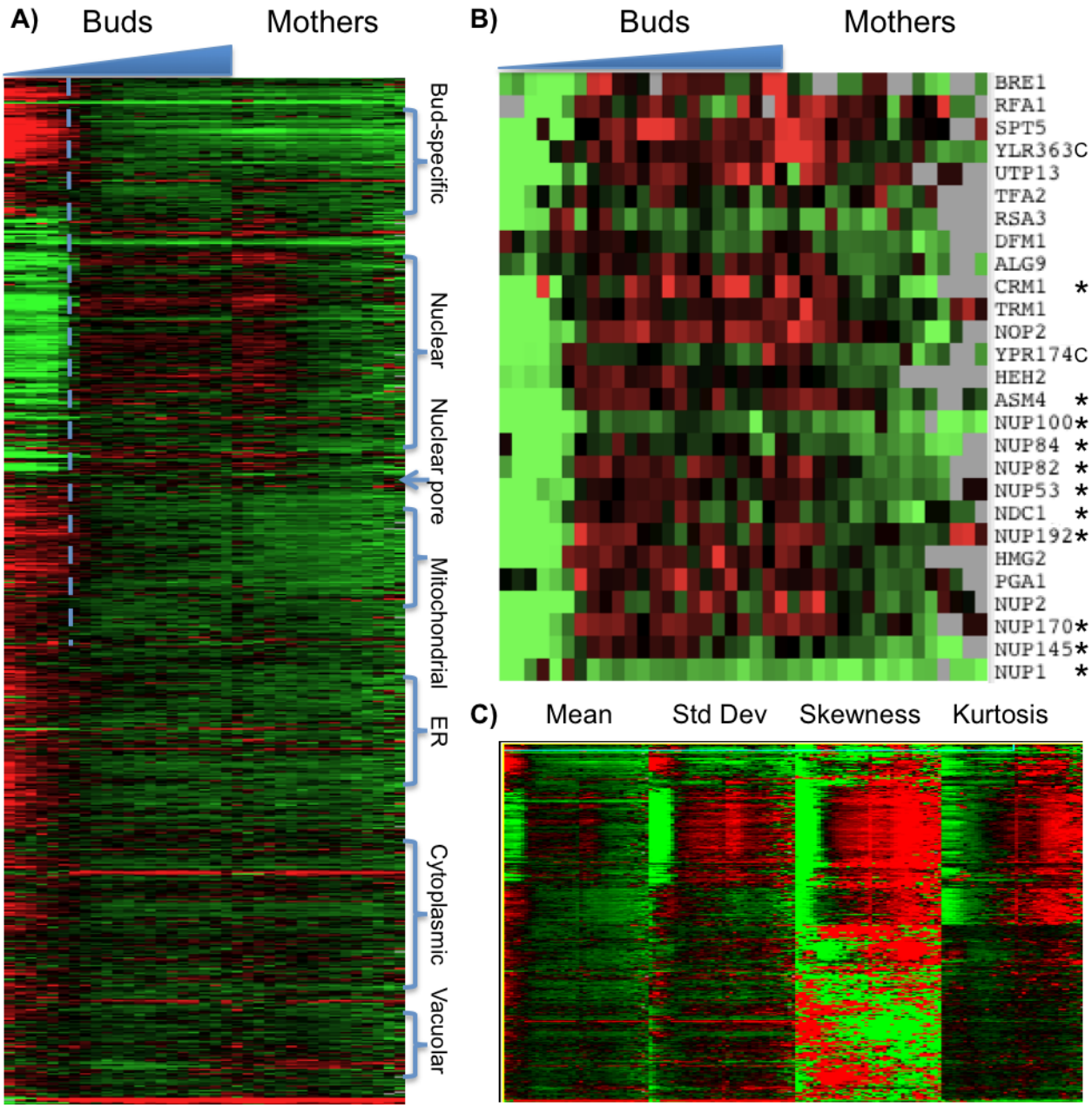


Figure II.6: **Hierarchical clustering of the binning of intensity distribution moments.** A) Heatmap of cell cycle deviation normalized and the mean level of GFP to RFP ratio B) Example of small cluster, showing nucleopore proteins (* GO annotation for Nucleopore [8]) C) Heatmap of the full cluster, which includes the higher intensity moment that were used by the hierarchical clustering.

2.2 Local Regression

As previously mentioned, we use a bud growth rate assumption to infer a quantitative representation of cell-stage. Under the assumption that the bud volume increases as a constant rate, we expect that time$^2 \sim |S|^3$, where $|S|$ is the number of pixel within the cell area. Hence, we define the numeric representation of cell-stage scales with $|S|^{\frac{3}{2}}$.

We then define a common basis to enable comparisons between expressions of different proteins in a similar manner to cell-stage bins. Each time series we define are expected feature values inferred from local regression for objects observed at 10 equidistant cell-stage keypoints $c_0, ...c_9 = \{0, 4465, 8930..., 37485\}$. We use local regression (LOESS) to infer the mean and variance at each keypoint (eq. 29). In addition, because we have developed a probabilistic cell confidence, which assigns each identified cell a posterior probability of being a cell (and not an artefact), we use the cell confidence to compute a weighted average, which is the expected profile conditioned on each identified object being drawn from the cell class:

$$\overline{F_j(c_j)} = \frac{\sum_{i=0}^{n-1} f_i P(Cell|\vec{q_i},|S_i|) K_h(|S_i|^{\frac{3}{2}} - c_j)}{\sum_{i=0}^{n-1} P(Cell|\vec{q_i},|S_i|) K_h(|S_i|^{\frac{3}{2}} - c_j)} \quad , \quad Var(F(c_j)) = \frac{\sum_{i=0}^{n-1} f_i^2 P(Cell|\vec{q_i},|S_i|) K_h(|S_i|^{\frac{3}{2}} - c_j)}{\sum_{i=0}^{n-1} P(Cell|\vec{q_i},|S_i|) K_h(|S_i|^{\frac{3}{2}} - c_j)} - \overline{F_j}^2 \tag{29}$$

where $\overline{F_j(c_j)}$ is feature value that is expected at the cell-stage keypoint c_j from feature values $\{f_0, f_1, ...f_n\}$, which are measured for the n identified object. $\{\vec{q_0}, \vec{q_1}, ...\vec{q_n}\}$ are the quality measures for each shape and $\{|S_0|, |S_1|, ...|S_n|\}$ are cell sizes. Finally, the kernel function is:

$$K_h(x) = e^{\frac{-x^2}{2h}} \tag{30}$$

The bandwidth parameter 'h' (in Eq 29) may either be fixed for all proteins, or be adapted for each protein. The selection of the bandwidth affects the sampling variance of inferred time series of mean and variance of feature measurement. If too small, the sampling variance increases so that time-profiles cannot be reproducible; if too large, the time dependence of protein expression may be obscured. The aim is to compare of protein expression independently on the amount of identified object and independently of the nature or class of the protein expression. As such, a fixed bandwidth value, which was selected to be equal to 1700, was selected for the Gaussian kernel in this chapter.

2.2.1 Sampling Variance

Single cell measurements, which refer to image features evaluated on individual identified foreground objects, are used to infer image features as a function of cell-stage. The parameterization of the function is subject to sampling variance, which is the uncertainty associated to the number of observations used in the estimation of a parameter. In this section, I show that the previously introduced use of cell confidence in time series estimation has lower sampling variance than a common alternative, which estimates parameters by ignoring any single cell measurements that are associated to objects recognized to be artefacts. Further, sampling variance will be shown to be far lower for certain parameters that describe a fraction of the cell-stage progression.

From Section 2.3, we noted that filtering objects that had a posterior probability below 0.8 detects the majority of manually labeled artefacts. When the cell confidence measure is used to weight the contribution of single cell measurements to cell-stage time-series of feature measurements, the time-series display lower levels of sampling variance than time-series inferred by simply filtering out single cell measurements from cells with confidence below a threshold value (Figure II.7A&B).

Instead of using a constant bandwidth value in order to define time-profiles, the bandwidth may be selected using the total number of identified cells. A closed from for the best bandwidth parameter exists when the background distribution of the cell-stages is known. Assuming that the distribution of cell-stage is the uniform distribution, this allows us to select the bandwidth as a decaying function in terms of the number of identified objects (Eq. 31; see Appendix 4). We note that the sampling variance inferred from jackknife resampling is lower, especially from protein profiles with few cell detected (Figure II.7C). Still, the fixed bandwidth is used the rest of in this chapter.

$$h = \frac{1}{12} \cdot \left(\frac{3n}{4}\right)^{\frac{-2}{5}} \tag{31}$$

It can be further shown that cell-stage specific bias in the estimation of cell confidence, which was noted in the last chapter (Figure I.17), causes higher sampling variance for corresponding cell-stages in time-profiles under the filtering scheme than under the weighting scheme. We observe that the cell confidence weighting reduces significantly the sampling variance in the first 4 cell-stage keypoints, where buds have low confidence (Table II.2). This is consistent with the observation that cell confidence is biased be low for small buds, since early cell-stage pairs would be preferentially excluded by the threshold method.

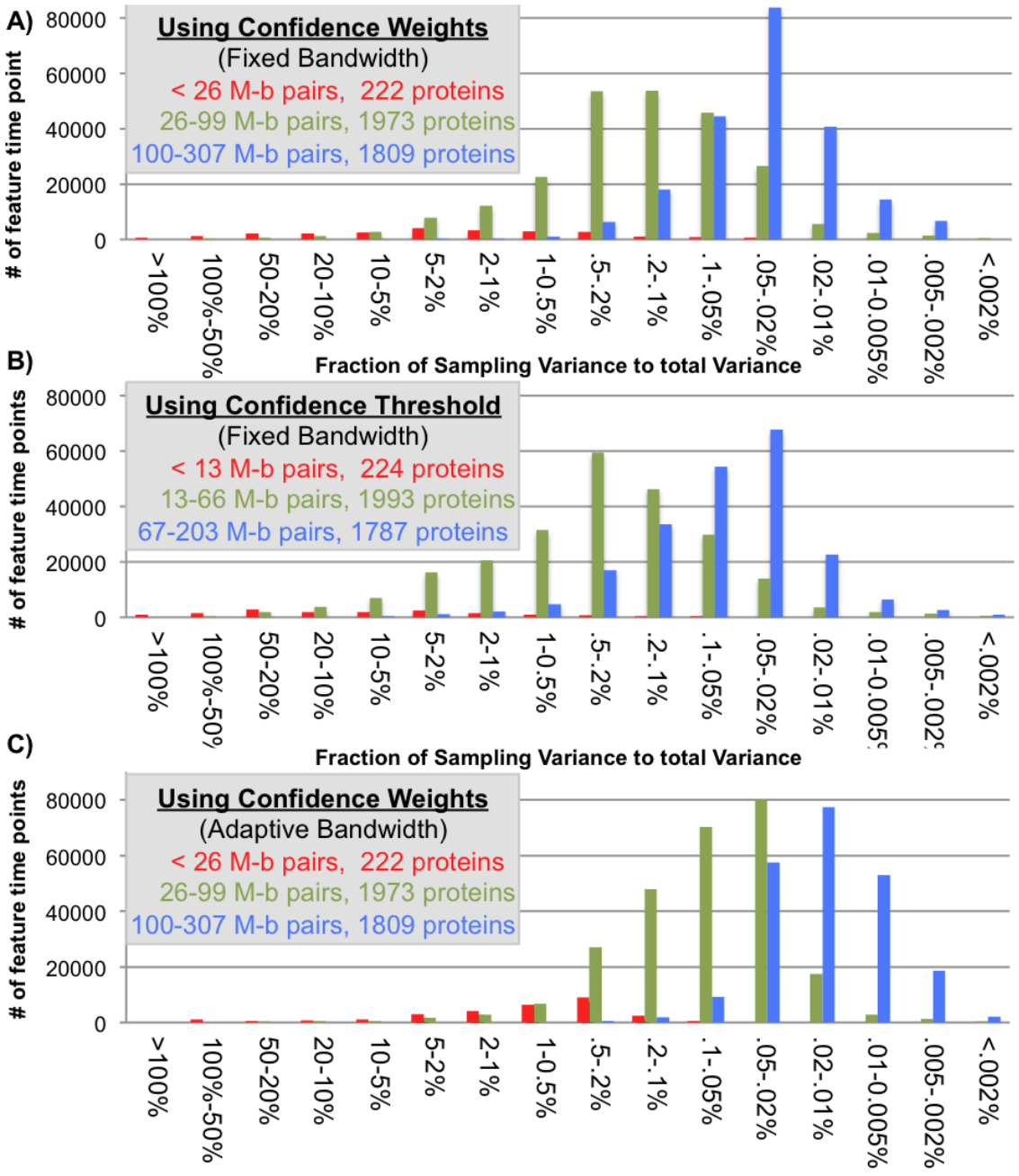


Figure II.7: **Global evaluation of the robustness of time-profiles.** A) We used the Jackknife [46] estimate of sampling variance on time-profiles that are computed from local regression (LOESS [95], eq. 29). The measured variances were normalized by the total cell-to-cell variance in the corresponding feature, so the robustness of all the 4004 x 10 x 6 x 2 time points is presented. The number of mother-bud pairs identified, which varies from protein to protein, affects the robustness of estimates. B) Instead of cell confidence weighting, any cell that had a cell probability below 0.8 was ignored from the analysis. Hence, all Mother-Bud pairs that have high enough confidence for both objects equally contribute to the time-profile estimation. The jackknife estimate reports slightly higher levels of sampling variance overall using the hard threshold. C) Adaptive bandwidth selection allows a significant reduction in sampling variance, even when little data is available.

Cell-stage	Intensity		Spread		MassC. D.		Perif. D.		Center D.		BudN. D.	
Key-point	Bud	M.	Bud	M.	Bud	M.	Bud	M.	Bud	M.	Bud	M.
1	5.78	5.07	3.31	4.36	3.63	4.10	3.53	4.25	3.44	4.32	4.08	4.55
2	3.05	2.70	2.72	2.31	2.46	2.12	2.01	2.07	2.29	2.05	2.67	2.81
3	2.04	1.91	1.89	1.78	1.86	1.84	1.76	1.80	1.85	1.77	1.50	1.45
4	1.57	1.72	1.52	1.54	1.54	1.65	1.54	1.64	1.59	1.64	1.31	1.41
5	1.62	1.71	1.52	1.54	1.54	1.61	1.51	1.58	1.52	1.59	1.48	1.45
6	1.49	1.60	1.43	1.32	1.43	1.36	1.37	1.33	1.43	1.33	1.41	1.41
7	1.39	1.51	1.06	1.35	1.11	1.39	1.13	1.41	1.27	1.42	0.73	1.23
8	1.52	1.57	1.28	1.24	1.30	1.28	1.23	1.38	1.27	1.37	1.29	1.25
9	1.57	1.40	1.46	1.27	1.46	1.45	1.40	1.36	1.37	1.24	1.31	1.08
10	1.73	1.71	1.63	1.58	1.62	1.65	1.56	1.57	1.56	1.55	1.51	1.45

Table II.2: **Comparison of sampling variance using Jackknife resampling.** Fold Difference in sampling variance from the use of confidence threshold (0.8) to confidence weighting. The sampling variances are estimated for each of the 10 cell-stage key-points; then, the average of ratios for the 1800 proteins that have the most identified cells are reported(≥ 100 cells for confidence weighted, ≥ 65 cells for threshold respectively). The compared sampling variance is extracted for each of the 10 cell-stage keypoints, and 6 feature measurements. A fixed bandwidth of 1700 is used in both cases. The first cell-stage key-point has sampling variances for threshold based profiles that are at least 3 times larger than for the confidence weighted time-profiles.

2.3 Detection of Cell Stage Dependencies

The image collection contains yeast cells that have a fluorescent marker that reports for the position of certain proteins of interest. For each of the 4000 protein types imaged, time series of average single cell measurements as a function of cell-stage estimate (referred to as time-profiles of protein expression) capture cell stage dependence of at least 8 proteins, whose abundance or subcellular localization have been previously identified (in biomedical literature) to be cell-stage dependent. In this section, I present these examples.

We examined the GFP intensity 'time series' (estimated as described above) for proteins whose quantity is known to vary over the cell cycle (Figure II.8). For example, Cdc6 [45], Sic1 [164] and Ash1 [94] have been reported to be targeted for degradation by the SCF, an ubiquitin ligase that degrades target proteins at the G1/S transition [89]. Remarkably, these three proteins show similar variation in their intensity profiles, supporting the idea that our estimates of GFP as a function of cell-stage are reflecting underlying biological variation in protein abundance. To test the statistical significance of these observations, we randomly permuted the cell-stage estimates and recomputed the 'time series'. We found that the coherent variation in the 'time series' estimated from the real data far exceeds what is typically observed in the permutations (Figure IV.3). For example, for Cdc6, of the 6 of 10 points in the bud 'time series' and 4 of the 10 points in the mother 'time series' fall within the 5% tail of the distribution observed in the permutations (compared to 1 expected to fall in the 5% tail by chance). In all, for these three proteins 26 of 60 time points fall in the 5% tail (compared to 3 expected by chance). This shows that for these proteins whose levels are known to

vary over the cell cycle, the variation observed in the 'time series' is statistically significant.

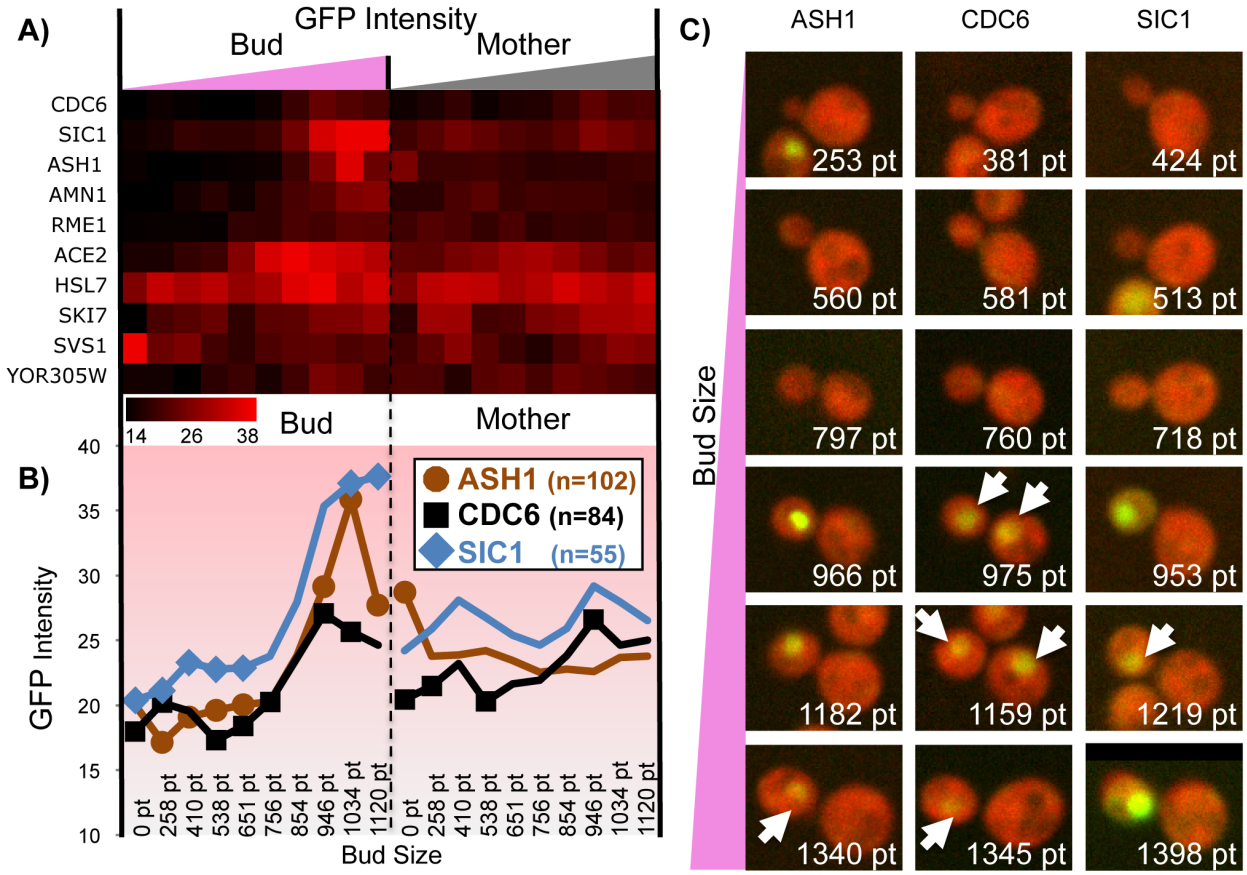


Figure II.8: **Intensity as a function of cell-stage estimate.** A) GFP intensity heatmap for several protein whose abundances are known to be cell cycle dependent. B) Profiles for 3 proteins showing significantly higher expression level in large buds. 'n' is the number of mother-bud pairs used to infer each time series. 26 out of the 60 time points (indicated with markers) show coherent cell-stage specific deviations (permutation test, See Fig IV.3). C) Examples of mother-bud pairs with the computed pixel size (pt) of the bud object (identical RFP/GFP intensity scale). The displayed cells were manually selected and then ordered by the computed bud size. Arrows indicate nuclear localization at lower intensity.

We estimated 'time series' for each of our 5 morphological distances and GFP intensity as described above for all of the bud and mother cell pairs. For each protein, we concatenate the 6 pairs of 'time series' into a 'time-profile', which is a vector of 120 values. An example of a striking cell cycle pattern is the profile observed for the subunits of the MCM complex (Figure II.9), which is known to be exported from the nucleus at a particular cell-stage by the activity of Clb/Cdc28 kinases [114]. This exclusion from the nucleus is captured by the distance features, since the protein gets closer to the cell periphery and, on the other hand, the average distance between proteins and to the cell centre increases. This exclusion is

observed in the mothers of small buds, so we can determine the size of the bud corresponding to the G2 to M transition, when the MCM complex nuclear localization signals are no longer specifically inhibited by Cdc28 (see figure 5B). Encouragingly, all 4 available members of this complex show this pattern (2 are missing from the GFP collection). This indicates that proteins displaying similar cell-stage variation can be identified from their time-profiles, despite the presence of noise in the images and heterogeneity in the distribution of identified cells on which the time-profiles are based. Remarkably, we observe that another protein with a similar stage-dependent morphological distance profile is also known to have its localization is modulated by Cdc28 (Whi5 [38], see figure 5). Upon examination of the images, we observe a very similar expression pattern in bud cells for Whi5 and the MCM subunits, but that (in contrast to the MCM subunits) Whi5 nuclear localization is only rarely found in mother cells (Figure II.9). This demonstrates the capacity of the generated profiles to capture cell cycle dependence of changes in localization. Furthermore, that these proteins are all substrates of Cdc28 suggests that similarity in our profiles of morphological measurements may indicate common mechanisms that control subcellular localization, just as similar mRNA expression profiles are often used as evidence for common mechanisms of transcriptional control [47, 149].

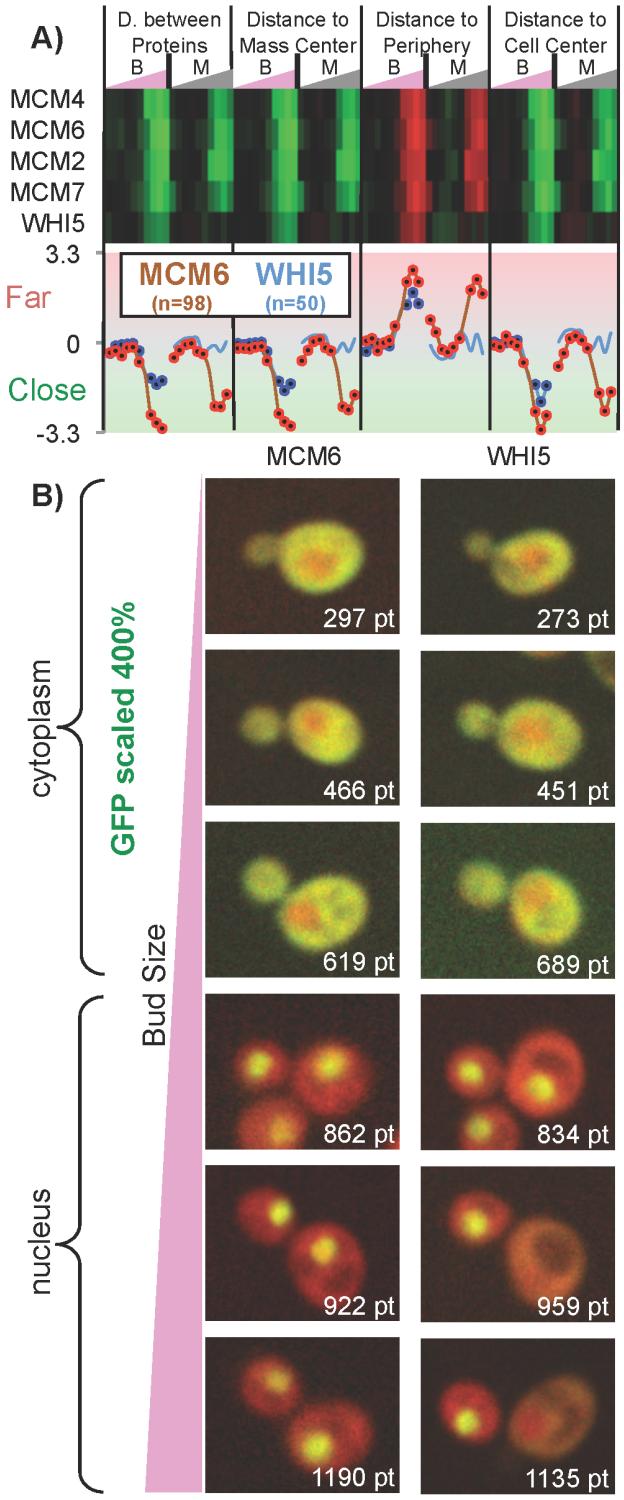


Figure II.9: **Time-profiles of morphological distances.** A) Top panel shows a heatmap of the morphological distances in bud and mother cells indicated as B and M, respectively. Bottom panel shows the data for two of these proteins as line graphs. The reported morphological distances are variance normalized. MCM complex subunits and Whi5 display a cell cycle dependent subcellular localization; cytoplasmic for small buds, nuclear for large buds. 'n' is the number of mother-bud pairs used to infer each time series. Out of the 80 timepoints for each protein, 34 for Whi5 (blue traces), and 72 for Mcm6 (red traces) show significant cell cycle variation ($P < 0.05$, indicated as dark dots). B) Examples of mother-bud pairs that were ordered by the computed bud size (pt). The GFP channel was scaled between images to more clearly illustrate the change in subcellular localization.

2.4 Biological Results

Time-profiles, here defined from 6 time series of single cell measurements, provide a quantitative representation of subcellular localization, which allows the definition of similarity metrics of protein expression that agree with the categorical representation from Huh et al. 2003 [76]. In the following sections, I present statistical support for this claim, and then describe the cell-stage progression patterns that are associated to subcellular localizations.

2.4.1 Similarity between profiles reflects biological relationships of subcellular localizations.

To get a global sense of whether the profiles in our biologically interpretable feature space reflect the biological similarity of protein expression patterns, we computed the average profiles for all the proteins within previously identified subcellular localization classes [76] (see 'Class profiles' in Methods). Because each profile represents a multivariate Normal distribution, where we estimate mean and standard deviation over the observed cells for 10 time points for each of the 6 features, for the mother and bud, we measure the similarity between the mean profiles for each localization class ('class profile') using the Bhattacharyya distance (Eq. 32). Consistent with their biological relationships, we observe that the class profiles representing nuclear proteins are much closer to nucleolar and nuclear periphery localized proteins (Bhattacharyya distance=5.41,2.39) than to the class profiles for cytoplasmic or cell periphery localized proteins (Bhattacharyya distance=34.20,21.16). Clustering of these class profiles placed several biologically related classes adjacent to each other in the hierarchy. For example, profiles for Golgi, Early Golgi and Late Golgi were clustered together (Figure II.10). To confirm this result, for each group of biologically related classes, we compared the average Bhattacharyya distances within the groups of related classes to the distances between the classes in each group all other classes. We found that the distances between biologically related classes were significantly smaller (6.14 vs. 15.44, $P = 0.00015$, permutation test, Fig II.10). Taken together, these results show that distances in this interpretable feature space recapitulate the known biological relationships between localization classes.

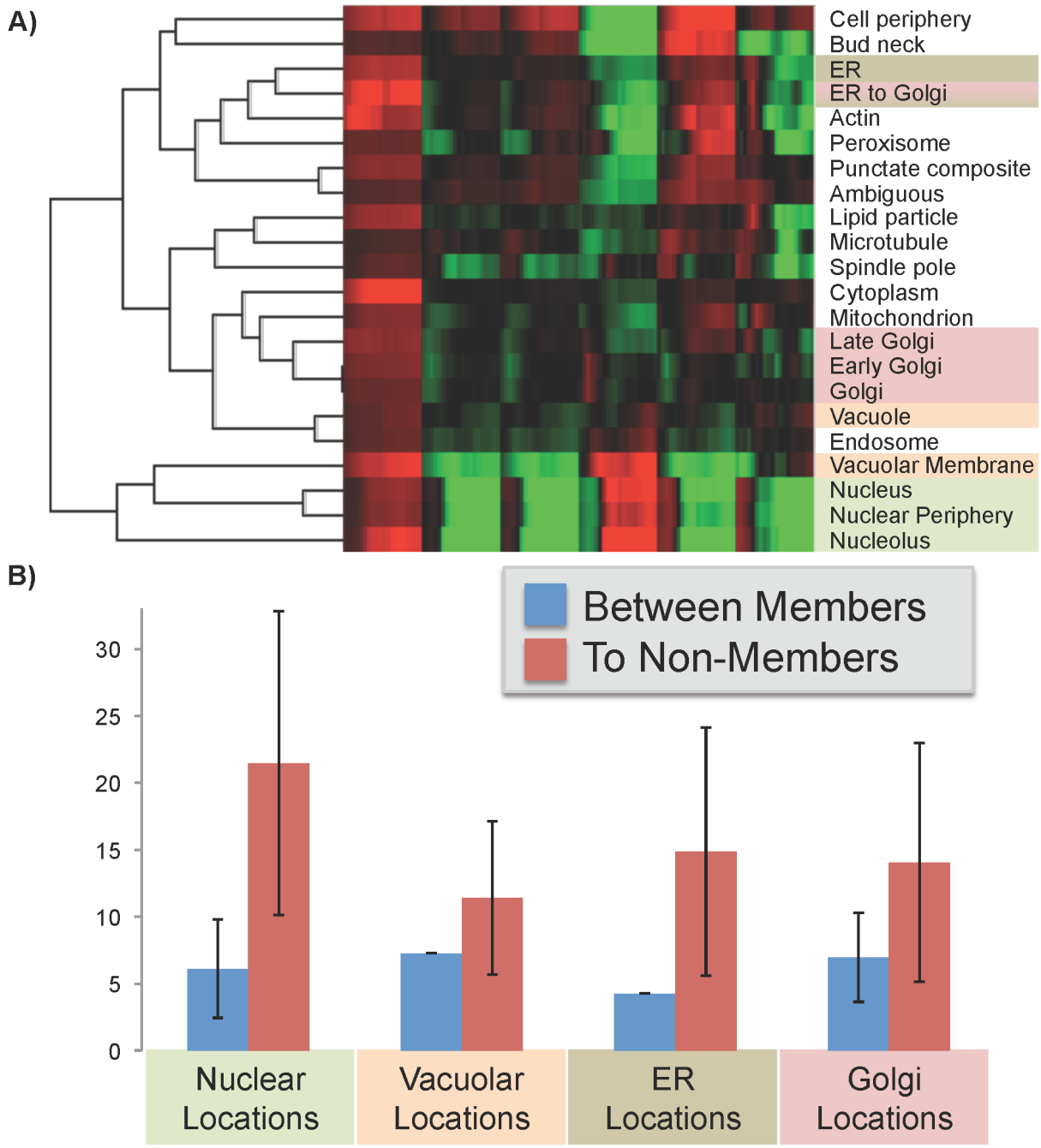


Figure II.10: **Comparison of time-profiles for different subcellular localizations.** a) Hierarchical Clustering of the class profiles based on Euclidean distance. Colours of location names indicate the 4 groups of subcellular localizations that were defined based on biological relationships. b) Average Bhattacharyya distance between subcellular localization class profiles within biologically related groups (between members, blue bars) is smaller than the average distances between these class profiles and those that are not biologically related (to non-members, red bars). We note that the sum of the difference in mean distance (difference between blue and red bars) is significantly lower than expected by chance ($P = 0.00015$, 10^6 permutations of the subcellular localizations that belong to each biological group)

2.4.2 Resolution of the ordering of inclusion into the bud for major organelles.

The unsupervised analysis of biologically interpretable features allows us to visualize a quantitative representation of protein localization over the cell cycle: we observe large clusters of proteins that appear in the bud sequentially. Most strikingly, in the clusters significantly enriched in nuclear proteins, protein expression is absent from the bud until approximately half-way through our time series. Other clusters also display cell cycle dependent variation in all morphological distances, which appears to be specific to subcellular localizations. For example, the three mitochondrion enriched clusters show signal unusually far from the bud neck at the same time. Interpreting this pattern, we predict the presence of punctae in small buds, and inspection of the images confirmed this prediction (Figure IV.4A).

In order to confirm that the observed trends in the protein profiles are truly linked to the subcellular localization of the proteins, we used the class profiles (see 'Class profiles' in Methods) for each subcellular localization (Figure II.11). We observe that proteins from the nucleus, nuclear periphery and nucleolus are the last to appear in the bud. This is explained by the fact that DNA replication is occurring within the mother cell, and that the new nucleus has yet to be included in the bud. We note that in the bud cells, mitochondrial and ER proteins show elevated distances from the bud neck at the time of nucleus inclusion, and that a subset of the mitochondrial proteins are found close to the bud neck in the smallest bud objects (Figure II.11B); this suggests that the mitochondria and ER may be included in the bud before the nucleus, and then pushed further from the bud neck as the nucleus occupies that position at the time of its entry into the bud. Interestingly, we also observe that the proteins of each organelle have typical distances in the mother cell to the current bud neck (Figure II.11C). For example, the ER has been previously reported to stay close to the nucleus [127], and we observe that both the ER proteins are closer to the bud neck than the mitochondrial proteins (-0.94 vs. 0.04, $P < 10^{-60}$, two-sample t-test) but not as close to the bud neck as the nuclear proteins (-0.94 vs. -2.25, $P < 10^{-53}$, two-sample t-test).

We also observe the motion of the actin proteins in both the bud and mother cells, which agrees with previous observations: actin proteins localize at the bud periphery and then at the bud neck [76, 137]. Since the polarity of yeast cells is determined by the cell-stage, and cell polarization is controlled via the action of the actin filaments [44], these results again indicate that our estimate of bud size is a good cell-stage indicator, and that the order of biological events may be extracted directly from the class profiles. Although these patterns were discovered through interactive exploration of a particular clustering result, we note that these patterns correspond to very strong signals in the data and were also easily identified in clustering results derived from alternative similarity metrics or alternative usage of the confidence measure (Figure IV.4B).

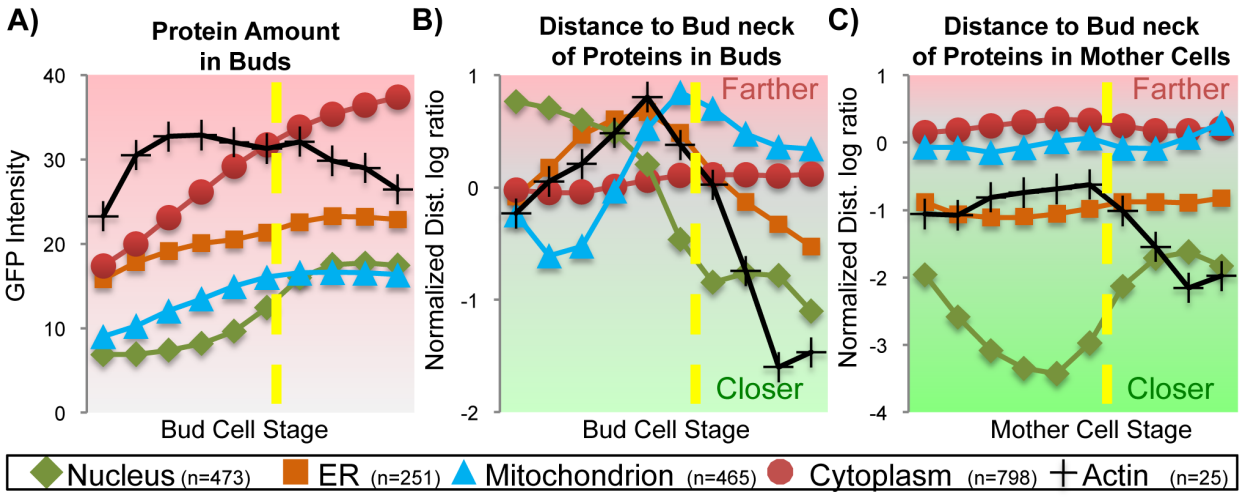


Figure II.11: **'Class profiles' for five subcellular localizations.** A) Time series for protein abundance in buds. Nuclear proteins are the last to appear in the bud (dashed line). B) The spatial distribution of protein expression is highly variable in the growing bud cell. Organelles appear to be pushed from the bud neck at the time of the nucleus inclusion (dashed line). Note that the absence of nuclear protein in the bud leads to irrelevant variations in the morphological distance features, perhaps due to auto-fluorescence captured in the GFP channel. Actin proteins migrate from bud tip to bud neck (black traces). C) In the mother cell, organelles appear to maintain a typical distance to the bud neck, expect for the nucleus. .

3 Unsupervised Analysis

Unsupervised methods offer an exploratory approach to high-throughput data analysis in which it is not necessary to predefine patterns of interest, and therefore can discover new patterns. This also enables the analysis of patterns that are very rarely observed, which typically are hard to capture in supervised analysis as a suitable training set for classification is difficult to construct [61]. Unsupervised analysis also has the advantage that it is unbiased by prior 'expert' knowledge, such as the arbitrary discretization of protein expression patterns into easily recognizable classes. For these reasons, unsupervised cluster analysis has become a vital tool of computational biology through its application to genome-wide mRNA expression measurements [47, 123, 149, 157], and protein-protein interaction data [9]. It has also been applied in automated microscopy image analysis [31, 33, 37, 50, 64] where it has been shown to provide complementary capabilities to supervised approaches.

I show that many previously defined subcellular localization patterns can be recognized in an unsupervised hierarchical cluster analysis. We find that protein complexes and small functional protein classes, which are not typically associated with their own subcellular localizations, cluster together in this analysis. Based

on these observations, I show that the resolution of the hierarchical clustering is significantly higher than previous manual subcellular localization assignments to discrete classes [76]. Further, we gain global insight into the cell-stage dependence of protein localization; for example, we find a large cluster of nuclear proteins that seem to appear in the bud at a clearly defined time, which we believe corresponds to the inclusion of the nucleus in the daughter cell. Finally, we identify groups of proteins that show complex, dynamic patterns of localization that cannot easily be predefined or described using simple localization classes; for example, many of the subunits of the exocyst complex are seen to localize to the bud periphery while the bud is small, but then move to the bud neck as the bud grows.

