Indirect learning of generative models for microtubule distribution from fluorescence microscope images

Aabid Shariff AABID@CMU.EDU
Gustavo K. Rohde GUSTAVOR@CMU.EDU
Robert F. Murphy MURPHY@CMU.EDU
Lane Center for Computational Biology, Carnegie Mellon University and Joint Carnegie Mellon University - University of Pittsburgh Ph.D. Program in Computational Biology, 4400 Fifth Ave, Pittsburgh, PA 15213, USA

Abstract

A detailed model of the components of the protein filament networks that constitute the cytoskeleton, and the ways in which these vary from cell type to cell type and under different conditions, will be necessary for systems biology efforts to understand complex cell behaviors. In this work we focus on extracting quantitative information related to microtubule networks automatically from fluorescence microscope images. While previous approaches have focused on direct estimation methods (tracing and tracking the position of single microtubules in time series images), these require specialized microscopy and are difficult to apply on a proteome scale. We describe an indirect method based on comparing features computed from images simulated with a generative model with those computed from real images. We show that the extraction of important parameters, such as the approximate number of microtubules, length distribution of microtubules, can be estimated from fluorescence microscope images without having to explicitly trace the position of each microtubule.

1. Background and motivation

Systems biology seeks to understand the structure and function of living systems through the application of engineering principles to model the large network of interacting molecules comprising complex biological assemblies. Quantitative analysis and modeling efforts have been described at various spatial scales and biological images can be a rich source of information for use in building cell simulations. An especially powerful illustration of this approach is the work of the Danuser group on inferring models of actin cytoskeleton dynamics directly from fluorescence microscope images (Ponti et al 2005). A complementary approach is our recent description of algorithms for building generative models of cell, nuclear and organelle structure from high-resolution confocal microscope images (Zhao & Murphy 2007). These models capture the variation in subcellular patterns across cell (image) populations, and can be used to generate new images that can be thought of as being derived from the same underlying population as the images used to train them. These models have been successfully developed for organelles consisting primarily of discrete objects (vesicles) that can learned directly from images, but we have not previously considered the learning of generative models for extensively connected networks (such as the cytoskeleton).

We present a method for constructing generative models of microtubule networks and indirectly estimating model parameters by comparison of synthesized images with real images obtained by fluorescence microscopy. Indirect estimation of tubulin distribution in the yeast spindle has previously been described by Pearson et al (2006).

2. Methods

2.1 Images

We used the collection of 8-bit 3D images of HeLa cells obtained previously by three-color confocal immunofluorescence microscopy (Velliste & Murphy 2002). Nuclear and cell membrane boundaries were segmented using an active contour approach (Chan and Vese, 2001).

2.2 Modeling approach

Our aim is to capture information related to the spatial organization and distribution of microtubules in a cell image or population of images. High resolution images capable of distinguishing individual microtubules and at the same time containing an entire cell in the field of view are difficult to obtain, and therefore researchers often settle for observing microtubules, and their dynamical properties, in a limited region of interest within the cell (Dorn et al 2005, Sargin et all 2007). Thus, information related to the global structural organization of microtubules (e.g., total number of microtubules, length

distribution, etc.) is often unavailable from such studies. We aim to circumvent this difficulty by developing a stochastic model for microtubules emanating from the cell's centrosome, generating images from such models, and optimizing for images obtained from our models that compare with real full cell images (collected at lower resolution than would typically be used for individual tubule tracking). The approach consists of the following modules: (1) generative growth modeling, (2) deriving an image from a model, (3) library generation, and (4) parameter search by image retrieval (optimization).

