Analyzing GFP-tagged Cytoskeletal Protein Colocalization in Human Carcinoma Cells

Stephanie M. Reed

Submitted to the Department of Mechanical Engineering
In Partial Fulfillment of the Requirements for the
Degree of
Bachelor of Science
at the
Massachusetts Institute of Technology
June 2007

© 2007 Massachusetts Institute of Technology
All Rights Reserved

Signature of Author

Department of Mechanical Engineering
May 11, 2007

Certified by

Paul T. Matsudaira
Professor of Biology and Bioengineering
Director of W1-MIT BioImaging Center
Thesis Supervisor

Accepted by

John H. Lienhard V
Chairman, Undergraduate Thesis Committee
Analyzing GFP-tagged Cytoskeletal Protein Colocalization in Human Carcinoma Cells

Stephanie M. Reed

Submitted to the Department of Mechanical Engineering
on May 11, 2007, in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science in Mechanical Engineering

Abstract

Cytoskeletal proteins function as dynamic, complex components involved in cellular structure and signaling. Characterizing the roles of such proteins would greatly benefit many research areas, including the study of cancer and protein-related diseases. There is currently no accurate, high throughput method of image analysis that clearly describes protein behavior within the cell. In addressing this problem, we chose to characterize proteins based on the colocalization parameter—the amount of overlap between two objects or signals. We aimed to create a single parameter that quantitatively defined colocalization yet complemented biological intuition about a complicated system. Cell culture techniques were used to transfect HeLa cells with four “marker” GFP-tagged protein constructs. Cells were fluorescently labeled in three channels—Hoechst for nucleus, Texas Red phalloidin for actin, and GFP for protein—and images were captured using Cellomics scanning microscopy. After collecting data and testing software applications, we analyzed our data with Definiens software and developed a flexible, comprehensible method of quantifying colocalization using minimal parameters.

Thesis Supervisor: Paul T. Matsudaira

Title: Professor of Biology and Bioengineering, Director of WI-MIT BioImaging Center
Introduction

Despite recent advances in microscopy targeting cellular migration and protein function, very little is known about the coordinated system of structural and signaling proteins. A complete assessment of the roles of cytoskeletal proteins would be a great asset for further developing research in cancer and protein-related disorders. We aimed to characterize the functionality of proteins in cancer cells, the results of which could contribute to blocking carcinoma metastasis or oncogenic pathways. It is advantageous to understand cytoskeletal protein behavior, as changes in expression of cytoskeletal proteins contribute to proper cellular function. Describing the interactions between proteins is merely the first step laying out the groundwork for a much larger scale effort—technology with potential for significant advancement in a variety of research areas.

Current protein interaction maps are complicated, visually difficult to decipher, and provide little spatial or temporal information. As a major assay parameter, we chose to focus on fluorescent signal colocalization: a value measuring the amount of spatial overlap between two signals. Colocalization is routinely used to determine potential interactions of two or more proteins. Images are typically presented without numerical data, and regions where red and green overlap to create yellow are considered colocalized. We hoped to take this analysis beyond solely images and incorporate quantitative biological measurements that intuitively corresponded to images. Quantifying a single colocalization parameter is attractive so that a variety of proteins can be quickly and easily compared without bias. Determining the colocalization between cytoskeletal proteins and subcellular components—such as nucleus, cytoplasm, plasma membrane, and actin network—will help define and simplify the proteins’ operations and locations. Future applications include protein interactions in the three dimensional cellular space over time, potentially providing greater understanding of a protein’s dynamic localization within the cell as it migrates, divides, or differentiates. Fluorescent imaging techniques allow for different components of the cell to be stained so that imaging software can separate each color into its own channel. Data from the distinct channels can then be manipulated and related during analysis to quantify colocalization between the proteins and subcellular components.
However, defining the proteome with a single colocalization parameter comes with many complications. We originally tested three different methods of analysis: CellProfiler using Matlab correlation module, Imaris colocalization application, and Cellomics co-occurrence parameter. The values from each of these analyses varied significantly given control input images, both synthetic and biological. CellProfiler implemented Pearson’s Coefficient to describe colocalization, a single value varying from $+1$ (entirely colocalized) to $-1$ (entirely anti-localized) with zero indicating random localization. The drawback with Pearson’s Coefficient lies in the vagueness of the output value. It is unclear how, for example, 0.5 or 0.1 relate to a biological system. Imaris and Cellomics provided more comprehensible values, for example 60% of protein X colocalizes with the nucleus, and 20% of the nucleus colocalizes with protein X. Both values are required to fully define the protein from a biological, and not merely numerical, perspective. Yet there were still other problems that all three applications shared including difficulty performing batch analysis and overall poor computational performance. Unresolved, these issues presented a huge obstacle if none of the applications were able to handle large amounts of data. We aimed to address two main goals: 1. to accurately and consistently quantify colocalization between a variety of proteins and subcellular objects and 2. ensure that this computational analysis can be performed in a scalable fashion. It was clear that a new method of analysis was needed to meet these objectives.