3.1 Metric based Hierarchical Clustering

Hierarchical clustering can be performed using various similarity metrics that are to compare time-profiles of protein expression. One peculiarity of this clustering problem is that each cell-stage keypoint has an associated mean and variance. I would like to test that the inferred variances are informative in evaluating protein expression similarity. To that aim, I am to compare metrics and clustering approaches that regroups proteins of similar subcellular localization and/or function.

3.1.1 Metric based Clustering

As previously introduced, I used 'C Clustering Library' in to render hierarchical clustering using both (One of the motivations to cover this aspect is the observed gain of using metrics that consider both the mean and cell-to-cell variability in measurements. To quantify this gain, I will present agglomerative clustering of time-profiles that are based on correlation and Euclidian distance. Then, I present clusters that are based of the Bhattacharyya metric, which is a distance measure for Normal distributions. The mean used to evaluate the quality of a cluster resides in evaluating the distribution of P-value obtained in the functional or localization annotation for proteins that are grouped. As the hierarchical clusters are to be compared, for each annotation I would find the best subpartition in the hierarchical cluster that as the best P-value.

$$D_B(\vec{\mu_1}, \Sigma_1, \vec{\mu_2}, \Sigma_2) = \frac{1}{8}(\vec{\mu_1} - \vec{\mu_2})(\frac{\Sigma_1 + \Sigma_2}{2})^{-1}(\vec{\mu_1} - \vec{\mu_2})^T + \frac{1}{2}log\frac{det(\frac{\Sigma_1+\Sigma_2}{2})}{\sqrt{det(\Sigma_1) \cdot det(\Sigma_2)}} \tag{32}$$

Equation 32:**Bhattacharyya distance.** Distance between multivariate normal distributions defined using 'μ' as means and 'Σ' as covariance matrices.

3.1.2 Maximum Likelihood Agglomerative Hierarchical Clustering

Each protein profile is a vector of means and variances of observations. We use the Maximum likelihood clustering criterion [86] (eq. 33) in order to agglomeratively join pairs of protein profiles, proteins to cluster profiles, or pairs of cluster profiles.

$$\begin{aligned} &\tfrac{n_1+n_2}{2}\log\left(1+\tfrac{n_1 n_2}{(n_1+n_2)^2}(\mu_1-\mu_2)^T(\tfrac{n_1\Sigma_1+n_2\Sigma_2}{n_1+n_2})^{-1}(\mu_1-\mu_2)\right) \\ &+\tfrac{n_1+n_2}{2}\log(\left|\tfrac{n_1\Sigma_1+n_2\Sigma_2}{n_1+n_2}\right|)-\tfrac{n_1}{2}\log(|\Sigma_1|)-\tfrac{n_2}{2}\log(|\Sigma_2|) \end{aligned} \tag{33}$$

where $|\Sigma|$ is the determinant of a covariance matrix. This criterion is the log-likelihood ratio for two protein expression groups of size n_1 and n_2 to be modeled as two multivariate Normal distributions (with their corresponding parameters μ and Σ), to a single multivariate Normal model explaining both expression groups.

The input to the clustering is a collection of time-profiles corresponding to the 12 concatenated time series of feature values. The initial covariance matrix Σ for each profile is a diagonal matrix whose values were estimated from the LOESS (see $Var(F(c))$ in eq. 29). New protein profile groups are characterized by multivariate normal distribution where the parameters are obtained from the two previous merged groups (eq. 34).

$$\begin{aligned} \mu' &\leftarrow \tfrac{n_1\mu_1+n_2\mu_2}{n_1+n_2} \\ \Sigma' &\leftarrow \tfrac{n_1(\Sigma_1+\mu_1\mu_1^T)+n_2(\Sigma_2+\mu_2\mu_2^T)}{n_1+n_2}-\mu'\mu'^T \end{aligned} \tag{34}$$

where $\{\mu_i, \Sigma_i, n_i\}$ are parameters for two normal distributions that each represents 'n_i' profiles, which are to be merged next in the agglomerative clustering step.

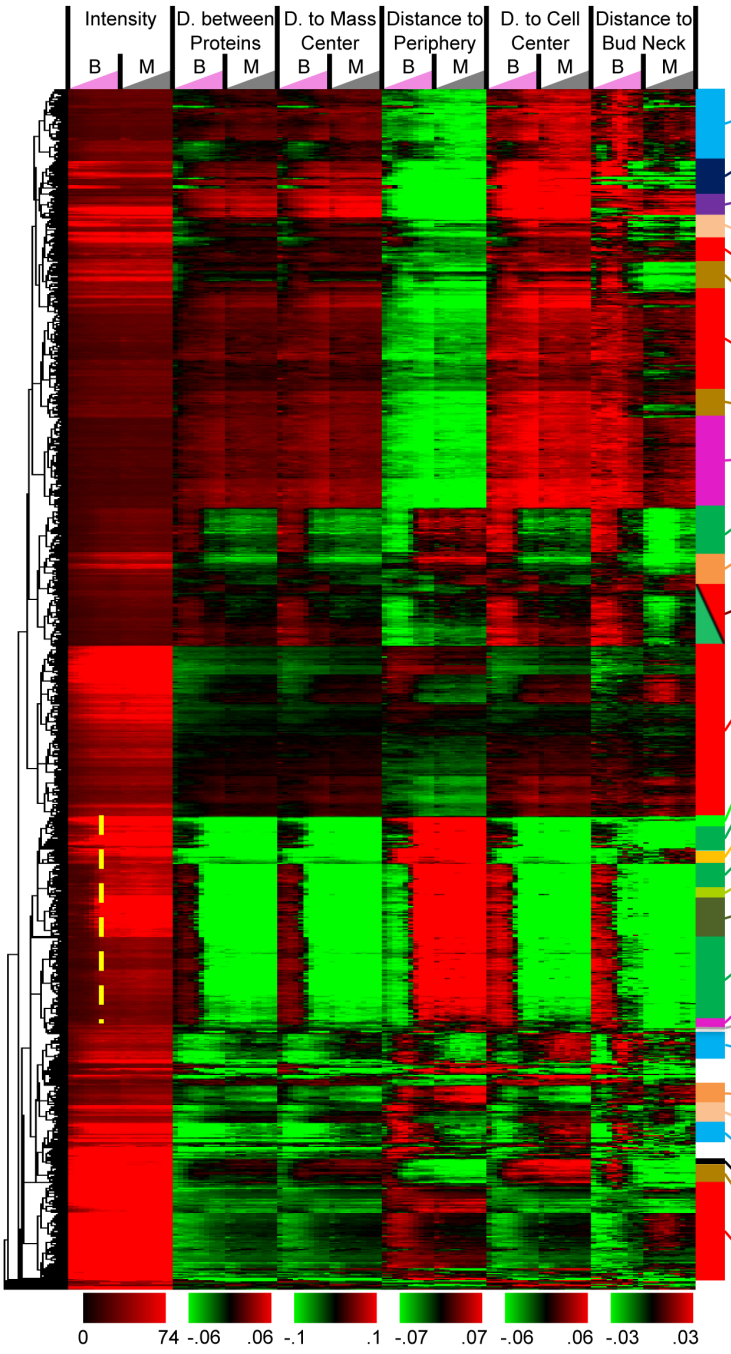
3.2 Results

Agglomerative hierarchical clustering of time-profiles produces a tree structure, where leaves of the tree are proteins. Proteins with identical subcellular localization or with similar biological functions, according to previous characterization of biological functions of yeast proteins (GO annotations [8], Pfam [13]), are significantly closer in the generated tree structure than would be expected by randomly assigning proteins to leaves. The next two sections contain the supporting evidence for subcellular localization and biological function respectively. The last section presents the clustering for an example group of proteins.

3.2.1 Enrichment of identically localized proteins

We performed a statistical enrichment analysis in order to compare our cluster analysis to previous knowledge about protein localization and function. We considered assignments of proteins to discrete localization classes from systematic manual assessments of the GFP collection [76] and GO annotations curated from the biological literature [8]. We found that many of our clusters were strongly enriched for GO annotations and previously identified subcellular localizations (Figure II.12). We note that these results were not dependent on the clustering parameters or algorithm used, as similar results were obtained using other hierarchical clustering methods (Table II.3).

In our global analysis, we also observed clusters that were statistically enriched in annotations that do not correspond to subcellular localization classes or compartments (Figure II.12). For example, translation is known to occur in the cytoplasm [39]. Nevertheless, we observe a cluster of 316 proteins where 86 (27%) correspond to structural components of the ribosome and a total of 121 (38%) are annotated as involved in translation. Consistent with the known cytoplasmic localization for the translational machinery, this cluster shows a similar overall pattern to cytoplasmic proteins, but can be distinguished because the average GFP intensity (presumably reflecting protein abundance) for these proteins is much higher than most other cytoplasmic proteins (Figure II.12). As another example, we also noticed a cluster where 16 of 43 (37%) of proteins were subunits of the proteasome. This cluster also contains 6 of 14 (43%) proteins annotated as vacuolar ATP-ases. The pattern associated with this cluster shows high levels of protein abundance and is similar to that of nuclear proteins, but this is not sufficient to explain why these complex subunits are distinguishable from the remainder of the highly expressed nuclear proteins. The localization pattern for these proteins is more compact than other nuclear proteins, and we speculate that these complex subunits display similar, typical levels of compactness and this is captured in our morphological distances (Figure II.12). These results suggest the possibility that a combination of a small number of interpretable features (e.g., cytoplasmic localization and high level of protein abundance) will define certain functional classes, which will be property evaluated in the next section.



Localization or GO Annotation (marked with *)	Fold Enrich.	Log_{10} Pvalue	Fraction
Mitochondria	6.15	-127	186/241
Bud Neck	19.21	-58.5	54/116
Cell Periphery	19.99	-63.4	55/72
Golgi	6.35	-6.3	18/97
Cytoplasm	1.86	-4.4	42/51
ER	12.05	-76.9	80/97
Cytoplasm	1.42	-6.1	162/258
ER	5.60	-14.6	35/89
Ambiguous	3.05	-10.4	53/302
Nucleus	2.32	-44.5	181/222
Vacuole	12.83	-38.1	35/70
Cytoplasm & nucleus	3.38	-38.6	113/167
Cytoplasm	2.18	-184	550/570
Proteasome*	66.51	-22.7	15/43
Nucleus	2.12	-6.6	47/63
Vacuolar Mem.	37.11	-33.8	26/51
Nucleus Only	5.93	-29.7	54/77
Nuclear Peri.	30.36	-18.6	17/38
Nucleolus	12.12	-52.5	62/128
Nucleus	2.59	-80.8	236/261
Spindle Pole	29.32	-15.7	15/32
Lipid Particle	116.0	-8.5	6/9
Mitochondrion	8.61	-81.4	87/87
Vacuole	11.45	-20.8	29/65
Golgi	23.95	-49.8	42/60
Mitochondrion	6.64	-35.1	56/73
Actin	69.03	-8.4	7/14
ER	11.39	-39.6	46/59
Ribosome*	6.15	-45.7	84/316

Figure II.12: **Time-profile clustering results.** A heatmap with 4004 GFP-tagged strains ordered using maximum likelihood agglomerative clustering based on the time-profiles of protein abundance and 5 morphological measures. Within manually selected clusters (coloured bars), the fraction of proteins in the cluster that have the same subcellular localization or GO Annotation (the latter indicated with stars) is listed under Fraction. Log-P-values were computed using the hypergeometric distribution to test against the null hypothesis that the cluster was drawn randomly from the protein annotations. Fold enrichment indicates the ratio of the Fraction of proteins in the cluster with each annotation compared to that in the protein collection. Nuclear proteins appear in the bud at a specific time (dashed line).

Localization	Confidence Weighted			Confidence Threshold (0.8)		
	MLAC	Euclidean	Correl.	MLAC	Euclidean	Correl.
ER	-79.19	-70.74	-85.83	-39.22	-66.29	-90.39
ER to Golgi	-4.23	-5.88	-5.73	-3.50	-5.73	-8.14
Golgi	-21.59	-21.64	-21.08	-9.03	-17.42	-12.27
actin	-11.95	-16.56	-14.89	-5.23	-16.60	-14.66
ambiguous	-38.70	-32.91	-31.23	-17.70	-24.13	-38.03
bud	-34.73	-22.85	-21.08	-19.06	-32.31	-31.55
bud neck	-63.20	-32.46	-23.32	-19.03	-32.33	-36.99
cell periphery	-103.04	-56.34	-24.17	-17.69	-78.74	-25.47
cytoplasm	-187.25	-75.38	-129.14	-135.21	-57.40	-111.60
early Golgi	-38.84	-19.91	-35.61	-17.51	-26.38	-17.18
endosome	-13.12	-21.85	-16.76	-4.72	-10.98	-17.97
late Golgi	-10.77	-8.11	-7.46	-12.94	-14.43	-11.67
lipid particle	-11.83	-9.34	-9.66	-4.79	-9.08	-9.50
microtubule	-5.59	-5.40	-5.85	-4.31	-9.27	-5.71
mitochondrion	-129.29	-74.39	-125.87	-41.98	-46.56	-104.22
nuclear periphery	-21.87	-17.19	-19.26	-9.06	-17.64	-26.22
nucleolus	-80.63	-84.78	-136.31	-37.36	-72.81	-125.63
nucleus	-109.27	-112.13	-209.03	-73.25	-124.37	-246.64
peroxisome	-14.66	-15.15	-14.66	-4.77	-7.20	-14.66
punctate composite	-15.06	-11.31	-11.31	-12.86	-8.90	-8.06
spindle pole	-19.04	-24.63	-37.12	-6.53	-21.25	-30.80
vacuolar membrane	-37.15	-52.45	-60.88	-17.86	-37.51	-60.57
vacuole	-34.83	-60.21	-37.96	-12.93	-35.95	-41.52
Sum of log-P-values	-1085.83	-851.58	-1084.23	-526.52	-773.27	-1089.47

Table II.3: **Log-P-value for subcellular localization enrichment.** Enrichments of proteins for subcellular localization labels, as defined by Huh et al. [76], for the best cluster within the hierarchical clustering. Maximum likelihood agglomerative Clustering (MLAC) is compared to agglomerative hierarchical clustering with complete linkage, which uses either Euclidean or correlation metric. Log_{10}-P-values are reported using the hypergeometric distribution directly (no multiple hypothesis correction applied). MLAC yeild enrichments that are comparable to the use of correlation metric, provided that time-profiles are inferred by weighing down single cell measurements by cell confidence (as opposed to filtering objects with confidence below a threshold).

3.2.2 Proteins in functional classes and complexes cluster together.

Groups of proteins have been previously characterized as sharing a biological function (2134 GO category groups, 761 Pfam groups). Given a list of protein groups, I define a statistic to evaluate the proximity of proteins from each protein group. In this section, I present a permutation test, where only leaves of agglomerative clustering generated tree structure are permuted, that indicates that the quantified proximity is negligibly likely to be measured by chance. I will further show that higher proximity than random is noted for all ranges of sizes of protein groups associated to biological function.

In order to report on the statistical significance of functional annotations in the hierarchical clusters,

Annotation		Confidence Weighted			Confidence Threshold (0.8)		
Set	#	MLAC	Euclidean	Correl.	MLAC	Euclidean	Correl.
GO	2134	-7078.582	-6636.108	-6903.946	-5228.228	-6451.375	-6711.712
Complexes	277	-988.918	-963.973	-992.304	-751.829	-926.786	-978.952
Complexes*	262	-932.945	-911.799	-943.106	-707.659	-871.988	-929.584
Pfam	761	-1661.314	-1545.283	-1562.141	-1317.446	-1585.944	-1566.065
Pfam*	688	-1490.510	-1386.628	-1396.513	-1192.673	-1355.214	-1385.060
GO 2	664	-827.963	-747.854	-805.486	-636.506	-740.678	-737.008
GO 3	337	-704.499	-674.288	-631.800	-565.320	-674.895	-681.155
GO 4-5	379	-1072.369	-1008.316	-1035.704	-846.634	-1000.144	-1015.949
GO 6-7	203	-789.636	-768.369	-747.943	-582.868	-747.549	-722.239
GO 8-9	113	-480.772	-482.741	-469.326	-352.638	-479.361	-468.683
GO 10-11	76	-360.014	-355.430	-383.526	-279.611	-363.421	-360.929
GO 12-14	82	-452.976	-412.640	-447.332	-348.773	-404.316	-422.178
GO 15-18	73	-417.326	-359.234	-420.572	-301.620	-366.033	-382.773
GO 19-24	71	-471.110	-471.681	-488.339	-332.840	-441.810	-475.002
GO 25-34	49	-340.913	-350.076	-386.002	-264.402	-327.498	-376.697
GO 35-49	32	-274.098	-244.916	-262.060	-183.540	-231.016	-254.584
GO 50-79	29	-296.519	-270.221	-286.405	-199.299	-243.172	-290.119
GO 80-200	18	-388.685	-359.921	-422.367	-194.838	-326.037	-409.087
GO 200+	8	-201.701	-130.421	-117.084	-139.340	-105.444	-115.310
Pure Loc.	22	-820.916	-719.995	-918.237	-440.193	-586.596	-878.358
Partial Loc.	23	-1085.825	-851.575	-1084.225	-526.518	-773.272	-1089.468

Table II.4: **Sum of log-P-values.** Global analysis of enrichment of functional or subcellular localization enrichment. Maximum likelihood agglomerative Clustering (MLAC) is compared to agglomerative hierarchical clustering with complete linkage, which uses either Euclidean or correlation metric. Note that only MLAC relies on variance estimates in time-profiles, which are inferred by local regression. MLAC overall yields higher levels of annotation enrichments; however, we note that filtering objects based on confidence threshold is detrimental to functional enrichments (compared to confidence weighting scheme), especially for MLAC.

for each of the 2134 GO annotations that are shared by at least two proteins, we found the cluster within the hierarchy that has the most significant P-value. We used the sum of the logarithm of these P-values as a summary statistic, S, for the enrichment of annotations. For the real data we obtained S = -7078. To test whether this value was more extreme than what would be expected if the clusters were random, we permuted the genes while conserving the hierarchical topology 10000 times, and obtained S on average to be -2746 ± 52 std. dev. Therefore, the observed value was 80 standard deviations away from the random expectation. Since we already have shown that the hierarchical clustering results contain clusters that are enriched in subcellular localizations, this strong statistical significance is expected, as subcellular localization and functional annotation of proteins are strongly connected. Therefore, we next tested whether functional annotations were enriched in our clusters beyond what could be explained from subcellular localization enrichments alone. To do so, we again generated the distribution of S, but this time constrained the permutation: proteins can be replaced only if they share the same set of discrete subcellular localization

annotations [76]. Even with this constraint on the permutations, we obtain a 32.1 std. dev. lower value of S than in the permutations, and note that none of the 10000 permutations showed a more extreme value of S ($P < 10^{-4}$).

3.2.3 Time-profiles better characterize protein function than subcellular localization.

The association of proteins to specific subcellular localization classes is known to be a major determinant of protein function, and therefore because our hierarchical tree structure reflects subcellular localization classes, we expect proteins with similar biological function to be close within the hierarchical tree. However, proteins of similar biological function are even closer in the hierarchical tree structure than can be expected based on their subcellular localization classes.

It is known that proteins associated to the same subcellular localization are more likely to share biological function than proteins of different subcellular localization. We expect proteins of similar biological function to be closer in the hierarchical tree when a given biological function is associated to subcellular localization. The proximity of proteins of similar function due to proximity of protein subcellular localization can be quantified by permuting proteins within the hierarchical clustering (leaves within the tree structure), where permutations are only allowed within protein groups that have been previously characterized to have identical subcellular localization. By definition, this constrained permutation preserves the proximity of proteins associated to subcellular localizations. The expected proximity of proteins sharing biological function can be quantified by computing the proximity statistic on the hierarchical tree structure obtained after the permutation, and we observe that the proximity statistic obtained in permutations is far lower that on the original hierarchical clustering (P-value = 0.001). As before, all ranges of group sizes show a proximity statistic which is unexpected under the distribution from the permuted hierarchical clusters (Table II.5). This analysis implies that the biological information in the hierarchical clusters cannot be fully explained by the subcellular localization annotations [76], and, more importantly, that this unsupervised analysis must be capturing finer similarities in temporal and spatial expression for many groups of functionally related proteins.

Further, we observe that the maximum likelihood linkage criterion for the hierarchical clustering of time-profiles shows a stronger association with previous knowledge of biological function than some other linkage methods (Complete linkage, with Euclidean or correlation metric). Stronger statistical association of proteins to subcellular localization is found within hierarchical clusters obtained by maximum likelihood agglomerative clustering compared to alternatives (Table II.4 & II.5). This result is likely to depend on the nature and quantity of single cell measurements, since the opposite statement would hold if a number of

objects are filtered out by a cell confidence threshold.

Annotation		Confidence Weighted			Confidence Threshold (0.8)		
Set	#	MLAC	Euclidean	Correl.	MLAC	Euclidean	Correl.
GO	2134	-31.907	-27.903	-26.964	-20.910	-28.160	-24.681
Complexes	277	-19.495	-18.627	-17.330	-13.754	-17.973	-17.209
Complexes*	262	-18.600	-17.810	-16.660	-13.095	-17.073	-16.489
Pfam	761	-17.128	-13.535	-13.625	-11.118	-14.146	-12.527
Pfam*	688	-14.496	-11.532	-11.537	-9.942	-11.306	-11.056
GO 2	664	-10.084	-7.776	-8.894	-6.223	-7.793	-6.323
GO 3	337	-9.276	-8.662	-6.320	-6.297	-9.104	-7.960
GO 4-5	379	-13.501	-11.519	-11.822	-9.199	-12.091	-11.095
GO 6-7	203	-14.718	-13.419	-11.675	-8.458	-13.122	-10.659
GO 8-9	113	-12.077	-12.433	-10.299	-6.112	-12.938	-10.292
GO 10-11	76	-11.958	-11.253	-12.814	-7.830	-13.285	-10.658
GO 12-14	82	-13.066	-10.426	-10.971	-9.267	-10.157	-9.849
GO 15-18	73	-12.294	-7.474	-11.194	-6.301	-8.896	-8.294
GO 19-24	71	-15.174	-15.944	-12.296	-9.512	-15.217	-12.021
GO 25-34	49	-10.119	-10.293	-8.843	-10.442	-10.370	-8.759
GO 35-49	32	-13.574	-10.807	-9.656	-9.205	-10.828	-9.774
GO 50-79	29	-13.287	-11.950	-10.768	-10.526	-10.518	-11.346
GO 80-200	18	-8.223	-7.479	-6.835	-5.179	-8.317	-6.885
GO 200+	8	-11.322	-6.272	-1.956	-9.079	-4.839	-0.675

Table II.5: **Z score of enrichments for a constrained permutation test.** Z score for the S statistic, where the variance in S statistic is inferred by permuting proteins in the hierarchical cluster that have identical subcellular localization characterized by Huh et al. [76].

3.2.4 Dynamic distinctions between bud neck classes.

Because our analysis explicitly models cell-stage, we can identify dynamic patterns where proteins move from one subcellular localization to another. For example, we identified a cluster of proteins that showed a large range of distances to the bud neck, and for many of them, the distance to bud neck varied over the cell-stage (Figure II.13). In this cluster, we find a group of proteins that first appears in the periphery of the bud, and then migrates at a particular cell-stage to the bud neck. Interestingly, these include Pkc1 and Lrg1 (Figure II.13), which are both in the cell-wall integrity pathway [97]. Another functionally related group of proteins that shows the same dynamic pattern are the subunits of the exocyst complex (e.g., Sec10 Fig II.13), but they appear to be more compact in small buds. This is in contrast to other profiles that represent proteins that always located at the bud neck. Unlike Pkc1, Lrg1 or the subunits of the exocyst, Bud3 shows a consistently small average distance to the bud neck (Figure II.13). It can therefore be considered a pure 'bud neck' localization pattern, as opposed to Pkc1, Lrg1 and the exocyst subunits that are cycling from the bud periphery to the bud neck analogous to the way the MCM subunits cycle from the cytoplasm to the nucleus. This suggests that these dynamically changing bud periphery to bud neck proteins have localization

that is targeted by a shared cell cycle regulatory mechanism. Yet another subtle variation on this theme is illustrated by proteins that are found specifically in the bud periphery, but do not migrate to the bud neck (e.g., Cla4, Fig II.13). We speculate that these proteins lack a specific portion of the cell cycle regulation shown by Pkc1, Lrg1 and the exocyst subunits.

We also find in the same cluster 23 proteins that were not previously annotated in systematic studies as being bud-specific or actin [29, 76]. We predict that these proteins show dynamic patterns within the bud during its growth, and were difficult to describe using discrete annotations. For these proteins, SGD [34] annotations mostly disagree with previous systematic annotations. Among the 23 proteins, we find proteins that have functional links to other proteins known to be bud-specific, such as Ack1/YDL203C which is thought to function upstream of Pkc1 [88]. Looking at the images, Ack1 shows a pattern similar to Pkc1, with the difference that the protein abundance in the bud is not strong relative to the basal cytoplasmic expression in the mother cell (Figure II.14). Similarly, Msb3, Lte1 and Zds1 have been previously reported to show bud-related patterns in low-throughput analyses [17, 135, 171] and are found in this cluster.

Hence, we hypothesize that proteins are found in this cluster because they are showing various dynamic localization patterns with respect to the bud. Indeed by further inspecting the images we found a protein of unknown function, YDR239C, which shows a dynamic bud pattern similar to Ack1. This protein has not been previously characterized to localize to the bud periphery or bud neck and therefore represents a new positive prediction obtained from the unsupervised analysis. In contrast, visual inspection of other images reveals that some proteins in this cluster do not show obvious dynamic bud patterns. For example, Tpo3 was characterized as a cell periphery [29, 76] and plasma membrane [34] protein. The subcellular localization of Tpo3 in our images is different than the dynamic bud patterns we previously described. Yet, it was clustered next to Rtk1, which appears in our images as a cell periphery protein that is partially localized at the bud neck at the expected cell-stage (Figure II.14). This inclusion of Tpo3 was likely due to the similarity in the pattern of Tpo3 and Rtk1. This is expected of hierarchical cluster analysis, in that there are no hard delineations between the quantitative patterns (see Discussion).

This cluster illustrates pattern discovery using biologically interpretable features. We identified a group of proteins showing complex expression patterns that have been difficult to define previously. We believe this is due in part to the higher resolution of our images, as well as our ability to assign dynamic, quantitative patterns to these proteins. We note that not every protein in this cluster actually shows (as far as we can tell by inspecting the images) a dynamic bud pattern (Figure II.14). Nevertheless, we could relate the consensus pattern in this cluster (variation in our measurement of 'average distance to bud neck') to cell

cycle dependent migration from bud periphery to bud neck.

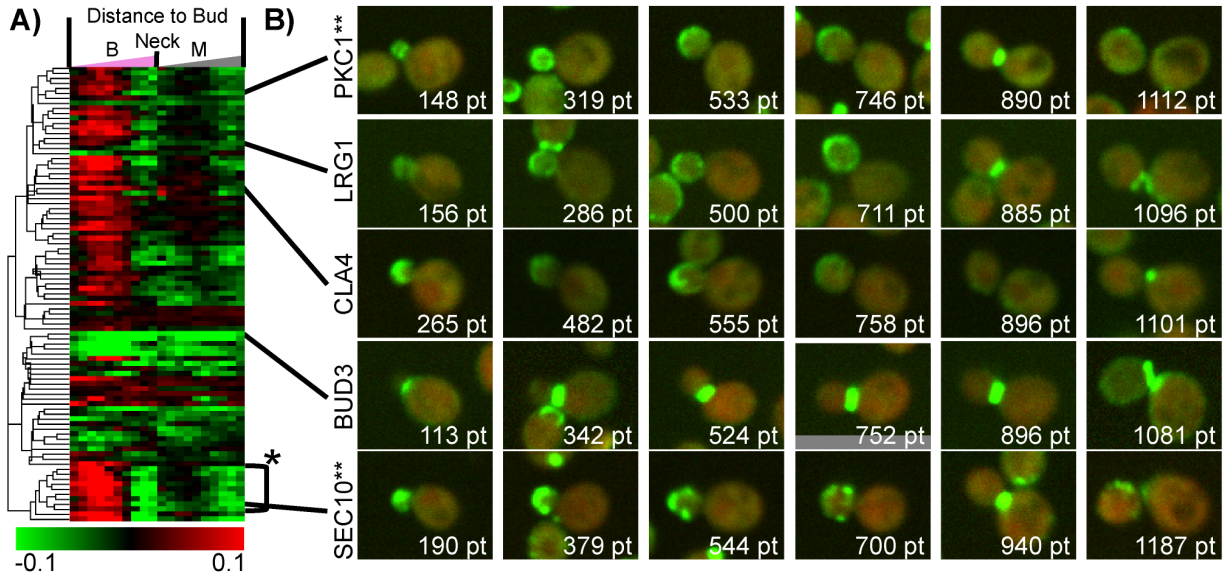


Figure II.13: **A cluster of 91 proteins displaying time-profiles with variable distances to the bud neck.** A) Heat map of the cluster displayed as in Figure 6. We observe several classes of dynamic patterns, which capture the localization to the bud neck and bud periphery. (*) 5 of the 8 subunits of the exocyst complex are found within 9 proteins. B) Examples of proteins with dynamic bud patterns. (**) The displayed GFP intensity was scaled down by 75%.

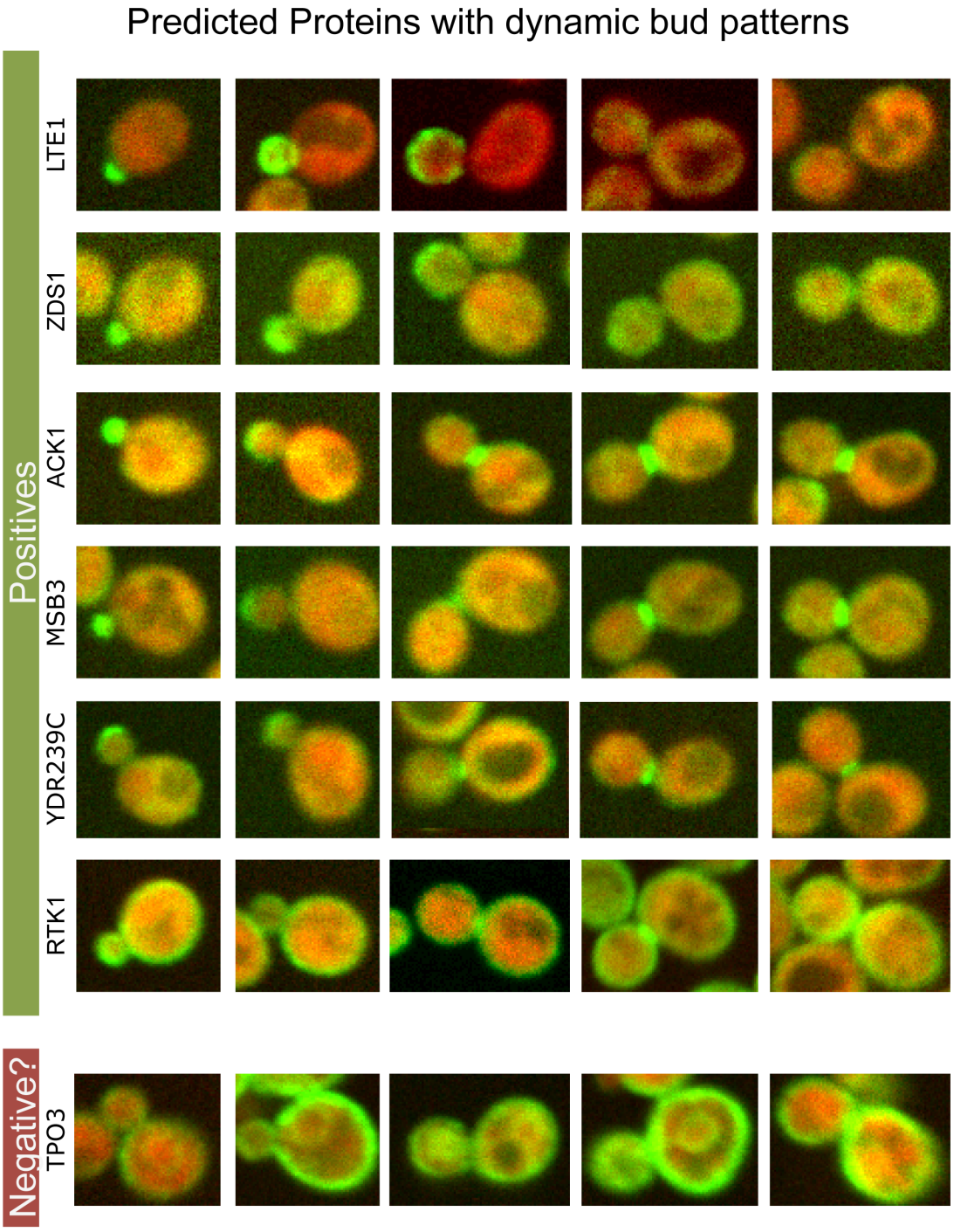


Figure II.14: **Examples of proteins in the dynamic bud cluster.** Images are representative of patterns for each protein. The contrast of each image has been enhanced to display patterns more clearly. These proteins were not previously annotated as showing bud-related patterns by Huh et al. [76] or Chen et al. [29]. The top 6 proteins (indicated using a green bar) are found localized to the bud tip and/or bud neck, so that they exhibit a dynamic bud pattern. For Tpo3 (indicated using a red bar), it is doubtful whether this is the case or not: Tpo3 typically appears in the cell periphery and nuclear periphery. Hence, Tpo3 is an example of a negative prediction of a dynamic bud protein.

4 Supervised Analysis

The previous section has described an unsupervised analysis of protein expression patterns. We now turn to the most prevalent method for the analysis of protein spatial expression: pattern recognition. In order to assess the predictive power of the 'time-profiles' described above, we next used these features for supervised classification of subcellular localization. Here we are given a 'gold standard' set of manually identified localization patterns, and the task is to recapitulate the labels.

Many image features have been considered for the classification of localization pattern, but with a success in budding yeast that is limited to major subcellular localizations. Proteins of mixed localizations are typically filtered out [29, 75] of classification analyses. In order to compare image features to time-profiles, I first replicate a previously reported classification experiments. Secondly, a different classification task is considered: identifying individual proteins from replicate experiments, a task that has not been attempted before (Section 4.2).

4.1 Pure subcellular localization Classification

Since the unsupervised analysis obtained functional annotation enrichments, it would be reasonable to assume that a metric based classification would exhibit a certain classification performance. One important difference between unsupervised and supervised approaches is that the image features ultimately used to train classifiers are typically first selected among a pool of possible features (feature selection). While the equivalent task could be performed on the time-profiles, the inner dependencies of key-point values are fully characterized from lowess inference: proper selection of simple feature measures and cell-stage keypoint should make this selection unnecessary. To test this, the previously introduced time-profiles are directed used without any filtering.

4.1.1 Support Vector Machine

For this supervised analysis, I will use two additional replicate experiments each containing about 4000 strains. Support vector machine (SVM) are used in order to learn subcellular localization from time-profiles of on experiment, and the classification performance is evaluated using a different image collection. Since a radial basis function (RBF) kernel is utilized for the SVM classification, a total of two parameters are required : a scale parameter and a penalty term for misclassified instances. Prior to the choice of scaling parameter, each of the 6 feature measurement has been rescaled to have a variance of 1 (all cell-stage keypoints confounded). Then, I selected a scale parameter of 0.1 and a penalty term of 1.0, and used svmlite [81].

This toolkit enables multi-class classification using one-against-the-rest SVM classification. Each subcellular localization class is then characterized by a binary classification and each time-profile is assigned to the class with largest distance to the classification boundary.

	# protein	Cytoplasm	Nucleus	Mitochondrion	ER	Vacuole	Nucleolus	Cell periphery	Vacuolar mem.	Nucl.Peri.	Spindle pole	Endosome	Late Golgi	Actin	Peroxisome	Lipid particle	Golgi	Bud neck	Early Golgi	Microtubule	ER to Golgi
Precision		.717	.850	.870	.794	.684	.978	.886	.857	.857	.952	.714	.700	.867	1.0	1.0	0	1.0	0	.667	0
Cytoplasm	791	.959	.005	.025	.009	.001	0	0	0	0	0	0	0	.001	0	0	0	0	0	0	0
Nucleus	471	.061	.906	.017	.004	.006	.002	0	0	.004	0	0	0	0	0	0	0	0	0	0	0
Mitochondrion	460	.162	.002	.829	.002	.002	0	.002	0	0	0	0	0	0	0	0	0	0	0	0	0
ER	240	.258	.008	.035	.696	0	0	.004	0	0	0	0	0	0	0	0	0	0	0	0	0
Vacuole	116	.361	0	.059	0	.546	0	0	.017	0	0	.008	.008	0	0	0	0	0	0	0	0
Nucleolus	69	.029	.304	.014	0	0	.652	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cell periphery	54	.161	0	.071	.054	.018	0	.696	0	0	0	0	0	0	0	0	0	0	0	0	0
Vacuolar mem.	50	.019	0	.019	0	.250	0	0	.692	0	0	.019	0	0	0	0	0	0	0	0	0
Nucl.Peri.	51	.040	.720	0	0	0	0	0	0	.240	0	0	0	0	0	0	0	0	0	0	0
Spindle pole	35	.108	.162	.081	.081	0	0	0	0	0	.541	0	0	0	0	0	0	0	0	.027	0
Endosome	31	.364	.030	0	0	.333	0	.030	.091	0	0	.152	0	0	0	0	0	0	0	0	0
Late Golgi	32	.606	0	.030	.121	0	0	0	0	0	0	0	.212	.030	0	0	0	0	0	0	0
Actin	24	.037	0	.037	.444	0	0	0	0	0	0	0	0	.481	0	0	0	0	0	0	0
Peroxisome	16	0	0	.167	.056	0	0	0	0	0	.056	0	0	0	.722	0	0	0	0	0	0
Lipid particle	19	.263	.105	0	.368	0	0	0	0	0	0	0	0	0	0	.263	0	0	0	0	0
Golgi	15	.867	0	0	.067	0	0	0	0	0	0	0	.067	0	0	0	0	0	0	0	0
Bud neck	15	.200	0	0	.067	0	0	.133	0	0	0	0	0	0	0	0	0	.600	0	0	0
Early Golgi	11	.727	0	0	0	0	0	0	.091	0	0	0	.091	0	0	0	.091	0	0	0	0
Microtubule	10	.400	.300	0	.100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.200	0
ER to Golgi	6	.333	0	0	.667	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table II.6: **Average of 6 confusion matrices for SVM classification.** Using 3 replicate image collections, 6 proteome-wide classifications are performed by selecting one as training set and another as test set. The average of the 6 confusion tables is reported. Overall, SVM Classification accuracies are lower than those reported by Chen et al. [75], which is not a fair comparison as Chen et al. only used subsets of a image collection as opposed to a full collection replicate to define test and training sets for the evaluation the accuracy.