2.2.1 GENERATIVE GROWTH MODELING

Our modeling approach creates a three-dimensional microtubule distribution given the positions of points defining the boundaries of the nuclear and cell membrane and a point X_0 defining the position of the centrosome. These are determined for the specific tubulin image whose model parameters are to be estimated. The centrosome location is determined by convolving the tubulin image with an averaging filter and choosing the maximum value. The modeling approach consists of "growing" microtubules in a random fashion, using the centrosome as a starting point. Assuming the centrosome to be a sphere, we fix the diameter of the centrosomal structure to be approximately 0.4 um. We model the microtubules from the centrosome by randomly picking starting points within the spherical volume and elongating them. Next, a uniformly distributed random direction is chosen, and the microtubule is grown in that direction with some length γ, which we denote as the stepsize. We denote the new point X_1. Given the two initial points X_0, and X_1, the subsequent point X_2 is chosen at random, using a uniform distribution, but constrained as follows:

$$0 < v_1 \bullet v_2 \le \cos\alpha \quad (1)$$

$$\text{with} \quad v_1 = \frac{X_1 - X_0}{\|X_1 - X_0\|}, v_2 = \frac{X_2 - X_1}{\|X_2 - X_1\|} \qquad (2)$$

and where α is the angle between $X_2 - X_1$ and $X_1 - X_0$.

Microtubules inside a cell vary in length. We model the length distribution as a normal distribution truncated such that there can be no negative lengths. This distribution is sampled N times, where N is the number of microtubules. The microtubule elongation procedure is iterated for each of *N* microtubules, until the sampled length of the microtubule polymer is satisfied. The model therefore has four parameters: number of microtubules, *n;* cosine of angle, *cos α*; mean of the normal distribution, μ; and standard deviation of the normal distribution, σ. The cosine of angle α determines the flexibility of each microtubule in the model. The points in each iteration are generated to satisfy the constraint in (1). Naturally, the growth process described above must be constrained within the confines of the cytoplasm in the cell. For this we make use of the segmented cell and nuclear boundaries of the given cell image. The growth model is constrained so that no point is chosen inside of the nucleus or outside of the cell.

2.2.2 DERIVING AN IMAGE FROM A MODEL

The parameters for the growth model described above can be estimated by choosing them so that they produce images that best match given real images of microtubules. To that end we now describe a method for estimating what a measured digital image would "look like" for a given fluorescence pattern emitted by a collection of microtubule tracks. In our methodology we model the image formation process of blurring due to the (incoherent) point spread function (PSF) of the imaging system. We accomplish this by convolving the model F with a 3D Gaussian G.

$$\mathrm{Im}age(n_1,n_2,n_3) = \sum_X \sum_Y \sum_Z F(k_1,k_2,k_3)G(n_1-k_1,n_2-k_2,n_3-k_3)$$

The parameters of the Gaussian were manually estimated for the HeLa dataset. Henceforth, we will refer to the images in the 3d HeLa dataset as "real images" and the images generated by the model as "synthetic images".

2.2.3 LIBRARY GENERATION

We generated a large library of synthetic images to implement an exhaustive search strategy for optimization by varying model parameters to retrieve an image that matches in content with the query image (3d HeLa image). The parameters varied took the following values:

n = 5, 25, 50, 75, 100, 125, 150, 175, 200, 250, 300, 350, 400

μ = 5, 25, 50, 75, 100, 125, 150 *microns*

σ = 1, 5, 10, 15, 20, 25 *microns*

cos α = 0.9, 0.95, 0.98

For a given cell morphology, a total of 1638 images were generated. Each of these images was generated with a different random number generator seed.

2.2.4 CONTENT BASED IMAGE RETRIEVAL

Given a query image, we would like to find an image from our library that is closest in content to it. This problem is similar to the traditional computer vision problem of content-based image retrieval. Thus we will motivate the set of image features used for content comparison and the distance metric for matching. To compare the real and synthetic microtubule distributions, we calculated thirteen 3D Haralick texture features for each image using the procedure described previously (Chen et al 2003). These features were also computed for images downsampled by a factor of two. Radial intensity features were computed as the total intensity in a discretized radial volume starting from the centrosome. Histogram features and the total intensity were also computed. A diagonal matrix D was computed that contain the variances of the features. This variance matrix

is then used to compute the Normalized Euclidean distance between a feature vector x_s computed from a set of simulated microtubules (synthetic image) and a feature vector x_r corresponding to the image based on which the microtubule simulation was computed (real image).

For any query image, we compute the Normalized Euclidean distances from it to each of synthetic images in the library. And find the parameters that minimize this distance.