A CellProfiler-based analysis process has been tested to discern between mitochondrial and non-mitochondrial proteins—whether or not various GFP-tagged proteins colocalize with mitochondria. However, no final colocalization values or coefficients were determined beyond the linear trends of the channel versus channel scatter plots. The selection and filter process was inconsistent, tedious, and possibly unreliable as it stemmed from only two parameters. Also, there was no method to compare levels of colocalization among the proteins. The goal of these mitochondrial measurements was to simply rank proteins, not numerically model a complex biological system, as is our goal. An underlying basis of comparison is imperative, especially if the relationship between protein and cellular region is not strictly linear. Often protein intensity will be drastically different than the subcellular components', and a nonlinear
(or even random) relationship exists. These trends must be easily recognizable and characterizable. Moreover, it is necessary to fully define the protein in terms of the components and also the components in terms of the protein, since the two relations differ. These definitions must be applicable to all cytoskeletal constructs, otherwise there is no foundation for analysis.

Even though an approach has been preliminarily implemented for GFP-tagged mitochondrial protein colocalization with mitochondria, it does not account for the complexity of GFP-tagged cytoskeletal protein colocalization with different areas of the cell. Depending on the type of cytoskeletal protein, image quality, and overall variation between cells, there is a large distribution in protein location, expanse of protein within the cell, fluorescent signal, cell viability, cell shape, and percentage of transfected cells. All of these factors make it difficult to create a standardized method of analysis. Further, current colocalization applications are inadequate for bulk analysis on large quantities of data, which will be generated by the magnitude of constructs (over 100 in our stock alone) and complexity of data gathered (multichannel 2D, 3D, 3D versus time). All of these problems needed to be improved before tackling any segment of the proteome characterization.

After extensive work with CellProfiler, Imaris, and Cellomics provided little progress, we tested Definiens Developer image analysis software in hopes of more conclusive and efficient colocalization results. We started protein colocalization analysis by first defining four basic constructs and controls: emptyGFP, vimentinGFP, actinGFP, tubulinGFP. For each construct, we selected representative cells using values of cell perimeter, area, aspect ratio, solidity, mean intensity, total intensity, and standard deviation intensity. Cells with parameter values outside preset thresholds were discarded prior to colocalization analysis. GFP-tagged proteins were compared with the various subcellular components by plotting the number of pixels in the protein channel against the number of pixels in the component channel. The constructs were further plotted using different combinations of parameters and clustered by their trend lines. In addition, the variation in GFP signal levels within a transfected population was used to determine potential expression-level dependent colocalization events.

Through the diverse tools available in Definiens, we expected to gather very
specific information at different levels within the cell that would resolve the
colocalization obstacles. Definiens handled the large amount of collected data yet still
analyzed the images at the pixel level. Images were segmented into small pixel clusters
as well as merged larger objects (nucleus, mitochondria, GFP, etc.), and myriad
parameters (intensity, distance to border, length to width ratio, etc.) were thresholded to
distinguish regions in the image. This ability to analyze images on both the clustered
pixel level and object level increased performance, provided insight, and improved
measurements that other software lacked. Definiens Developer’s flexibility allowed us to
adapt self-constructed equations and variables to any combination of objects. The visual
interface in Definiens was a critical improvement over Matlab-based analysis and
enhanced the prototyping process. In addition, Definiens has a client-server architecture,
unlike desktop-restricted applications such as Imaris. Our plan was to lay down the
colocalization groundwork in Definiens using a small sample set of the four control GFP
constructs, with hopes of creating a working method of analysis that can be easily scaled
to large quantities of data.

Methods
The following experimental procedures were each carefully optimized to yield the
best possible results. The overall order of methods followed the standard order widely
used in bioimaging: seeding and transfection of cells, fluorescent staining, two-
dimensional imaging, and analysis. Our goal was to maintain consistency with the
general process while improving upon techniques by either introducing new elements or
modifying existing protocols. The details of each step are described below.

Optimization of Transfection
The initial cell lines chosen to study included human cervical carcinoma cells
(HeLas), mouse fibroblasts (NIH3T3s), and rat fibroblasts (SRs). Src oncogene
transformed mouse fibroblasts (3T3-527s) were included as an additional cancerous line
and contained introduced podosomes. These cells lines were selected for their migratory
behavior, versatility, vitality, and familiarity. However, this procedure is easily adaptable
to other mammalian cell lines as well. Dulbecco’s Modified Eagle’s Medium (DMEM)
used to sustain the cells contained 10% Fetal Bovine Serum and 1% Penicillin-
Streptomycin antibiotics. Cells were cultured as normal—maintained at 37 C and 5% CO2, washed with phosphate buffer solution (PBS), and detached with trypsin. Cells were seeded into 96-well plates at 10,000 cells per well using standard cell suspension and hemocytometer counting techniques.