Since 3 replicate proteome-wide collections are available, it was possible to perform 6 classification experiments (Table II.6). We note that the classification accuracy (77.6%) is lower than what was previously reported (87.8%), even if the comparison is not completely fair as we are having the same sets of protein in the test and training set while it is never the case for Chen et al. [75]. The same biases in classification are observed: larger subcellular localization classes are better classified. This result indicates that the feature selection and proper parameter selection, which may need to be adaptive [75], are critical for improving the classification accuracy.

4.1.2 Nearest Neighbour Classification

In this section, I show that nearest-neighbour classification of time-profiles, using a pair of image collections to define test and training set respectively, can outperform SVM classification for the prediction task of subcellular localization pattern (as characterized by Huh et al. 2003). I report on the average classification performance for 6 experiments (Table II.7). The classification accuracy increased to 81.9%, which is still lower than previous reports [75]. The main difference is that small classes have better recall than what was reported using SVMs (Figure II.6). The recall for all subcellular localization is always above 42%, and it can be up to 40% above previous reports ('Early Golgi', 'Golgi', 'Bud neck' and 'Microtubule').

Nearest Neighbour and SVM classification have some methodological similarities, but one difference is that there are similarity metric that cannot be utilized in a SVM framework, such as the Bhattacharyya metric (Eq. 32), which violates the triangle inequality. The Bhattacharyya metric allow each data point (time-profile) to possess its own covariance matrix, while standard SVM essentially constrain each data point to have the same kernel function. If an adaptive scale is required to better capture similarities of protein expression, cell-to-cell differences in the time-profiles may inherently report of the amount of variance which is typical of each subcellular localization, for each of the feature of interest. It also accounts for correlation for feature measurement, which may be class specific.

Interestingly, the Nearest Neighbour classification is typically inferior SVMs [18]; partly because the Nearest Neighbour classification does not learn any parameters from training data (the choice metric alone defines the classifier). In the next section, I present another result that is critical to properly interpret this unexpectedly higher accuracy for Nearest Neighbour over SVM.

	# protein	Cytoplasm	Nucleus	Mitochondrion	ER	Vacuole	Nucleolus	Cell periphery	Vacuolar mem.	Nucl.Peri.	Spindle pole	Endosome	Late Golgi	Actin	Peroxisome	Lipid particle	Golgi	Bud neck	Early Golgi	Microtubule	ER to Golgi
Precision		.878	.859	.887	.688	.726	.807	.904	.725	.588	.786	.364	.556	.864	.800	.615	.733	1.0	.500	.455	1.0
Cytoplasm	791	.881	.018	.020	.057	.005	0	.004	0	.001	.003	.004	.003	.003	0	.001	0	0	0	.001	0
Nucleus	471	.017	.892	.017	.006	.002	.023	0	.002	.036	0	.002	0	0	0	.002	0	0	0	0	0
Mitochondrion	460	.081	.009	.841	.037	.004	0	0	0	0	.007	.009	.007	0	0	0	.002	0	.002	.002	0
ER	240	.067	0	.029	.846	.025	0	.004	0	0	0	0	.013	0	.004	.004	.004	0	0	.004	0
Vacuole	116	.095	.034	.052	.052	.595	0	0	.069	0	0	.095	0	0	0	.009	0	0	0	0	0
Nucleolus	69	0	.290	.014	0	0	.667	0	0	.014	0	0	0	0	.014	0	0	0	0	0	0
Cell periphery	54	.037	0	.019	.056	0	0	.870	0	0	0	0	.019	0	0	0	0	0	0	0	0
Vacuolar mem.	50	0	.020	0	0	.080	0	0	.740	0	0	.160	0	0	0	0	0	0	0	0	0
Nucl.Peri.	51	.020	.314	0	0	0	0	0	0	.588	.020	.020	0	0	0	0	0	0	0	.039	0
Spindle pole	35	.057	.114	.114	.029	.029	0	0	0	0	.629	0	0	0	0	0	0	0	0	.029	0
Endosome	31	.065	0	0	.032	.161	0	.032	.161	.032	0	.516	0	0	0	0	0	0	0	0	0
Late Golgi	32	.250	.063	.031	.031	0	0	0	0	0	0	0	.469	.031	.031	.031	0	0	.063	0	0
Actin	24	.120	0	0	.120	0	0	0	0	0	0	0	0	.760	0	0	0	0	0	0	0
Peroxisome	16	0	0	.167	0	0	0	0	0	0	0	0	0	0	.667	0	.056	0	.111	0	0
Lipid particle	19	.053	.053	.053	.316	.105	0	0	0	0	0	0	0	0	0	.421	0	0	0	0	0
Golgi	15	.067	0	0	.067	.067	0	0	0	0	0	0	.067	0	0	0	.733	0	0	0	0
Bud neck	15	.200	0	.067	0	0	0	0	0	0	0	0	0	0	0	0	0	.733	0	0	0
Early Golgi	11	.182	0	0	.091	0	0	0	0	0	0	0	.182	0	0	0	.091	0	.455	0	0
Microtubule	10	0	.300	0	.100	0	0	0	0	.100	0	0	0	0	0	0	0	0	0	.500	0
ER to Golgi	6	0	0	0	.500	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.500

Table II.7: **Average of 6 confusion matrices for 'Nearest Neighbour' classfication.** Using 3 replicate image collections, 6 proteome-wide classifications are performed by selecting one as training set and another as test set. The average of the 6 confusion tables is reported. Classification recall is still lower than Chen et al. [75], but even classes with little representation have a recall above 42%.

4.2 Individual Protein Recognition

When using Nearest Neighbour classification, which simply maps instances from the test set to their closest instances found in a training set under the Bhattacharyya metric (eq. 32), a number of associations made may be pairs of cell populations that both have the same protein tagged. Since ∼4000 different proteins are found in both training and test set, such pairing of cell populations by chance has a low probability; nevertheless, I observe that several hundreds of proteins are corrected recognized from time-profiles. In the 3 wild type image collections that we previously analyzed ('HOwt','ura3','rap0')), 7% to 12.6% of the proteins are correctly identified (Table II.8). These percentages are much higher than the 0.03% that is expected by chance.

This explains the unexpected success of Nearest Neighbour in recognizing subcellular localizations (previous section 4.1.2) as this fraction of proteins identified are guaranteed to properly propagate their associated class labels. This guaranty holds for any classification problem where dataset replicates are used as test and training set. Interestingly, 55 proteins were always properly paired in 6 protein recognition experiments (Table II.9), which suggests that imaged cell populations contains sets of characteristics that uniquely describe many proteins.

Training Set		Test Set					
	# proteins	HOwt	ura3	rap0	alp1	alp2	alp3
HOwt	3967		488/3913	389/3914	273/3407	211/3885	200/3894
ura3	4035	496/3913		323/3975	265/3451	241/3930	182/3939
rap0	4046	338/3914	266/3975		269/3484	254/3958	226/3967
alp1	3518	245/3407	234/3451	270/3484		995/3489	543/3488
alp2	4002	196/3885	199/3930	251/3958	949/3489		1253/3962
alp3	4013	150/3894	128/3939	222/3967	516/3488	1259/3962	

Table II.8: **Fraction of proteins recognized by 'Nearest Neighbour' classification.** Fraction of the proteins from the set that are correctly assigned to the same protein strain within the training set. The denominator represents the number of proteins that are available in both the test and training set. No instances in training set are filtered: so the expected number of proper identifications from random labeling is in fact lower than one.

I further analyze 3 other image collections. In these 3 new collections, yeast cultures were exposed to a mating pheromone (alpha factor), which prevents the production of new buds and causes the morphology of 'lone' cells to change (Figure II.15). The 3 image collections were imaged 15, 30 and 45 minutes after the introduction of the mating pheromone. Morphology changes may significantly perturb image feature measurements and cell identification accuracy, which may limit the classification accuracy.

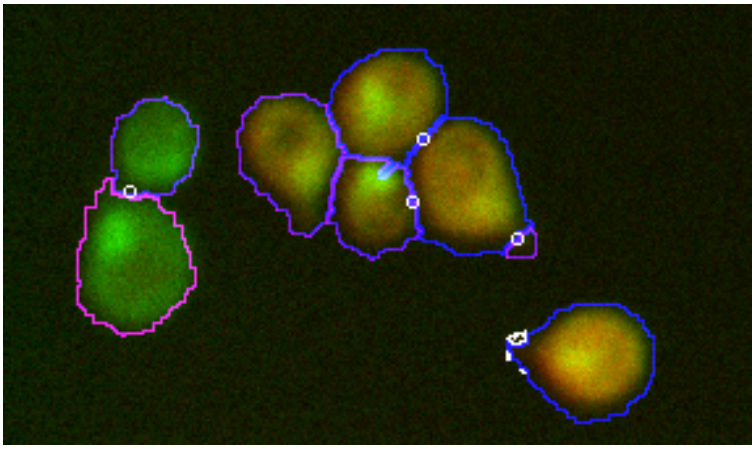


Figure II.15: **Morphology change under mating pheromone.** Cells were in a media with alpha factor added for 45 minutes. Such cells grow a mating projection tip, which does not grow any further. The cell segmentation may or may not detect the projection tip as a bud.

The distribution of object sizes significantly differ globally from previous the wild type (WT) image collections. Nevertheless, a fair fraction of proteins can be identified, even if the morphology of a fraction of the cell has significantly changed (Table II.8). Further, the number of proteins that are correctly identified across the 6 inner protein recognition experiments, which are defined using image collection ('alp1','alp2','alp3') alone, increased to 144; showing that disagreements are lower in time-series of experiments, where environmental and experimental conditions are most similar, even if cells are reacting to a given drug.

ALG6	BNI5	COS4	EDC3	FTR1	HXT3	LYS21	PEX25	RTN1	SLY1	YEH2
AQR1	BUD3	COS5	ENA1	GAS1	INP1	MCD1	PGA1	SAC6	TFP1	YLR407W
ATP2	CBF1	CRH1	ENA5	GID8	IST2	NCE102	RFS1	SGT1	TIM17	YLR413W
ATR1	CDC11	DCW1	FCY2	HOR7	IVY1	NSR1	RPN3	SKG1	VPH1	YMR295C
AXL2	CDC28	DFG5	FLC1	HXT2	LAP4	PEA2	RSN1	SLG1	WHI5	YNL134C

Table II.9: **Example of 55 Proteins recognized by 'Nearest Neighbour' classification.** Using 3 replicate image collections, 6 proteome-wide classification are performed by selecting one as training set and another as test set. In the 6 classification, the above proteins were correctly recognized using the Bhattacharyya metric.

This is further indication that time-profiles captures rich characteristics of protein spatial and temporal expression, which are often specific enough to identify proteins uniquely. This is in agreement with the previous observation that time-profile capture protein similarity beyond what can be explained by subcellular localization (Section 3.2.2). What remains to show is that there is a significant gain of learning cell-to-cell difference in feature measurements. This untold premise that motivated the use of both the maximum likelihood agglomerative clustering, as opposed to a metric based hierarchical clustering, and Nearest Neighbour classification with Bhattacharrya distances, as opposed to support vector machine (SVM). In the last part of

this thesis, I will verify if differences in feature measurement can be related to biological processes, or if they are purely generated from experimental and sampling variations. Interestingly, many of the proteins that are here shown to be robustly identifiable from Nearest Neighbour classification (Table II.9) will be shown to have unusual stochastic properties.

Part III

Cell-to-Cell Variability

Overview

The aim of research in biology is to describe and understand biological phenomena. More recently, a specialization of the field arose under the name 'systems biology', whose purpose is to further describe biological observations and measurements so they could be understood as the outcome of mathematic and/or probabilistic models. This approach proved to be successful in the characterization of protein production within cells (gene expression), which could be shown to fluctuate in predicable way from either changing the number of promoters or by changing environmental factors that induce the gene expression [168]. In defining such mathematical models, it was discovered that protein production models allow cells with identical genomes to be in different states and maintain their current state even though they are in identical conditions [146]. It was observed that this heterogeneity is not an artefact of the mathematical formalism, but that not all cells may exhibit the same phenotype in identical conditions, so phenotypes can be 'selected' to be heterogeneous in a cell population (incomplete penetrance) [128].

A natural extension of the mathematical model, which typically describes continuous variables as opposed to discrete variables (whole numbers), is to model the exact molecule abundance as opposed to a concentration value. Chemical events hence can no longer be simply described as occurring at a given rate; for example, two reactions that consume the same molecule cannot both occur if a single molecule exists in the initial state. Such system is stochastic (or non-deterministic) as one of the two reactions is selected randomly. Extending the mathematical model to account for stochasticity allowed the comprehension of the diversity of phenotypic expressions that are often observed in populations of identical organisms [150]: the many states predicted by mathematical models (steady states) could be all reached, each with a certain probability that relates to stochastic fluctuation in the protein abundance [83].

In this last chapter, I combine the measurements made on imaged cell populations so to quantify cell-to-cell heterogeneity in feature values. Doing so, I will verify that a defined measure characterizes biological variation, as opposed to sampling variance or fluctuations expected from cell cycle progression. In the 1^{st} section, cell-to-cell variability in protein abundance is evaluated specifically; in the 2^{nd} second section, cell-to-cell variability in subcellular localization of protein is evaluated, which is a task that was never attempted before on the whole yeast proteome. The main result is that a single method is shown to be applicable to both problems; the method is specifically described in section 1.5.

1 Stochasticity in Protein Abundance

1.0 Background

Regulatory network topologies generate diverse stochastic properties; for example, negative feedback regulation lowers the stochastic noise [154]. It is possible to characterize the stationary distribution of molecule abundance for simple systems (three-stage model) [140]. More complicated regulation may induce a variable number of steady states, which may be reached in a switch like manner. It has been proposed that stochasticity is a desirable feature that modulates the fitness of the overall population, under fluctuating evolutionary pressure [2, 120]. Stochasticity has been shown to be critical for the understanding of certain processes such as cell differentiation [60]. As such, noise and stochasticity are topics of major current interest in biology [129].

***Sources of Variability*:** It is possible to measure the stochasticity in time series of protein abundances [113] or in protein spatial concentrations [24] in live cells. Swain et al. [152] indicated that previous studies only modeled the intrinsic source of variability, and ignored or poorly modeled 'extrinsic' sources of variability. They define such 'extrinsic' variability to be induced from sources that are independent of the protein of interest, but influence protein expression. Number of RNA polymerases, room temperature, nutriment abundance or even density of cells in a media are all potential source of 'extrinsic' variability. One important extrinsic source of variability that can be controlled in experiments [113], and in this work, is the cell-stage dependence for protein expression.

1.0.1 Quantification of Intrinsic Noise from Fluorescence Microscopy

The quantification of intrinsic noise can be performed using several approaches. Regulatory elements that lead to the production of a protein of interest may be replicated and re-inserted in a cell genome or on a plasmid, so that at least two fluorescent reporters that use the same regulatory elements report simultaneously for protein production rates [48]. By observing how the two reporters are correlated under different conditions or within a time course, it is possible to quantify the fraction of the total variance η^2_{tot} that comes from extrinsic factors η^2_{ext}, and that is intrinsic variance η^2_{int} ($\eta^2_{tot} = \eta^2_{int} + \eta^2_{ext}$). It has been noted that the proportion of intrinsic variability and extrinsic variability varies from protein to protein (Newman et al. [113]).

***Quantification of Stochastic Properties*:** More detailed analyses of intrinsic noise are possible if the regulatory components that generate the extrinsic variability are adequately modeled. For example, starting a cell culture from a single cell, differences are observed in the expression of proteins for its progeny [134]. The time scale of autocorrelation for the intrinsic noise can be captured from computing the correlation of between time points within time series. The resulting autocorrelation function was best fitted by a sum

of two exponential functions, so that factors contributing to the total variability may be distinguished and are here characterized as a fast ($\leq$ 5 minutes) or a slow component ($\sim$ 40 minutes). Having a fluorescent reporter for the activity of an upstream regulation factor, it could be determined that the intrinsic noise can be measured in the fast component only, so that the slow component is then explained by extrinsic factors. One of the drawback of the analysis of reporter fluorophore is that the replicating the regulatory element do not guaranty that the deviation in expression of protein is due to intrinsic variability: the location of the ORF, on a construct or integrated into the genome, may also influence the expression of the gene.

***High-throughput Assessment of Intrinsic Noise*:** The yeast collection of GFP tagged proteins has the advantage that modified genes are retained in genome at the endogenous locus, so that the production rate of protein should be preserved. Newman et al. [113], measured the fluorescence in billions of cells that were scanned using flow cytometer [143]. To remove extrinsic sources of variability, they filtered out cells based to on the cell size of scanned objects, as they are being captured by the flow cytometer for measurements of the scattered light in a side and forward detectors (flow cytometer have no spatial resolution). By selecting certain ranges of allowed 'scatter parameters', they filtered out 99% of the cells that were scanned, so that the remaining 1% is presumably cells of similar sizes.

Unless diploid cells are used, having a replicate marker in the indigenous locus is impossible by definition, hence the separation of extrinsic to intrinsic noise was reported for 4 proteins only, through the use of diploid cells (as opposed to haploid cells in the GFP collection). The few examples of two-coloured reporters allowed us to show that the extrinsic sources of variability were damped by filtering the majority of the objects, while the intrinsic noise level remained unchanged. Further, Newman et al. showed that Hhf2, which is a nuclear protein, had a bimodal histogram of protein abundance, which is caused by the doubling in abundance that systematically occurs in the cell cycle. By filtering 99% of the cells, the bimodal histogram became unimodal, hinting that the cells remaining were in a similar cell-stage. Nevertheless, filtering cells does not remove all the extrinsic sources of variability: levels comparable to the intrinsic noise level were reported for the four examples. The reported noise level hence still includes an unknown level of extrinsic noise, but a fair fraction of the extrinsic noise could be shown to be filtered (different fractions for the given examples).

Once the objects are filtered, the protein abundance is measured and normalized to account for the basal autofluorescence level. Then, the variance in the protein abundance is computed. Newman et al. showed that of coefficient of variation (CV) were highly reproducible using replicate experiments for a fraction of the protein collection (2008 total). As to further define proteins of 'low' and 'high' intrinsic noise, the difference of the CV to the local median of CVs of proteins of similar abundance was utilized (deviation to the median;

DM). This normalization allows finding functional enrichment protein the protein that were said to have high or low variability level. This showed that variability levels of protein abundance are inherent properties of protein expression that can be robustly measured and that relates to protein function.

1.1 Approach

The first task for this chapter is to quantify cell-to-cell variability in protein abundance, which is captured from the mean GFP intensity over the cell area. To validate the results of these approaches, I will compare my results against the variability measurements of Newman et al. [113], which reported variability levels using the same budding yeast GFP-collection. The microscopy approach differs from the use of flow cytometry, in that (i) the total number of imaged cells is order of magnitudes fewer from what Newman et al. scanned and that (ii) I have spatial resolution of the protein while it was unavailable to Newman et al.

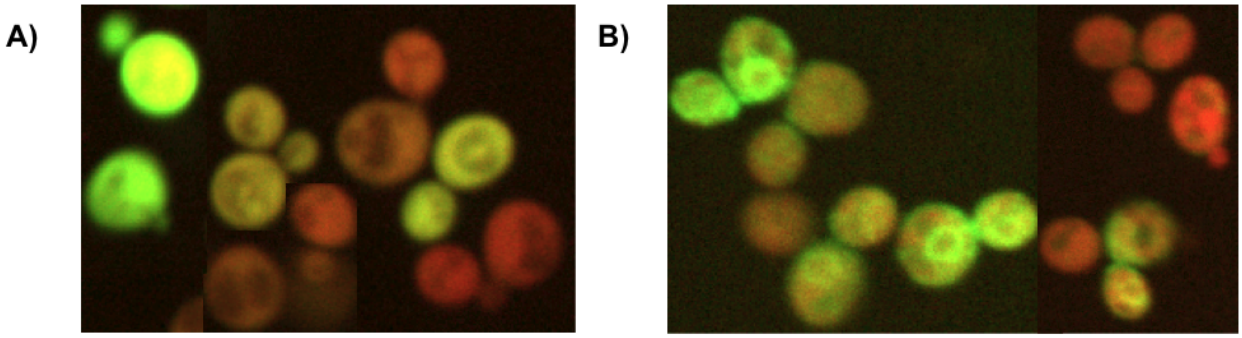


Figure III.1: **Proteins with high cell-to-cell variability in abundance.** A) Cells of a yeast strain with fluorescently tagged Tim17, which is the protein detected as most variable by Newman et al., and ranked as the 4th most variable using 'relative variability' (our proposed method for quantifying cell-to-cell variation). B) Cells with tagged Pir1. While Newman et al. ranked Pir1 as the 3rd most variable (out of 2008), its relative variability rank is 570 (out of 1909). Visual inspection suggests they both show very high cell-to-cell variability levels. Our analysis indicates that much of the variability in Pir1 can be explained by the cell cycle, which is not the case for Tim17. See text and figure for details.

In my case, filtering the vast majority of the objects will leave little data to evaluate the variance in measurements, which implies that the sampling variance will significantly contribute in variability levels. Hence, the goal is to utilize all the available data to measure the variability level. On difficulty that arises is that cells are unsynchronized, so a fraction of the variance measured can be explained by cell-stage. We observe that selecting mother-bud pairs that have buds of similar size reduces the variance in feature measurement for both the mother and bud object (Figure III.2). This shows that a fraction of the total cell-to-cell variability

can be explained by cell-stage. This component of the variance is an 'extrinsic' source of variability, which we want to truncate from the total variability so to better bound the 'intrinsic' source of variability. We note that we cannot fully distinguish between stochastic (intrinsic) and environmental (extrinsic) sources of variability in our analysis, therefore we will not refer to our estimates as 'intrinsic variability' or 'noise' in protein abundance.

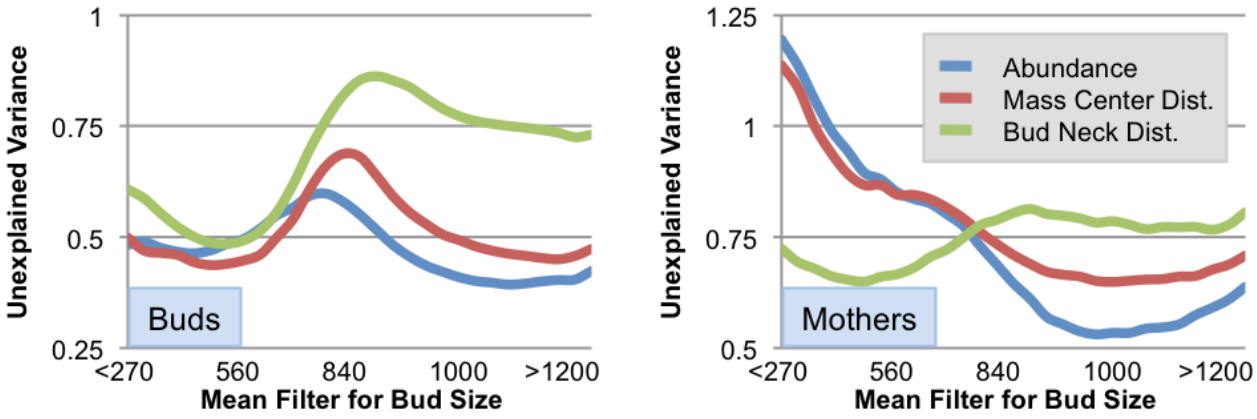


Figure III.2: **Change in variance from filtering using bud size.** Ratio of the variance in feature measurements for buds or mothers that remain after filtering out 87.5% of the identified cells to the total variance. By varying the interval of bud size allowed, which is adapted to the background frequency of cell sizes so that the total number of cell filtered is constant, we observe that ranges of the cell cycle display more variance than others. In buds, the variance is higher at the time of the inclusion of the nucleus within the bud.

1.2 Results

1.2.1 Method Comparison for Variability Level Estimation

In order to measure cell-to-cell variability, I proposed two approaches that report coefficient of variation (CV) for protein abundance (Gaussian Process and Linear regression; section 1.4.2 and 1.4.1) and one method that aims at characterizing proteins that have 'low' or 'high' relative variability level (RV; section 1.5). Our new measure, the relative variability in protein abundance (based on mean pixel intensity over the foreground object area), has a higher correlation with a previous report than 2 methods that evaluate coefficient of variance on our data (Gaussian process, geometric mean of CV in cell-stage bins). I report the global correlation for variability levels that are estimated by the above three methods, and also report their individual agreement with the variability levels reported by Newman et al. [113] (Table III.1). The statistical significance of the difference of correlation levels of RV to DM can be evaluated by using the Fisher transformation [52], which assumes transformed correlations have a normal distribution with a variance that depends on the sample

size. Hence, the correlation for RV is higher than CV with a P-value less than 0.0002 for both methods that report CV, on either mother or bud objects.

Needless to say, no method managed to fully recapitulate global variability levels that were previously reported. This though is expected, I measure an average for variability level over the cell cycle, as opposed to measuring the variability for cells of similar size. Filtering out 99% of the cells in a manner similar to Newman et al. is not possible, because 102 Mother-bud pairs are identified on average per proteins(Figure II.5).

		Newman et al.		Gaussian P. (CV)		Linear Regr. (CV)		Rel. Var. (RV)	
		(CV)	(DM)	B	M	B	M	B	M
Newman et al.	(CV)		.719	.082	.152	.113	.163	.119	.192
	(DM)			.139	.307	.236	.296	.309	.379
Gaussian Process	Bud				.568	.627	.375	.382	.414
	Mother					.388	.456	.365	.430
Linear Regression	Bud						.831	.545	.514
	Mother							.486	.597
Relative Variability	Bud								.779
	Mother								

Table III.1: **Correlation between of cell-to-cell variability estimates.** Correlation between different cell-to-cell variability levels measurement for protein abundance. The three presented measures are compared from the analysis of the same collection of images. Cell-to-cell variability measures are reported for 'bud' and 'mother' objects. The significance of the correlation of DM to RV has a P-value smaller than 10^{-45} (Assuming normal distribution, 2008 proteins compared).

The three methods I present strongly agree for their report on stochasticity level, which is partly explained by the fact that the same cell measurements are utilized. On the other hand, mild agreements are observed with the two variability measures reported by Newman et al. [113]: coefficient of variation (CV), which is the standard deviation divided by the mean, and deviation from median (DM), which is the difference between a given CV to the median of CVs of proteins of similar mean protein abundance. Interestingly, the global correlations to Newman et al. are lower for CV than for DM for all of my variability level estimates, which includes the two methods that similarly report CVs. The purpose of computing the deviation to the median (DM) is to identify proteins that have 'low' or 'high' variability in comparison to the theoretical expectation. Coincidentally, the method I present with highest agreement evaluates 'relative variability' (RV), which similarly attempts to identify protein of unexpected stochasticity level when compared to 'similar'

proteins. These two observations motivates the use of methods that rank protein stochasticity levels, so that the inherent scale of variability level reported does not depend on the normalization procedures for protein abundance measurement, which in turn depends on the nature of biases associated to experimental measurement (such as auto-fluorescence level). Additional observations that motivate the use of 'relative variability' as a mean to report of a stochasticity level are described in section 1.4.

1.2.2 Proteins with High Variability Level

ORF	Name	N. et al	Rel. Var.	Gaussian P.	Linear Regr.
YJL143W	TIM17	1	4	13	15
YBR009C	HHF1	2	34	95	161
YKL164C	PIR1	3	570	32	49
YER103W	SSA4	4	53	2	2
YEL065W	SIT1	5	16	46	38
YNL134C	YNL134C	6	8	18	17
YELO058W	PCM1	26	5	28	31
YKL103C	LAP4	29	6	34	32
YHR136C	SPL2	126	1	5	3
# Ranked		2008	1909	3871	2008

Table III.2: **Most variable proteins in abundance.** Ranking of proteins based on measurement of variability level for protein abundance. Lists of proteins with reported variability level differ in the four cases: the total number of ranked proteins differs from one method to another.

Many proteins appear to have the largest values for variability level according to all approaches (Table III.2). Nevertheless, each method appears to detect different proteins as being the most variable. Among the disagreements, relative variability level does not identify Pir1 has exhibiting 'high' variability, while the three other methods do. Visual inspection of the images suggests that Pir1 has extremely high stochasticity level is comparable to Tim17 (Figure III.1). Can the visual inspection be trusted? For this case, it appears that it cannot: Tim17 does not have a strong cell-stage dependency while Pir1 does (Figure III.3). Since the final report on variability level averages over the whole cell-stage, the high variability level will be scaled down by the occurrence frequency of cells at that cell-stage. This is consistent with the fact that Newman et al. filtered budded objects and reported variability level in unbudded cells of a certain size. As such, the disagreement is explained by the difference in what is measured. Since the underlying motivation is to factor out any variability that can be associated to cell-stage, this case is a prime example of a protein of complex stochastic properties, which are best characterized while considering cell-stage dependency for

protein expression.

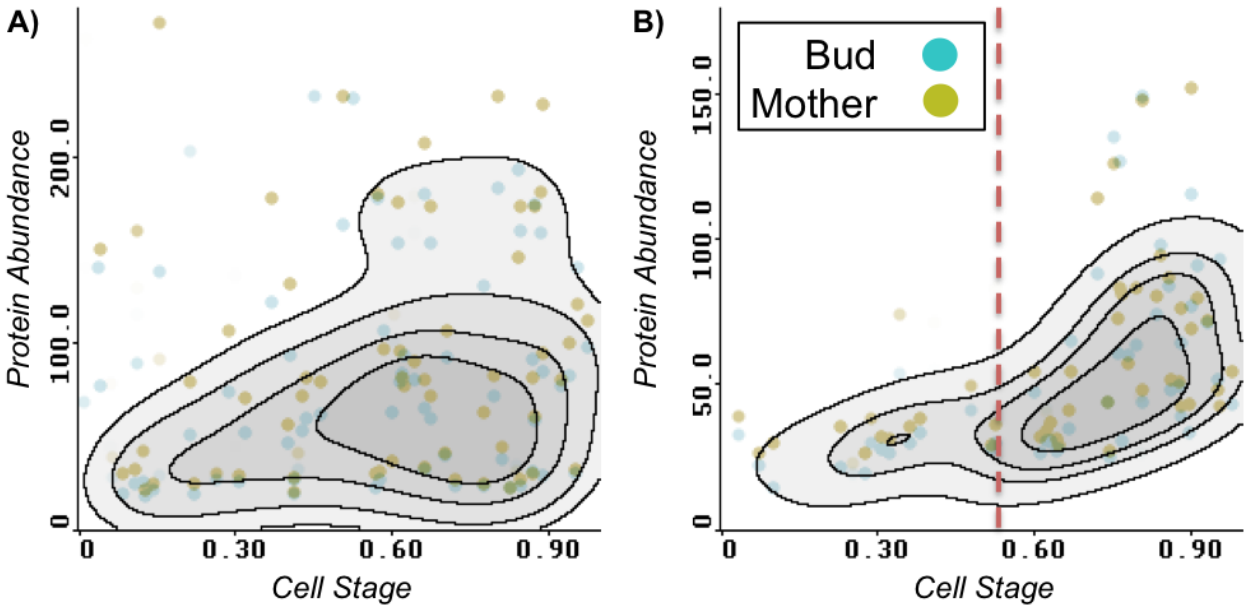


Figure III.3: **Cell-stage dependency for stochasticity in protein abundance.** A) Tim17 displays highly variable levels of GFP intensity throughout the whole cell cycle in both buds and mother cells. B) Pir1 displays low levels of cell-to-cell variability when the bud is small, and the variability dramatically increases once bud reaches a certain size (red dotted line). This is observed in both the bud and mother cell, which suggests that Pir1 displays high level of in cell at the G1 cell-stage exclusively.

ORF	Name	N. et al	Rel. Var.	Gaussian P.	Linear Regr.
YLR390W-A	CCW14	1916	2	12	16
YFR031C-A	RPL2A	1907	7	19	141
YDL226C	GCS1	1835	11	65	199
YBL027W	RPL19B	1992	14	271	240
YPL079W	RPL21B	1989	17	3	119
# Ranked		2008	1909	3871	2008

Table III.3: **Protein strongly disagreeing for cell-to-cell variability level estimates.** Ranking of protein based of measurement of variability level for protein abundance. Lists of proteins with reported variability level differ in the four cases: the total numbe ranked protein differs from a method to another.

I also noted the opposite scenario; several proteins with the highest relative variability (RV) measurement

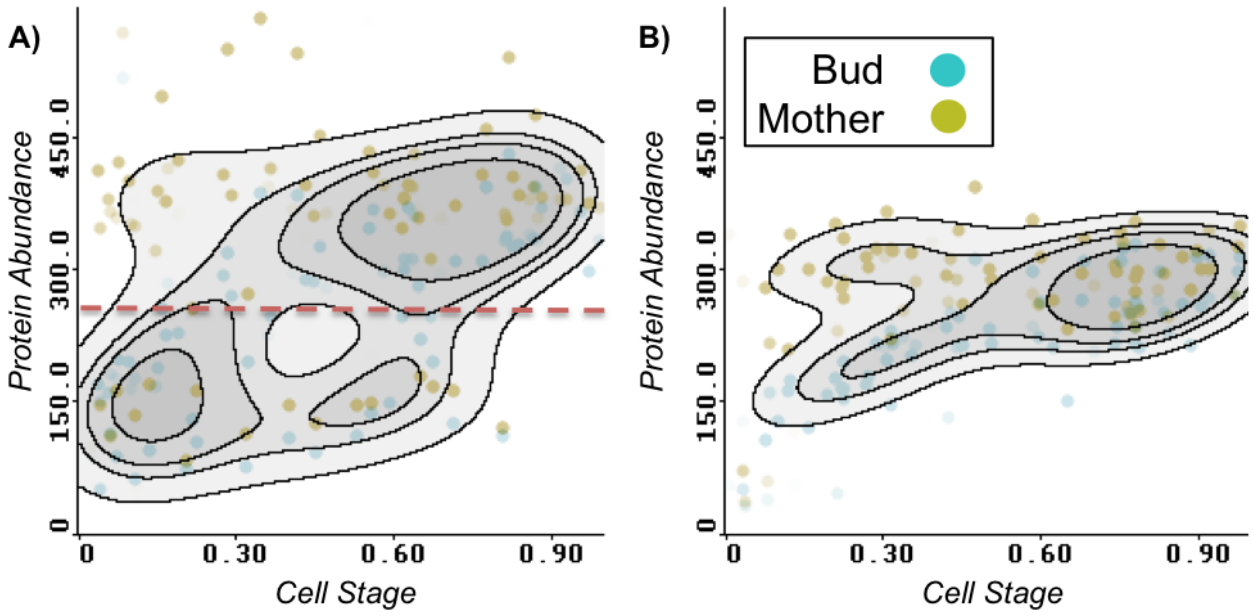


Figure III.4: **Cell-stage and stochasticity for ribosome proteins.** A) Rpl2A display cell-to-cell variability that cannot be fully explained by cell-stage: both mother and bud exhibit two modes of protein abudance at any given cell-stage (shown separated by red dashed line). The contour lines represent the probability density of observed protein abundance, bud and mother cell confounded. B) In the case of Rpl35B, a single mode for the protein abundance is observed instead.

appear among the least varying by Newman et al. only (Table III.3). Three of these proteins (Rpl2A, Rpl21B and Rpl19B) are ribosomal 60S subunits. Similar to other 60S subunits, they have high protein abundance throughout the cell-stage. Visual inspection reveals that cell population exhibit large cell-to-cell variability in protein abundance when compared to other ribosomal 60S subunits (Figure III.5). The other 60S subunits are not reported to 'high' relative variability level; visual inspection confirms that these 3 proteins are the only proteins with such cell-to-cell variability level. In turn, Newman et al. reported low variability level to all 60S subunits, which is indicative of the fact that their normalization procedure for defining 'low' and 'high' variability level differ from 'relative variability', and shows that disagreements to Newman et al. are not necessary false-positives or false negatives, even though they could show that their report on variability level are highly reproducible. In my case, having microscopy images allows visual inspection to confirm that the cell-to-cell variability measure is real. Note that we cannot tell whether Newman et al. failed to detect such variability, as it is possible that environmental conditions alter protein expression or that the images I analyzed are contaminated with other GFP-tagged yeast strains.

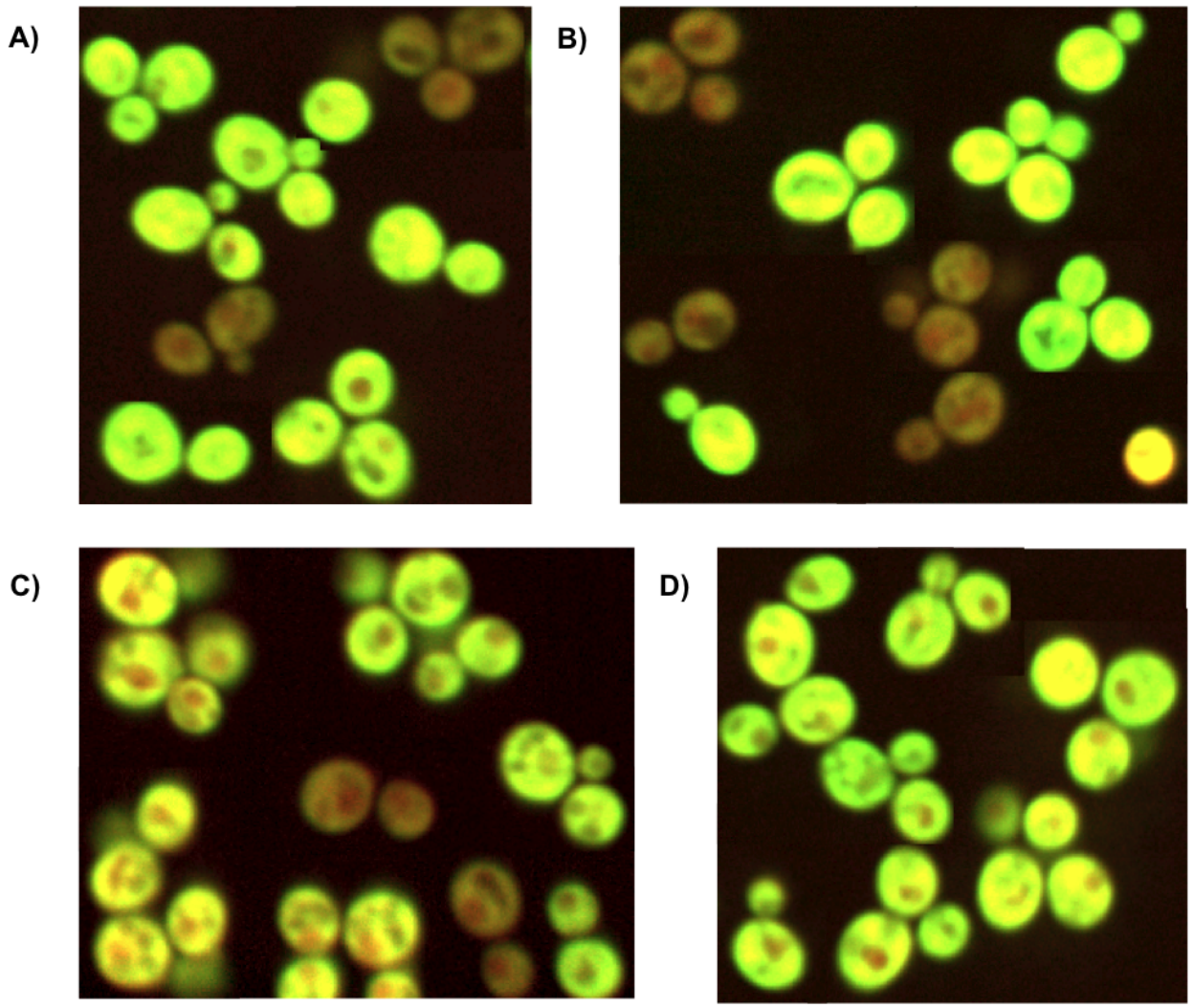


Figure III.5: **Variability in protein abundance for ribosome subunits.** RPL2A, RPL21B and RPL19B (A,B and C respectively) are ribosomal 60S subunits, which all have 'low' cell-to-cell variability in protein abundance according Newman et al. [113] and 'high' cell-to-cell variability using the 'relative variability' measure. Visual inspection and intensity measurement indicates that cells exhibit two ranges of intensities, which cannot be related to cell-stage. D) RPL35B is another ribosomal 60S subunits, detected as the 1966th (out of 2008) most varying protein in abundance by Newman et al., while it is 1326th (out of 1909) for relative variability level. Hence, relative variability level reports that this protein differs from the previous 3 subunits for stochasticity in abundance, while it is not detected by Newman et al.. Cells with 'low' protein abundance could not be found in any images for this strain.

