Published in final edited form as:
Proc IEEE Int Symp Biomed Imaging. 2010 April ; 2010: 1037–1040. doi:10.1109/ISBI.2010.5490167.

A Graphical Model to Determine the Subcellular Protein Location in Artificial Tissues

Estelle Glory-Afshar[1], **Elvira Osuna-Highley**[1,2], **Brian Granger**[3], and **Robert F. Murphy**[1,2,3,4]

[1]Center for Bioimage Informatics and Department of Biomedical Engineering, Carnegie Mellon University, Pittsburgh, PA, USA

[2]Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA, USA

[3]Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA, USA

[4]Department of Machine Learning, Carnegie Mellon University, Pittsburgh, PA, USA

Abstract

Location proteomics is concerned with the systematic analysis of the subcellular location of proteins. In order to perform comprehensive analysis of all protein location patterns, automated methods are needed. With the goal of extending automated subcellular location pattern analysis methods to high resolution images of tissues, 3D confocal microscope images of polarized CaCo2 cells immunostained for various proteins were collected. A three-color staining protocol was developed that permits parallel imaging of proteins of interest as well as DNA and the actin cytoskeleton. The collection is composed of 11 to 21 images for each of the 9 proteins that depict major subcellular patterns. A classifier was trained to recognize the subcellular location pattern of segmented cells with an accuracy of 89.2%. Using the Prior Updating method allowed improvement of this accuracy to 99.6%. This study demonstrates the benefit of using a graphical model approach for improving the pattern classification in tissue images.

Index Terms

subcellular protein location; artificial tissue; machine learning; graphical model; prior updating

1. Introduction

Extensive work on automated methods for determining the subcellular location of proteins from microscope images has been described previously [1-3]. Most of the work in location proteomics has centered on single-cell analysis or sparse multi-cell images. This is the logical starting point for these types of analyses as it provides a simpler problem. Although the study of cell behavior in a single-cell environment is informative and useful, analysis of subcellular location is ultimately needed at the tissue level. Databases exist that contain images of tissues that have been stained for various proteins. The most notable of these is the Human Protein Atlas (HPA) [4]. It contains millions of images of various normal and cancerous human tissues. Immunohistochemistry was used to stain for various proteins and visual examination was used to annotate the images. Automating the annotation process is an important goal, since automated methods have the potential of performing as well as, or better than, visual examination [3]. In fact, encouraging results have been described for the automated classification of major subcellular location patterns in the immunohistochemically stained HPA images [5]. However, this work analyzed patterns at the level of the entire tissue (due to the difficulty of automatically segmenting tissue images into single cell regions) and therefore did not take full advantage of the methods that have been developed for single-cell images. Furthermore, there may be multiple cell types

contained in one tissue sample and the subcellular location of the protein of interest may depend on the cell type. To develop and evaluate automated methods for such complex situations, tissue images in which the boundaries of each cell and its true cell type and subcellular pattern are known can be expected to be of great utility.

Graphical models are algorithms of growing interest, particularly to design machine learning algorithms for complex systems. They are particularly well suited to analyze tissues where multiple cells are organized in structures capable of fulfilling a function determined by the physical proximity and phenotype of cells. We therefore sought to distinguish major subcellular protein patterns in individual cells of a tissue using graphical models. Our first step was to use a supervised classifier to estimate the pattern probabilities for each cell (using the Random Forest algorithm). Then, we determined whether the classification accuracy could be improved by allowing these probabilities to be influenced by the probabilities of neighboring cells. We have previously showed that the implementation of graphical models can significantly improve classification performance on multi-cell synthetic images [6]. In particular, an approximation of loopy belief propagation with a novel voting potential (a combination termed Prior Updating) was observed to be much faster than other approaches while achieving high accuracies [7]. We present here, for the first time, to our knowledge, the application of a graphical model to improve the determination of the subcellular pattern in tissue images.

2. Artificial Tissue Generation

We chose to use CaCo2 cells as a model system to extend current single-cell methods to tissue. CaCo2 is a human colon cancer cell line which grows in a tight monolayer that simulates the density of tissue. When allowed to grow several days past confluency, CaCo2 cells will form tight junctions between cells and establish a polarized morphology. The protocol used for plating, fixing, permeabilizing, and staining CaCo2 cells will be described elsewhere [8]. Table 1 summarizes the 9 subcellular patterns for which we collected images, along with their respective antibodies or probes.