A common method of introducing GFP-tagged protein constructs into a cell is through transfection. Transfection is a cell culture technique in which DNA is mixed with a transfection reagent, and this cocktail is pipetted directly onto cells in media adhered to a plate. The transfection reagent alters the pores in the plasma membrane, allowing material to pass through more easily. The cells then take up the introduced DNA construct via endocytosis, the DNA is delivered to the nucleus, and the GFP-tag is transferred to the specific protein encoded by the construct. For this preliminary optimizing experiment, transfection was only tested on emptyGFP, a control construct that has a uniform distribution throughout the cell. EmptyGFP consisted of enhanced GFP under control of a CMV promoter.

Transfection was performed on the plated cells under varying conditions in order to determine the optimal protocol. Variables that we optimized included the amount of transfection reagent and the concentration of DNA. By varying these two conditions over a defined range, we found the best combination that produced the highest transfection efficiency. Transfection efficiency was measured by imaging the cells at each condition and comparing the number of transfected cells (colored green) to non-transfected cells (no color). We tested a cationic lipid reagent, Lipofectamine2000 (Invitrogen), and another lipid reagent, Fugene (Roche), by varying their percentage dilutions in PBS between 1% and 2%. DNA concentrations of emptyGFP ranged from 8ng/ul, 10ng/ul, to 12ng/ul. Combinations of these two parameters were tested in separate wells of a 96-well plate. The graph below (Figure 1) illustrates how transfection efficiency fluctuated with the ratio of DNA concentration to percentage of lipid transfection reagent.
Figure 1 The top graph displays the efficiencies of Lipofectamine200 (LF2K) and Fugene at various percentages of lipid transfection reagent and concentrations of DNA. The transfection efficiency is plotted against the ratio of DNA concentration to percent lipid.

The highest efficiency, at 8ng/ul DNA with 2% Lipofectamine2000 solution, showed significant cell death in images. Thus the second highest efficiency, 10ng/ul with 2% Lipofectamine2000, maintained healthy cells and was chosen for optimization. The striking increase in cell health was worth the tradeoff for the small decrease in transfection efficiency. Figure 2 illustrates the drastic difference in cell vitality between the top two most efficient set of conditions.
Figure 2 The image on the left shows a field of sickly or dead SR cells at 8ng/ul DNA and 2% LF2K. Healthy cells were scarce throughout the entire well at these conditions. The image on the right shows a field of healthy SR cells at 10ng/ul and 2% LF2K. The entire well exhibited healthy, living cells comparable to those in this field. Both images were taken at 20X with Cellomics Target Activation.

Further proof of the discrepancy in cell vitality is shown by the total cell count in each well. Figure 3 illustrates the number of cells that survived the transfection process, starting at 100,000 cells per well prior to transfection, fixing, and staining. Lipofectamine2000 at 2% solution coupled with 10ng/ul DNA yielded ~25% transfection of all cell lines while retaining nearly 40% cell vitality. Lipofectamine at 2% solution and 8ng/ul DNA produced ~26% transfection for all cell lines but exhibited much greater toxicity with only~8-23% cell vitality.
Helas, 3T3s, 527s, and SRs all demonstrated the same transfection behavior; the best yield occurred at a 5:1 ratio of DNA concentration to transfection reagent. Fugene transfection reagent returned significantly lower transfection results and considerable cell death, so we did not continue working with Fugene. Another method of introducing GFP-tagged constructs takes advantage of viruses' natural ability to infect cells and bring in anything appended to them. However, the time overhead in prepping the samples is
costly and undesirable, and the automated alternative, a robot-controlled device, would require purchase, laboratory space, and setup.

Once the optimal parameters were determined, the bulk of the experiment characterizing colocalization proceeded. Using the said Lipofectamine solution and DNA concentration, transfection of HeLa cells was performed with a collection of 23 common GFP-tagged cytoskeletal constructs. HeLa cells were chosen over other lines because they are a routinely used, highly studied cancerous cell line. According to the 96-well plate layout in Figure 4, one protein construct was transfected in each well, with four duplicate wells to ensure accuracy in data analysis and improve statistical significance.

*Optimization of Fluorescent Staining*

After the media change 6 hours post transfection, the 96-well plate was prepared for fluorescent staining. Cells were fixed with 3.7% formaldehyde for 30 minutes, washed with PBS, and stained for an hour. We established a protocol for fluorescent staining to minimize bleedthrough between channels and maximize signal. Optimization included four channels to allow for the potential of a far red tubulin (or other protein) stain. Fluorescent dyes including Hoechst33342 staining the nucleus, 568 maleimide staining the whole cell, and AlexaFluor 647 phalloidin staining the actin, comprised three channels blue, red, and far red, respectively. The green channel where GFP usually presents was left empty, as there was no need to transfect cells for this experiment.