1.2.3 Robustness of Variability Estimates

Buds \ Mothers	HOwt	ura3	rap0	nctr	nhu	nmms
HOwt		.428	.383	.456	.456	.447
ura3	.451		.482	.324	.360	.370
rap0	.376	.430		.360	.419	.392
nctr	.360	.231	.246		.719	.701
nhu	.386	.259	.336	.666		.757
nmms	.373	.260	.291	.679	.739	

Table III.4: **Correlation in 'Relative Variability' Levels in protein abundance.** Correlation for 'relative variability' (RV) measured on six different image collections. The upper triangle reports of correlation relative variability in Mother cells, while the lower triangle reports correlation of relative variability in bud objects. The first three image sets (HOwt, ura3 & rap0) differ in the segmentation approach to the last three image collections (nctr, nhu & nmms; from Tkach et al. [156]).

In this section, I show that RV estimates for protein abundance are reproducible between image collections. I computed the relative variability (RV) levels using replicate experiments ('ura3' and 'rap0'), which also have RFP staining the whole cell area (as 'HOwt'). Further, I also report RV levels in image collections that use a RFP that is instead tagging Nup49, and effectively report for the nuclear envelope (Imaged by Tkach et al. [156]. Two image collections represents cells treated with either hydroxyurea ('nhu') or methyl methanesulphonate ('nmms'), while one image collection is a control ('nctr'). It proved to be often feasible to identify cells using the autofluorescence in the GFP channel (Suppl. Fig IV.7), and still use that information to measure the size of the bud cells as a cell-stage indicator. While the accuracy of the segmentation was not evaluated, close to 1 million of mother-bud pairs were identified for each image collection, whose shape could be assessed for cell confidence using the same parameters defined for the HOwt collection (except for mean RFP level).

The correlation level of RV between experiments effectively shows that RV are reproducible (Table III.4). Mild correlations levels can be explained by 1) an uneven number of identified cells between collections for each protein, 2) differences in experimental conditions and 3) biases induced by microscope settings and methodology for cells identification or protein abundance measurement. The highest correlation levels are reported between image collections rendered by Tkach et al. [156], which indicates that the treatment with DNA damage agents does not perturb RV values globally. Sampling variance appears the factor limiting the reproducibility of RV value: on average 101.2 mother-bud pairs are identified for the first 3 image collections, while 237.3 mother-bud pairs are identified in Tkach et al. [156] image collections.

Mean protein abundance is highly reproducible between experiments, one could hypothesize that biases

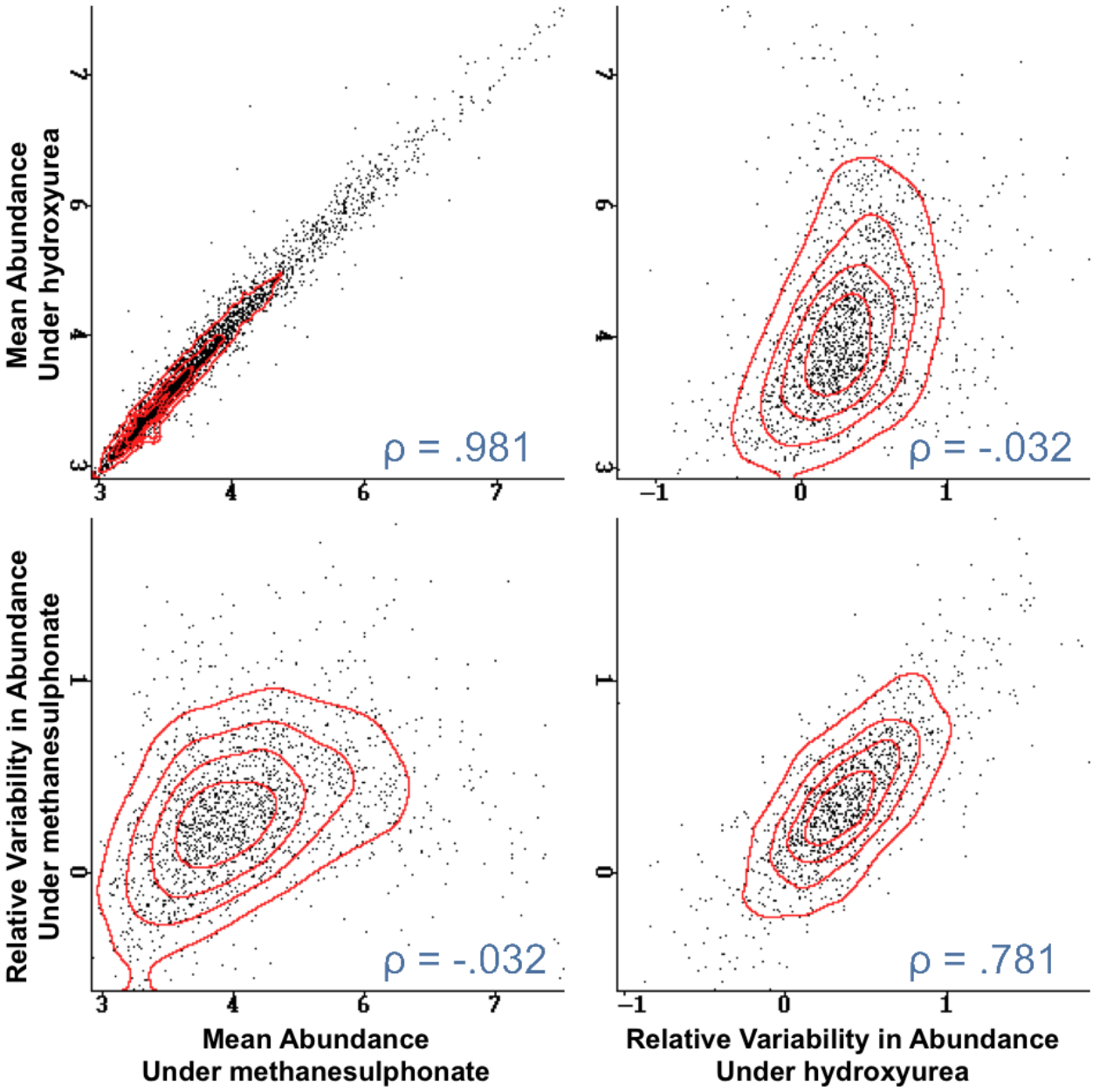


Figure III.6: **Reproducibility of 'Relative Variability' in protein abundance.** Comparison of mean abundance and 'Relative Variability' level between two experiments. Protein abundances (top-left graph, $\rho = 0.981$) and RV (bottom-right, $\rho = 0.781$) are highly correlated. On the other hand, the mean protein abundance is weakly correlated to the RV from the same experiment in both cases (top-right, $\rho = -0.03274$; bottom-left, $\rho = -0.03279$).

allows in the RV calculation are highly dependent on mean protein abundance, which explains the observed reproducibility. In order to illustrate that it is not the case, we observe that relative variability level and

mean protein abundance exhibit a lower correlation in the same experiment, than both the correlation of mean protein abundance and relative variability level across experiments (Figure III.6). This indicates that a fraction of the covariance in relative variability between experiments cannot be explained by reproducing mean protein abundance.

1.3 Discussion

I have shown that it is possible to measure cell-to-cell variability in protein abundance from microscope images, even in the event that cells in images are unsynchronized. Defining mean variability level over the cell cycle allows producing fairly reproducible levels of relative variability (RV), even certain situations where experiments are not exact replicates. I have shown that allowing the methodology to accounts for cell-stage specific changes in variability level allowed us to distinguish between proteins whose variability is explained by cell-stage (Pir1) or not (Tim17).

Flow cytometry allows measuring a far larger quantity of single cell measurements, which is appealing for accurately measuring cell-to-cell variance in protein abundance. One gain from developing a mean to detect variability from microscopy images is that it allows visual inspection to confirm the validity of predictions: reproducibility alone cannot guaranty correctness of a measure. Still, the real motivation for the use of microscopy images is the acquisition of the subcellular localization of proteins at the single cell level. For that matter, the methodology used to measure cell-to-cell variability, which is presented in the next two sections, will be shown to be applicable for variability in subcellular localization of proteins (Section 2).

1.4 Methods for Coefficient of Variation Measurement

In this section, I describe two methodologies to estimate the coefficient of variation of protein abundance (CV, standard deviation divided by the mean). While methods could be shown to capture cell-to-cell variability, I also present the some improper assumptions or undesirable properties of the two variability estimates introduced, since an additional purpose for this section is illustrate why I next consider reporting ratios of standard deviations (Relative Variability; section 1.5) as opposed to CV. The usage of 'relative variability' is a novelty of this work; hence, a first motivation for its use is presented in this section.

1.4.1 Inference based on Gaussian Process

A first approach was the use a Gaussian Process (GP [130]) to infer the level of variability in protein expression. Under such a model, an observation for protein abundance 'y' at cell-stage 'x' is generated by

an unknown function and noise parameter σ_n^2:

$$y = w^T \cdot \vec{\phi}(x) + \epsilon \quad \text{where} \quad w \sim N(\vec{0}, \Sigma) \quad \text{and} \quad \epsilon \sim N(0, \sigma_n^2) \tag{35}$$

GP is non-parametric model, which utilizes the list of observations to construct $\vec{\phi}(x)$ and Σ, which defines the unknown function. The major appeal of this method for our task is that the level of cell-to-cell variability that we seek is a parameter of the model and therefore can be obtained directly from the σ_n^2 in the maximum likelihood parameterization for a Gaussian process that explains the data. One disadvantage though is that the model assumes the noise is identical for all observation, which in turn implies that the level of variability is constant through-out the cell cycle. It also ignores uncertainty in the cell-stage assessment. However, because this bias will systematically contribute to variability estimates for every protein, the overall ranking of variability levels might still be correct.

The Gaussian process uses as prior that the mean of observation is zero, which is not the case for protein abundance. Instead, I compute the mean of the protein abundance, and use Gaussian Process to model deviation to this sample mean, as oppose to model protein abundance directly. Doing so, the CVs are reported by dividing σ_n by the square root of that sample mean. For datapoints are sampled using equation 35, the covariance between any two observations is:

$$E(y_i) = 0 \quad , \quad Cov(y_i, y_j) = E(y_i y_j) = \vec{\phi}(x_i)^T \Sigma \vec{\phi}(x_j) \quad \forall i \neq j \tag{36}$$

Instead of characterizing $\vec{\phi}(x)$, which is a priory could be a function spawning a vector of infinite length, a prior is utilized to characterize the covariance of data points:

$$k_{i,j} = Cov(y_i, y_j) = \sigma_f^2 e^{-(\frac{x_i - x_j}{L})^2} + \sigma_n^2 \delta_{\{i=j\}} \tag{37}$$

Equation 37:**Covariance function.** Hyper-prior model for the prior model of covariance of measurement (y_i) as a function of their state (x_i). Under the prior model, data points displays a certain level of variance explained by a function defined on the state σ_f^2 and explained by noise σ_n^2. The parameter L relates to the distance in state variable x_i needed for to data points to become uncorrelated. Since these three parameters describe the GP covariance prior K, we consider them hyperparameters. This characterization has a finite (3) number of parameters, which are optimized instead.

Given a list of observation and cell-stage estimates, the covariance matrix 'K' between all observations is

defined from equation 37. The likelihood '$p(y|K)$' of a set of observations can be evaluated directly from the formulation of K:

$$p(y|K) = -\frac{1}{2}y^T K^{-1} y - \frac{1}{2}log(|K|) - \frac{1}{2}log(2\pi) \tag{38}$$

$$\frac{dlog(p(y|K))}{d\alpha} = \frac{1}{2} \cdot tr\left((K^{-1}yy^T K^{-1} - K^{-1})\frac{dK}{d\alpha}\right) \tag{39}$$

Equation38&39:**Likelihood function under GP prior.** Likelihood of a vector of observations '$\vec{y}$', under a prior covariance matrix 'K'. Estimation of the maximum likelihood parameters for Gaussian Processes

requires parameter updates that uses gradient ascent on the likelihood surface (see eq 38 & 39). I use Newton's method in order to update the three hyper-parameters (L, σ_n, σ_f). One concern is that there may be an arbitrary number of local maxima in the likelihood function. In order to obtain robust estimate of CV and avoid randomly selecting a local optima, the iterative procedure needs to be performed using a large number (1000) of initial guesses, so that the parameter set with the highest likelihood is utilized. CVs are obtained from normalizing the noise term σ_n by the mean abundance (Table III.5).

As expected, the Gaussian process uncovers the best noise term that explains the observations through the cell cycle (Figure III.7). Nuclear proteins show lower variance in early bud than in larger buds, which relates to the inclusion of the nucleus into the bud at a specific time. As previously mentioned, the noise level is not allowed to vary throughout the cell cycle, which results in an overestimation of the true variance. A second undesirable property is that CV may differ from each other due to the numerical procedure, whose robustness is hindered by sampling variance that changes in the likelihood surface (Figure IV.6). In order to account for cell-stage dependent variability level, I next consider an approach that allows levels to fluctuate within the cell cycle.

ORF	Name	subcellular Localization	Bud Var Coef	Mother Var Coef
YBR115C	LYS2	cytoplasm	0.89235	0.85623
YHR136C	SPL2	punctate composite	0.84707	0.77240
YER103W	SSA4	cytoplasm	0.77369	0.82136
YPL079W	RPL21B	cytoplasm	0.74864	0.80684
YPL081W	RPS9A	cytoplasm	1.15269	0.45937
YJL200C	ACO2	mitochondrion	0.56992	0.80600
YPL160W	CDC60	cytoplasm	0.60610	0.71057
YML121W	GTR1	vacuolar membrane	0.73579	0.54725
YER102W	RPS8B	cytoplasm	0.76233	0.50761
YHR144C	DCD1	cytoplasm	0.61698	0.61485
YLR390W-A	CCW14	ER	0.55021	0.53453
YDL181W	INH1	mitochondrion	0.56553	0.49383
YKR094C	RPL40B	cytoplasm	0.47586	0.55978
YFR031C-A	RPL2A	cytoplasm	0.53254	0.47804
YJL143W	TIM17	cytoplasm,nucleus	0.49236	0.51645
YCL009C	ILV6	mitochondrion	0.51919	0.48877
YJL153C	INO1	cytoplasm	0.42467	0.59318
YOR210W	RPB10	nucleolus,nucleus	0.59052	0.36483
YNL134C	YNL134C	cytoplasm,nucleus	0.43932	0.48071
YPL093W	NOG1	nucleus	0.58000	0.36391

Table III.5: **Most variable under Gaussian Process model.** Proteins ordered by the geometric mean of the coefficient of variation for the bud and mother cell.

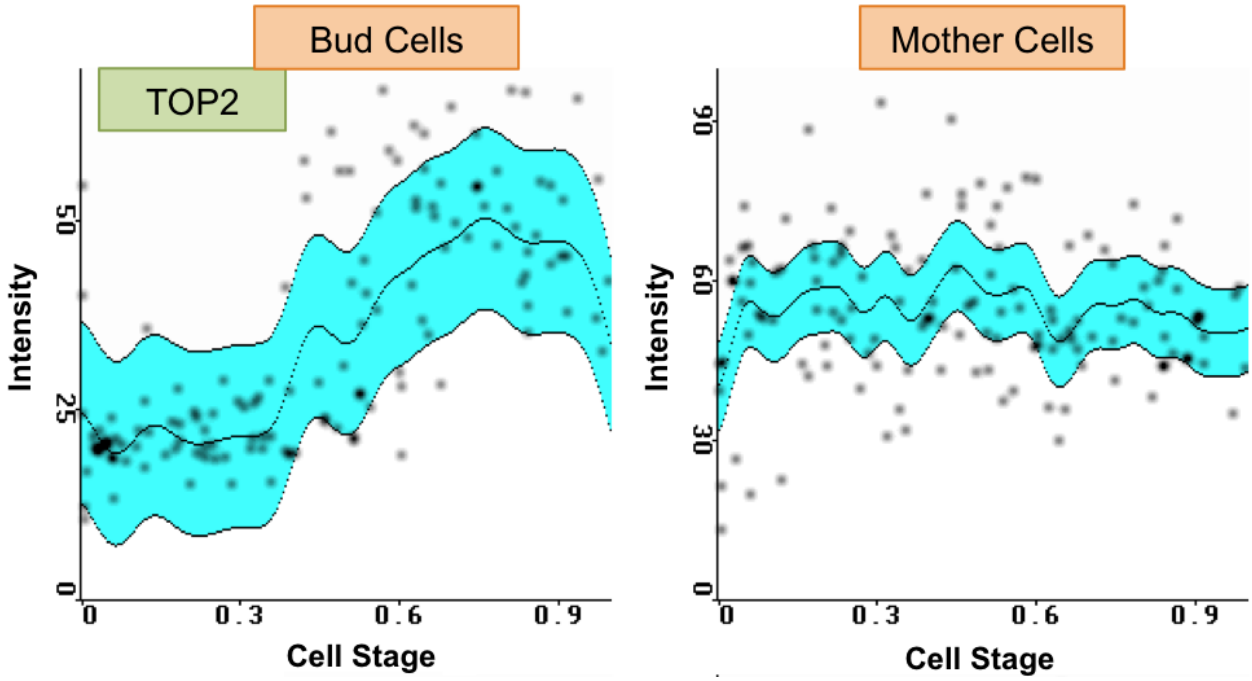


Figure III.7: **Gaussian process fit.** Representation of the Gaussian Process fit to the TOP2 protein abundane in Bud and Mother cells, as a function of bud size (cell-stage). The cyan area represents where 50% of observations of protein abundance occurs, under the Gaussian Process model.

1.4.2 Linear Regression in Cell Stages Bins

In this work, the quantification of protein abundance has been performed by reporting the average GFP intensity over the cell area. The need of correcting for autofluorescence level in the calculation of CVs was noted by Newman et al. [113]. Any additive term to all protein abundance measured would scale down CVs for protein of low abundance more than for high abundance ones. One puzzling observation specific to this set of microscopy images is that there appears to be a significant global correlation between RFP levels and GFP levels from cell-to-cell, which justified the normalization of GFP intensities by mean RFP intensity. The level of correlation varies in a non-trivial manner depending on the size of objects and the subcellular localization of the GFP tagged protein (Table III.6). Autofluorescence in the GFP channel may be a major contributing factor to this observation, but there is no clear explanation for a correlation between RFP to either GFP or autofluorescence in images. Typically, protein of low abundance show higher levels of correlation, but these correlations differ from strain to strain significantly (0.34 ± 0.19 std. dev. for the 1000 proteins with lowest intensity; correlation levels evaluated from a minimum of 10 Mother-bud pairs).

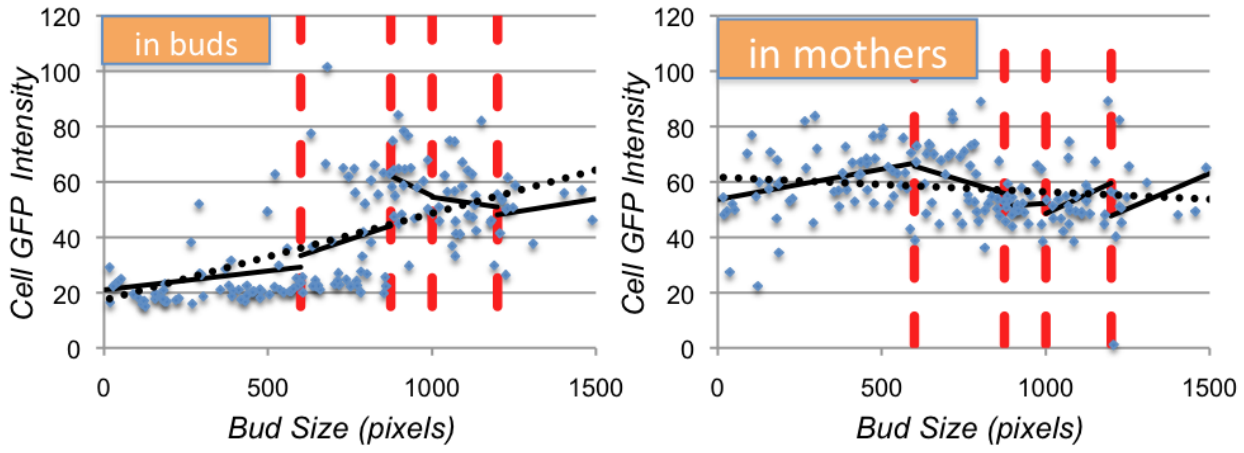
	Nucleolus	Nucleus	Mitochondrion	Golgi	Cell periphery	Cytoplasm	Vacuole	Actin	Bud neck	Lipid particle	Punctate	Endosome	Microtubule	Peroxisome	Ambiguous	ER
Buds	.082	.118	.170	.158	.186	.202	.162	.107	.203	.186	.180	.188	.216	.176	.237	.179
Mothers	.186	.291	.282	.250	.309	.303	.245	.203	.354	.262	.283	.339	.390	.308	.363	.293

Table III.6: **Correlation of fluorophore intensities.** Average of correlation level between RFP intensity and GFP intensity, as dependent of the localization of the GFP-tag (Huh et al. [76]). Proteins of low abundance, such as ambiguous proteins, show high level of correlation.

Since no prior knowledge appears to explain the observed correlations, an alternative is to explicitly quantify and control for the fraction of variance that can be explained by either cell size or RFP level. This task is equivalent to use linear regression to model their specific contributions to the raw GFP intensities. Similar to the method I previously proposed for 'cell confidence' (section 2.2), this measurement of covariance between GFP and RFP are be performed in 5 cell-stages bins, so to capture covariances levels that have a complex dependency on cell size. The bin ranges for bud size were selected so to partition the 400k identified buds equally.

For each bin, I computed the coefficient of variation, where the fraction of the variance explained by bud size and RFP intensity has been regressed out. Then I use the geometrical average to reduce the influence

of sampling variance on the reported quantity, and note that this average agrees with the stochasticity levels that were reported in Newman et al. [113]. Among the list of 50 most variable proteins reported by geometrical average of the coefficient of variation and those reported in Newman et al., 20 proteins fall in both lists. Also, a global correlation of 0.303 is observed on the collection of 2008 yeast strains for which variability levels were available in both cases. (Figure III.8).



Variance	1st bin	2nd bin	3rd bin	4th bin	5th bin
Total	85.3	301.0	136.8	92.2	41.2
Residual	83.1	286.2	136.4	89.7	40.8

Variance	1st bin	2nd bin	3rd bin	4th bin	5th bin
Total	191.6	247.6	129.4	76.0	232.4
Residual	191.3	246.5	113.1	66.8	228.5

Figure III.8: **Coefficient of variation from cell-stage bins.** Plot the intensity of the GFP tagged TOP2 protein within identified buds (left) and Mother (right) cells. The red dashed line represents the bins threshold, so that the linear dependencies of GFP intensities and bud size are reported with black lines. If the cells are not binned, the linear fit to the cell is shown using the black dashed line. The table at the bottom report on total variance of abundances and the residual variance from the linear regression in each bin.

The effort to account for particular biases in observations comes at the cost of an increased sampling variance: each linear regression required at least four mother-bud pairs per bin to measure any non-zero residual. As such, quantification of absolute variability level is difficult from microscopy image exhibiting such correlation levels. As previously noted in table III.1, coefficients of variation produced show a stronger correlation to the deviation to the median (DM) that were also reported by Newman et al. than to the confident of variability (CV) [113] (correlation: 0.308 vs. 0.144; spearman rank correlation: 0.243 vs. -0.06). The fact that local comparison (differences) of variability estimate agree more than the absolute quantifies suggests that the scale of variability estimates strongly disagrees between the two experiments. This suggests that normalization of intensity data disagree between the two experiments, which hinder the reproduction

variability levels. To test this hypothesis, I next introduce a measure explicitly defines 'relative' levels of variability from comparing variability levels, which is a framework that does not rely on a proper normalization of the intensity data for protein abundance quantification.

1.5 Methods for 'Relative Variability'

Due to similitudes in the calculation, ratio of standard deviations could be alternatively used instead of differences in coefficient of variations utilized by Newman et al. [113] (DM). While they compare coefficients of variation by looking at their differences (subtraction), one could report ratios of CV instead. If this alternative is considered, normalizing the variance by the mean abundance (in the CV calculation) may have no effect on ratios of CVs: variability levels can be compared directly by assuming that mean abundance and local mean abundance are nearly identical:

$$\text{if} \quad E(X)(1-\delta) \leq E(Y_i) \leq E(X)(1+\delta) \quad \forall i \quad \text{then} \tag{40}$$

$$(1-\delta)\frac{\sqrt{Var(X)}}{\frac{1}{n}\sum_{i=1}^{n}\sqrt{Var(Y_i)}} \leq \frac{\sqrt{Var(X)}}{E(x)} \Bigg/ \frac{1}{n}\sum_{i=1}^{n}\frac{\sqrt{Var(Y_i)}}{E(Y_i)} \leq (1+\delta)\frac{\sqrt{Var(X)}}{\frac{1}{n}\sum_{i=1}^{n}\sqrt{Var(Y_i)}} \tag{41}$$

where X is the protein abundance for a protein for interest, while Y_i is the protein abundance of proteins with similar mean abundance. The argument is that if all proteins compared have similar mean abundance (equation 40), then ratio of coefficient of variation (middle of the equation 41) is also similar to ratios for variances (lower and upper bounds). For this reason, I next report ratios of variance to a 'local' variance estimate directly (relative variability; RV). Assuming that ratios of intrinsic to total variance are locally constant, this ratio also represents fold change in stochasticity level.

1.5.1 Modeling Deviation to Expectation

Since I have shown that cell-to-cell variability may be cell-stage dependent, the same 5 cell-stage bins will be utilized (Section 1.4.2). Instead of residuals to linear regressions, I compute the deviations for each abundance in 'Bud' and 'Mother' to a local mean $F(c)$ that is expected using local regression on all mother-bud pairs identified (not bin specific), as it was defined in equation 29. Deviations to the time-profile $F(c)$ are modeled using a multivariate normal distributions with mean μ^*_{pj} and covariance Σ^*_{pj} in the j^{th} bin:

$$\begin{aligned} \mu^*_{pj} &= \tfrac{1}{w_{pj}} \sum_{i \in B_{pj}} c_i(\vec{x_i} - F(|S_i|^{\frac{3}{2}})) \quad , \quad \mu_{pj} = \tfrac{1}{w_{pj}} \sum_{i \in B_{pj}} c_i \vec{x_i} \\ \Sigma^*_{pj} &= \tfrac{1}{w_{pj}} \sum_{i \in B_{pj}} c_i(\vec{x_i} - F(|S_i|^{\frac{3}{2}}))(\vec{x_i} - F(|S_i|^{\frac{3}{2}}))^T - \mu^*_{pj}\mu^{*T}_{pj} \quad , \quad \text{where} \quad w_{pj} = \sum_{i \in B_{pj}} c_i \end{aligned} \tag{42}$$

where $F(|S_i|^{\frac{3}{2}})$ is the local regression evaluated at the cell-stage estimate $|S_i|^{\frac{3}{2}}$ for the i^{th} cell. The set of mother-bud pairs assigned to j^{th} cell-stage bin is indicated using B_{pj}; the cell confidence of the i^{th} cell is indicated using c_i (previously described as '$P(Cell|\vec{q_i}, |S_i|)$'). Computing deviations to a local regression (here denoted with '*') does not guaranty that the mean deviation in the any bin μ^*_{pj} is zero. The cell-to-cell variance in deviations to the time-profile in the 5 bins are (Σ^*_{pj}), which are 2 by 2 matrices that incorporate potential covariance for 'Bud' and 'Mother' protein abundances. Mean and variances are concatenated, so to define 'time-profiles' for deviations (denoted with a single subscript):

$$\mu_p = \begin{vmatrix} \mu_{p1} \\ \mu_{p2} \\ \mu_{p3} \\ \mu_{p4} \\ \mu_{p5} \end{vmatrix} \quad , \quad \mu^*_p = \begin{vmatrix} \mu^*_{p1} \\ \mu^*_{p2} \\ \mu^*_{p3} \\ \mu^*_{p4} \\ \mu^*_{p5} \end{vmatrix} \quad , \quad \Sigma^*_p = \begin{vmatrix} \Sigma^*_{p1} & 0\ 0 & 0\ 0 & 0\ 0 & 0\ 0 \\ 0\ 0 & \Sigma^*_{p2} & 0\ 0 & 0\ 0 & 0\ 0 \\ 0\ 0 & 0\ 0 & \Sigma^*_{p3} & 0\ 0 & 0\ 0 \\ 0\ 0 & 0\ 0 & 0\ 0 & \Sigma^*_{p4} & 0\ 0 \\ 0\ 0 & 0\ 0 & 0\ 0 & 0\ 0 & \Sigma^*_{p5} \end{vmatrix} \tag{43}$$

1.5.2 Local Regression for Variability level

In order to find proteins that are highly variable, we then first need to characterize the 'local' expected level of variability, which changes depending on the mean protein abundance at the different cell-stages. For example, we know that protein of different subcellular localization appears in the bud at different times, which may result in different level of variability at different cell-stages. Hence, the proteins that have a similar cell-stage variation in their abundance can be used to characterize expected variability levels, as protein of similar abundance may have different variability level at a given cell-stage from their subcellular localization alone. For example, nuclear protein are absent in small bud, it may be undesirable to compare their variability level at that cell-stage to other cytoplasmic proteins of small abundance. Instead, their variability should be compared to proteins with the most similar abundance and subcellular localization pattern if possible.

To define 'similar' protein, I use the mean abundance in each bin to form 'time-profiles' of protein abundance, which can be compared using the Euclidean metric. An alternative approach would be to simply compare variability for protein with the same subcellular localization (e.g. nuclear proteins), but ranking of only the nuclear protein by their variability identified protein that mislabelled; upon inspection, they were not nuclear proteins (see section 2.2). Further, comparing variability based on the time-profiles allows noise levels in the quantification protein abundance (e.g. autofluorescence) to be comparable between proteins, as

this contribution in the estimate of the variance depends on the mean protein abundance in each bin. Hence, the similarity measure will be based exclusively on time series of protein abundances (no morphological measurements). We use local regression to estimate level of variability from a collection of proteins 'Ω' for a given time-profile:

$$\hat{\mu}^*_{pj} = \frac{\sum_{q \in \Omega - \{p\}} w_{qj} \mu^*_{qj} \cdot e^{-b_p (\mu_q - \mu_p)^T (\mu_q - \mu_p)}}{\sum_{q \in \Omega - \{p\}} w_{qj} e^{-b_p (\mu_q - \mu_p)^T (\mu_q - \mu_p)}} \tag{44}$$

$$\hat{\Sigma}^*_{pj} = \frac{\sum_{q \in \Omega - \{p\}} w_{qj} (\Sigma^*_{qj} + \mu^*_{qj} \mu^{*T}_{qj}) \cdot e^{-b_p (\mu_q - \mu_p)^T (\mu_q - \mu_p)}}{\sum_{q \in \Omega - \{p\}} w_{qj} e^{-b_p (\mu_q - \mu_p)^T (\mu_q - \mu_p)}} - \hat{\mu}^*_{pj} \hat{\mu}^{*T}_{pj} \tag{45}$$

where $\hat{\mu}^*_{pj}$ and $\hat{\Sigma}^*_{pj}$ are local estimates of protein abundance for each bin j based on similar proteins. In contrast to the local regression used to define 'time-profiles', this local regression (denoted with ' ˆ ') uses similarity between protein time-profiles (μ_p) as opposed to a numerical representation of cell-stage. μ^*_{pj} and Σ^*_{qj} correspond to deviation to local regression (see equation 42 and 43). In addition, the local regression required distances are scaled by a factor 'b_p' that is specific to the protein 'p' of interest. This bandwidth parameter is iteratively updated so that the denominator of the above two equations converges to a constant. Newman et al. [113] defined fixed size windows of mean protein abundance to extract the median coefficient of variation; here, the windows are replaced by a kernel function that weights the relevance of measured variability level based to the distance to the protein profile of interest. In analogy with the definition of a window for computing a median, we enforce a standardized number of time-profiles (1% of the time-profiles) that is used to evaluate the expected level of variance (Figure III.9). The rationale is that similarity of time-profiles may differ by folds depending on the type of cell-stage dependencies and/or subcellular localization pattern: larger bandwidths are needed for actin proteins relative to nuclear or cytoplasmic proteins, because there are few actin protein and because they are associated with variable morphological structures and are therefore much further apart from each other in the feature space (Figure II.11 A).

Finally, we estimate variability levels using the log-ratio of the determinants of the observed sample covariance matrix by the covariance matrix predicted by local regression (Eq. 46). This defined 'relative variability' is a summary statistic of the cell-to-cell variability: it essentially is equivalent to report the geometric average of the variance ratios from each of the 5 cell-stage bins.

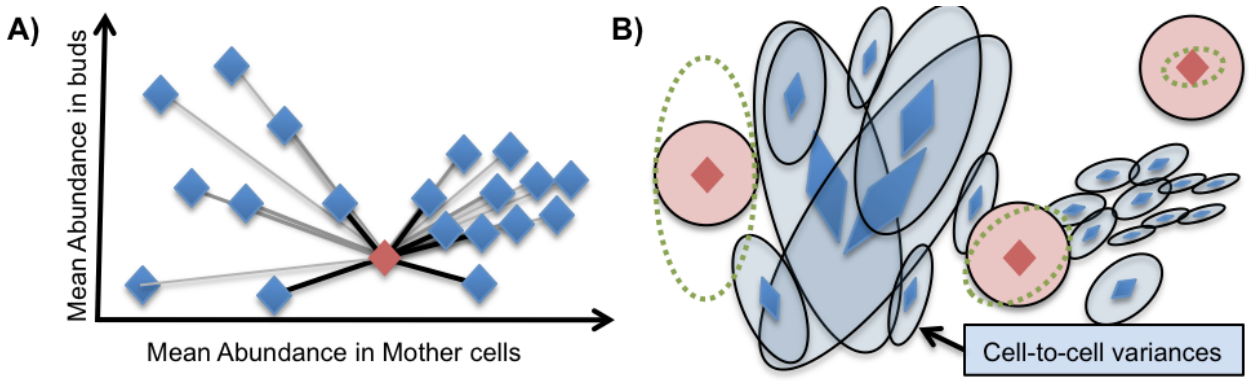


Figure III.9: **Schema for the definition of 'Relative Variability' level.** A) Proximity of protein time-profiles to a query point (red lozenge). Local regression on variability level associate weights to neighbouring protein profiles (gray scale lines), so infer a variability level based on mean level of protein abundance (simplified to a single cell-stage bin). B) For the same point collection, visualization of the cell-to-cell variance in time-profiles, including covariance for bud and associated mother cell measurements. Local regression of variabilities at three query points are shown using green dotted ellipses. In a cell-to-cell variance of a protein corresponds to the red circle, comparison to the local variability level identifies this protein as exhibiting 'low', 'average' or 'high' relative variability level (left to right respectively).

$$\text{Relative Variability}_p = \frac{1}{2 \cdot Rank(\Sigma_p^*)} \cdot \left(log(|\Sigma_p^*|) - log(|\hat{\Sigma}_p^*|) \right) \tag{46}$$

1.5.3 Significance of Local Differences in Variability level

The number of identified cells per protein varies significantly throughout the image collection. Under the multivariate Gaussian assumption, I can define likelihood ratio tests to evaluate statistical significance of unequal variances between groups of observations. The probability that high RV measurements are due to sampling variance is shown to be negligible.

We model deviations to the local regression mean using multivariate normal distribution, so that the significance of deviations from two collections of observation can be evaluated using a Log-likelihood ratio test. In particular, we evaluate the probability that deviations level of variance in protein abundance are caused by sampling variance. Note that both the local regression and the cell confidence down acts as a weight in the estimation of the variance, so that they both may be included so to evaluate significance of deviations; the maximum likelihood parameter of a local regression of multivariate normal is inferred under

the three possible assumptions 1) equal mean and covariance, 2) non-equal mean and equal covariance 3) unequal mean and covariance (Appendix 6). As such, we can test whether the protein of interest specifically has a different covariance than 'nearby' proteins, while all protein mean expression are allowed to differ but have identical covariance structure. The null hypothesis we attempt to reject allows neighbouring protein to have different mean abundance; otherwise the divergences in mean abundance will dominate in the statistical test (Figure III.10).

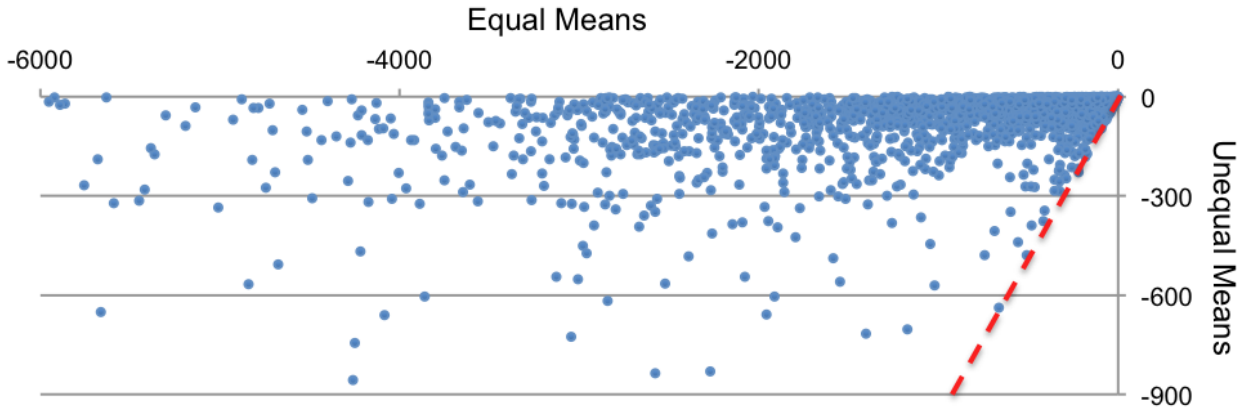


Figure III.10: **Log-P-value for likelihood ratio tests for variability in protein abundance.** Log_{10}-P-values for rejecting the local abundance model (inferred using local regression) from difference in covariances when the local model for the null hypothesis assumes all neighbouring protein have either identical mean abundance or different means. The dotted red like indicates that rejecting equal mean is always more significant than if the null hypothesis allows unequal mean.