3. Results

A sample image from the 3D HeLa cell dataset described above is shown in Figure 1. A sum projection in the *x-y* plane is shown. Although the image acquisition procedure is capable of providing information about the global distribution of microtubules in this particular cell, individual microtubules are not easily distinguishable, eliminating the possibility of direct estimation of microtubule positions, lengths, etc.

3.1 Generated Images

The indirect method is based on a generative modeling and simulation approach whereby microtubules are grown from the centrosome location using the method described above. In addition to the position of the centrosome, the boundaries of the nucleus as well as the cell membrane are needed to constrain the simulated growth of the microtubules. A three-dimensional rendering of the segmented nucleus and simulated microtubule tracks is shown in Figure 2. As described above, models such as that shown in Figure 2 can be converted to an image with similar voxel resolution as the original image. An example of the 2D projection of the image for the model shown in Figure 2 is shown in Figure 3.

3.2 Sensitivity Analysis

Validation of the model can be a tricky problem to solve when the ground truth is not known. We describe here a method to test the robustness of the approach by using a synthetic image as the query image instead of a real image. An image library was created for various combinations of parameters, and 400 parameter combinations were chosen at random as query images. These images were generated with a different random number generator seed as that of the images in the image library. The formal goal of this brute force search was: For each of 400 query images, output parameters that minimizes the Normalized Euclidean distance between it and the images in the library, and compute a measure of the error between the known parameters of the query and the parameters of the best match from the library. The error metric used was the absolute percentage error (APE).

$$APE = \frac{1}{400}\sum_{t=1}^{400}\left|\frac{R_t - S_t}{R_t}\right|$$

where R_t , S_t are the parameters for the query and the estimate image.

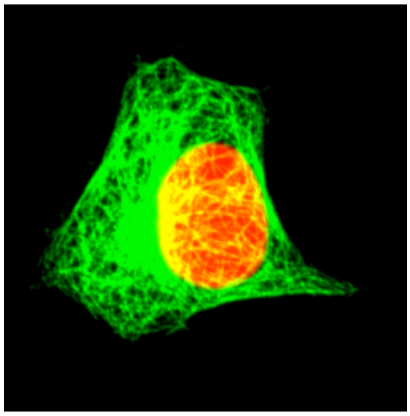

Figure 1. Example image of microtubule distribution from the collection used in model building. The image is a summed projection onto the x-y plane. The distribution of DNA is in red, and tubulin in green.

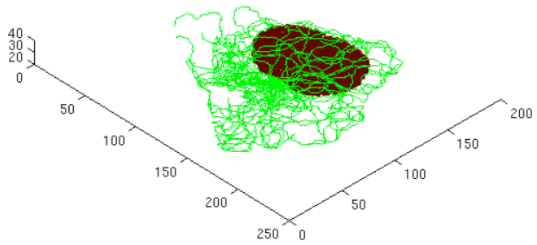

Figure 2. Example synthesized microtubule network. This sample was generated using the model and the nuclear and cell boundaries of the cell shown in Figure 1.

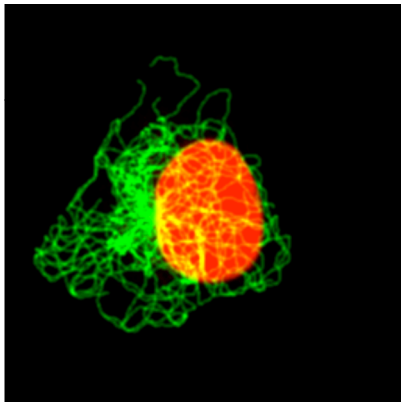

Figure 3. An example of a simulate image derived from synthesized microtubule network model shown in Figure 2. The image is a summed projection onto the x-y plane

Table 1 reports the APE values for each of the model parameters.