Different concentrations of maleimide and phalloidin dyes were tested to determine which provided the strongest signal with the least bleedthrough into other channels of surrounding wavelengths. Maleimide dilutions ranged from 1:1000 to 1:5000 in methanol at 1.135 uM, while phalloidin dilutions were varied between 1:25, 1:50, 1:100, and 1:200 in methanol at 0.5128 uM. Hoechst staining was kept constant at the manufacturer-specified dilution 1:1000. Fixed 527s were stained for an hour with all dyes simultaneously. Sample images from this four channel staining are shown in Figure 4 below. All images were collected on the Widefield 200M microscope at 63x magnification. Channel 1 displayed Hoechst 33342 (shown in blue) at BP475/40 with a 50 ms exposure time. Channel 2, exposed for 100 ms, was empty and would have contained GFP in actual experimental setting with protein constructs. Channel 3 showed
568 maleimide (shown in green) at BP570/60 with a 100 ms exposure. Channel 4 contained 647 phalloidin (shown in red) at BP670/50 with a 200 ms exposure. The 1:25 dilution of 647 phalloidin emitted the brightest and clearest fluorescence, and the 1:5000 dilution of 568 maleimide resulted in the least bleedthrough while maintaining sufficient intensity.
## Controls

<table>
<thead>
<tr>
<th></th>
<th>Hoechst 33342</th>
<th>Hoechst 33342</th>
<th>Hoechst 33342</th>
<th>Hoechst 33342</th>
<th>Hoechst 33342</th>
<th>Hoechst 33342</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1:1000</td>
<td>1:1000</td>
<td>1:1000</td>
<td>1:1000</td>
<td>1:1000</td>
<td>1:1000</td>
</tr>
<tr>
<td>647 Phalloidin</td>
<td>1:25</td>
<td>1:50</td>
<td>1:100</td>
<td>1:200</td>
<td>568 Maleimide</td>
<td>568 Maleimide</td>
</tr>
<tr>
<td>1:10000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 4** Representative images of channels 2, 3, and 4 during four channel staining. The top row of images contains channels 1, 3, and 4 overlaid. Specifically, bleedthrough from maleimide and phalloidin in channels 3 and 4 was studied. Controls are in the top matrix, while variations of dilutions are in the bottom matrix. Images were taken at 63x on the Widefield.
Histograms of the above images reveal the intensity values of all the pixels in each image. The ranges help explain the relationship of the fluorescent objects to the image background as well as provide a quantitative basis for comparison between images. Histograms are shown below in Figure 5.
Figure 2 Histograms of each image showing the intensity value of every pixel. The top matrix consists of controls, while the bottom matrix contains various dilutions of maleimide and phalloidin.
Further staining experiments were performed to find the bottom limit of the 568 maleimide; the less volume of dye used, the less bleedthrough into other channels, which may skew data analysis. 568 maleimide was diluted to 1:5000 and 1:25000, while 647 phalloidin was diluted to 1:25 and 1:50. Because the whole cell stain may have interfered with the actin staining, we adapted a consecutive staining protocol and compared it to the previously used simultaneous staining. 527s were stained with Hoechst and phalloidin first for one hour, and after aspiration, maleimide stain was applied for one hour. The same imaging techniques were used as previously stated. Figure 6 illustrates the difference between simultaneous and consecutive staining. The optimally balanced conditions with the least bleedthrough and strongest uniform signal resulted from consecutive staining of 1:25 647 phalloidin and 1:5000 568 maleimide. Respective histograms for the images in Figure 6 are seen below in Figure 7.
Representative images of channels 2, 3, and 4 during four channel staining. The top row of images contains channels 1, 3, and 4 overlaid. Specifically, bleedthrough from maleimide and phalloidin in channels 3 and 4 was studied for both simultaneous and consecutive staining. Controls are in the top matrix, while variations of dilutions are in the bottom matrix. Images were taken at 63x on the Widefield.
Figure 7 Histograms of each image showing the intensity value of every pixel. The top matrix consists of controls, while the bottom matrix contains various dilutions of maleimide and phalloidin for simultaneous and consecutive staining procedures.
Next, we included the 23 common cytoskeletal protein constructs and stained a 96-well plate of transfected HeLa cells. Despite optimizing fluorescent staining for four channels, we opted to begin with three-channel staining to simplify analysis and ascertain the behavior of the system without additional unnecessary complications. The procedure for three channel staining was already established: 1:1000 Hoechst 33342 nuclear stain in channel 1 (blue), GFP protein constructs in channel 2 (green), and 1:200 Texas Red phalloidin actin stain in channel 2 (red). Staining was simultaneously performed for one hour, followed by washing and image analysis.

**Imaging**

Cellomics vHCS:Scan V Target Activation application was used to scan the 96-well plate. The XF93 filter set was chosen to capture the Hoechst, GFP, and Texas Red channels. Snapshots within each well were taken at 20x magnification, totaling 25 fields per well. After manipulating the application's threshold for cell detection, exposure time was set to 0.02 s for all channels. Data from the scanned plate was exported as 16 bit TIF image files. The total number of files collected was substantial—2,400 for one 96-well plate—so we reduced the data size by two, excluding half of the duplicate wells for each construct, easing data handling issues and decreased export time.