The likelihood ratio test controls for proteins that do not have enough identified cells such that differences to neighbouring proteins are due to sampling variance. For any protein where the null hypothesis can only be rejected with probability smaller than 0.1%, we do not report their relative variability levels. Note that the distribution of the feature measurements is not necessary multivariate normal distribution, hence the sole purpose of the likelihood ratio test is to account for sampling variance, by evaluating the probability that high or low relative variability levels are due to sampling variance, under the multivariate Gaussian assumption and further assuming measurements are independent, which is violated if the foreground/background separation differs between images for example. One interesting observation is that the statistical significance of protein of 'high' and 'low' relative variability is radically different (Figure III.11). This is due to the fact that the number of samples needed to reach a certain statistical significance is far higher when claiming that measured variability is lower than expected compared to claiming that it is higher than expected. The fraction of the protein that succeeds to reject the null hypothesis with the required probability varies from image collection to another (Table III.7), and proteins that have 'high' RV are more frequently reported.

	HOwt	ura3	rap0	nctr	nhu	nmms
#proteins	1895	2787	2314	2345	2389	2324

Table III.7: **Number of proteins with significant 'Relative Variability' levels.** Number of 'Relative Variability' measurements whose value cannot be explained by sampling variance under the multivariate Gaussian assumption (P-value $\leq .001$).

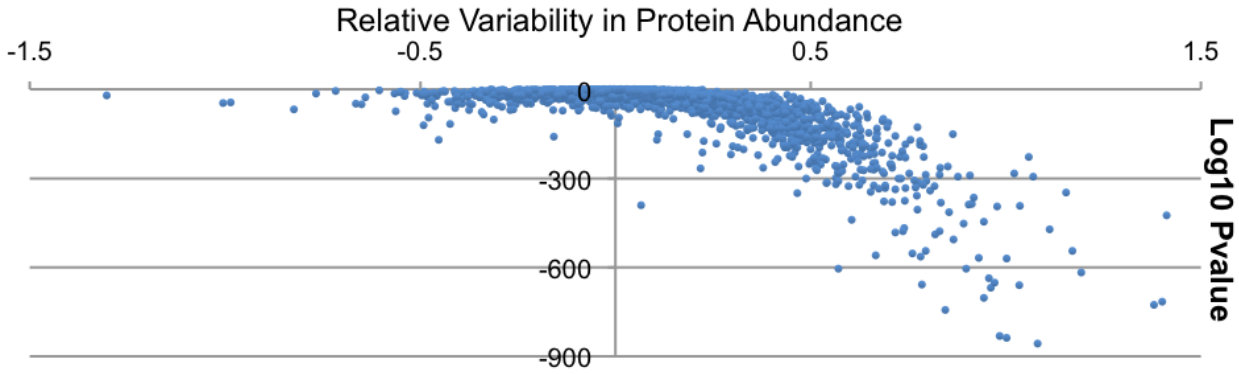


Figure III.11: **Significance of deviations for measured relative variability level.** Log_{10}-P-values for rejecting the local abundance model, as a function of relative variability. Since the significance scales with number of identified cells, we note that more identified cells are required to find proteins of 'low' relative abundance compared to finding significant proteins with 'high' variability levels.

2 Spatial Variability

2.0 Background:

Cell-to-cell variability in protein spatial spread has not been characterized by high-throughput methods from fluorescence microscopy. We want to test if we can quantify variability in spread and relate it to protein function. There are known examples of proteins that show stochastic changes in subcellular localization, such as CRZ1 [24] (Figure III.12), but it remains unknown how many proteins show these type of variable spatial patterns, or whether some proteins have more variability than others. A means to quantify of the spatial variability has not been previously proposed as a global measure on yeast proteome. Methodology to define heterogeneity in subcellular localization for yeast cell populations inherently depends on the choice of methodology for recognizing subcellular localization. In the case of supervised approaches, decision boundaries might vary depending on the choice of image feature and the training data utilized. Although no

systematic reports of cell-to-cell variability in subcellular localization have appeared, the current opinion in the field is that much biological insight is expected from a better quantitative characterization of protein expression and its variability [107, 121].

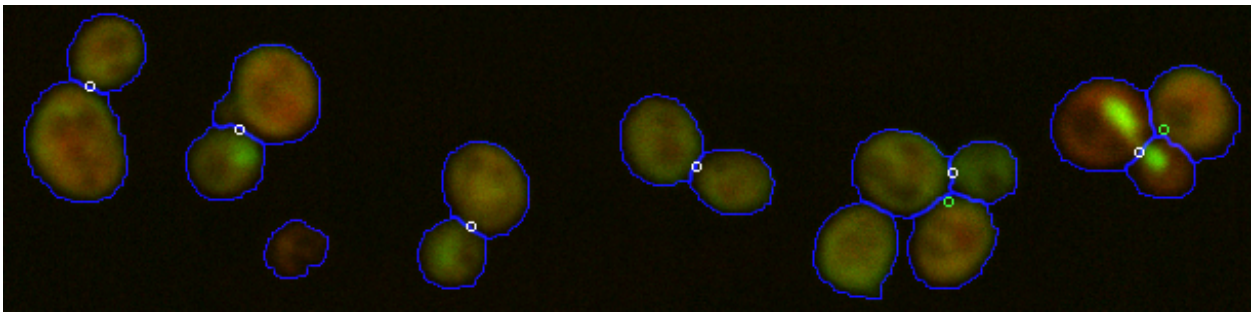


Figure III.12: **Previously known spatially variable protein.** Microscope image of CRZ1, which has been previously characterized as a protein that sporadically localizes in the nucleus. In the image collection, the above contains the three most compelling cases of nuclear localization, out of that 111 identified mother-bud pairs.

We therefore sought to develop statistical methods to search through our large image datasets to systematically identify new candidate spatially variable proteins. As this has not been attempted before, it is necessary to show that any quantitative measure of spatial variability meets several requirements: First, it is reproducible across different experiments, and second, is not simply the result of sampling variance or abundance variability. Here I address all of these issues, and provide (to our knowledge) the first systematically defined list of proteins whose expression patterns are the most spatially variable. One of the results from the unsupervised analysis from the previous chapter, is there are protein of similar function that display similarity in their expression that is not characterized by subcellular localization alone. In this section, I show that quantified variability in morphological measure are reproducible, which indicate that levels spatial variability is a property that can be attributed to proteins. The reuse of the methodology previously introduced for protein abundance allows the identification of a number of proteins that exhibit cell-to-cell heterogeneity in subcellular localization.

2.1 Approach

Proteins known to display spatial variability change subcellular localization from cell to cell, so it is natural to first characterize variability by recognition of subcellular localization at the single cell level. As previously discussed (Section 0.2), classification of subcellular localization does not intrinsically define continuous

lineation for intermediate subcellular localization, unless specific methods need to be considered to that aim [124]. Since proteins that change localization change in proportions for two or more subcellular localizations, quantification of spatial variability will suffer from the same limitations that classification methods for characterizing fraction of mixed localization have.

One possible approach to detecting proteins with cell-to-cell variability in subcellular localization patterns is to use a supervised classification system to put each cell into a localization class, and then look for images that contain cells in multiple localization classes. However, since a number of proteins appears be uniquely recognized (Section 4.2), proteins with cell-stage specific deviations and atypical mixtures of subcellular localizations are likely to be detected as exhibiting high spatial variability due to inaccuracies of the recognition of subcellular localization at the single cell level. One clear example of such proteins are the MCM subunits (Figure II.9), which systematically change their subcellular localization from nuclear to cytoplasmic for the portion of the cell cycle corresponding to the budding process. While the supervised method are powerful method so determining the subcellular localization, it is hard to distinguish between heterogeneity caused by inaccuracies in class label recovery and by systematic biological variation in subcellular localization (such as through the cell cycle) and true 'intrinsic' cell-to-cell variability.

Hence, the challenge in defining a measure for variability in subcellular localization, as it may need to combine a number of image features. Quantification of spatial variability is a hard problem because spatial image features have no natural scale (so a simple CV cannot be used) and summary statistics based on these features cannot necessarily be compared between images with very different feature distributions. I am not aware of any attempts to report level of spatial variability in a high-throughput manner. Here I will measure spatial variability using 'time-profiles' of simple image features, which reflect the hierarchy of subcellular localization (Section 2.4.1). Similar to the supervised approach, the interpretation of variance in feature measurement is difficult: the scale of each feature measure differs from subcellular localization to another, as proteins are typically constrained in compartment, whose sizes are also variable.

In order to infer what is the scale for typical deviation in feature measurement, I use the same approach as I used for protein abundance (Section 1.5), which does not require the knowledge of the natural scale for the variance in the feature measurement. By comparing proteins that are similar in their morphological measurements and their corresponding cell-stage dependence, we derive a measure of variability that is typical for that subcellular localization. In the next section, I apply the methodology to several image feature of interest, and compare obtained relative variability levels.

2.2 Results

2.2.1 Image Features and Spatial Variability

In this section, I show that the choice of the image feature (among 4 features considered) used to quantify spatial variability has a limited impact on RV measurements for the analyzed image collection. Relative variability (RV) evaluated using 4 different image features have high global correlation (Figure III.13). I will use a single morphological feature proved to distinguish certain subcellular localizations, so it alone could predict change in subcellular localization:

$$E(Dist_{\text{Proteins to GFP Mass Center}}) = \sum_{\vec{x} \in S} ||\vec{x} - \left(\sum_{\vec{x} \in S} \vec{x} \frac{G(\vec{x})}{T_G} \right) || \frac{G(\vec{x})}{T_G} \tag{47}$$

The above measure, which was referred to as distance of center of mass, will be referred to as 'subcellular spread' in this chapter.

Note that this result may be due to the fact 3 out of 4 image features considered are highly correlated (or anti-correlated). Still, in the case of distance to the bud-neck, the correlation in unexpectedly higher for the relative variability level than the mean feature measurement (Figure III.13). While better choices for image feature to define spatial variability may be suggested, this indicates that the RV measures captures similar deviations, even if the measurement slightly changes.

A high correlation (0.614) is observed between RV levels for protein abundance and subcellular spread (Figure III.14). Protein abundance is expected to influence image feature measurements, since the noise level to pixel intensity is relatively high due to auto-fluorescence in images. Still, the proteins within highest RV in subcellular spread do not appear to be variable in protein abundance from visual inspection (see Section 2.2.3). Conversely, TIM17 is known to be highly variable in protein abundance (see Section 1.2.2) and does have a high RV in subcellular spread, which agrees with the fact that TIM17 is systematically be localized to the cytoplasm in microscope images (Figure III.1 A).

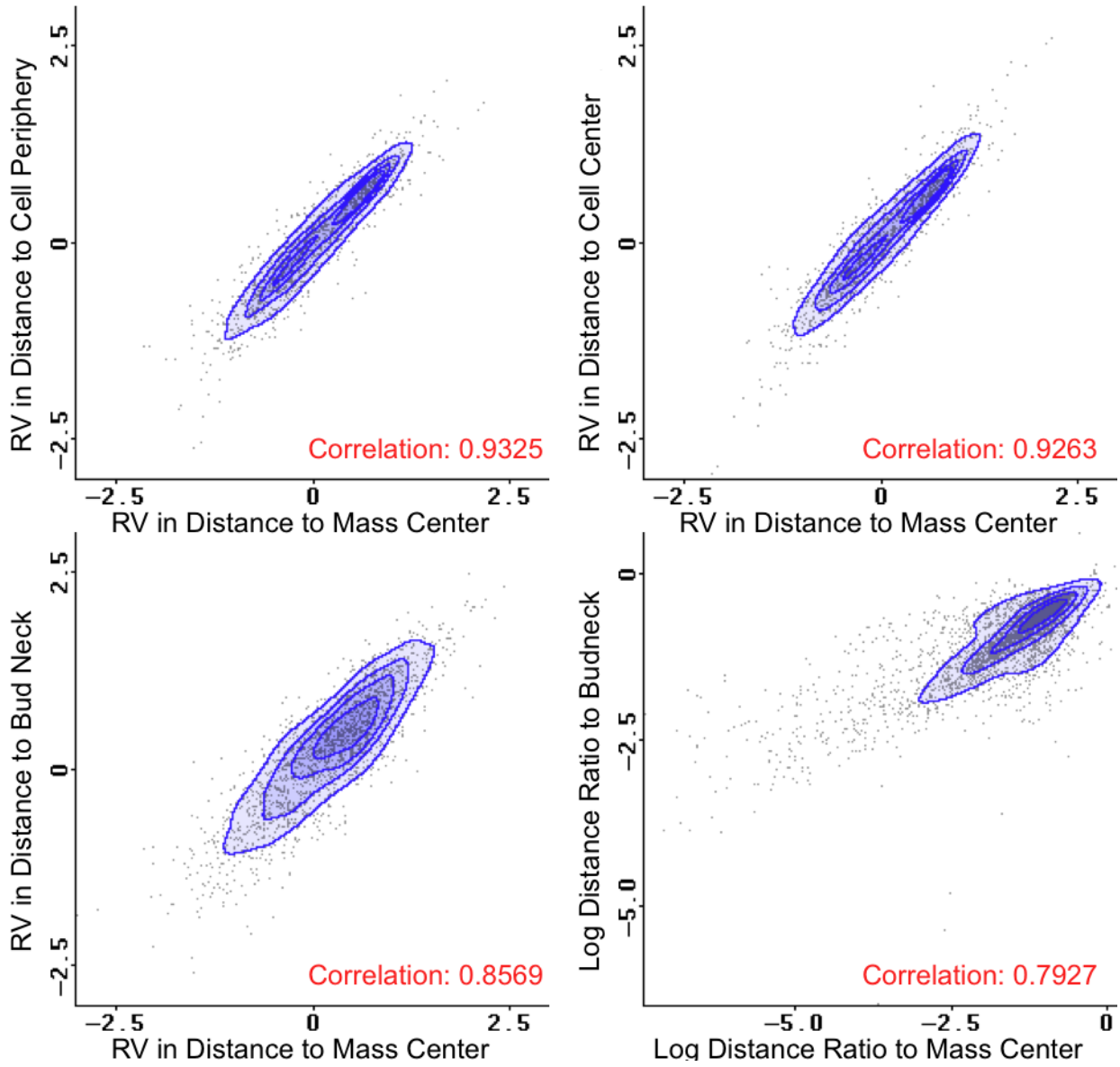


Figure III.13: **Comparison of relative variability using different image features.** Relative variability (RV) levels measured on 1296 proteins generates relative variability levels that are highly correlated (correlation displayed in red). Feature measurements themselves are highly correlated or anti-correlated; still, the correlation of RV measures is higher than the correlation between image features for the specific comparison of distances to budneck and mass center (0.8569 and 0.7927 respectively).

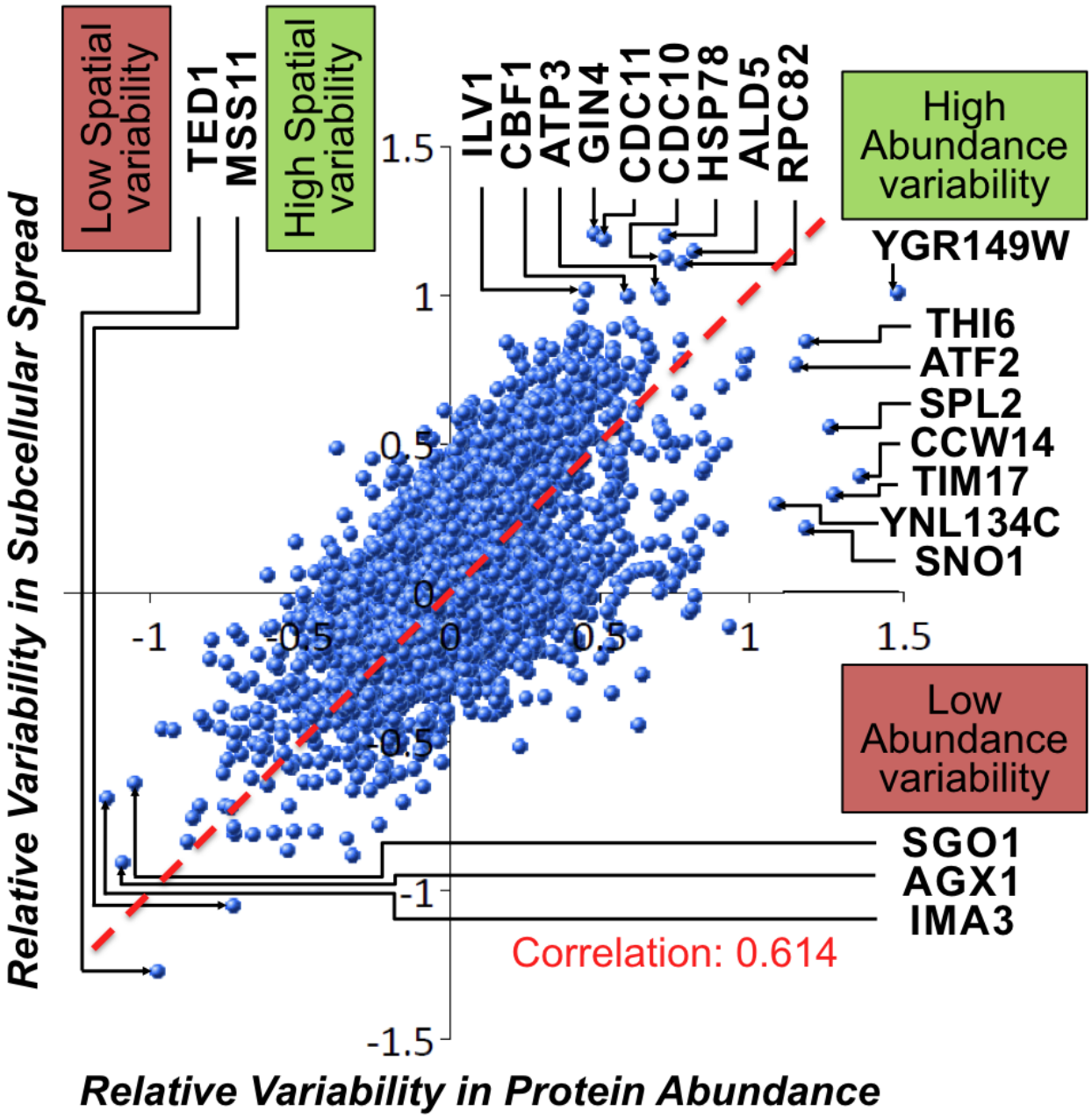


Figure III.14: **Relative variability in protein abundance and subcellular spread.** While a global correlation between variability level is observed, proteins that are known the be stochastic in protein abundance only, such as TIM17 which is localized to the cytoplasm, appear to have little spatial variability. Protein can be said to have high or low variability a given component, by reporting the larger variability level (above or below the dashed red diagonal).

2.2.2 Spatial variability within cell population and between cell populations

RV measurements are reproducible between image collections for the selected image feature, which is referred to as subcellular spread. High global correlations between RV levels are reported between image collections (Table III.8).

Buds \ Mothers	HOwt	ura3	rap0	nctr	nhur	nmms
HOwt		.686	.706	.337	.453	.358
ura3	.635		.896	.338	.447	.371
rap0	.622	.864		.333	.481	.407
nctr	.274	.355	.324		.655	.604
nhur	.386	.423	.465	.642		.668
nmms	.298	.333	.376	.598	.665	

Table III.8: **Correlation in 'Relative Variability' levels in 'Subcellular Spread'.** Analysis of the agreement of relative variability level defined on six different sets of microscopy images. The upper right triangle reports on correlation of subcellular spread in Mother cell, and the lower left triangle reports for Bud objects. The first three image sets (HOwt, ura3 & rap0) differ in the segmentation approach to the last three image collections (nctr, nhu & nmms; from Tkach et al. [156]).

Since the number of cells per image may vary for an experiment to another, the list of proteins considered and utilized in the local regression differ from one experiment to another. As before, this implies that a certain level of disagreement is expected from differences in the protein lists. Still, I observe that the amount of correlation between replicate experiments is greater than the correlation between mean subcellular spread and its corresponding relative variability level (Figure III.15).

A number of proteins can be shown to be inherently exhibiting high relative variability in several experiments. To show this, I rank proteins according to their RV level in each experiment (Table III.9). I note that a number of proteins systematically appear to be most variable, and hence have low ranks. In order to evaluate the statistical significance of these occurrences, I evaluate the exact probability for the sum of 'n' random ranks to be smaller or equal to the observed sum of 'n' ranks, where 'n' is less than 6 when RV values are not reported. A total of 190 proteins exhibit high variability across experiments (with P-value < 0.001 for low sum of ranks) and 66 proteins exhibit low variability across experiments (with P-value < 0.001 for high sum of ranks).

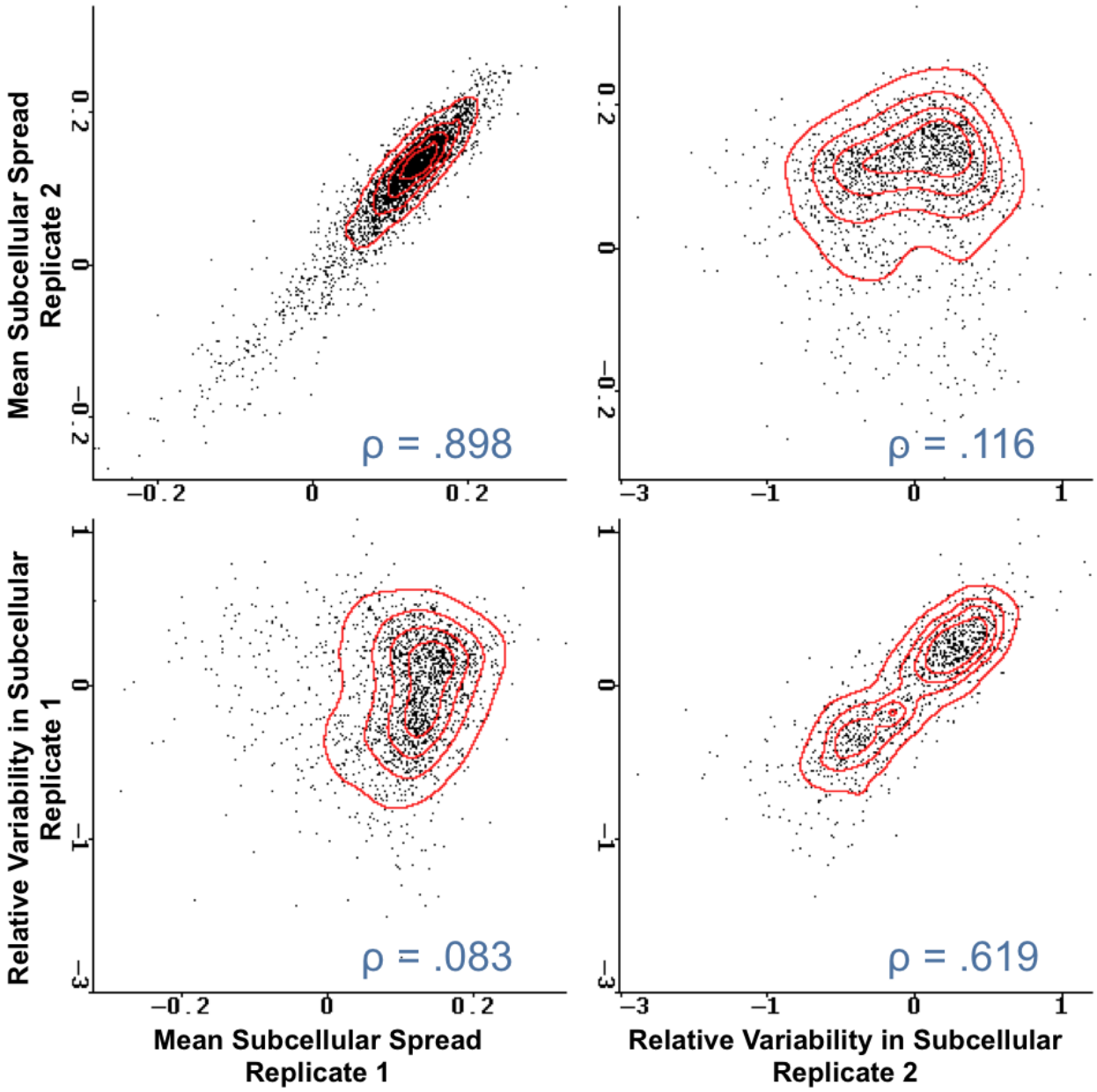


Figure III.15: **Reproducibility of 'Relative Variability' is subcellular spread.** Reproducibility levels for the mean compactness measure and relative variability levels (correlation of .898 and .619 respectively). In both replicate, mean compacness and variability levels are mostly uncorrelated. Density of point is displayed using a contour plot (in red).

ORF	Name	'Relative Variability' rank						Rank	log_{10}
		HOwt	ura3	rap0	nctr	nhur	nmms	sum	P-value
YJR076C	CDC11	2	1	2	26	1	4	36	-13.5
YDR258C	HSP78	1	7	17	12	21	11	69	-11.8
YJR060W	CBF1	5	3	9	4	38	29	88	-11.1
YCR002C	CDC10	10	5	37	21	17	3	93	-11.0
YOL147C	PEX11	N/A	N/A	1	3	11	7	22*	-9.09
YBR039W	ATP3	6	15	11	25	94	67	218	-8.8
YDL225W	SHS1	11	12	5	187	24	23	262	-8.32
YDR040C	ENA1	13	173	77	47	54	197	561	-6.34
YNL312W	RFA2	217	18	169	41	142	38	625	-6.06
YPR190C	RPC82	3	6	108	71	N/A	N/A	188*	-5.88
YPL139C	UME1	58	33	589	170	137	163	1150	-4.47
YJL173C	RFA3	219	147	138	46	361	141	1052	-3.87
YAR007C	RFA1	185	N/A	N/A	31	155	156	527*	-3.40
#Rank		2355	3324	3531	1449	1900	1252	25267	

Table III.9: **Examples of proteins with high 'Relative Variability' levels in 'Subcellular Spread' for 6 experiments.** Analysis of the agreement of relative variability ranks defined on different sets of microscopy images. Ranks are summed; then, the exact probability for such for sums to have a value smaller or equal to the observation to be generated by chance is evaluated (P-value). The number of proteins ranked in from each experiment is shown on the bottom row of the table. Note that sums with missing ranks are accounted for in the P-value calculation (marked with *).

2.2.3 Protein with population heterogeneity for subcellular localization.

In this section, I show that many images associated with high RV in subcellular spread display foreground objects that can be recognized to originate from two different classes, which are subcellular localization patterns. Several classes of heterogeneity (two different subcellular localization patterns in one image) are detected by a single measure, without prior knowledge of protein association with subcellular localization, or prior knowledge about what the heterogeneity classes might be. From the list of 190 proteins that systematically exhibit high variability across experiments (with P-value < 0.001 for low sum of ranks), I selected examples of proteins that represent 5 heterogeneity classes in subcellular location.

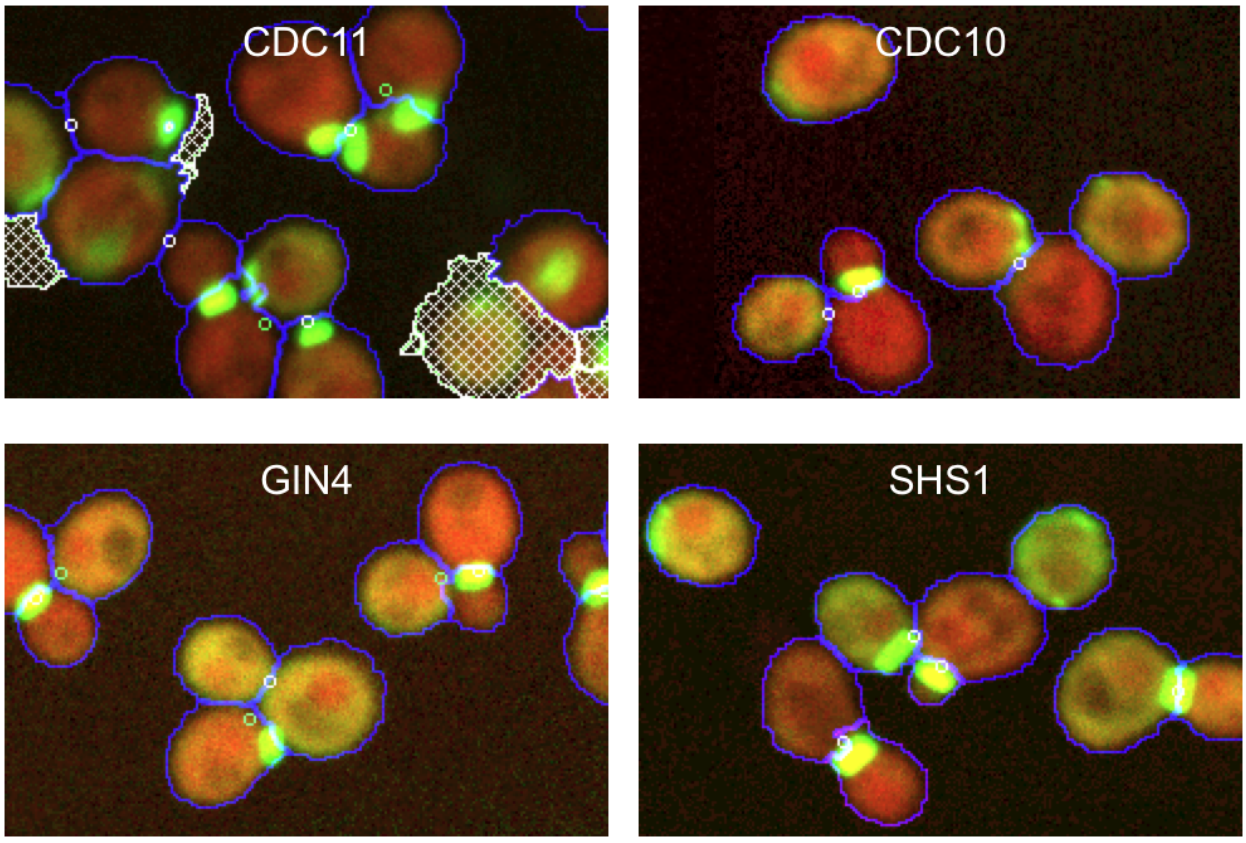


Figure III.16: **Bud neck proteins with high relative variability.** Protein localized to the bud neck or cytoplasm for a fraction of the cell imaged. In rare occurrences, Cdc11 exhibits punctae patterns, but it is difficult to assess the significance of such observations given the complexity of the bud-neck pattern.

Among the most variable proteins, four proteins are localized to the bud-neck (Figure III.16). We note that these proteins share a similar pattern, in that they localize to either the bud-neck or the cytoplasm.

While this is indeed a case of cell-to-cell variability in subcellular localization, the change in localization is inherently determined by the cell-stage. These proteins are either subunits of the septin ring or are contributing to its assembly, so that overexpression or loss of these proteins can be shown to causes abnormal cell-stage progression or changes in the bud morphology [100]. Why were these detected as spatially variable, despite our attempts to control for cell cycle variability? One possibility is that the cell-stage dependency is poorly captured, due to uncertainty in cell-stage assessment or the cell-stage resolution modeled, but I note that all the MCM subunits are not among spatiality variable proteins (ranked 546, 662, 1336 and 1461 out of 2355). Another explanation is that the bud-neck subcellular localization is problematic, as the cell segmentation may assign the entirety of the bud-neck area to either the mother or bud cells area, so that the distance to the center of mass deviates significantly depending on the inclusion or exclusion of the of the bud-neck area. As such, this observation is difficult to interpret, as it is likely that divergences in are measured for protein of rare subcellular localization, and we previously noted that many bud-neck proteins have dynamic patterns with differ from each other (Figure II.13).

Next, we have proteins that have been characterized to be localized to the nucleus and cytoplasm [76]. These proteins are clear cases of cell-to-cell variability in the protein expression (Figure III.17). In all cases, the proteins are most often localized to the nucleus, but a fraction of the cells lose protein localized to the nucleus. In the case of Cbf1, that fraction clearly localize the protein to the cytoplasm, as the basal level of intensity for the cytoplasm is lower for cell exhibiting nuclear localization. For Rpc87, the cytoplasmic fluorescence level does not appear to change in cells with no nuclear fluorescence, so it is possible that the protein is degraded as opposed to have a change in subcellular localization. Since autofluorescence generates a background level of intensity in the cytoplasm, the change nuclear intensity has a similar effect on protein subcellular spread measurement as for Cbf1. Visual inspection for Ume1 and Rpb2 suggest that stochastic changes subcellular localization are occurring, but it is unclear if protein degradation is also a contributing factor to the observed cell-to-cell variability. Nevertheless, this shows that measuring relative variability level allows the detection cell-to-cell deviation in subcellular localization.

Among the most variable proteins, several proteins are previously described to localize to the mitochondria (Figure III.18). Mitochondrial proteins have a complex subcellular spread, as the shape and number of mitochondrion vary inherently between cells. Since there is a large number of mitochondrial proteins (527), it is unlikely that outstanding deviations are measured due to poor estimation of the basal cell-to-cell variability for this subcellular localization. Indeed, visual inspection reveals that these proteins appear to further localize to some punctae, which are situated on mitochondria. GFP intensities in punctae are 3-8 folds higher than intensities within the cell that can be associated to a more common mitochondrial subcel-

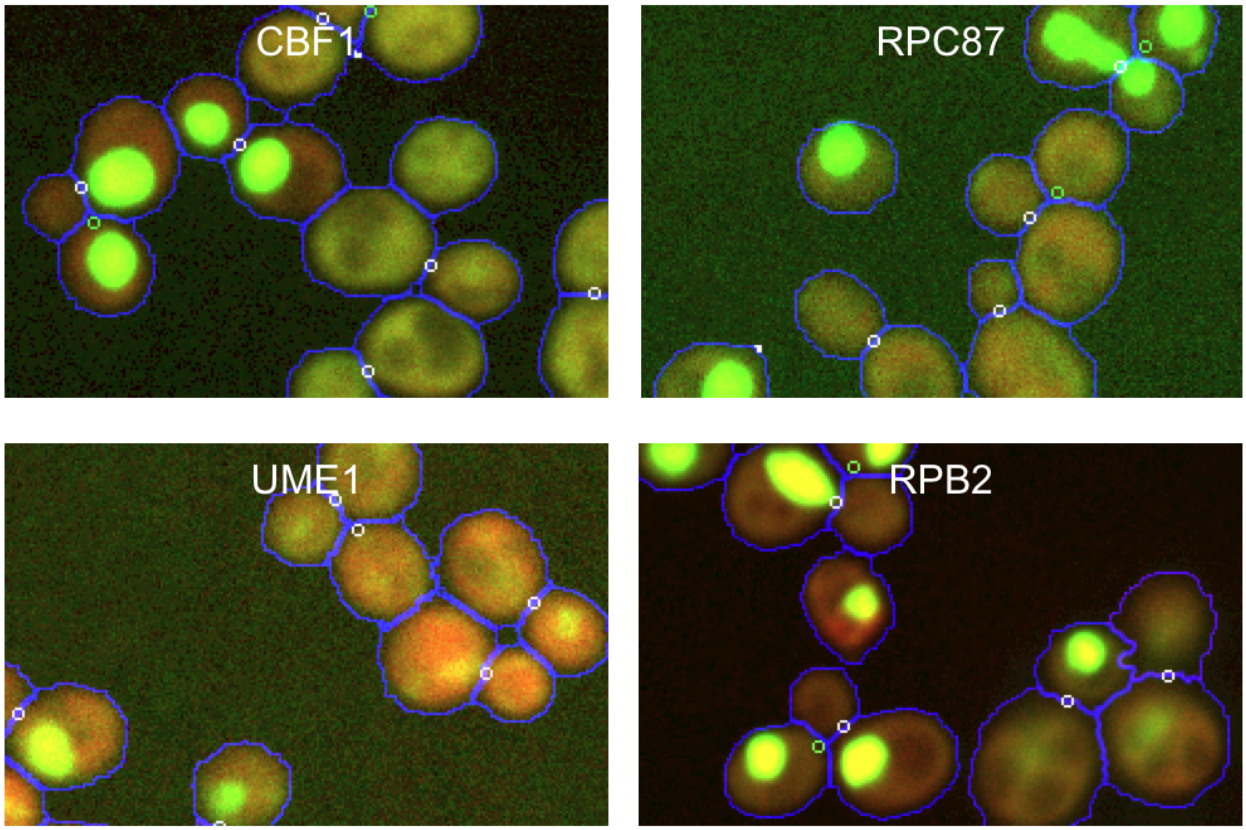


Figure III.17: **Nuclear proteins with high relative variability.** Proteins exhibit nuclear localization in a subpopulation of the imaged cells. The intensity of the GFP was adjusted to show that the Cbf1 basal cytoplasm changes with the occurrence of nuclear localization, while no adjustment can for Rpc87. Hence, Cbf1 is a clear example of protein with stochastic subcellular localization.

lular localization pattern. The occurrence of one or several punctae is found in a fraction of the cells: many cells do not display any punctae. This shows that deviation in mixture of complex subcellular localization such as punctae and mitochondria can render measures that allow the detection of cell-to-cell heterogeneity in the fraction for each subcellular localization.

Thus, relative variability in subcellular spread detects several classes of heterogeneity in specific subcellular localization pairs. In addition to the two clear cases of subcellular localization variation previously presented, Ena1 appear to be either localized to the cell periphery or to punctae inside the cell (Figure III.19). It has been previously observed that Ena1 localized in punctate bodies intra-cellular membrane [15]. Crz1 is known to activate the production of Ena1 proteins [102], so that stochasticity in the expression of Ena1 could be expected, but stochasticity in subcellular localization has not been previously reported.