Table 1. Average Percent Error for matching of synthetic images to an image library

N	μ	σ	COS α
10.8%	22.1%	189.9%	1.33%

3.3 Real Images

We finally used the model to test whether sensible parameters can be estimated from real images. The parameters estimated for the real image shown in Figure 1

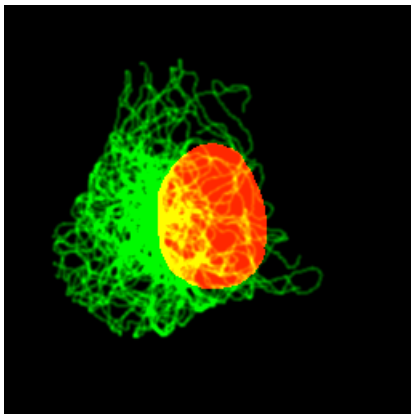

Figure 4. A sum projected image generated with parameters estimated. Number of microtubules = 125, Mean of length distribution = 25 um, Std Dev = 10 um, cos α = 0.9.

are: Number of microtubules = 125, Mean of length distribution = 25 um, Std Dev = 10 um, cos α = 0.9. An image generated with these parameters is shown in Figure 4.

4. Discussion

We have described a method to estimate microtubule growth parameters by an indirect approach using stochastic generative growth modeling and image matching. Our method simulates images of microtubule networks based on an individual cell's geometric configuration. The approximately correct model parameters can be found by maximizing the similarity (minimizing a matching function) between the simulated and real images. Due to the imprecise localization of microtubules in real fluorescence microscope images, we match these indirectly by comparing rotation invariant features using the Normalized Euclidean distance. The results demonstrated that the generative model proposed is capable of producing images of good visual correspondence with real images.

The parameters estimated are quantitative information about microtubules such as the number and the length distribution. These parameters are "sensible" parameters, and are often difficult to estimate with a direct approaches such as tracing and tracking due to the nature of the data. Traditionally, parameters of a physical model often require only some "relative" measurement such as kinetic parameters or ratios that can lead to many possible solutions for "sensible" parameters, and hence cannot be estimated. Our model includes parameters that can directly explain the structure and appearance of microtubules in three dimensional fluorescence microscopy images.

Acknowledgements

This work was supported in part by National Science Foundation grant EF-0331657 (R.F.M.), and National Institutes of Health grants R01 GM075205 (R.F.M.) and U54 RR022241 (Alan Waggoner).

References

T. Chan and A. Vese. Active contours without edges. IEEE Transactions on Image Processing, 10:266–276, 2001.

X. Chen, M. Velliste, S. Weinstein, J.W. Jarvik and R.F. Murphy (2003) Location proteomics - Building subcellular location trees from high resolution 3D fluorescence microscope images of randomly-tagged proteins. *Proc. SPIE* 4962:298-306.

J. F. Dorn, K. Jaqaman, D. R. Rines, G. S. Jelso, P. K. Sorger and G. Danuser (2005) Yeast Kinetochore Microtubule Dynamics Analyzed by High-resolution Three-Dimensional Microscopy, *Biophys. J.* 89:2835-2854.

C. G. Pearson, M. K. Gardner, L. V. Paliulis, E. D. Salmon, D. J. Odde, and K. Bloom (2006) Measuring Nanometer Scale Gradients in Spindle Microtubule Dynamics Using Model Convolution Microscopy, *Mol. Biol. Cell* 17:4069-4079.

A. Ponti, A. Matov, M. C. Adams, S. Gupton, C. Waterman-Storer, and G. Danuser (2005) Periodic patterns of actin turnover in lamellipodia and lamellae of migrating epithelial cells analyzed by Quantitative Fluorescent Speckle Microscopy, *Biophys. J.* 89:3456-3469.

M.E. Sargin, A. Altinok, E. Kiris, L. Wilson, S. Feinstein, K. Rose and B.S. Manjunath (2007) Tracing Microtubules in Live Cell Images. *Proc. 2007 IEEE Intl. Symp. Biomed. Imaging (ISBI 2007),* pp. 296-299.

M. Velliste and R.F. Murphy (2002) Automated Determination of Protein Subcellular Locations from 3D Fluorescence Microscope Images. *Proc. 2002 IEEE Intl. Symp. Biomed. Imaging (ISBI 2002),* pp. 867-870.

T. Zhao and R.F. Murphy (2007) Automated Learning of Generative Models for Subcellular Location: Building Blocks for Systems Biology. *Cytometry* 71A:78-990.