**Analysis**

Imaris

The exported tifs were read directly into Imaris, a powerful image analysis program that efficiently handles 2D, 3D, and 4D data. We attempted to use Imaris' Coloc feature in hopes of finding an existing method of colocalization analysis. Imaris masked regions of the cell based on an intensity threshold in the corresponding channel. Nuclei were identified with the Hoechst stain, cytoskeletal proteins were identified with GFP, and cells were identified using the Texas Red actin stain. The output provided values for parameters such as the channel correlation, % region of interest (ROI) colocalized, % volume A above threshold colocalized, % volume B above threshold colocalized, % ROI material A colocalized, and % ROI material B colocalized. These initially seemed like informative parameters that could contribute to colocalization analysis. However, the numbers Imaris generated did not make sense in a biological context. The Coloc feature did not allow cell-based analysis or bulk analysis, nor could it process more than one image at a time.
Cellomics

Analysis using Cellomics vHCS:View software was tested, eliminating the need to export the data to tif image files. Nuclei, cells, and GFP spots were identified during analysis with an isodata thresholding factor in each channel. Cellomics offered a wide range of parameters on the well, field, and cell levels, but the % co-occurrence feature yielded seemingly random values. Co-occurrence percentages for similar cells in the same well were inconsistent, and the equation for this parameter was inadequately defined. Also, Cellomics could not analyze colocalization on a subcellular level, making it nearly impossible to adequately define protein localization within the cell.

CellProfiler

We attempted to create a Matlab-based colocalization script using CellProfiler modules. CellProfiler analyzed entire batches of data at a time, but the processing time was extremely inefficient. CellProfiler claimed to allow flexibility, advertising that written Matlab scripts could be run using CellProfiler. However, while we weren’t limited computationally in Matlab, the CellProfiler user interface was poorly constructed, and it was difficult to bridge compatible code between Matlab and CellProfiler. The correlation module internal to CellProfiler used Pearson’s correlation coefficient to relate two channels. The correlation coefficient between red and green channels is shown below in Equation 1.

\[ r_p = \frac{\sum (R_i - R_{avg})(G_i - G_{avg})}{\sqrt{\sum (R_i - R_{avg})^2 \sum (G_i - G_{avg})^2}} \]  

(Eq. 1)

The total intensity for all red and green pixels is given by \( R_i \) and \( G_i \), respectively. The average intensity values were subtracted from the total intensity to normalize the contributions from both signals. CellProfiler provided correlations between blue, green, and red channels within different cellular objects. Correlation coefficient values varied from -1 (anti-colocalized) to +1 (colocalized), where 0 corresponded to random colocalization. This coefficient range offered very little information about the biological behavior of the system, and distinguishing proteins based on 0.1 or 0.2 seemed vague.

It became clear that we needed to compare software by applying all analyses on the same image sets. We secluded images of single cells transfected with three control constructs—emptyGFP, actinGFP, and tubulinGFP—along with a zero transfection control. Images were captured using Cellomics vHCS:Scan at 20x magnification and high resolution setting with 1x1 binning. These images were analyzed in Imaris (% ROI
colocazlied), CellProfiler (Pearson’s coefficient), and Cellomics (% co-occurrence) with their respective colocalization applications. Figure 8 illustrates the control images and the theoretical colocalization results for each method of analysis. The expected correlation coefficients for the red and green channel were 0.5, 1.0, -1.0, and 0 for emptyGFP, actinGFP, tubulinGFP, and no GFP, respectively. EmptyGFP is diffuse throughout the cell, so there should be some colocalization with actin. ActinGFP should entirely colocalize with actin in the cell, whereas tubulinGFP should be anti-colocalized with actin. The non-transfected sample should have zero colocalization.

Figure 8  Expected correlation coefficient values for control constructs. Single cell images were captured on Cellomics vHCS:Scan at 20x and cropped. Texas Red phalloidin and GFP channels were theoretically compared. Our experimental method involved analyzing the red versus green and blue versus green signals in all regions: actin, nucleus, whole cell, and entire image. The correlation coefficient was calculated between the red and green channels and blue and green channels in each location. Figure 9 displays the actual experimental results for each method of analysis in a variety of cellular regions.
Figure 9  Experimental correlation coefficient values for control constructs. Single cell images were captured on Cellomics vHCS:Scan at 20x and cropped. Texas Red phalloidin, Hoechst, and GFP channels were compared at a variety of locations within the image.