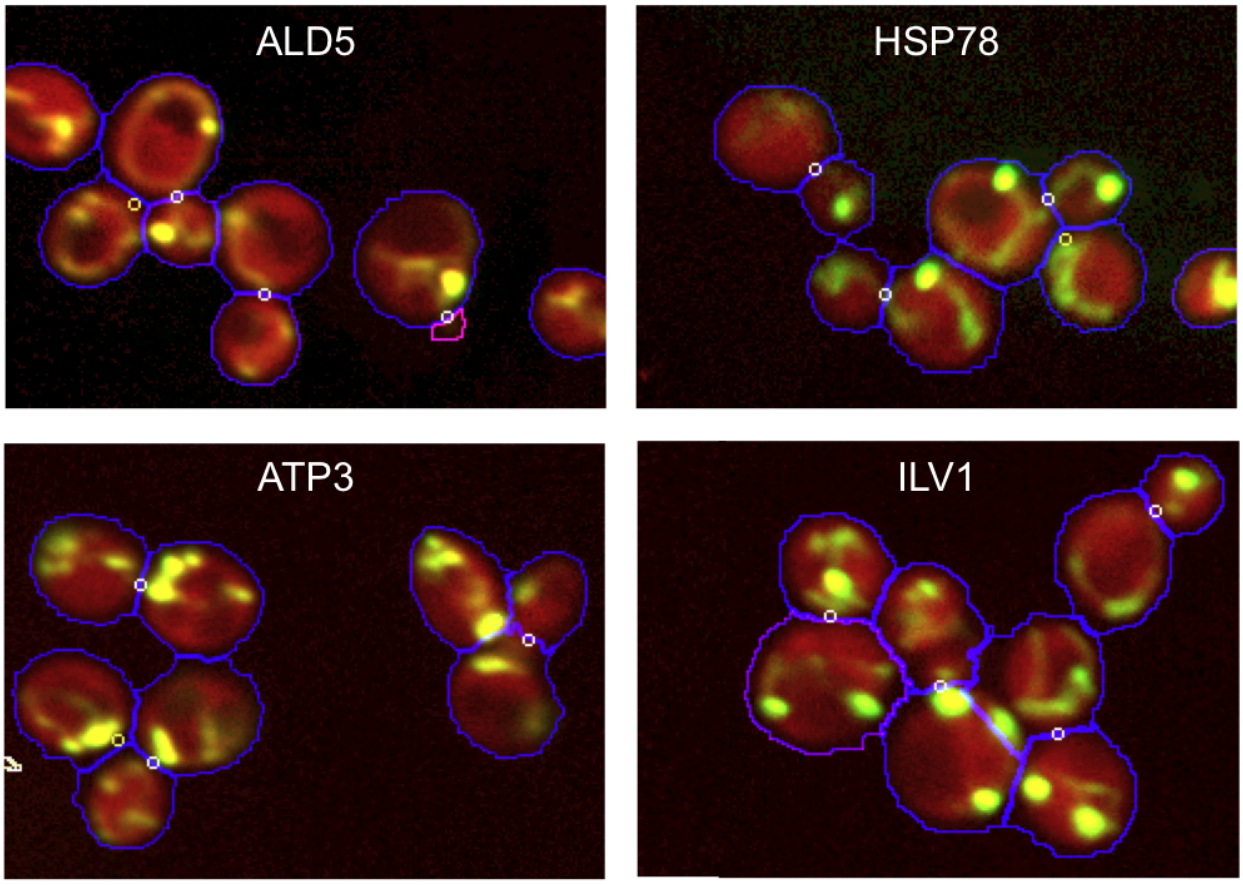


Figure III.18: **Mitochondrial proteins with high relative variability.** Example of mitochondrial proteins with high relative variability level in protein subcellular spread. Such proteins often exhibit punctate patterns atop of the basal mitochondrial expression.

One last class of variability in subcellular localization is defined for 3 subunits of a protein complex that are typically localized to the nucleus, but sometimes localize in punctae outside the nucleus (Figure III.19). In the unsupervised analysis of time-profiles (Section 3.2), the 3 subunits were perfectly clustered; this suggested that one feature of the protein expression set them apart from the other nuclear proteins. Replication proteins (RFAs) are essential protein required for DNA repair and DNA replication [96]. Only Rfa1 was previously known to relocalize to the cytoplasm in response to hypoxia (according to SGD [34]), visual inspection suggests that all three factors stochastically change their subcellular localization, which is captured by the relative variability level measured albeit the low frequency of cells exhibiting punctate patterns.

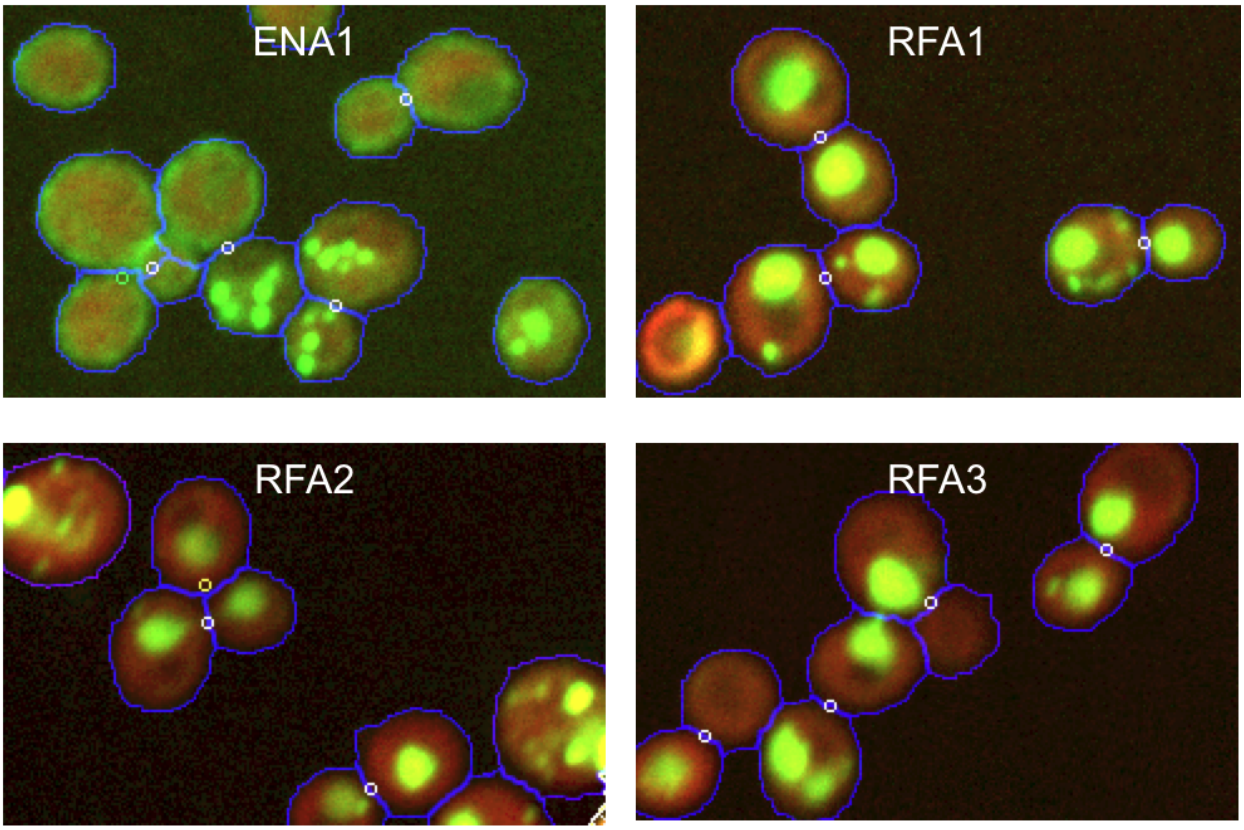


Figure III.19: **Punctae proteins with high relative variability.** Some proteins often exhibit punctate patterns as an alternative to the basal subcellular localization, which is cell periphery for ENA1 and nuclear for the 3 replication factor A proteins.

2.3 Discussion

A single measure pulled from the list of yeast proteins several classes (5) of variability in spatial subcellular localization without predefining the classes of subcellular localization that are to be recognized. The cases presented only account for cases where visual inspection is sufficient to confirm for heterogeneity in subcellular localization; for example, visual inspection cannot explain why Pex11 and Pex25 systematically have extremely high relative variability while Pex1 and Pex4 do not. These 4 proteins have a complex subcellular spread that is due to their localization to the 'peroxisome' [76]. Nevertheless, the success of the RV in the detection of various heterogeneities in subcellular localization suggests that either the abundance or subcellular localization of Pex11 and Pex25 is more variable than expected relative to the other proteins localized to the peroxisome.

It can be shown that several classes of previously characterized biological functions can be associated with

proteins exhibiting high RV. The group of 190 proteins with high RV contains subgroups of proteins that are associated to major subcellular localization classes. Such subgroups are more likely to share a biological function (as previously characterized) than by forming a random group of proteins that are associated to the same subcellular localization classes. The statistical significance is evaluated using the hypergeometric distribution, which is corrected for multiple hypotheses testing (Bonferroni correction) (Table III.10). This suggests that cell-to-cell variance associated to single cell measurements may further characterize protein expression, which likely contributed the result introduced in the second chapter that showed that similarity in protein time-profile often imply similarity in protein function. While the methodology differs, the same information is used to compare time-profiles of protein expression or relative variability levels.

While several classes of heterogeneity are uncovered, some protein known to stochastically change subcellular localization, such as Crz1 [24] or Msn2 [98], are not detected. For such cases, the frequency of nuclear localization is low, suggesting that RV estimate may fail to detect low entropy heterogeneities. Further, the 'subcellular spread' measure may fail to capture many class of subcellular localization changes.

While a number of proteins appear to systematically have high relative variability (Table III.9), a global correlation of 0.60 to 0.70 for RV between replicate experiments indicates that the detection of protein of high variability is difficult from a single experiment. The robustness of reported RV is limited by sampling variance, even though it is accounted for by using a significance threshold for deviations in variability level (Section 1.5.3). The issue is that the representation of each protein depends on the number of cells imaged, so this sampling varianace changes the expectation of the variance from an experiment to another. For example, if 10 proteins have an unusual pattern, which have high variance when compared to other 10 similar proteins, the RV level will be modulated by the proportion of cells identified in each of the two groups.

location (# High RV in Location)	#	Annotation (# in Location)	Description	Log_{10} P-value
mitochondrial	4	GO:0043644 (12)	ATP synthesis	-3.97
(25 out of 313)	6	GO:0015986 (13)	mitochondrial nucleoid	-8.03
nucleolus	12	GO:0003723 (14)	RNA binding	-18.31
(62 out of 157)	23	GO:0005515 (35)	protein binding	-32.94
nucleus	6	GO:0006337 (10)	nucleosome disassembly	-4.64
(100 out of 1022)	14	GO:0003723 (76)	RNA binding	-5.08
	9	GO:0030687 (22)	preribosome	-6.35
	18	GO:0003677 (92)	DNA binding	-8.66
peroxisome	2*	GO:0005779 (4)	peroxisomal membrane	-3.89
(5 out of 17)	2*	GO:0005515 (4)	protein binding	-3.89

Table III.10: **Enrichment of annotations for proteins with high spatial variability.** Enrichment of functional annotations for the 190 proteins showing high RV in protein subcellular spread. Enrichments are computed within subcellular localization classes, as defined by Huh et al. [76]. Bonferroni correction for multiple hypothesis testing was utilized in the P-value calculation. * Only one protein shared.

Part IV

Conclusion

Material from: Handfield et al. 2013 [66]

Within Sections: 1.1.1, 1.1.2, 1.1.3, 1.2.2

1 Summary

1.1 Biological Contributions

1.1.1 In silico synchronization of yeast cells.

Previous studies have demonstrated the feasibility of uncovering cell-stage from images of unsynchronized cell populations, either from time lapse movies [148] or from still images [21]. We apply this approach to high-throughput still images of budding yeast. To do so, we devised a segmentation method to identify and separate the bud and mother cells, and uncover the cell-stage based on measurement of the bud size. Our method depends critically on our estimates of bud size, and I show that the automatically estimated sizes were comparable to those obtained from manually identified cells. Several parts of the analysis may be improved. For example, since the bud-site selection is predetermined by the position of the preceding daughter cell [54], it could be used to help determine the correct mother-bud assignments. Similarly, a better model for the relation between daughter cell size and the cell cycle could be used to infer a more accurate estimate of cell-stage.

1.1.2 Quantitative descriptions of subcellular expression patterns.

Typically, spatial patterns of protein expression are described by assigning labels [76] or functional annotations [10]. Such discrete classes are not sufficient to fully describe a protein's expression if it is present in quantitatively different localizations or abundances at different cell-stages, or if a protein is simultaneously present in several locations with quantitatively different fractions [174]; because our approach assigns a quantitative expression profile to each protein, we can characterize protein expression at a finer scale than the resolution currently achieved by discrete classes. Approximating protein expression patterns as discrete classes has also led to challenges for computational analysis. For example, in previous work based on discrete classes [29, 75] many proteins are often filtered out because they have either been annotated as 'ambiguous' or are reported to be located in several localization classes.

Because we treat expression patterns quantitatively, our analysis identifies clusters of proteins that are significantly enriched in 'ambiguous' proteins and proteins that were manually annotated [76] as localized in multiple compartments. Furthermore, our analysis identified and organized a group of proteins that show

complex patterns relating to the growth of the bud, which were not consistently annotated previously using discrete categorizations. To our knowledge no previous genome-scale analysis of still microscopy images has identified groups of proteins with subcellular localization patterns that change as a function of cell-stage, such as the MCM and exocyst complex subunits discussed above, although recent work on smaller collections of time-lapse images has demonstrated that functionally related proteins can be identified in unsupervised analysis of dynamic protein expression profiles [50].

1.1.3 Clustering protein expression patterns.

One limitation of cluster analysis is that the members of each cluster identified are not always consistent between different parameter settings, or different clustering methods. Indeed, the remarkably specific groupings corresponding to specific regulatory mechanisms (such as the clustering of all 4 of MCM complex subunits and of all 3 DNA replication factor A complex subunits) were not always observed when we varied the distance metric or clustering method used.

Despite these limitations, our analyses consistently identified clusters that were enriched in functional groups of proteins (Ribosome, Proteasome, DNA-damage pathway, exocyst complex, etc.) that are not usually associated with their own subcellular compartments. Because we used hierarchical clustering of interpretable features, we could see that these functional groups of proteins showed patterns of localization similar to those localized in the same compartment, but in each case showed subtle differences in pattern that allowed them to be distinguished. These results suggest that high-resolution images could be used directly for functional discovery as has been reported for mammalian cells [37].

This work demonstrates that accurately identifying large numbers of cells for each protein allows quantitative characterization of spatial and temporal characteristics of protein expression patterns and permits direct interpretation of image-based measurements without requiring human inspection of large numbers of images to train classifiers. Our analysis gives new insight into the relationship between protein function and protein expression patterns inferred from high resolution microscope images.

1.1.4 Protein-level classification from protein expression patterns.

In agreement the previous result, I have shown that the quantitative characterization further allows the identification of a large number of proteins, whose expression is likely to significantly differ from all other proteins. I have shown that time-profiles allow the definition of a simple Nearest Neighbour classifier, whose accuracy is unexpectedly high given that no parameters are learned so to improve the classification power.

Further, changes in cell morphology have a limited influence on the time-profiles; many proteins are uniquely identified by comparing time-profiles. This further indicates that fluorescence microscopy of unsynchronized yeast population exhibit rich information about the protein expression, which is minimally requires several hundred to one thousand classes to describe.

1.1.5 Cell-stage dependency of cell-to-cell variability.

I have shown that it is not necessary to filter cells in order to eliminate the cell-stage induced cell-to-cell of variability, and that this extrinsic source of variability may be captured even in relatively small sample of observations ($\sim$ 50). In comparing many approaches, proteins with similar 'time-profiles' exhibited reproducible levels of stochasticity, which show a certain agreement for protein of high cell-to-cell variability with Newman et al. [113]. This approach allowed us to detect ribosomal proteins showing high cell-to-cell variability (Figure III.5) that were missed by Newman et al, which I hypothesize is due the differences in the normalization and interpretation of variability levels.

It is known that cell-to-cell variability may inherently depend on external factors [24]; since the cell morphology fluctuates within the cell-stage, it can be expected that some stochasticity properties are modulated by cell-stage. The analysis of the level of variability showed that proteins may change their stochasticity properties within the cell cycle (PIR1). The methodology introduced allows the specific quantification of such cell-stage dependency, which was never attempted before to my knowledge.

1.1.6 High-throughput quantification of spatial variability.

Finally, I have shown that the local comparison of stochastic properties answers one of the concerns for a high throughput analysis of cell-to-cell variability in subcellular localization, which is the heterogeneity of subcellular localization in the protein collection and the absence of a natural scale or distribution for image feature measurements required to characterize subcellular localization. The method allowed the identification of many classes of cell-to-cell variability in subcellular localization (cytoplasm to nucleus, mitochondria to punctate, nucleus to punctate, etc.). While the reproducibility of the reported level of variability is limited to the most variable proteins, no prior work attempted to quantify spatial variability on the whole yeast proteome.

1.2 Computational Contributions

1.2.1 Morphology based cell segmentation

One core contribution of my work is the implementation of an image analysis pipeline, which is specialized in the recognition of yeast cells from unsynchronized populations. I fully implemented this pipeline (C++ source code available, see [66]), which has no external dependency. The segmentation accuracy was only reported for a small image collection of manually labeled cells; the accuracy was shown to be superior to existing methods that partition clumped cells, for both artefact recognition and yeast shape parameterization accuracy.

The pipeline can be adapted to different types of microscopy images, though the accuracy of cell identification is not guaranteed. For instance, segmenting cells using autofluorescence (Figure IV.7) wield a lower accuracy due to a smaller signal to noise ratio, a higher cell density (cells concentration in an image), and a dependence on the signal intensity coming from GFP tagged protein. Nevertheless, downstream analyses and suggest that cell segmentation is robust; I report reproducibility levels for means (Figure III.6) and variances (Table III.4 & III.8) in image feature measurements. The cell segmentation is also possible when cells are undergoing morphology changes that are induced by a mating pheromone (Figure II.15). The use of a proper shape model for determining the size of joined 'catchment basins' was previously proposed as an alternative to the use of seeds [41]; I have shown that morphological models can also be used to determine the association of 'catchment basins' into cell areas.

1.2.2 Probabilistic model yields confidence estimates.

We presented a cell identification pipeline that includes a confidence measure, which summarizes the probability that an object identified in our images is actually a correctly identified cell. To do so, we characterized the deviation of real cells from an elliptical model using several quality measures whose distribution for real cells were inferred from ellipses that had been manually fit to cells by eye. Our confidence measure allows us to distinguish correctly identified cells from artefacts and misidentified objects, without specifying what the nature of artefacts might be (Figure I.16). We believe that this type of approach for measuring the confidence of automatically identified objects in image analysis will be generally useful, because artefacts tend to vary between microscope, experiments and computational methods, whereas cell shapes are expected to be much more consistent. In addition, this confidence measure is explicitly defined as a posterior probability of an identified object to be a properly identified cell. This allows us to weight probabilistically data points according to the posterior probability. For classes of cells where our model does not fit as well, such as very early non-ellipsoidal buds, we expect to weight down all the data points, but we can still include information

from these data points in our analysis. This is in contrast to the situation where we used a hard threshold to exclude artefacts. In that case, certain classes of cells are preferentially excluded (Figure I.17B), and the statistical significance of downstream analyses is reduced (Table II.4 &II.5).

1.2.3 Maximum likelihood agglomerative clustering

In this work, the number of registered cells per yeast strain is relatively high (about 100 or more). By acknowledging that all cells conjointly report for what is the expected configuration of proteins through the cell cycle, is it possible to infer what would be the mean and variance of any feature measurement at any given cell-stage. I have shown that empirically inferred variances can better identify similarities in the expression of proteins. For instance, I show one instance where using variances recorded in time-profiles allows better recognition of subcellular localization between replicate experiments better than SVM [151], which solely uses means recorded in time-profiles (Section 4.1.2). SVM finds the best kernel function to define what data points are 'similar' to the training set that should be used for class label prediction; in comparison, the cell-to-cell variances may define a better kernel, since its parameters are allowed to vary from data point to another.

I proposed the use of the maximum likelihood criterion as a mean to cluster protein expression time-profiles. As opposed to the previously proposed divisive use [86], the agglomerative use locally compares variances recorded in time-profiles independently of the global spread of mean time-profile in the protein collection. This allows the robust detection of small cluster of similar protein expressions, better than hierarchical clustering with complete linkage (Table II.4). The hierarchical organization generated could be shown to match more closely the expert defined protein categorizations than other clustering methods, independently of the size of protein classes.

One important note is that the above result relies on the robustness of variances recorded time-profiles. I could show that weighting down instances of cells that have segmentation of lower quality, as opposed to the use of a filter based on detection of artefacts, produces time-profiles that are more robust (Figure II.7). Robust profiles are critical for the maximum likelihood agglomerative clustering because functional enrichments in hierarchical clusters are inferior to standard hierarchical clustering methods if cell are filtered out using a confidence threshold (Table II.4).

1.2.4 Local analysis of variance

Protein similarity in function or in subcellular localization defines classes of protein that are subsets of all sizes from the protein collection. As such, the associated image collection contains feature measurements whose distribution is inherently heterogeneous. Detecting a protein with outstanding variance in a feature measurement is difficult as a background model would need to account for this heterogeneity. In this work, I defined a local comparison of variances that reports a variance ratio for a protein of interest to the local background model, which inferred from proteins that are 'similar' to the protein of interest. Proteins that have the highest relative variance (RV) matches previously identified proteins (TIM17, HHF1, SSA4, SIT1 [113]) or identify proteins with heterogeneity in subcellular localization (ENA1, HSP78, RFA1). This shows that outstanding variances may be detected without prior knowledge of the distribution for feature measurements, where the scale of feature measurement is highly heterogeneous.

One limitation of the method is that the local inference of the background model is not robust to uneven representation of cell per imaged strain across replicate experiments. While RV level globally agree between replicate (Table III.4 and III.8), list of proteins exhibiting the highest RV differ between experiments. This issue was partially resolved by accounting for sampling variance. Any reported RV level was required to rejecting the null hypothesis, where a measured variance level is identical to the local expectation of for that variance level.

2 Future Work

2.1 Characterization of cell-stage dependence for protein expression.

This work proved to be successful in estimating cell-stage directly from an accurate measurement of bud size. By focusing of mother-bud objects, 'lone' cells were mostly ignored from the analysis. The proximity of a lone cell to a mother-bud pair may hint that this cell used to be the bud of that mother cell. The manually identification of cells (Section 1.3.2) was allowed to identify such cells (labeled as 'daughter' cells). Such cells could be utilized to better describe early stages of a small G1 cell that recently detached from its mother.

The main reason such 'daughter' cells were ignored is that the accuracy in recovering the cell-stage is unknown; one major concern in this work has been that the available data did not contain clear means evaluate the accuracy of the proposed quantitative representation of cell-stage. In appendix 5, it was attempted to capture the "true" cell-stage by inference using a cell-stage model, whose dependence on bud size is learned from the changes in protein expression within the image collection. While the approach proved to be un-

successful, ongoing research has shown selecting an array of cell cycle markers can help the quantification of cell-stage progression (ISMB 2013 [30]), and shown that the higher dimension of spawned by the array of markers allowed us to observe a continuous path, which relates to cell-stage progression.

In this case, I could note (Figure III.2) that the bud size is less informative of cell-stage once it reaches a certain size (about 850 pixels). At this point, several cell-stage checkpoints are temporally stopping the cell cycle progression. The accuracy in determining the cell-stage would greatly benefit from the use of a marker protein, such as one of the many bud neck proteins that would allow us to monitor the formation and diffusion of the bud neck. The marker protein is often used to help the segmentation, using more than this one marker make the task of building a large Yeast strain collections more complex. I noted that the autofluorescence in images can be used to segment cells. Since that additional information may not be required for proper segmentation, it would be possible to use marker protein in double mutant yeast stain collection for cell-stage monitoring instead. A marker with dynamic bud localization (Figure II.13) may we ideal to span a high dimensional feature space that would allow us to monitor cell-stage progression in a similar way than previously proposed [30].

Having a more accurate cell-stage predictor is critical for the interpretation of variability in measurements; for example, Newman et al. and this analysis claims that Pir1 is a highly stochastic protein (Figure III.1), but inaccuracy in capturing cell-stage from object size or bud size may occlude the fact that Pir1 may be a protein whose abundance is not stochastic and is simply determined by cell-stage (Figure III.3).

2.2 Detection of differential expression.

This work shows that both the unsupervised and supervised analysis of time-profiles can organize proteins of similar function or subcellular localization. The main appeal of a crude representation of subcellular localization is that it allows the quantification of deviations in both protein abundance and subcellular localization, which is critical to evaluate stochastic properties for protein expression (chapter III). Stochasticity levels, which need to be comparable for their interpretation, can be shown to be significantly different by rejecting the hypothesis that measured fluctuations are induced by sampling variance. The method described inherently defines a mean to evaluate the statistical significance of deviations in mean and variability level between two collections of observations (Section 1.5.3). Instead of comparing a set of observations to the set of observations produced by a local regression, two sets of observations could be directly compared. If the observation sets comes for the same yeast strain under different environment conditions, one could evaluate if the cell populations differ significantly in protein abundance or subcellular spread, in both mean and/or cell-to-cell variability level (Figure III.10).

Undergoing research attempts to utilize cell microscopy for identification of differential expression [156]. The analysis of change in subcellular localization is typically performed by manual annotation of the subcellular localization. One reason is that the detection of deviations is difficult for automated supervised analysis, as delineations between subcellular localizations are not learned from training sets made of pure subcellular localizations. The use of likelihood ratio tests can specifically address the significance of deviations in the set of observations, which may also include specific changes in level of cell-to-cell variability (Appendix 6.1). This defines an automated means to detect which proteins deviate in its expression among proteins of identical subcellular localization. For example, a preliminary analysis the image collection from Tkach et al. [156] supports that DNA damage agents may increase (HSP42, EDC3) and decrease (RPC82) cell-to-cell variability levels in subcellular localization, which agrees with visual inspection of images. This may provide further insight on the mode of operation of elements within biological pathways.

Further, I have shown that changes in cell morphology have limited impact on the detection protein of similar function (Table II.8). This is an incentive to further investigate commonalities for protein spatial expression for different organisms, which may significantly differ in their morphologies. For example, certain strain of Yeast grow buds that are elongated and highly ellipsoidal [41]; characterizing few image measurements, whose dependence on cell morphology can be understood, may be sufficient of recognize a large number of protein expression profiles shared by the two organisms. This would open the path to comparison protein spatial expression of organisms of varying morphologies, so to find proteins whose localization may have adapted and differs from an organism to another.

2.3 Other applications

This work sought to better acquire biological knowledge from microscopy of GFP-tagged strain collections. This mainly required reassembling all observation into a probabilistic model that allows comparison of the protein expression and the detection of similarities and deviations systematically on the whole Yeast proteome. As previously discussed, means to improve of the methodology in the specific case of Yeast exists, but the present success of the methodology suggests that the approach may be applicable for complex analyses with large number of observations with a heterogeneous context. This work showed that a non-parametric modeling of observations allows the control for non-homogeneity in the representation of protein classes (and inherent hierarchies) and to control for raw image feature measurements, whose natural probability distributions and cell-stage dependencies are unknown.

Other problems that have similar difficulties as in this analysis may be tackled be this approach. Such

problems are not limited to biological problems, as I could use the idea of maximum likelihood clustering to segment a color photograph of a complex scene (Figure IV.8). As for biological problems, the vast majority of the many means to acquire biological data requires harvesting or killing biological samples (In situ mRNA hybridization, RNA [153] or DNA [110] sequencing, mass spectrometry and tumour imagery); interestingly GFP imagery is one of the rare exceptions that allows movies of live cell to be imaged. Hence, for many biological problems, the extraction of time-dependence for observations can only be obtained from comparing different samples at different times. To do so, variability between samples needs to be factored into time-dependence and inherent variability, which could be also include other external factors.

For instance, advancement in the sequencing allows the measure of mRNA levels at the single cell level. While lacking spatial resolution, the dimensionality of the data is immense: a report for 7K difference mRNA levels and some alternative splicing for a single cell is possible [153] (as opposed to a single protein measure per cell). The characterization of co-expression of mRNA at the single cell level and inherent cell-to-cell variability would be possible. In order to make sense of such large covariance matrices that could be generated, the knowledge of potential cell states (cell-stage and/or cell types or tumour cell type) may be required so to understand the basal variance attributed to measures coming from heterogeneous cell populations. Under a proper characterization such states, the methodology for comparison of cell populations and detection deviation that cannot be attributed to sampling variance could be predicted and assessed for statistical significance under this framework, which inherently acknowledges that every protein has unique characteristics that may impact of every experimental measurement in a peculiar and unpredictable way.

Appendix

1 Inclusion of outlier detection in mixture of model

1.1 Problem definition

We model a set of data points using a finite number of probability distributions, but a fraction of the input may be erroneous values, which are prone to affect significantly the parameter estimation. The variety of classes of undesirable data is so that the only valid assumption is that they represent a limited fraction of the inputs. We define a probability distribution for the artefact class, which allows us to use the EM paradigm. If $Z = 0$ for the artefact class, its class conditional probability may be rewritten as:

$$P(Z = 0|X) = \frac{1}{\frac{\sum_{k=1}^{m} P(Z=k|Z\neq 0)L(\theta_k|X)}{\frac{P(Z=0)}{1-P(Z=0)}L(\theta_0|X)} + 1} \tag{48}$$

If we desire that any datapoint is as likely under the outlier class then the probability density function is a constant, which is similar to the uniform distribution. We can also include the prior probability of belonging the outlier class in not defined; the presented bound is defined on the posterior probability for the outlier class. Since we use a uniform distribution as prior for the prior probability of belonging to the outlier class, we do not need to define the probability density for the outlier class and the prior outlier class separately. In a maximization step, we are to find the constant 'K_o' that satisfy the introduced constraint:

$$\frac{1}{n}\sum_{i=0}^{n-1} \frac{1}{1 + K_o \sum_{k=1}^{n} P(Z_i = k|Z_i \neq 0)L(\theta_k|X_i)} \leq B \tag{49}$$

The likelihood maximization step for K_o reduces to find the smallest positive value for which the above inequality holds. In the maximization, the parameters are updated based on the expected latent variables. As such, the fraction of datapoint in the outlier class do not change based on K_o. Then, the derivative of the total likelihood function given latent variables is:

$$\begin{aligned} \frac{d \log L(K_o|X_i \ldots X_{n-1})}{dK_o} &= \sum_{i=0}^{n-1} P(Z_i = 0|X_i) \frac{\log dL(\theta_0|X_i)}{dK_o} \\ &= \frac{-(1-P(Z=0))}{K_o \cdot P(Z=0)} \sum_{i=0}^{n-1} P(Z_i = 0|X_i) \end{aligned} \tag{50}$$

Since the above is a negative for any positive 'K_o', the lower the 'K_o' the higher the likelihood. If no bounds were defined on the posterior probability, then every datapoint could be assigned to the outlier class, so that the total likelihood may be infinite. On the other hand, lowering 'K_o' also increases the fraction of data point that will be assigned to the outlier class on the following Expectation step. Hence, we cannot freely

maximize the likelihood if we introduce a hard bound on a posterior probability. Instead, we want estimate a 'K_o' that will assign a fraction of the input corresponding to the bound to the outlier class, and ultimately terminate the EM procedure upon convergence and satisfaction of the bound.

1.2 Numerical updates

In order to update 'K_o', we could be find the exact value maximize the likelihood while obeying the bound, while all the parameters for the other classes are fixed. This requires the update of 'K_o' using a numerical updates, which further needs to the class conditional probability densities of every pixels. Since we number of datapoint may be large, such an operation may be slow when embedded in an EM procedure. For that reason, 'K_o' are updated using an estimation. If the current guess for 'K_o' is off by magnitudes, we get that the fraction of objects assigned to the outlier class may be close to 0 or 1 and have small derivative. For this reason, I introduced a numerical update for 'K_o'.

On the first EM-step, we set $K_o = \infty$, so that no data point will be assigned to the artefact class. Concurrently to the E-step, the fraction of the datapoint that would be assigned to the outlier class with hypothetical 4 'K_o' values is also computed. We choose the 4 K_o to be equi-spaced in magnitude, which initially covers the full range of positive floating points, so that the next guess for 'K_o' become one of the root of the cubic polynomial $P(\log K_o)$ that fit the 4 observations. Once the new K_o is found, we execute the M-step and move to the next EM iteration. The spacing of the hypothetical 4 'K_o' is to decrease or increase based on the current the agreement of the posterior fraction and the bound. In the event that the fraction is larger by a magnitude to the bound, we skip the maximization step and preserve current class parameters. Once the 4 K_o to be equi-spaced in magnitude manages to fit within a magnitude of 2, we instead use 4 equi-spaced points, and find the roots of the polynomial $P(K_o)$.

Updating the parameters of other classes may drastically change the class conditional probability densities, so that convergence of the EM procedure is not guaranteed. In this case though, it successfully converged for all the 48K image considered, by carefully choosing the rules that increases or decrease the range of hypothetical 'K_o' value queried may in practice guaranteed to allow a desired fraction of the image to the artefact class. Formally showing that such an approach allows the measure of the maximum likelihood parameters as inferred by an EM procedure would require the knowledge what is real probability distribution for the outlier class. This framework attempts to make no assumption on the nature of the outlier beyond their frequency in a given dataset, they could be even be missing value or erroneously computed quantities. In this case, the purpose is to ensure that the decision boundary between foreground and background is robust to the occurrence artefacts, such as stains that had arbitrary intensities and sizes.

2 Ellipse from Coordinate Statistics

Given an arbitrary shape that is defined by a set of pixel coordinates, we want to deterministically fit an ellipse to the shape. A theatrical ellipse requires 5 parameters; by defining a function that takes the value D inside an ellipse and zero otherwise, we obtain a function with a total of 6 parameters:

$$F_{\vec{c},A,r,D}(\vec{x}) = \begin{cases} D & (\vec{x}-\vec{c})^T A(\vec{x}-\vec{c}) \leq r^2 \\ 0 & (\vec{x}-\vec{c})^T A(\vec{x}-\vec{c}) > r^2 \end{cases} \tag{51}$$

Assuming that we have an ellipse centered at the origin that has its focal points on the X-axis, we have that the ellipse is defined to be the set of points for which:
$(\frac{x}{a})^2 + (\frac{y}{b})^2 \leq 1$ where $a > b > 0$
We have that:

$$\begin{array}{ll} \int\int D * I[(\frac{x}{a})^2 + (\frac{y}{b})^2 \leq 1]dxdy = ab\pi D & \int\int x^2 D * I[(\frac{x}{a})^2 + (\frac{y}{b})^2 \leq 1]dxdy = \frac{a^3 b\pi D}{4} \\ \int\int xD * I[(\frac{x}{a})^2 + (\frac{y}{b})^2 \leq 1]dxdy = 0 & \int\int y^2 D * I[(\frac{x}{a})^2 + (\frac{y}{b})^2 \leq 1]dxdy = \frac{ab^3 \pi D}{4} \\ \int\int yD * I[(\frac{x}{a})^2 + (\frac{y}{b})^2 \leq 1]dxdy = 0 & \int\int xyD * I[(\frac{x}{a})^2 + (\frac{y}{b})^2 \leq 1]dxdy = 0 \end{array} \tag{52}$$

Supposing that we recorded the integrals in an unknown orientation, we have that:

$$\begin{array}{ll} \int\int D * I(\frac{x}{a})^2 + (\frac{y}{b})^2 \leq 1)drds = V_0 & \int\int r^2 D * I(\frac{x}{a})^2 + (\frac{y}{b})^2 \leq 1)drds = V_1 \\ \int\int s^2 D * I(\frac{x}{a})^2 + (\frac{y}{b})^2 \leq 1)drds = V_2 & \int\int rsD * I(\frac{x}{a})^2 + (\frac{y}{b})^2 \leq 1)drds = V_3 \\ r = x \cdot cos(\theta) + y \cdot sin(\theta) & s = -x \cdot sin(\theta) + y \cdot cos(\theta) \end{array} \tag{53}$$

We can derive that:

$$\begin{array}{rl} V_1 = & \int\int (x^2 \cos^2(\theta) + y^2 \sin^2(\theta) + xy \sin\cos) * I(\frac{x}{a})^2 + (\frac{y}{b})^2 \leq 1)dxdy \\ & = \frac{a^3 b\pi}{4} \cos^2(\theta) D + \frac{ab^3 \pi}{4} \sin^2(\theta) D \end{array} \tag{54}$$

When normalized with the theoretical area, we obtain:
$\frac{V_1}{V_0} = \cos^2(\theta)\frac{a^2}{4} + \sin^2(\theta)\frac{b^2}{4}$
$\frac{V_2}{V_0} = \sin^2(\theta)\frac{a^2}{4} + \cos^2(\theta)\frac{b^2}{4}$
$\frac{V_3}{V_0} = \sin(\theta)\cos(\theta)(\frac{a^2}{4} - \frac{b^2}{4})$
We want to recover the coordinates of the focal points in the unknown orientation, as well as the major axis length: $2a$.

$\frac{V_1}{V_0} = \frac{\cos 2\theta+1}{2}\frac{a^2}{4} + \frac{1-\cos 2\theta}{2}\frac{b^2}{4}$
$\frac{V_2}{V_0} = \frac{1-\cos 2\theta}{2}\frac{a^2}{4} + \frac{\cos 2\theta+1}{2}\frac{b^2}{4}$
$\frac{V_3}{V_0} = \frac{\sin 2\theta}{2}(\frac{a^2}{4} - \frac{b^2}{4})$
$\frac{8V_1}{V_0} = a^2 + b^2 + \cos 2\theta(a^2 - b^2)$
$\frac{8V_2}{V_0} = a^2 + b^2 - \cos 2\theta(a^2 - b^2)$
$\frac{8V_3}{V_0} = \sin 2\theta(a^2 - b^2)$
Hence we have that:
$\frac{4(V_1+V_2)}{V_0} = a^2 + b^2$
$\frac{4(V_1-V_2)}{V_0} = \cos 2\theta(a^2 - b^2)$
$\frac{8V_3}{V_0} = \sin 2\theta(a^2 - b^2)$
$\frac{16((V_1-V_2)^2+4V_3^2)}{V_0^2} = (a^2 - b^2)^2$
$\frac{4\sqrt{(V_1-V_2)^2+4V_3^2}}{V_0} = a^2 - b^2$ since $a > b > 0$
Hence:

$$\begin{aligned} a &= \sqrt{\frac{2*(V_1+V_2+\sqrt{(V_1-V_2)^2+4V_3^2})}{V_0}} \\ b &= \sqrt{\frac{2*(V_1+V_2-\sqrt{(V_1-V_2)^2+4V_3^2})}{V_0}} \\ \theta &= \frac{1}{2}acos\left(\frac{V_1-V_2}{\sqrt{(V_1-V_2)^2+4V_3^2}}\right) \end{aligned} \tag{55}$$

One parameterization of the ellipse uses the coordinates of the two focal points of an ellipse, and the diameter (largest width) of the ellipse. The diameter of the ellipse is 'a', and the two focal points have the following position relative to the ellipse center:
$(\cos(\theta)\sqrt{a^2 - b^2}, \sin(\theta)\sqrt{a^2 - b^2}))$ and $(-\cos(\theta)\sqrt{a^2 - b^2}, -\sin(\theta)\sqrt{a^2 - b^2})$
Finally, the density parameter D is defined as the ratio of the amount of pixel to the theoretical ellipse area:
$D = \frac{V_0^2}{4\pi\sqrt{V_1 \cdot V_2 - V_3^2}}$
By substituting in the area coordinate statistics, we get the following expression:

$$D = \frac{n^2}{4\pi\sqrt{\left(\sum_{i=0}^{n-1} x_i^2 - \frac{1}{n}(\sum_{i=0}^{n-1} x_i)^2\right)\left(\sum_{i=0}^{n-1} y_i^2 - \frac{1}{n}(\sum_{i=0}^{n-1} y_i)^2\right) - \left(\sum_{i=0}^{n-1} x_i y_i - \frac{1}{n}(\sum_{i=0}^{n-1} x_i)(\sum_{i=0}^{n-1} y_i)\right)^2}} \tag{56}$$

3 Algebraic Ellipse fitting

Several methods exist for finding local minima in functions. The most commonly known are the gradient descent (or method of steepest descent) and Newton method. Starting from an initial guess "x", gradient descent updates requires evaluating the derivative of the objective function at the current point, and moving the current guess in the direction with maximum gradient, using a defined step-size.
$x_{i+1} = x_i + \alpha_i \cdot F'[x_i]$

One issue with this approach arises when the magnitude of the curvature (second derivative) varies in magnitudes is different orientation close to a local minima, the iterative solution tend to oscillate about the local minima, and slowly converge if the α_i are chosen to decay to zero.