The single cell images showed unclear results, and a less complicated set of images was needed to produce straightforward, comparable values. We decided to create artificial data resembling the same spatial distribution that proteins have within cells. Simple synthetic data was tested using Imaris and CellProfiler, as seen in Figures 10 and 11. Red, green, and blue squares were arranged and overlapped to simulate protein colocalization with actin and the nucleus. Green and red overlapped to produce yellow, and green and blue overlapped to produce aqua. Still, analysis demonstrated glaring errors, mostly in the anti-colocalized area. It was difficult to trust Imaris and CellProfiler because of the spread in data and lack of intuitive, palpable values. Defining colocalization using a ratio of intensities, without regard to the location or area of the pixels, did not fully describe what was happening in the biological system. Additional factors other than intensity contributed to colocalization, and we hoped to specify the minimum parameters required.
Considering these inaccurate results along with all of the drawbacks of each application, we found Cellomics, Imaris, and CellProfiler to be inadequate methods of analysis. We moved on to Definiens, a software package that analyzed images at both the pixel and object levels. The major benefit of Definiens was that it provided the same functionality as Matlab but had a pre-configured set of tools to eliminate complicated code. Using Definiens Developer, we could create a rule set to identify objects that involved self-written equations, loops, and self-adapting variables.
We initially focused on the material parameter to quantify colocalization. Material encompassed both the area and intensity of an object and served as a common method of comparison between protein constructs.

\[
material = \frac{\text{area}}{\text{intensity}} \quad \text{(Eq. 2)}
\]

It was useful to know the location of the GFP objects within the cell, so we defined a parameter for relative distance within the cell (Equation 3). Relative distance, \(d\), depended on the distance to the nucleus, \(r_{\text{nuc}}\), and the distance to the cell border, \(r_{\text{cell border}}\).

\[
d = \frac{r_{\text{nuc}}}{r_{\text{nuc}} + r_{\text{cell border}}} \quad \text{(Eq. 3)}
\]

First images were divided into clusters of pixels using Definiens multiresolution segmentation based on color and shape. Second, nuclei and GFP spots were identified by setting an intensity threshold in the corresponding channel, and pixel clusters were merged to form the objects. Cells were distinguished using nuclei as seeds and building cell area radially outward until the background intensity threshold was reached. The cytoplasm was defined as the region difference between the cell and nucleus. The nucleus and cell border were defined as well. Definiens provided the ability to set up multiple levels in a hierarchy and specify certain objects to the desired level. In our ruleset, cells were on the highest level, with nuclei, cytoplasm, and GFP spots as subobjects on the lower organelle level. Figure 12 illustrates this hierarchy as well as the Definiens segmentation and merging process.
Figure 12 Matrix showing the steps of Definiens analysis for each of four cytoskeletal constructs: emptyGFP, vimentinGFP, actinGFP, and tubulinGFP. Raw images were loaded in and divided using multiresolution segmentation. Thresholds were set to isolate specific objects, whose segmented clusters were merged to form one solid body. Nuclei (green), cytoplasm (yellow), and GFP (orange) were all on the organelle level below the cell level, making them subobjects within the cell (aqua). The red outlined cell in each field identifies the segmented cell, shown at 500% zoom. The top left image measures 340x340 um.
We used multiresolution segmentation again on GFP objects alone. We defined ranges of intensities within the whole GFP object, and each bracket of GFP intensity was assigned a class. This allowed us to look at low intensities with respect to area, or high intensities with respect to location in the cell, for example. Figure 13 is an image from Definiens visual interface showing the GFP classes, ranging from dark red to light pink as intensity increases, overlaid on the Texas Red phalloidin stained channel. The division of the GFP group into smaller sections allowed for more diverse and informative analysis. Data collected in Definiens was exported as spreadsheets and handled in Matlab and Excel for plotting and calculations. Using a series of loops, we wrote a Matlab script to read in all 400 files, separating variables based on the construct and parameter of interest.

Figure 13 Image from Definiens Developer with GFP objects overlaid on the actin Texas Red phalloidin channel. HeLa cells were transfected with emptyGFP. GFP objects change from dark red to light pink as their intensity increases. Multiresolution segmentation shows the pixel clusters used to form larger objects. The image measures 340x340 um.

Results

We chose a few key parameters to describe the protein’s intensity and location with respect to its surrounding cellular objects. These values were averaged over both duplicate wells, all 25 fields in each well, with around 140 cells per field. Thus the number of cells sampled during analysis was approximately 7,000 for each protein construct, and this large set of data contributed to precise medians with minimal deviations. We determined the intensity, area, and material for each construct in order to compare both the concentration and distribution in parallel. The material put the data in perspective by providing a density-like parameter. The area of each compartment—nucleus, whole cell, and GFP—
was measured in pixels, and the intensity of each compartment was measured in light intensity units (LIU). Average values for intensity, area, and material are shown below in Table 1 for each control construct. Uncertainty values were determined using the t-statistic method and propagation of error. Looking at Table 1, the material of the GFP constructs was much lower than the corresponding cell material because of the cell's expansive area and low intensity actin stain. Certain GFP constructs were brighter and more concentrated than others, as shown by materials values less than 1.

**Table 1** Average material, area, and intensity values per cell for emptyGFP, vimentinGFP, actinGFP, and tubulinGFP with precision uncertainty. Average intensity included that of the green channel for GFP objects, red channel for cell, and blue channel for nucleus.

<table>
<thead>
<tr>
<th>Material [pixel/LIU]</th>
<th>Area [pixel]</th>
<th>Intensity [LIU]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleus 2.2067±0.0294</td>
<td>1232.42</td>
<td>558.50</td>
</tr>
<tr>
<td>Cell 51.1423±0.0253</td>
<td>6447.38</td>
<td>126.07</td>
</tr>
<tr>
<td>GFP 1.6412±0.0250</td>
<td>1817.70</td>
<td>1107.56</td>
</tr>
<tr>
<td>Nucleus 2.5070±0.0232</td>
<td>1429.59</td>
<td>570.25</td>
</tr>
<tr>
<td>Cell 54.9913±0.055</td>
<td>6362.83</td>
<td>115.71</td>
</tr>
<tr>
<td>GFP 0.3241±0.0326</td>
<td>372.71</td>
<td>122.70</td>
</tr>
<tr>
<td>Nucleus 2.5164±0.0186</td>
<td>1427.43</td>
<td>567.24</td>
</tr>
<tr>
<td>Cell 52.1136±0.0820</td>
<td>6394.30</td>
<td>122.70</td>
</tr>
<tr>
<td>GFP 0.1127±0.0417</td>
<td>120.81</td>
<td>1072.27</td>
</tr>
<tr>
<td>Nucleus 2.5075±0.0180</td>
<td>408.75</td>
<td>163.01</td>
</tr>
<tr>
<td>Cell 54.313±0.0107</td>
<td>6464.19</td>
<td>119.02</td>
</tr>
<tr>
<td>GFP 0.0975±0.0104</td>
<td>100.89</td>
<td>1035.05</td>
</tr>
</tbody>
</table>

In an attempt to further define the GFP-tagged proteins, we took advantage of the wide range of intensities exhibited within each GFP object and divided the total object into distinct subgroups. Each GFP subgroup was specific to a range of intensity values, and the resulting cascade of intensities is shown below in Figure 14, along with the results from Table 1. The intensity thresholds were the same across all constructs, as seen by the consistency at each GFP subgroup (GFPO through GFP6). By dividing the whole GFP spot into smaller sections, we were able to see if a protein existed at many intensities through the cell, one average intensity for the whole object, or a very high intensity at one location with a very low intensity at another. We hoped to capture this intensity distribution of the cytoskeletal proteins and apply it to characterizing their behavior.
Intensity of Objects for GFP Constructs

Figure 14  Average intensity per cell for each designated object, including the array of intensity-thresholded GFP subgroups. Intensity includes that of the green channel for GFP objects, red channel for cell, and blue channel for nucleus.

Next, we spatially determined the location the GFP spot in the cell. Relative distance in cell was a parameter that defined GFP location with respect to the nucleus and cell edge borders. An object on the nuclear border would have a relative distance approaching zero, while an object on the cell periphery would have a relative distance approaching 1. The nucleus itself did not display a value of zero since the majority of the pixels do not lie on the nucleus border. Figure 15 shows how emptyGFP increased with intensity as it approached the nucleus. ActinGFP also displayed this trend. VimentinGFP showed a less dramatic intensity increase as the protein’s location drew closer to the nucleus. However tubulinGFP remained level throughout most of the cell with a sharp drop off in intensity toward the cell edge.
Relative Distance in Cell for Objects

Figure 15 Relative distance of objects in cellular space from the nucleus.

By plotting the relative distance against the intensity ranges of the GFP subgroups, we observed trends in the GFP breakdown for different cytoskeletal constructs (Figure 16).