Newton method used for finding local minima requires evaluating both the first and second derivative, at the current guess point, and finding the local extremum of the quadratic equation, which may also be a local minimum, maxima or a saddle-point. Methods were proposed to handle the events type mismatches for guess extremum to the desired extremum, but is no completely satisfactory rationale was proposed [62]. We consider identifying the center of cellular objects by minimizing the algebraic error of a parameterized ellipse to the pixels found on the contour of cell clusters. Many contour pixels are expected to me missing between adjacent cells, hence we need to identify subsets of the given contour pixels, which locally resemble arcs that are explained by a single ellipse.

A parametric representation of the ellipse is commonly used [59]:

$$Err(\vec{x}|r,\vec{c},A) = r + (\vec{x}-\vec{c})^T A(\vec{x}-\vec{c}) \quad \text{where A is a positive definite matrix} \tag{57}$$

The set of vector for which the evaluated function is zero may be a line, a circle, an ellipse or a hyperbola. In our case, we want to enforce that the object is an ellipse, can constrain the eccentricity. The squared eigenvalues of the matrix A, ($\lambda_1^2 > \lambda_2^2$) scales with the length of the major and minor axis of the ellipse, so that:
$\frac{\lambda_2^2}{\lambda_1^2} = \frac{minoraxis}{majoraxis}$
We have that:
$\frac{\lambda_2}{\lambda_1} + \frac{\lambda_1}{\lambda_2} = \frac{(Trace(A))^2}{Det(A)} - 2$
If we define 'A' such that

$$A = \begin{vmatrix} \frac{1}{2} + \frac{\cos\theta}{\alpha}\frac{1}{1+e^\tau} & \frac{\sin\theta}{\alpha}\frac{1}{1+e^\tau} \\ \frac{\sin\theta}{\alpha}\frac{1}{1+e^\tau} & \frac{1}{2} - \frac{\cos\theta}{\alpha}\frac{1}{1+e^\tau} \end{vmatrix} \quad \text{where} \quad \alpha \geq 2 \tag{58}$$

Then

$$Trace(A) = 1 \quad \text{and} \quad Det(A) = \tfrac{1}{4} - \tfrac{1}{\alpha^2}\cdot\tfrac{1}{1+e^\tau} \tag{59}$$

Hence:

$$\tfrac{1}{4} - \tfrac{1}{\alpha^2} \leq det(A) \leq \tfrac{1}{4} \quad , \quad 4-2 \leq \tfrac{\lambda_2}{\lambda_1} + \tfrac{\lambda_1}{\lambda_2} \leq \tfrac{4}{1-\frac{4}{\alpha^2}} - 2 \tag{60}$$

By choosing $\alpha = 6$, we are guaranteed that the minor to major axis length ratio is at least $\frac{1}{2}$. In order to avoid thin ellipse, I chose $\alpha = 10$, so the lowest aspect ratio allowed is $\frac{2}{3}$.

4 Kernel Density Estimation

While displaying the complex relationship between quantities from a large collection of data points, a common approach would be to make a heat-map (2D histogram). In doing so, one needs to decide on the sizes of the bins, which are critical since patterns might be lost completely with poorly chosen bins. An alternative is to consider the convolved empiric density map generated from the collection of points. While several types of kernel may be considered for the convolution and might be more proper if prior knowledge about the spread of the data exists, I here consider using of the Gaussian function for kernel for which the parameterization has been studied [167]. For the rendered graphs, I will use the bandwidth parameter for multivariate normal distribution derived by Wand [167]:

$$H_{AMISE} = (\frac{4}{2+d})^{\frac{2}{4+d}} \Sigma \cdot n^{\frac{-2}{4+d}} \tag{61}$$

where H the covariance matrix of the Gaussian convolution kernel, d is the number of dimensions, Σ is the sample covariance matrix of the data and 'n' the number of data points. If the data is Normal distributed, then the obtained density is expected to be closest to the Normal distribution density function.

5 Modeling of Cell Cycle from Protein Expression

A first approach is to define a mixture model for inferring the distribution of bud sizes that are typical of a defined cell-stage. We obtain such distributions by finding the maximum likelihood parameters that best fits the collection of images, under a model that can be represented as a Bayes network (Figure IV.1). The expectation is that we have a sufficient number of proteins that changes in intensity or localization at different cell-stages, so that the hidden cell-stage variable would learn what the typical sizes of objects that show little variability throughout the protein collection are.

One shortcoming of this approach is that this formulation does not disallow multiple hidden states to the same range of bud sizes; the uncovered maximal likelihood parameters do not appear to learn specific cell-stage transition (Figure IV.2). The likelihood formulation is so that it is best to partition identified cells into 6 groups independently of bud size in order to minimize the variance of each normal distribution. This reflects the fact that bud size is not an accurate estimate for cell-stage throughout the cell cycle. For this reason, a simpler approach is considered: cell-stage bins are defined from bud size thresholds that partition

the collection of bud sizes into 5 groups with equal representation.

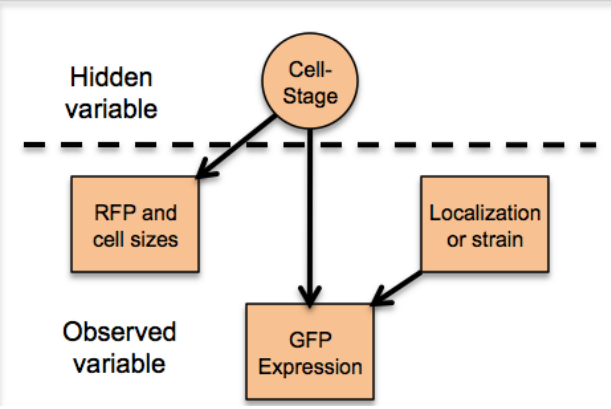


Figure IV.1: **Hierarchical clustering of protein expression stage profiles.** Each identified cell was assigned to a cell-stage using its size and protein expression.

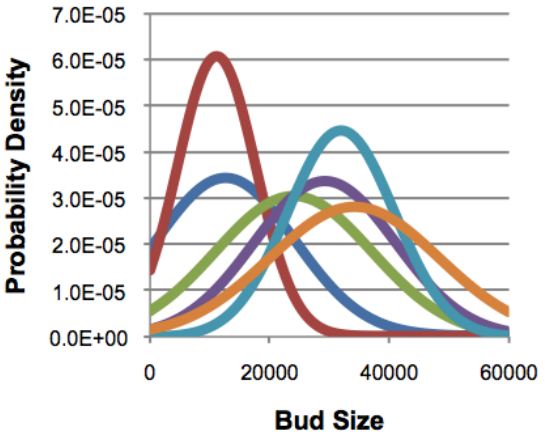


Figure IV.2: **Distribution of bud sizes in 6 cell-stage bins.** Maximum likelihood was in each identified cell was assigned to a cell-stage using its size and protein expression. The bud volume (x-axis) is estimated using $area^{\frac{3}{2}}$.

6 Likelihood ratio test with weighted observations

Extracting vectors of image features from each cell, we want to determine whether two populations of cells have the same mean vector or not, and/or if their covariance and significantly different or not. For that, I use the likelihood ratio test. In order for Wilks' theorem to be valid, the parameter space for the null hypothesis needs to be a contained subset of the parameter space for the alternative hypothesis. For example, we can reject the null hypothesis that two cell populations have the same mean and covariance matrix by uncovering the maximum likelihood to the cell population when, first, the mean and covariances of the two populations are constrained to be equal ($\mu_A = \mu_B$ and $\sigma_A = \sigma_B$), and when both mean and covariance are unconstrained. In this case, each datapoint comes with a given weight, so that we use the following sufficient statistic for uncovering maximum likelihood parameters:

$$\begin{aligned} d & \quad (Dimension\ of\ feature\ vectors) \\ W = \sum_{\forall c} w_c & \quad (Sum\ of\ the\ weights) \\ \vec{V} = \sum_{\forall c} w_c \cdot \vec{f_c} & \quad (weighted\ sum\ of\ the\ feature\ vectors) \\ O = \sum_{\forall c} w_c \cdot \vec{f_c}\vec{f_c}^T & \quad (weighted\ sum\ of\ the\ feature\ outer\ products) \end{aligned} \tag{62}$$

The likelihood function for weighted data points can be expressed by the above quantities, even when certain parameters are constrained to be equal between populations. The weighted log-likelihood (λ) is:

$$\lambda = -\frac{W}{2} \cdot (d \cdot log(2\pi) + log(|\Sigma|)) - \frac{1}{2} \sum_{\forall c} w_c (f_c - \mu)^T \Sigma^{-1} (f_c - \mu) \tag{63}$$

Using the property that $Tr(AB^T) = Tr(B^T A)$ and that the trace of a scalar is the scalar itself, we get that:

$$\begin{aligned} & \sum_{\forall c} w_c (f_c - \mu)^T \Sigma^{-1} (f_c - \mu) \\ = & \sum_{\forall c} w_c tr[(f_c - \mu)^T \Sigma^{-1} (f_c - \mu)] \\ = & \sum_{\forall c} w_c tr[\Sigma^{-1} (f_c - \mu)(f_c - \mu)^T] \\ = & tr[\Sigma^{-1} (w_c \sum_{\forall c} (f_c - \mu)(f_c - \mu)^T)] \\ = & tr[\Sigma^{-1} (O - \mu \vec{V}^T + (W\mu - \vec{V})\mu^T)] \end{aligned} \tag{64}$$

Hence, the likelihood function can be expressed as a function of the sufficient statistics and the mean and variance parameters for that multivariate normal distribution. Interestingly, for any fixed Σ, the maximum likelihood μ always matches the weighted sample mean:

$$\frac{d\lambda}{d\mu} = -\frac{1}{2} tr[\sigma^{-1} 1 \frac{d}{d\mu} (O - \mu \vec{V}^T + (W\mu - \vec{V})\mu^T)] \tag{65}$$

Assuming Σ is not singular, the derivative is zero if and only if:

$$\begin{aligned} 0 &= (-2\vec{V}^T + 2W\mu^T) \\ \mu &= \frac{\vec{V}}{W} \end{aligned} \tag{66}$$

If we want to test that two populations of cells have a distribution of features that significantly differ in the mean value, while the covariance matrix is assumed to be the same, we get that the maximum likelihood

mean for each population always matches their sample mean. Hence, the maximum likelihood covariance is uncovered by measuring the deviation to the corresponding sample mean:

$$\begin{aligned}\Sigma &= \frac{\sum\limits_{\forall c\in A} w_c(f_c-\mu_A)(f_c-\mu_A)^T + \sum\limits_{\forall c\in B} w_c(f_c-\mu_B)(f_c-\mu_B)^T}{\sum\limits_{\forall c\in A\cup B} w_c} \\ &= \frac{O_A-\mu_A\vec{V}_A^T+(W_A\mu_A-\vec{V})\mu_A^T+O_B-\mu_B\vec{V}_B^T+(W_B\mu_B-\vec{V})\mu_B^T}{W_A+W_B} \\ &= \frac{O_A-\frac{\vec{V}_A\vec{V}_A^T}{W_A}+O_B-\frac{\vec{V}_B\vec{V}_B^T}{W_B}}{W_A+W_B}\end{aligned} \tag{67}$$

Hence, we can test if two populations of feature measurement are better explained by independent parameter, by comparing their log-likelihood based on the closes form for their maximum likelihood parameters:

$$\begin{aligned}&\mu_A=\mu_B=\frac{\vec{V}_A+\vec{V}_B}{W_A+W_B} \quad \Sigma_A=\Sigma_B=\frac{O_A+O_B-\frac{(\vec{V}_A+\vec{V}_B)(\vec{V}_A+\vec{V}_B)^T}{W_A+W_B}}{W_A+W_B} \\ &\mu_A=\frac{\vec{V}_A}{W_A},\mu_B=\frac{\vec{V}_B}{W_B} \qquad \Sigma_A=\Sigma_B=\frac{O_A-\frac{\vec{V}_A\vec{V}_A^T}{W_A}+O_B-\frac{\vec{V}_B\vec{V}_B^T}{W_B}}{W_A+W_B} \\ &\mu_A=\frac{\vec{V}_A}{W_A},\mu_B=\frac{\vec{V}_B}{W_B} \qquad \Sigma_A=\frac{O_A-\frac{\vec{V}_A\vec{V}_A^T}{W_A}}{W_A},\Sigma_B=\frac{O_B-\frac{\vec{V}_B\vec{V}_B^T}{W_B}}{W_B}\end{aligned} \tag{68}$$

With the above, we can evaluate the significance of the deviation in mean and/or covariance using Wilks' theorem, which states that the background distribution for the log-likelihood ratio converges to the chi-square distribution as the number of datapoint increases (with a factor of 2). The number of degree of freedom is determined by the difference in number of free parameters, which here could be either $d, d+\frac{d\cdot(d+1)}{2}$ or $\frac{d\cdot(d+1)}{2}$ depending on the selected null and alternative hypothesis. For example, we can reject that two populations have equal means, while their covariance matrix are assumed to be identical, by comparing the first two log-likelihoods. In practice, we observe that this test generates log-P-values that are highly correlated (0.99984, 19K hypotheses) with the T^2-hotelling test log-P-values; the main difference though is that the T^2-hotelling test is not inherently defined on weighted datapoints.

6.1 Maximum Likelihood Covariance in constrained a subspace

Suppose that we want to verify that the covariances of two populations differ significantly in a selected axis or set of axes, can we find a closed form for the two covariances? This concept of freeing a portion of the dimensions to increase the likelihood of the data has been performed before under the setup of a Maximum Likelihood Local Regression (MLLR) [57,58]. In contrast to this, the task here is to measure the significance of deviations, so that the axes are defined by the hypothesis and not by the data. Formally, we predefine a constraint that forces the covariance matrices Σ_A and Σ_B to be so that:

$$R\Sigma_A R^T = D^{(W_A+W_B)}R\Sigma_B R^T D^{(W_A+W_B)} \tag{69}$$

Where R is a predefined orthogonal matrix, and D is diagonal that has a subset of its diagonal as free parameters, while all other entries are 1.

$$-2\lambda = W_A log(|\Sigma_A|) + W_B log(|\Sigma_B|) + tr[\Sigma_A^{-1}(O_A - \frac{\vec{V_A}\vec{V_A}^T}{W_A}) + \Sigma_B^{-1}(O_B - \frac{\vec{V_B}\vec{V_B}^T}{W_B})] \tag{70}$$

Let $\Sigma = D^{\frac{W_B}{2}} R \Sigma_A R^T D^{\frac{W_B}{2}}$, then:

$$\begin{gathered} -2\lambda = W_A log(|\Sigma D^{-W_B}|) + W_B log(|\Sigma D^{W_A}|) + tr[(...)] \\ -2\lambda = (W_A + W_B) log(|\Sigma|) + tr[(...)] \end{gathered} \tag{71}$$

The matrix within the trace becomes:

$$D^{W_B}\Sigma^{-1}D^{W_B}R(O_A - \tfrac{\vec{V_A}\vec{V_A}^T}{W_A})R^T + D^{W_B}\Sigma^{-1}D^{-W_A}R(O_B - \tfrac{\vec{V_B}\vec{V_B}^T}{W_B})R^T \tag{72}$$

Hence, we are to maximize the likelihood by scaling the sample variance in the direction defined by the orthogonal matrix R. The above may be rewritten in term of sample variances that have been transformed by R:

$$-2\lambda = (W_A + W_B) log(|\Sigma|) + tr[\Sigma^{-1}(D^{W_B} S_A D^{W_B} + D^{-W_A} S_B D^{-W_A})] \tag{73}$$

The sum of positive define matrices is positive definite, and by the spectral theorem, there exist a square root to any positive definite matrix. By performing a change of variables, we get:
$B = \sqrt{D^{W_B} S_A D^{W_B} + D^{-W_A} S_B D^{-W_A}} \Sigma^{-} 1 \sqrt{D^{W_B} S_A D^{W_B} + D^{-W_A} S_B D^{-W_A})}$
Then:

$$\begin{gathered} -2\lambda = (W_A + W_B) log(|D^{W_B} S_A D^{W_B} + D^{-W_A} S_B D^{-W_A}|) \\ -(W_A + W_B) * log(|B|) + tr[B] \end{gathered} \tag{74}$$

Hence, the matrix B can be determined independently of the other matrices, and can be shown to be a unit matrix scaled by $W_A + W_B$. Hence, the maximization of the likelihood reduces to find the best allowed diagonal matrix D that minimizes the determinant of the following matrix:

$$|D^{W_B} S_A D^{W_B} + D^{-W_A} S_B D^{-W_A}| \tag{75}$$

The derivative of the logarithm of determinant of an invertible matrix is so that:

$$\frac{d}{d\alpha} log|A| = tr[A^{-1} \frac{d}{d\alpha} A] \tag{76}$$

In order to uncover the best matrix D, we then need to perform gradient descent using the above equation. Finally, we reconstruct the matrices Σ_A and Σ_B based on the found matrix D.

7 Supplementary Tables and Figures

	Mother cells					
Buds	HOwt	ura3	rap0	alp1	alp2	alp3
HOwt		.9665	.9663	.9732	.9594	.9484
ura3	.9652		.9736	.9729	.9666	.9531
rap0	.9662	.9686		.9737	.9635	.9537
alp1	.9692	.9661	.9680		.9868	.9737
alp2	.9548	.9580	.9583	.9851		.9794
alp3	.9382	.9420	.9429	.9653	.9745	

Table IV.1: **Correlation for protein abundance in time-profiles between experiments.** Correlation at the 4^{th} cell-stage key-point for protein abundance time series, for either Mother or Bud cells (Upper and lower triangle in table respectively). Experiments are either replicates (HOwt,ura3,rap0) or a time lapse of cells treated with alpha factor (alp1,2,3).

	Mother cells					
Buds	HOwt	ura3	rap0	alp1	alp2	alp3
HOwt		.9811	.9675	.9692	.9604	.9495
ura3	.9428		.9635	.9686	.9624	.9499
rap0	.9372	.9326		.9737	.9489	.9427
alp1	.9326	.9375	.9258		.9862	.9705
alp2	.9203	.9216	.9029	.9752		.9802
alp3	.8964	.8998	.8960	.9230	.9455	

Table IV.2: **Correlation for subcellular spread measure in time-profiles between experiments.** Correlation at the 4^{th} cell-stage key-point for subcellular spread time series, for either Mother or Bud cells (Upper and lower triangle in table respectively). Experiments are either replicates (HOwt,ura3,rap0) or a time lapse of cells treated with alpha factor (alp1,2,3).

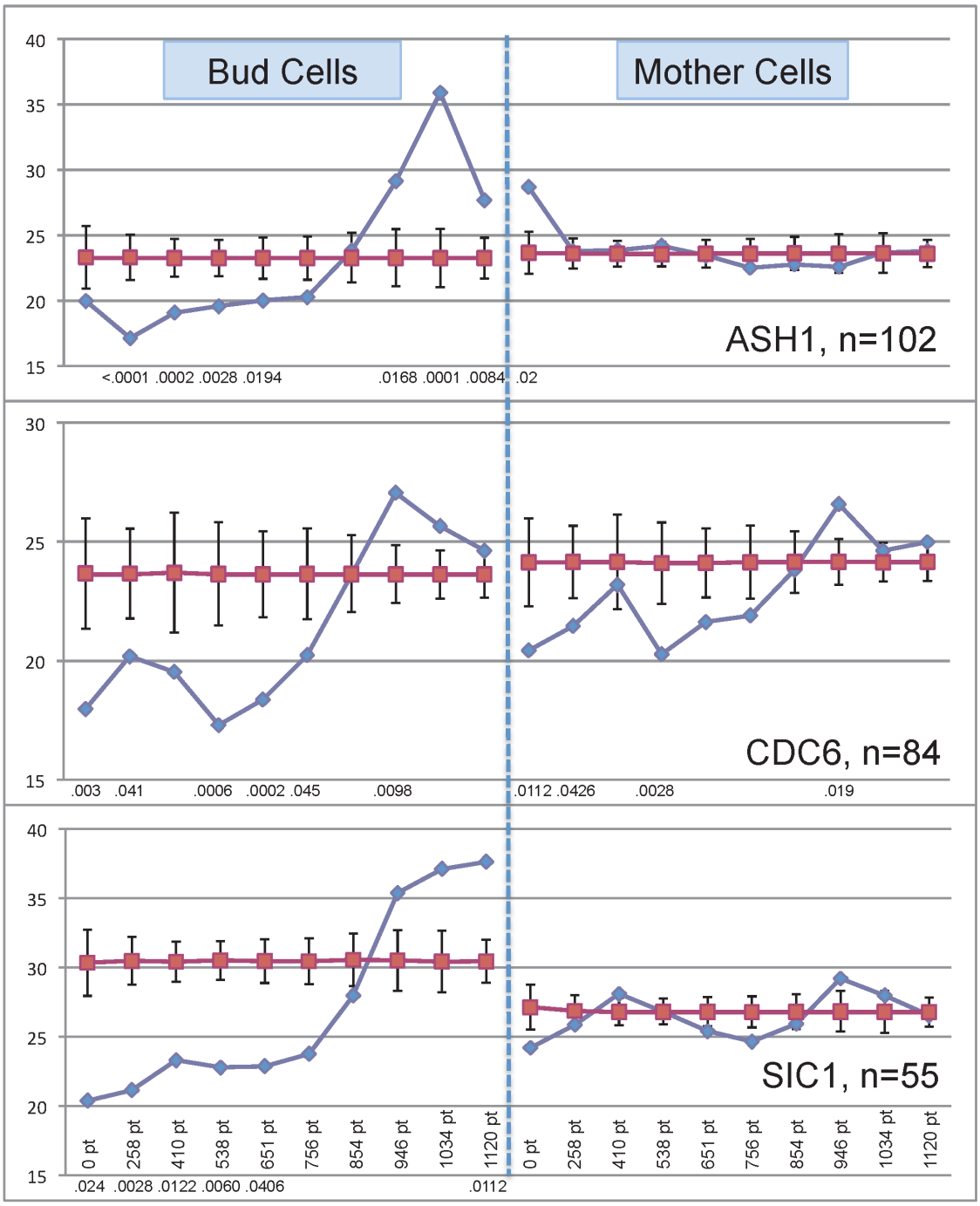


Figure IV.3: **Evaluation of significance of cell-stage deviations in protein expression.** Display of the local regression time-profile for the intensity of the proteins Ash1, Cdc6 and Sic6 (blue traces and symbols). The background distribution of intensity estimated at each time point is produced by permuting the cell-stage estimates for each identified mother-bud pair 10000 times (red traces and symbols. Error bars represent the standard deviation of the empirical distribution of the permutations). Numbers below the time points display P-values for the deviation of the time point from the real data (positive and negative deviations in the 2.5% tails of the empirical distribution of the permutations are reported).

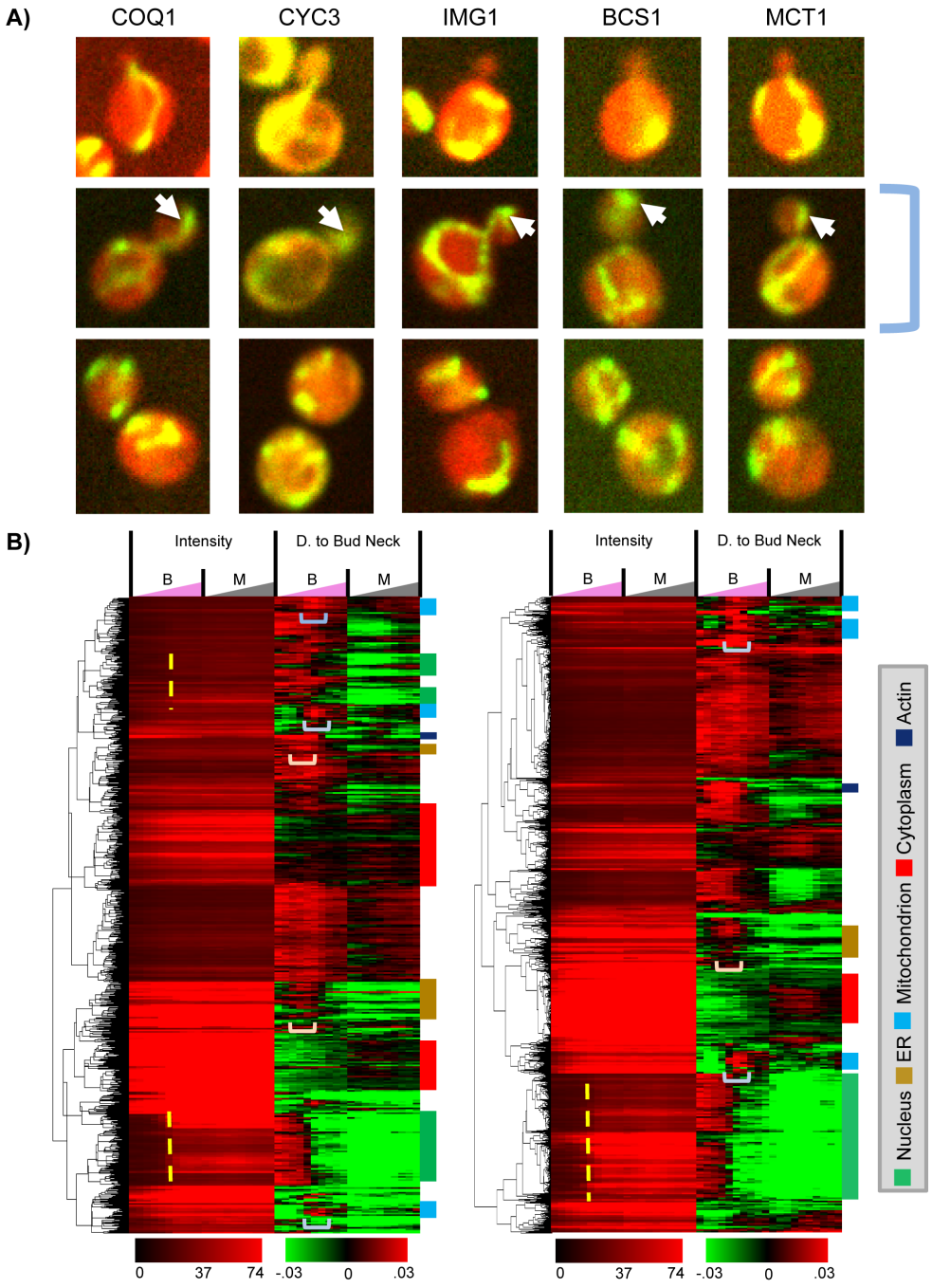


Figure IV.4: **Clustering visualization.** A) Visual inspection of the cell populations of 5 randomly chosen mitochondrial proteins allows us to identify mother-bud pair examples that appeared to correspond to our expectation (punctae indicated with arrows). For comparison we include mother-bud pairs with smaller or larger buds (top and bottom rows, respectively). B) On the left, the hierarchical clustering was performed on time-profiles that used a cell confidence threshold (of 0.8). On the right, the correlation metric and complete linkage hierarchical clustering was used. The inclusion of the nucleus in the bud is indicated with the dotted yellow line, and the characteristic time for proteins to reach their maximum distance to the bud neck is shown in light blue braces for Mitochondrion, and light orange brace for ER.

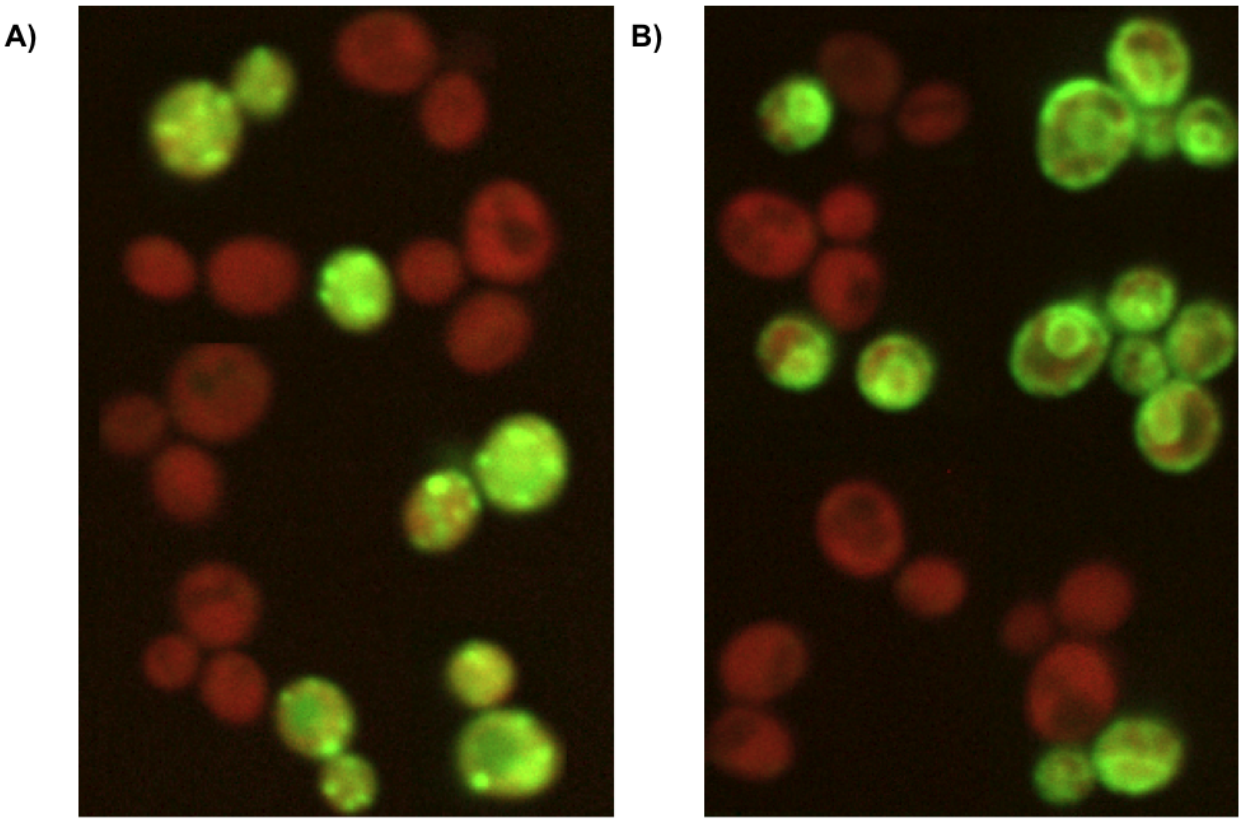


Figure IV.5: **Example with highest coefficient of variation.** A) SPL2 IS THE MOST VARIABLE IN PROTEIN ABUNDANCE. (COEF. OF VAR. = .811) B) CCW14 IS THE SECOND MOST VARIABLE. INTERESTINGLY, IT IS THE 1915^th IN THE RANKING OF PROTEIN BY VARIABILITY ACCORDING TO NEWMAN ET AL. [113](OUT OF 2008). (COEF. OF VAR. = .599)

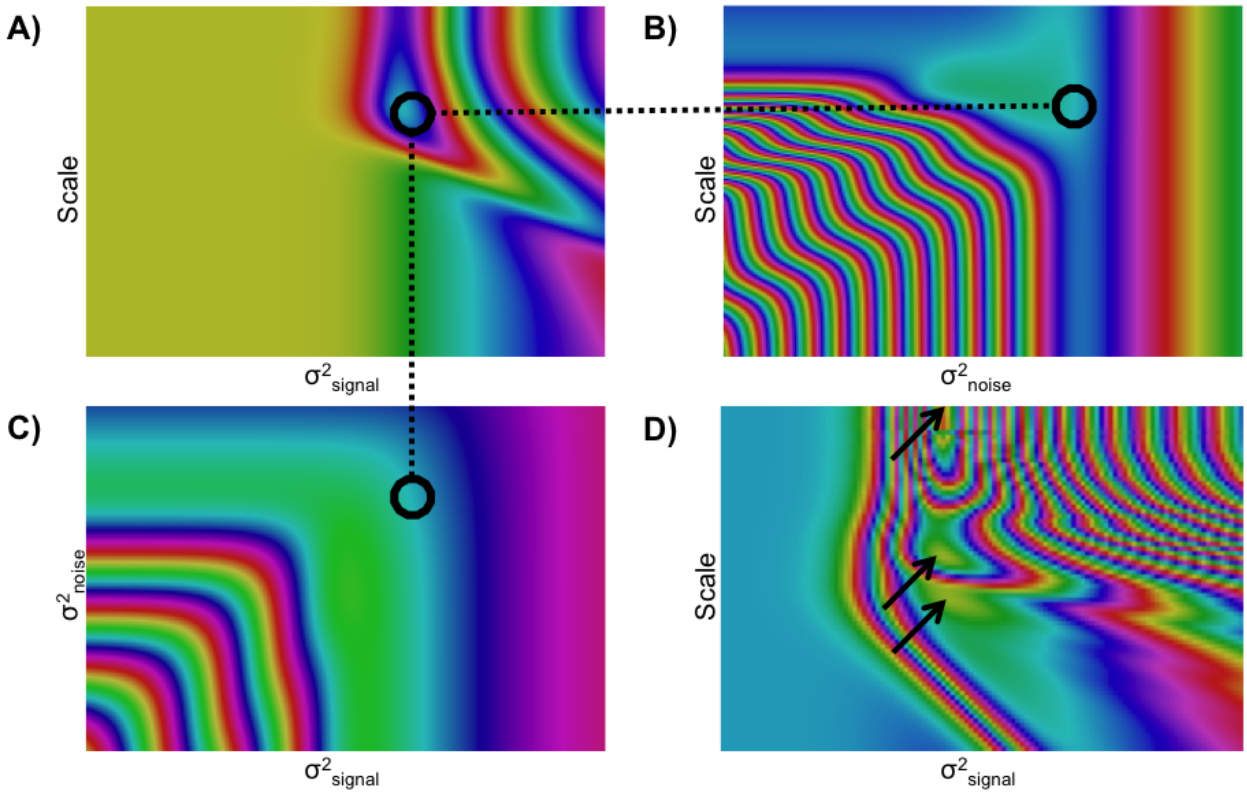


Figure IV.6: **GP likelihood landscape.** A,B,C) 2D slices of the 3 dimensional likelihood function, close to maximum likelihood parameters that model 6 data points. The color encoding renders a contour plot for the likelihood surface. D) Likelihood function coloured contour plot where the GP model 60 data points. We observe that 3 local maxima exist for the likelihood function (black arrows). The global maximum corresponds to the top arrow, which is reached when the scale is infinity. This indicates that the mean value for the datapoints is significantly different than zero, and that datapoint are correlated independently of their distances.