Looking at the intensity of a compartment in one channel with respect to the intensity of the cell in that same channel, we determined the percent an object occupies within the cell for that channel. Using intensity values alone did not capture the full system behavior; the product of intensity and area gave a volume-like parameter that
completely described the amount of signal emitted for all objects in all channels. For each fluorescent stain—Hoechst, Texas Red phalloidin, GFP—a specific percentage was found in each cellular compartment. Similarly based on the directionality of colocalization, for each compartment—nucleus, cell, protein—a specific percentage contained each stain. These two values were not necessarily the same, and herein lies the complexity of colocalization. The four plots in Figure 17 show colocalization values for all stains, compartments, and GFP constructs. Solid bars designate the stains with respect to the compartments, specifically what percent of the stain is in each compartment. Diagonal patterned bars designate the compartments with respect to the stains, specifically how much of the compartment contains each stain. TubulinGFP and actinGFP showed 100% colocalization of GFP in the nuclear region. In reality neither of these proteins exists in the nucleus, but the 2D nature of the images superimposes the actin and tubulin above and below the nucleus to look as if it were inside. Surprisingly, vimentinGFP showed 0% of GFP within the nucleus. This can be attributed to the very high intensity of the vimentinGFP spots around the nucleus, so any GFP within the nuclear region was not detected.
Figure 17 Plots for the four control constructs showing 1) where the fluorescent stains colocalize with the cellular compartments (given by solid bars) and 2) how the cellular compartments colocalize with the fluorescent stains (given by diagonal patterned bars).
Table 2 below shows a simplified chart of these colocalization values for the control constructs. The rows, as denoted on the left, are the three stains, while the columns, as denoted at the top, are the three compartments. Complete colocalization (100% from both angles) only existed between a fluorescent dye and the object it was intended to stain. For example, 100% of Hoechst was found in the nucleus, and 100% of the nucleus contained Hoechst. These pairings with complete colocalization are shown as diagonals in Table 2. For pairings with directional colocalization, the lower left value corresponds to the percentage of stain in that compartment, whereas the upper right value corresponds to the percentage of the compartment containing that stain. It can be discerned that 24.55% of emptyGFP was located in the nucleus, and 100% of the nucleus contained emptyGFP. Texas Red phalloidin colocalization with the nucleus should have been relatively constant despite the protein construct analyzed, but the values varied from 12.9% to 29.86% which can be attributed to inconsistency in nuclear area. Variation in the size of the nucleus was unavoidable when averaging ~7,000 cells per construct, but the protein constructs also may have affected the size and shape of the transfected nucleus depending on the toxicity of the of the construct.

Table 2 Colocalization values for Hoechst, Texas Red phalloidin, and GFP fluorescent stains versus nucleus, cell, and protein compartments. The constructs are colored to match previous bar graphs. For directional colocalization, lower left values indicate the amount of stain in the compartment. Upper right values indicate the amount of the compartment that possesses stain.
Colocalization values were ranked in a heat map in Table 3. The highest values with 100% colocalization were assigned bold colors, and decreasing colocalization values were assigned proportionally lighter colors. We can compare constructs and conclude trends quickly with this illustrative device.

Table 3 Heat map demonstrating trends in colocalization values. Bold colors indicate 100% colocalization, while decreasing hues indicate decreasing colocalization.

<table>
<thead>
<tr>
<th></th>
<th>% Nucleus</th>
<th>% Cell</th>
<th>% Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>emptyGF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Hoechst</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% TR Phalloidin</td>
<td>12.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% GFP</td>
<td>24.58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HIMGF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Hoechst</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% TR Phalloidin</td>
<td>29.86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% GFP</td>
<td>25.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>edonGF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Hoechst</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% TR Phalloidin</td>
<td>15.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% GFP</td>
<td>16.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tulinGF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Hoechst</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% TR Phalloidin</td>
<td>20.45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% GFP</td>
<td>32.05</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Discussion**

This method of analysis, while only tested on four basic constructs, lends itself to much more complicated analysis of innumerable protein constructs, not merely cytoskeletal proteins. Having accurately defined a single parameter to characterize colocalization, as well as other parameters that help describe the system, our experimental methods and analysis can be altered to target specific needs. To continue work in characterizing cytoskeletal proteins and determining protein interactions, additional compartments could easily be incorporated into the Definiens ruleset. It would be useful to add a compartment at the nucleus-cytoplasm border to determine what proteins are perinuclear or to categorize other subcellular components such as the Golgi apparatus. Similarly, adding a compartment at the cell periphery to characterize colocalization with the plasma membrane would be very beneficial as well. Any additional compartments like these would further describe the biological system and allow more flexibility in analysis. Another step to take before proceeding would be to create an actin compartment on the
organelle level. Currently the cell compartment exists on the cell level, which is higher than the organelle level so that cells are superobjects to all organelle objects. Texas Red phalloidin staining actin was used to create the cell mask, so a higher threshold would need to be set to only capture bright or dense regions of actin and exclude the empty spaces in between actin fibers.

In general, better organization of the objects and levels would have made data analysis easier to understand. Creating GFP objects that both included and excluded the nuclear region would have made the comparison of areas simpler. Also, having GFP objects on a separate level would have eased calculations such as intensity overlap and area overlap of GFP with other compartments. GFP could be divided using thresholds for other parameters besides intensity, such as the relative distance in cell, brightness, area, length to width ratio, main direction, etc. Analyzing GFP sections instead of GFP as a lump object will provide refined colocalization data that reflects intensity, relative distance in the cell, or any parameter used to threshold the classes.

After refining the Definiens ruleset, we next plan to analyze the set of 100+ cytoskeletal proteins, which have already been transfected into HeLa cells and imaged using Cellomics. This preliminary work is useful for determining trends and classes of proteins, and we hope to group our set of constructs into categories based on their colocalization measurements. Certain constructs might be signaling proteins while others might be structural, so making this distinction (among others) is critical for characterizing protein interactions. The potential of this analysis platform is far-reaching even beyond proteome characterization, as applications to computational biology and engineering are enhanced by its highly adaptable, quantitative yet intelligible output.