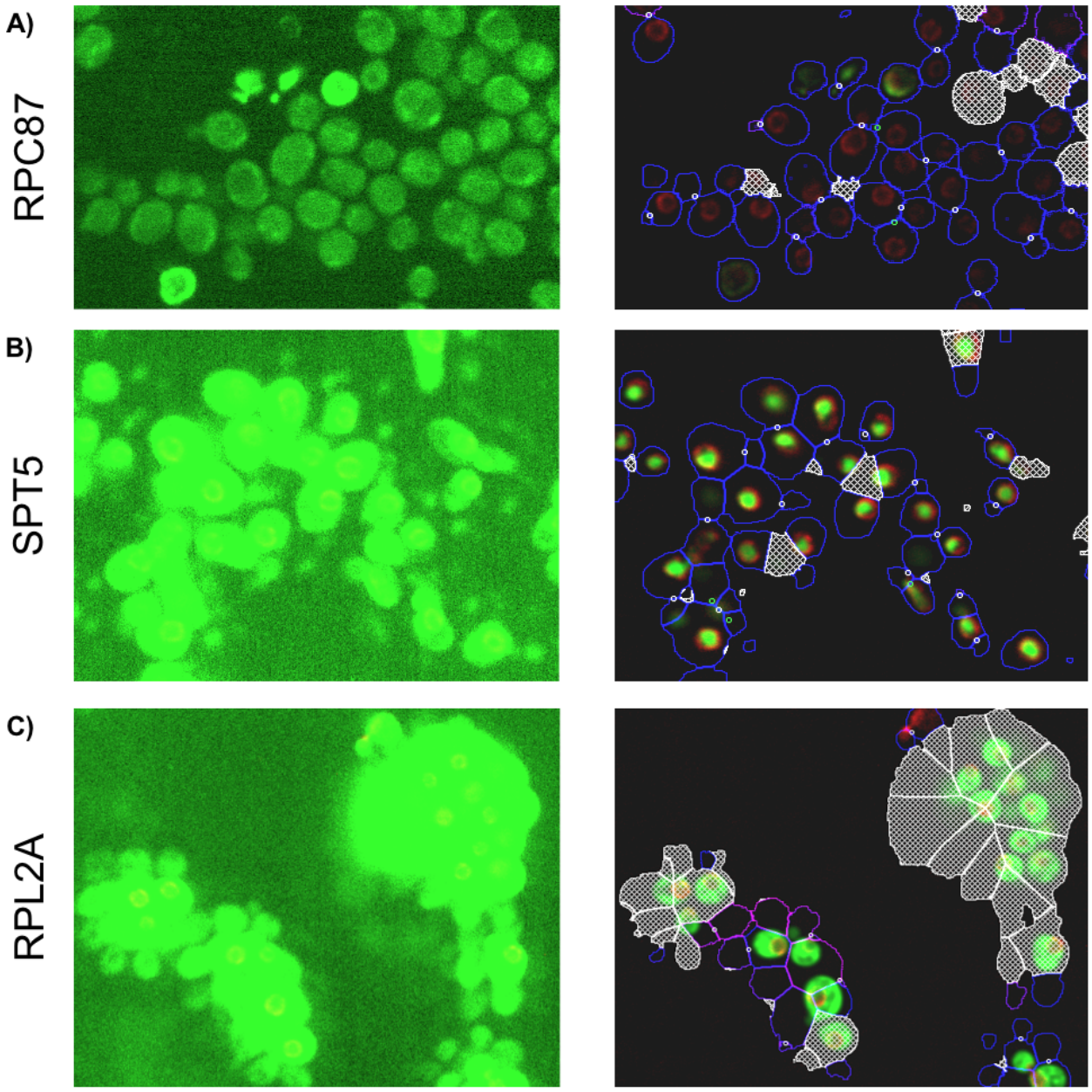


Figure IV.7: **Cell segmentation from autofluorescence.** A) Low levels of intensities are recorded in the GFP is due to autofluorescence in cells (left image). It can be used to report as a cell contour marker (right image). B) If the GFP intensity is high, the background distribution captures diffraction pattern generated by neighboring GFP signal. This generates erroneous lone cells, but the morphology is well captured. C) Extremely high intensities may overshadow the signal from autofluorescence, which appear to disallow the segmentation for the 100 most abundant proteins. (Images from Tkach et al. [156])

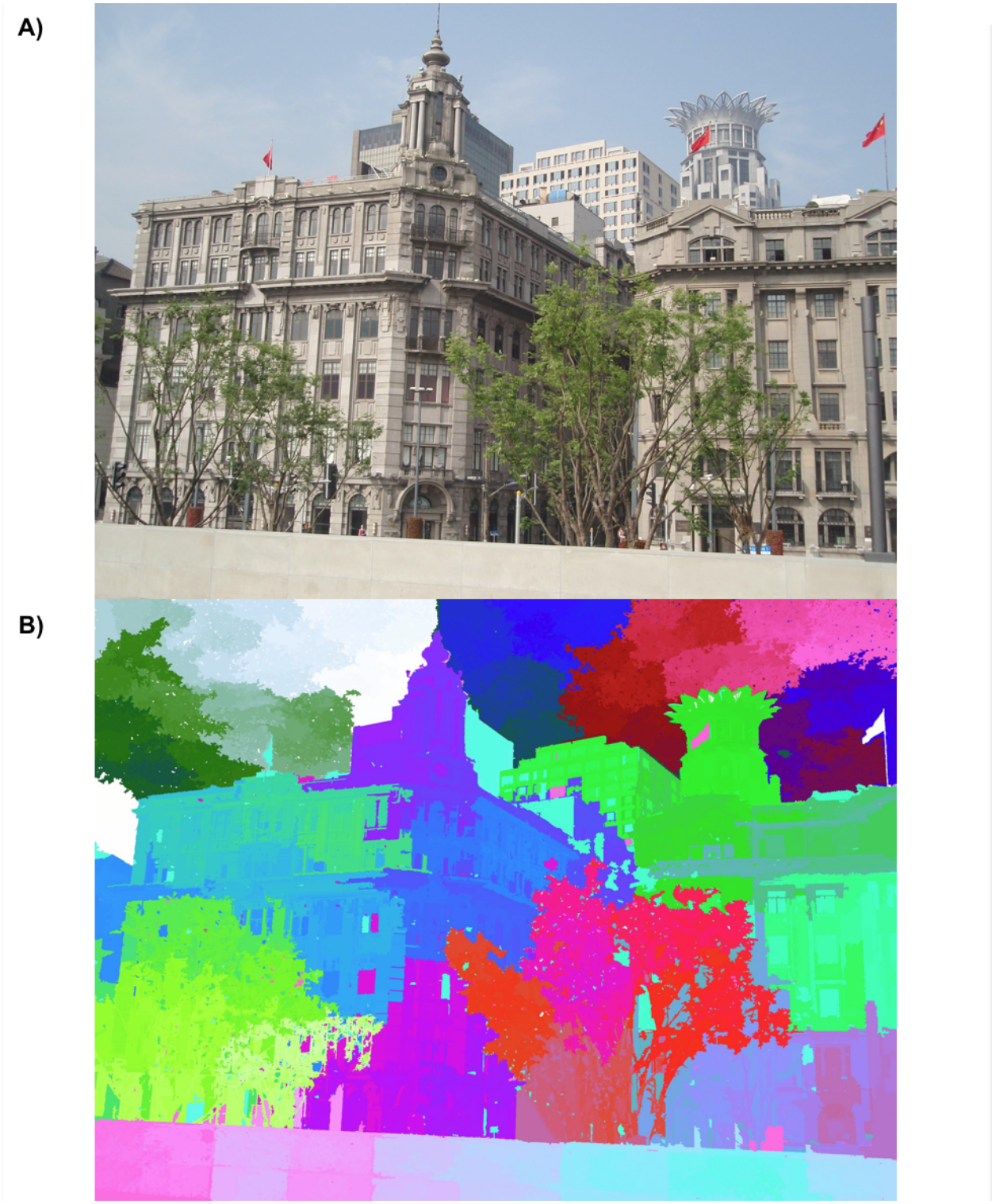


Figure IV.8: **Maximum likelihood clustering segmentation.** A) Color photograph of the bund, in Shanghai B) Maximum likelihood hierarchical clustering of the photograph pixels, where each pixel is a vector of the made of red, green, blue, x and y coordinates. The hierarchichal clustering is displayed as an image, where pixel color relates to the traversal ordering of the defined by the hierarchical clustering.

Bibliography

[1] M.D. Abràmoff, P.J. Magalhães, and S.J. Ram. Image processing with imagej. *Biophotonics international*, 11(7):36–42, 2004.

[2] M. Acar, J.T. Mettetal, and A. van Oudenaarden. Stochastic switching as a survival strategy in fluctuating environments. *Nature genetics*, 40(4):471–475, 2008.

[3] Y. Al-Kofahi, N. Dowell-Mesfin, C. Pace, W. Shain, J.N. Turner, and B. Roysam. Improved detection of branching points in algorithms for automated neuron tracing from 3d confocal images. *Cytometry Part A*, 73(1):36–43, 2007.

[4] R.W. Allen and M. Moskowitz. Arrest of cell growth in the g1 phase of the cell cycle by serine deprivation. *Experimental Cell Research*, 116(1):127–137, 1978.

[5] B. Amos. Lessons from the history of light microscopy. *Nature cell biology*, 2(8):E151–E152, 2000.

[6] P. Anandan. A computational framework and an algorithm for the measurement of visual motion. *International Journal of Computer Vision*, 2(3):283–310, 1989.

[7] D. Anguita, S. Ridella, and D. Sterpi. A new method for multiclass support vector machines. In *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, volume 1. IEEE, 2004.

[8] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25, 2000.

[9] S. Asur, D. Ucar, and S. Parthasarathy. An ensemble framework for clustering protein–protein interaction networks. *Bioinformatics*, 23(13):i29–i40, 2007.

[10] A. Bairoch. Uniprotkb/swiss-prot: New and future developments. In *Data Integration in the Life Sciences*, pp. 204–206. Springer, 2008.

[11] A. Bar-Even, J. Paulsson, N. Maheshri, M. Carmi, E. O'Shea, Y. Pilpel, and N. Barkai. Noise in protein expression scales with natural protein abundance. *Nature genetics*, 38(6):636–643, 2006.

[12] R.W. Barnard, K. Pearce, and L. Schovanec. Inequalities for the perimeter of an ellipse. *Journal of mathematical analysis and applications*, 260(2):295–306, 2001.

[13] A. Bateman, L. Coin, R. Durbin, R.D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E.L.L. Sonnhammer, et al. The pfam protein families database. *Nucleic acids research*, 32(suppl 1):D138–D141, 2004.

[14] Richard Ernest Bellman. Dynamic programming. *Rand Corporation*, 1957.

[15] Begoña Benito, Francisco J Quintero, and Alonso Rodrıguez-Navarro. Overexpression of the sodium atpase of¡ i¿ saccharomyces cerevisiae¡/i¿: conditions for phosphorylation from atp and p¡ sub¿ i¡/sub¿. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1328(2):214–225, 1997.

[16] S Beucher, M Bilodeau, and X Yu. Road segmentation by watershed algorithms. In *PROMETHEUS Workshop, Sophia Antipolis, France*, 1990.

[17] E. Bi, J.B. Chiavetta, H. Chen, G.C. Chen, C.S.M. Chan, and J.R. Pringle. Identification of novel, evolutionarily conserved cdc42p-interacting proteins and of redundant pathways linking cdc24p and cdc42p to actin polarization in yeast. *Molecular biology of the cell*, 11(2):773–793, 2000.

[18] Oren Boiman, Eli Shechtman, and Michal Irani. In defense of nearest-neighbor based image classification. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8. IEEE, 2008.

[19] M.V. Boland, M.K. Markey, and R.F. Murphy. Classification of protein localization patterns obtained via fluorescence light microscopy. In *Engineering in Medicine and Biology Society, 1997. Proceedings of the 19th Annual International Conference of the IEEE*, volume 2, pp. 594–597. IEEE, 1997.

[20] H. Breit and G. Rigoll. Improved person tracking using a combined pseudo-2d-hmm and kalman filter approach with automatic background state adaptation. In *Image Processing, 2001. Proceedings. 2001 International Conference on*, volume 2, pp. 53–56. IEEE, 2001.

[21] T.E. Buck, A. Rao, L.P. Coelho, M.H. Fuhrman, J.W. Jarvik, P.B. Berget, and R.F. Murphy. Cell cycle dependence of protein subcellular location inferred from static, asynchronous images. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pp. 1016–1019. IEEE, 2009.

[22] B. Bunow, J.P. Kernevez, G. Joly, and D. Thomas. Pattern formation by reaction-diffusion instabilities: Application to morphogenesis in drosophila. *Journal of Theoretical Biology*, 84(4):629–649, 1980.

[23] D. Cai, X. He, and J. Han. Speed up kernel discriminant analysis. *The VLDB Journal*, 20(1):21–33, 2011.

[24] L. Cai, C.K. Dalal, and M.B. Elowitz. Frequency-modulated nuclear localization bursts coordinate gene regulation. *Nature*, 455(7212):485–490, 2008.

[25] M.E.K. Calvert and J. Lannigan. Yeast cell cycle analysis: combining dna staining with cell and nuclear morphology. *Current Protocols in Cytometry*, pp. 9–32, 2010.

[26] A.E. Carpenter, T.R. Jones, M.R. Lamprecht, C. Clarke, I.H. Kang, O. Friman, D.A. Guertin, J.H. Chang, R.A. Lindquist, J. Moffat, et al. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome biology*, 7(10):R100, 2006.

[27] K.C. Chen, A. Csikasz-Nagy, B. Gyorffy, J. Val, B. Novak, and J.J. Tyson. Kinetic analysis of a molecular model of the budding yeast cell cycle. *Molecular biology of the cell*, 11(1):369–391, 2000.

[28] S.C. Chen, T. Zhao, G.J. Gordon, and R.F. Murphy. A novel graphical model approach to segmenting cell images. In *Computational Intelligence and Bioinformatics and Computational Biology, 2006. CIBCB'06. 2006 IEEE Symposium on*, pp. 1–8. IEEE, 2006.

[29] S.C. Chen, T. Zhao, G.J. Gordon, and R.F. Murphy. Automated image analysis of protein localization in budding yeast. *Bioinformatics*, 23(13):i66–i71, 2007.

[30] Tiffany Chen. De novo reconstruction of cell cycle progression using tour-recovered automatic models for cellular continuums (tracc) on multiparameter flow cytometry data. In *ISMB/ECCB 2013, late breaking research*, 2013.

[31] X. Chen, R.F. Murphy, et al. Objective clustering of proteins based on subcellular location patterns. *Journal of Biomedicine and Biotechnology*, 2:87, 2005.

[32] X. Chen, M. Velliste, and R.F. Murphy. Automated interpretation of subcellular patterns in fluorescence microscope images for location proteomics. *Cytometry Part A*, 69(7):631–640, 2006.

[33] X. Chen, M. Velliste, S. Weinstein, J.W. Jarvik, and R.F. Murphy. Location proteomics-building subcellular location trees from high resolution 3d fluorescence microscope images of randomly-tagged proteins. In *Proc sPie*, volume 4962, pp. 298–306, 2003.

[34] J.M. Cherry, C. Adler, C. Ball, S.A. Chervitz, S.S. Dwight, E.T. Hester, Y. Jia, G. Juvik, T.Y. Roe, M. Schroeder, et al. Sgd: Saccharomyces genome database. *Nucleic acids research*, 26(1):73–79, 1998.

[35] J.M. Cherry, C. Ball, S. Weng, G. Juvik, R. Schmidt, C. Adler, B. Dunn, S. Dwight, L. Riles, R.K. Mortimer, et al. Genetic and physical maps of saccharomyces cerevisiae. *Nature*, 387(6632 Suppl):67, 1997.

[36] K.C. Chou and Y.D. Cai. Predicting protein localization in budding yeast. *Bioinformatics*, 21(7):944–950, 2005.

[37] A.A. Cohen, N. Geva-Zatorsky, E. Eden, M. Frenkel-Morgenstern, I. Issaeva, A. Sigal, R. Milo, C. Cohen-Saidon, Y. Liron, Z. Kam, et al. Dynamic proteomics of individual cancer cells in response to a drug. *Science*, 322(5907):1511–1516, 2008.

[38] M. Costanzo, J.L. Nishikawa, X. Tang, J.S. Millman, O. Schub, K. Breitkreuz, D. Dewar, I. Rupes, B. Andrews, and M. Tyers. Cdk activity antagonizes whi5, an inhibitor of g1/s transcription in yeast. *Cell*, 117(7):899–913, 2004.

[39] J.E. Dahlberg, E. Lund, and E.B. Goodwin. Nuclear translation: What is the evidence? *Rna*, 9(1):1–8, 2003.

[40] A. Danckaert, E. Gonzalez-Couto, L. Bollondi, N. Thompson, and B. Hayes. Automated recognition of intracellular organelles in confocal microscope images. *Traffic*, 3(1):66–73, 2002.

[41] M.A.G. De Carvalho, R.A. Lotufo, and M. Couprie. Morphological segmentation of yeast by image analysis. *Image and Vision Computing*, 25(1):34–39, 2007.

[42] Michiel de Hoon, Seiya Imoto, and Satoru Miyano. The c clustering library. *Institute of Medical Science, Human Genome Center, University of Tokyo*, 2003.

[43] H. De Jong. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of computational biology*, 9(1):67–103, 2002.

[44] D.G. Drubin and W.J. Nelson. Origins of cell polarity. *Cell*, 84(3):335–344, 1996.

[45] L.S. Drury, G. Perkins, and J.F.X. Diffley. The cdc4/34/53 pathway targets cdc6p for proteolysis in budding yeast. *The EMBO journal*, 16(19):5966–5976, 1997.

[46] B. Efron and B. Efron. *The jackknife, the bootstrap, and other resampling plans*, volume 38. SIAM, 1982.

[47] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863, 1998.

[48] M.B. Elowitz, A.J. Levine, E.D. Siggia, and P.S. Swain. Stochastic gene expression in a single cell. *Science Signalling*, 297(5584):1183, 2002.

[49] O. Emanuelsson, H. Nielsen, S. Brunak, and G. von Heijne. Predicting subcellular localization of proteins based on their n-terminal amino acid sequence. *Journal of molecular biology*, 300(4):1005–1016, 2000.

[50] S. Farkash-Amar, E. Eden, A. Cohen, N. Geva-Zatorsky, L. Cohen, R. Milo, A. Sigal, T. Danon, and U. Alon. Dynamic proteomics of human protein level and localization across the cell cycle. *PloS one*, 7(11):e48722, 2012.

[51] L. Firestone, K. Cook, K. Culp, N. Talsania, and K. Preston. Comparison of autofocus methods for automated microscopy. *Cytometry*, 12(3):195–206, 2005.

[52] Ronald Aylmer Fisher et al. On the" probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32, 1921.

[53] L.J. Foster, C.L. de Hoog, Y. Zhang, Y. Zhang, X. Xie, V.K. Mootha, and M. Mann. A mammalian organelle map by protein correlation profiling. *Cell*, 125(1):187–199, 2006.

[54] D. Freifelder. Bud position in saccharomyces cerevisiae. *Journal of bacteriology*, 80(4):567, 1960.

[55] I. Friedberg. Automated protein function predictionthe genomic challenge. *Briefings in bioinformatics*, 7(3):225–242, 2006.

[56] J.G. Fujimoto and D. Farkas. *Biomedical optical imaging*. Oxford University Press, USA, 2009.

[57] Mark JF Gales. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer speech and language*, 12(2), 1998.

[58] Mark JF Gales and PC Woodland. Mean and variance adaptation within the mllr framework. *Computer Speech and Language*, 10(4):249–264, 1996.

[59] W. Gander, G.H. Golub, and R. Strebel. Least-squares fitting of circles and ellipses. *BIT Numerical Mathematics*, 34(4):558–578, 1994.

[60] T.S. Gardner, C.R. Cantor, and J.J. Collins. Construction of a genetic toggle switch inescherichia coli. *Nature*, 403:339–342, 2000.

[61] E. Glory and R.F. Murphy. Automated subcellular location determination and high-throughput microscopy. *Developmental cell*, 12(1):7–16, 2007.

[62] S.M. Goldfeld and R.E. Quandt. Nonlinear simultaneous equations: estimation and prediction. *International Economic Review*, 9(1):113–136, 1968.

[63] G. Gong. Cross-validation, the jackknife, and the bootstrap: excess error estimation in forward logistic regression. *Journal of the American Statistical Association*, 81(393):108–113, 1986.

[64] N. Hamilton and R. Teasdale. Visualizing and clustering high throughput sub-cellular localization imaging. *BMC bioinformatics*, 9(1):81, 2008.

[65] N.A. Hamilton, R.S. Pantelic, K. Hanson, and R.D. Teasdale. Fast automated cell phenotype image classification. *BMC bioinformatics*, 8(1):110, 2007.

[66] Louis-François Handfield, Yolanda T Chong, Jibril Simmons, Brenda J Andrews, and Alan M Moses. Unsupervised clustering of subcellular protein expression patterns in high-throughput microscopy images reveals protein complexes and functional relationships between proteins. *PLoS computational biology*, 9(6):e1003085, 2013.

[67] R.M. Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5):786–804, 1979.

[68] JA Hartigan and MA Wong. A k-means clustering algorithm. *Journal of the Royal Statistical Society C*, 28(1):100–108, 1979.

[69] J. Heitman, N.R. Movva, and M.N. Hall. Targets for cell cycle arrest by the immunosuppressant rapamycin in yeast. *Science*, 253(5022):905–909, 1991.

[70] W. Hilbe, A. Gächter, H.C. Duba, S. Dirnhofer, W. Eisterer, T. Schmid, A. Mildner, J. Bodner, E. Wöll, et al. Comparison of automated cellular imaging system and manual microscopy for immunohistochemically stained cryostat sections of lung cancer specimens applying p53, ki-67 and p120. *Oncology reports*, 10(1):15, 2003.

[71] P.W. Holland and R.E. Welsch. Robust regression using iteratively reweighted least-squares. *Communications in Statistics-Theory and Methods*, 6(9):813–827, 1977.

[72] K. Höllig. *Finite element methods with B-splines*, volume 26. Society for Industrial Mathematics, 2003.

[73] A.D. Hoppe, S.L. Shorte, J.A. Swanson, and R. Heintzmann. Three-dimensional fret reconstruction microscopy for analysis of dynamic molecular interactions in live cells. *Biophysical journal*, 95(1):400–418, 2008.

[74] Y. Hu, E. Osuna-Highley, J. Hua, T.S. Nowicki, R. Stolz, C. McKayle, and R.F. Murphy. Automated analysis of protein subcellular location in time series images. *Bioinformatics*, 26(13):1630, 2010.

[75] S. Huh, D. Lee, and R.F. Murphy. Efficient framework for automated classification of subcellular patterns in budding yeast. *Cytometry Part A*, 75(11):934–940, 2009.

[76] W.K. Huh, J.V. Falvo, L.C. Gerke, A.S. Carroll, R.W. Howson, J.S. Weissman, E.K. O'Shea, et al. Global analysis of protein localization in budding yeast. *Nature*, 425(6959):686–691, 2003.

[77] S. Hunter, R. Apweiler, T.K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, et al. Interpro: the integrative protein signature database. *Nucleic acids research*, 37(suppl 1):D211–D215, 2009.

[78] A. Hyvärinen, J. Hurri, and P.O. Hoyer. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision.*, volume 39. Springer, 2009.

[79] C. Jackson, E. Glory, RF Murphy, J. Kovačević, et al. Model building and intelligent acquisition with application to protein subcellular location classification. *Bioinformatics*, 27(13):1854–1859, 2011.

[80] C. Jackson, R.F. Murphy, and J. Kovacevic. Intelligent acquisition and learning of fluorescence microscope data models. *Image Processing, IEEE Transactions on*, 18(9):2071–2084, 2009.

[81] Thorsten Joachims. Svmlight: Support vector machine. *SVM-Light Support Vector Machine http://svmlight. joachims. org/, University of Dortmund*, 19:4, 1999.

[82] T. Jones, A. Carpenter, and P. Golland. Voronoi-based segmentation of cells on image manifolds. *Computer Vision for Biomedical Image Applications*, pp. 535–543, 2005.

[83] Mads Kærn, Timothy C Elston, William J Blake, and James J Collins. Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics*, 6(6):451–464, 2005.

[84] Z. Kam et al. Microscopic imaging of cells. *Q. Rev. Biophys*, 20:201–259, 1987.

[85] M. Kanehisa, S. Goto, M. Hattori, K.F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics: new developments in kegg. *Nucleic acids research*, 34(suppl 1):D354–D357, 2006.

[86] A. Kannan, M. Ostendorf, and JR Rohlicek. Maximum likelihood clustering of gaussians for speech recognition. *Speech and Audio Processing, IEEE Transactions on*, 2(3):453–455, 1994.

[87] R. Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International joint Conference on artificial intelligence*, volume 14, pp. 1137–1145. Lawrence Erlbaum Associates Ltd, 1995.

[88] S.A. Krause, H. Xu, and J.V. Gray. The synthetic genetic network around pkc1 identifies novel modulators and components of protein kinase c signaling in saccharomyces cerevisiae. *Eukaryotic cell*, 7(11):1880–1887, 2008.

[89] W. Krek. Proteolysis and the g1-s transition: the scf connection. *Current opinion in genetics & development*, 8(1):36–42, 1998.

[90] M. Kvarnström, K. Logg, A. Diez, K. Bodvard, and M. Käll. Image analysis algorithms for cell contour recognition in budding yeast. *Optics express*, 16(17):12943–12957, 2008.

[91] K.Y. Lee, D.W. Kim, D.K. Na, K.H. Lee, and D. Lee. Plpd: reliable protein localization prediction from imbalanced and overlapped datasets. *Nucleic acids research*, 34(17):4655–4666, 2006.

[92] T.W. Lee and M.S. Lewicki. Unsupervised image classification, segmentation, and enhancement using ica mixture models. *Image Processing, IEEE Transactions on*, 11(3):270–279, 2002.

[93] H.N. Lin, C.T. Chen, T.Y. Sung, S.Y. Ho, and W.L. Hsu. Protein subcellular localization prediction of eukaryotes using a knowledge-based approach. *BMC bioinformatics*, 10(Suppl 15):S8, 2009.

[94] Q. Liu, B. Larsen, Mhuh2003global. Ricicova, S. Orlicky, H. Tekotte, X. Tang, K. Craig, A. Quiring, T. Le Bihan, C. Hansen, et al. Scfcdc4 enables mating type switching in yeast by cyclin-dependent kinase-mediated elimination of the ash1 transcriptional repressor. *Molecular and Cellular Biology*, 31(3):584–598, 2011.

[95] C. Loader. *Local regression and likelihood*. Springer Verlag, 1999.

[96] Maria Pia Longhese, Paolo Plevani, and Giovanna Lucchini. Replication factor a is required in vivo for dna replication, repair, and recombination. *Molecular and cellular biology*, 14(12):7884–7890, 1994.

[97] A. Lorberg, H.P. Schmitz, J. Jacoby, and J. Heinisch. Lrg1p functions as a putative gtpase-activating protein in the pkc1p-mediated cell integrity pathway in saccharomyces cerevisiae. *Molecular Genetics and Genomics*, 266(3):514–526, 2001.

[98] Isabel Mayordomo, Francisco Estruch, and Pascual Sanz. Convergence of the target of rapamycin and the snf1 protein kinase pathways in the regulation of the subcellular localization of msn2, a transcriptional activator of stre (stress response element)-regulated genes. *Journal of Biological Chemistry*, 277(38):35650–35656, 2002.

[99] J.P. McIntire, L.K. McIntire, and P.R. Havig. A variety of automated turing tests for network security: Using ai-hard problems in perception and cognition to ensure secure collaborations. In *Collaborative Technologies and Systems, 2009. CTS'09. International Symposium on*, pp. 155–162. IEEE, 2009.

[100] Michael A McMurray, Aurelie Bertin, Galo Garcia III, Lisa Lam, Eva Nogales, and Jeremy Thorner. Septin filament formation is essential in budding yeast. *Developmental cell*, 20(4):540–549, 2011.

[101] P. Meer, D. Mintz, A. Rosenfeld, and D.Y. Kim. Robust regression methods for computer vision: A review. *International journal of computer vision*, 6(1):59–70, 1991.

[102] I Mendizabal, A Pascual-Ahuir, R Serrano, and I de Larrinoa. Promoter sequences regulated by the calcineurin-activated transcription factor crz1 in the yeast ena1 gene. *Molecular Genetics and Genomics*, 265(5):801–811, 2001.

[103] T.E. Merryman and J. Kovacevic. An adaptive multirate algorithm for acquisition of fluorescence microscopy data sets. *Image Processing, IEEE Transactions on*, 14(9):1246–1253, 2005.

[104] RS Michalski, A. Rosenfeld, Z. Duric, M. Maloof, and Q. Zhang. Learning patterns in images. *Machine Learning and Data Mining: Methods and Applications*, pp. 241–268, 1998.

[105] D. Michie, D.J. Spiegelhalter, C.C. Taylor, and J. Campbell. *Machine learning, neural and statistical classification*. Ellis Horwood London, 1994.

[106] A. Mitiche and P. Bouthemy. Computation and analysis of image motion: A synopsis of current problems and methods. *International journal of computer vision*, 19(1):29–55, 1996.

[107] Brian Munsky, Gregor Neuert, and Alexander van Oudenaarden. Using gene expression noise to understand gene regulation. *Science*, 336(6078):183–187, 2012.

[108] R.F. Murphy, M.V. Boland, M. Velliste, et al. Towards a systematics for protein subcellular location: quantitative description of protein localization patterns and automated analysis of fluorescence microscope images. In *Proc. Int. Conf. Intell. Syst. Mol. Biol*, volume 8, pp. 251–259, 2000.

[109] R.F. Murphy, M. Velliste, and G. Porreca. Robust numerical features for description and classification of subcellular location patterns in fluorescence microscope images. *The Journal of VLSI Signal Processing*, 35(3):311–321, 2003.

[110] Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, et al. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90–94, 2011.

[111] V. Neduva and R.B. Russell. Dilimot: discovery of linear motifs in proteins. *Nucleic acids research*, 34(suppl 2):W350–W355, 2006.

[112] J. Newberg, J. Hua, and R.F. Murphy. Location proteomics: systematic determination of protein subcellular location. *Methods in Molecular Biology, Systems Biology*, 500:1–20, 2009.

[113] J.R.S. Newman, S. Ghaemmaghami, J. Ihmels, D.K. Breslow, M. Noble, J.L. DeRisi, and J.S. Weissman. Single-cell proteomic analysis of s. cerevisiae reveals the architecture of biological noise. *Nature*, 441(7095):840–846, 2006.

[114] V.Q. Nguyen, C. Co, K. Irie, and J.J. Li. Clb/cdc28 kinases promote nuclear export of the replication initiator proteins mcm2-7. *Current Biology*, 10(4):195–205, 2000.

[115] A. Niemistö, M. Nykter, T. Aho, H. Jalovaara, K. Marjanen, M. Ahdesmäki, P. Ruusuvuori, M. Tiainen, M.L. Linne, and O. Yli-Harja. Computational methods for estimation of cell cycle phase distributions of yeast cells. *EURASIP Journal on Bioinformatics and Systems Biology*, 2007:2–2, 2007.

[116] P. OBrien and JR Haskins. High content screening: A powerful approach to systems cell biology and drug discovery ed. *Taylor et al*, pp. 415–425, 2007.

[117] Y. Ohya, J. Sese, M. Yukawa, F. Sano, Y. Nakatani, T.L. Saito, A. Saka, T. Fukuda, S. Ishihara, S. Oka, et al. High-dimensional and large-scale phenotyping of yeast mutants. *Proceedings of the National Academy of Sciences of the United States of America*, 102(52):19015, 2005.

[118] C Ortiz de Solorzano, E Garcia Rodriguez, A Jones, D Pinkel, JW Gray, D Sudar, and SJ Lockett. Segmentation of confocal microscope images of cell nuclei in thick tissue sections. *Journal of Microscopy*, 193(3):212–226, 1999.

[119] N. Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.

[120] J. Paulsson. Summing up the noise in gene networks. *Nature*, 427(6973):415–418, 2004.

[121] Lucas Pelkmans. Using cell-to-cell variabilitya new era in molecular biology. *Science*, 336(6080):425–426, 2012.

[122] H. Peng. Bioimage informatics: a new area of engineering biology. *Bioinformatics*, 24(17):1827–1836, 2008.

[123] H. Peng, F. Long, J. Zhou, G. Leung, M. Eisen, and E. Myers. Automatic image analysis for gene expression patterns of fly embryos. *BMC Cell Biology*, 8(Suppl 1):S7, 2007.

[124] T. Peng, G.M.C. Bonamy, E. Glory-Afshar, D.R. Rines, S.K. Chanda, and R.F. Murphy. Determining the distribution of probes between different subcellular locations through automated unmixing of subcellular patterns. *Proceedings of the National Academy of Sciences*, 107(7):2944–2949, 2010.

[125] T. Peng and R.F. Murphy. Image-derived, three-dimensional generative models of cellular organization. *Cytometry Part A*, 2011.

[126] M. Piel and P.T. Tran. Cell shape and cell division in fission yeast. *Current Biology*, 19(17):R823–R827, 2009.

[127] D. Preuss, J. Mulholland, C.A. Kaiser, P. Orlean, C. Albright, M.D. Rose, P.W. Robbins, D. Botstein, et al. Structure of the yeast endoplasmic reticulum: localization of er proteins using immunofluorescence and immunoelectron microscopy. *Yeast (Chichester, England)*, 7(9):891–911, 1991.

[128] Arjun Raj, Scott A Rifkin, Erik Andersen, and Alexander van Oudenaarden. Variability in gene expression underlies incomplete penetrance. *Nature*, 463(7283):913–918, 2010.

[129] Arjun Raj and Alexander van Oudenaarden. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216–226, 2008.

[130] C.E. Rasmussen and C.K.I. Williams. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, MA, 2006.

[131] Michael Riffle and Trisha N Davis. The yeast resource center public image repository: A large database of fluorescence microscopy images. *BMC bioinformatics*, 11(1):263, 2010.

[132] J.R. Robbins, D. Monack, S.J. McCallum, A. Vegas, E. Pham, M.B. Goldberg, and J.A. Theriot. The making of a gradient: Icsa (virg) polarity in shigella flexneri. *Molecular microbiology*, 41(4):861–872, 2001.

[133] A. Rodriguez, D.B. Ehlenberger, D.L. Dickstein, P.R. Hof, and S.L. Wearne. Automated three-dimensional detection and shape classification of dendritic spines from fluorescence microscopy images. *PLoS One*, 3(4):e1997, 2008.

[134] Nitzan Rosenfeld, Jonathan W Young, Uri Alon, Peter S Swain, and Michael B Elowitz. Gene regulation at the single-cell level. *Science Signaling*, 307(5717):1962, 2005.

[135] V. Rossio and S. Yoshida. Spatial regulation of cdc55–pp2a by zds1/zds2 controls mitotic entry and mitotic exit in budding yeast. *The Journal of cell biology*, 193(3):445–454, 2011.

[136] D. Sage, M. Unser, P. Salmon, and C. Dibner. A software solution for recording circadian oscillator features in time-lapse live cell microscopy. *Cell division*, 5(1):17, 2010.

[137] I. Sagot, S.K. Klee, and D. Pellman. Yeast formins regulate cell polarity by controlling the assembly of actin cables. *Nature cell biology*, 4(1):42–50, 2001.

[138] T.L. Saito, M. Ohtani, H. Sawai, F. Sano, A. Saka, D. Watanabe, M. Yukawa, Y. Ohya, and S. Morishita. Scmd: Saccharomyces cerevisiae morphological database. *Nucleic acids research*, 32(suppl 1):D319–D322, 2004.

[139] M. Séguin and B. Villeneuve. *Astronomie et astrophysique: cinq grandes idées pour explorer et comprendre l'univers*. Éditions du renouveau pédagogique, 2002.

[140] V. Shahrezaei and P.S. Swain. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences*, 105(45):17256–17261, 2008.

[141] L. Shamir, J.D. Delaney, N. Orlov, D.M. Eckley, and I.G. Goldberg. Pattern recognition software and techniques for biological image analysis. *PLoS Computational Biology*, 6(11):e1000974, 2010.

[142] N.C. Shaner, R.E. Campbell, P.A. Steinbach, B.N.G. Giepmans, A.E. Palmer, and R.Y. Tsien. Improved monomeric red, orange and yellow fluorescent proteins derived from discosoma sp. red fluorescent protein. *Nature biotechnology*, 22(12):1567–1572, 2004.

[143] H.M. Shapiro and R.C. Leif. *Practical flow cytometry*, volume 736. Wiley Online Library, 2003.

[144] H. Shen, B. Roysam, C.V. Stewart, J.N. Turner, and H.L. Tanenbaum. Optimal scheduling of tracing computations for real-time vascular landmark extraction from retinal fundus images. *Information Technology in Biomedicine, IEEE Transactions on*, 5(1):77–91, 2001.

[145] Sae Shimizu-Sato, Enamul Huq, James M Tepperman, and Peter H Quail. A light-switchable gene promoter system. *Nature biotechnology*, 20(10):1041–1044, 2002.

[146] Sae Shimizu-Sato, Enamul Huq, James M Tepperman, and Peter H Quail. A light-switchable gene promoter system. *Nature biotechnology*, 20(10):1041–1044, 2002.

[147] O. Shimomura, F.H. Johnson, and Y. Saiga. Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan, aequorea. *Journal of cellular and comparative physiology*, 59(3):223–239, 1962.

[148] A. Sigal, R. Milo, A. Cohen, N. Geva-Zatorsky, Y. Klein, I. Alaluf, N. Swerdlin, N. Perzov, T. Danon, Y. Liron, et al. Dynamic proteomics in individual human cells uncovers widespread cell-cycle dependence of nuclear proteins. *Nature Methods*, 3(7):525–531, 2006.

[149] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell*, 9(12):3273–3297, 1998.

[150] J.L. Spudich, DE Koshland Jr, et al. Non-genetic individuality: chance in the single cell. *Nature*, 262(5568):467, 1976.

[151] J.A.K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[152] P.S. Swain, M.B. Elowitz, and E.D. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences*, 99(20):12795, 2002.

[153] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377–382, 2009.

[154] Y. Tao, X. Zheng, and Y. Sun. Effect of feedback regulation on stochastic gene expression. *Journal of theoretical biology*, 247(4):827–836, 2007.

[155] I.V. Tetko, D.J. Livingstone, and A.I. Luik. Neural network studies. 1. comparison of overfitting and overtraining. *Journal of chemical information and computer sciences*, 35(5):826–833, 1995.

[156] Johnny M Tkach, Askar Yimit, Anna Y Lee, Michael Riffle, Michael Costanzo, Daniel Jaschob, Jason A Hendry, Jiongwen Ou, Jason Moffat, Charles Boone, et al. Dissecting dna damage response pathways by analysing protein localization and abundance changes during dna replication stress. *Nature cell biology*, 14(9):966–976, 2012.

[157] P. Tomancak, B.P. Berman, A. Beaton, R. Weiszmann, E. Kwan, V. Hartenstein, S.E. Celniker, and G.M. Rubin. Global analysis of patterns of gene expression during drosophila embryogenesis. *Genome biology*, 8(7):R145, 2007.

[158] A.H.Y. Tong, G. Lesage, G.D. Bader, H. Ding, H. Xu, X. Xin, J. Young, G.F. Berriz, R.L. Brost, M. Chang, et al. Global mapping of the yeast genetic interaction network. *Science*, 303(5659):808–813, 2004.

[159] Amy Hin Yan Tong, Marie Evangelista, Ainslie B Parsons, Hong Xu, Gary D Bader, Nicholas Page, Mark Robinson, Sasan Raghibizadeh, Christopher WV Hogue, Howard Bussey, et al. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science Signaling*, 294(5550):2364, 2001.

[160] C.L. Tsai, C.V. Stewart, H.L. Tanenbaum, and B. Roysam. Model-based method for improving the accuracy and repeatability of estimating vascular bifurcations and crossovers from retinal fundus images. *Information Technology in Biomedicine, IEEE Transactions on*, 8(2):122–130, 2004.

[161] R.Y. Tsien. The green fluorescent protein. *Annual review of biochemistry*, 67(1):509–544, 1998.

[162] Johannes Tuikkala, Laura L Elo, Olli S Nevalainen, and Tero Aittokallio. Missing value imputation improves clustering and interpretation of gene expression microarray data. *BMC bioinformatics*, 9(1):202, 2008.

[163] M. Valinluck, S. Ahlgren, M. Sawada, K. Locken, and F. Banuett. Role of the nuclear migration protein lis1 in cell morphogenesis in ustilago maydis. *Mycologia*, 102(3):493–512, 2010.

[164] R. Verma, RM Feldman, and R.J. Deshaies. Sic1 is ubiquitinated in vitro by a pathway that requires cdc4, cdc34, and cyclin/cdk activities. *Molecular biology of the cell*, 8(8):1427, 1997.

[165] Carolina Wählby, I-M SINTORN, Fredrik Erlandsson, Gunilla Borgefors, and Ewert Bengtsson. Combining intensity, edge and shape information for 2d and 3d segmentation of cell nuclei in tissue sections. *Journal of Microscopy*, 215(1):67–76, 2004.

[166] H. Wallrabe, A. Periasamy, et al. Imaging protein molecules using fret and flim microscopy. *Current opinion in biotechnology*, 16(1):19–27, 2005.

[167] MP Wand. Error analysis for general multtvariate kernel estimators. *Journal of Nonparametric Statistics*, 2(1):1–15, 1992.

[168] Jessica B Warner and Juke S Lolkema. Lacz-promoter fusions: the effect of growth. *MICROBIOLOGY-READING-*, 148(5):1241–1243, 2002.

[169] P. Xu and A.K. Chan. Support vector machines for multi-class signal classification with unbalanced samples. In *Neural Networks, 2003. Proceedings of the International Joint Conference on*, volume 2, pp. 1116–1119. IEEE, 2003.

[170] Q. Xu, D.H. Hu, H. Xue, W. Yu, and Q. Yang. Semi-supervised protein subcellular localization. *BMC bioinformatics*, 10(Suppl 1):S47, 2009.

[171] S. Yoshida, R. Ichihashi, and A. Toh-e. Ras recruits mitotic exit regulator lte1 to the bud cortex in budding yeast. *The Journal of cell biology*, 161(5):889–897, 2003.

[172] C.S. Yu, Y.C. Chen, C.H. Lu, and J.K. Hwang. Prediction of protein subcellular localization. *Proteins: Structure, Function, and Bioinformatics*, 64(3):643–651, 2006.

[173] X. Yuan, J.T. Trachtenberg, S.M. Potter, and B. Roysam. Mdl constrained 3-d grayscale skeletonization algorithm for automated extraction of dendrites and spines from fluorescence confocal images. *Neuroinformatics*, 7(4):213–232, 2009.

[174] T. Zhao, M. Velliste, M.V. Boland, and R.F. Murphy. Object type recognition for automated analysis of protein subcellular location. *Image Processing, IEEE Transactions on*, 14(9):1351–1359, 2005.