3. Image Acquisition

Imaging was performed on a Zeiss LSM 510 Meta NLO Confocor 3 Inverted Spectral Confocal Microscope using a 63×/1.4NA oil-immersion objective. Three-color, 3D images were acquired simultaneously using 488 nm (FITC-Phalloidin, actin channel), 543 nm (Alexa-568, protein channel) and 633 nm (DRAQ-5, DNA channel) excitation and appropriate emission filters (see Figure 1). Images were acquired in 1024×1024 pixels each with a resolution of 0.09 μm/pixel. The z-slices were separated axially by 0.33 μm. The number of images and segmented cells of each pattern are reported in Table 1.

4. Segmentation

The determination of subcellular patterns for each cell in 3D tissue images requires a delineation of the volume occupied by each cell. Taking advantage of the monolayer structure of the epithelial tissue formed by CaCo2, i.e. confluent but non-overlapping polarized cells elongated orthogonally to the filter plane, the segmentation was performed by projecting (average intensity) half of the slices from the middle of each 3D stack into 2D images. Only the central slices were used to avoid the regions of dense actin networks found in the apical and basal extremities of CaCo2 cells. The projection was performed independently for the DNA and the actin channels (see Figure 1).

We used Ridler-Calvard thresholding on projected DNA channels to generate nuclear seeds, as described previously [3]. The seeded watershed algorithm was performed on the projected

actin channel using nuclear seeds. The actin distribution gives an approximation to the location of cell membranes. Finally, the extent of each cell was defined as the orthogonal projection in 3D of the 2D masks resulting from the watershed algorithm.

5. Feature Extraction

For each segmented cell, the feature set SLF19 was calculated to describe the 3-dimensional distribution of the protein within cells [9]. SLF19 is composed of 14 morphological features for 3D objects (number of objects, relative volume of objects, etc.), 14 DNA-related features (average distance from the nucleus, overlapping volume with the nucleus, etc.), 2 edge features, and 26 Haralick texture features. As a result, each cell is described by a vector of 56 features.

6. Classification

A supervised machine learning approach was used to train a classifier to distinguish the 9 major patterns (classes). We used a random forest implementation in R (http://www.stat.berkeley.edu/~breiman/RandomForests) [10] for this purpose. The random forest is a collection of decision trees grown independently. Each tree is trained with a random subset of labeled samples (cells whose class is known) among the entire dataset. Each node of a decision tree uses a random selection of features in order to create a collection of decision trees with controlled variations. The remaining set of data (cells whose class is unknown), called the out-of-bag portion, is used to test the tree. Finally, a class is assigned to a cell by determining the majority voting of classes provided by individual trees for which the cell belonged in the out-of-bag portion. The agreement of this assignment with the true label can be reported in a confusion matrix and the overall accuracy of the random forest can be calculated. With this approach, there is no need of cross-validation because each tree has unique training and testing sets.

7. Graphical Model

7.1. Physical and feature space models

The first step of the graphical model consists of building a graph where nodes represent cells and edges link nodes that are allowed to exchange information. We created two versions of this graph. In the physical space model, cells were linked based on the Euclidean distance between their centroids in each image. Edges longer than a certain distance cutoff (d_{cutoff}) were removed to limit the size of the graph. In the feature space model, a graph is created in the feature space where the Mahalanobis distance between features of every cell in the dataset determines the length of edges.

7.2. Voting potential

The voting potential function is designed to influence the evidence $P(x_i)$ of node i for every possible label x with a voting potential φ which reflects the evidence of all of the neighboring i's into a summary vote. The overall probability of a vector of classes x is:

$$P(x)=\frac{1}{Z}\prod_{\text{nodes i}} P(x_i)\phi(x_i, v_1, v_2, \ldots, v_m)$$

where Z is a normalizing constant to ensure the sum of the vector P(x)=1. The output of the random forest can be converted into probabilities $P(x_i)$ that classes x are assigned to a cell i by calculating the proportion of predicted classes x assigned to i by the forest (considering

only trees for which *i* belongs to the out-of-bag portion). The voting potential φ is defined as:

$$\phi(x_i, v_1, v_2, \ldots, v_m) = \frac{\lambda / n + \sum_{k \in N(i)} I(v_k, i)}{m + \lambda}$$

where v_1, v_2, …v_m are the *m* neighbors of *i*, *n* is the number of classes $N(i)$ is the set of neighbors of node *i*, and λ is a smoothing parameter that controls the influence of the neighbor at each iteration. The smaller λ is, the more strongly neighbors will influence the cell classification. *I(v_k,i)* is a probability vector initialized to equal values for each class (as there are 9 classes, the initial probability *I(v_k,i)* of the class *x* for an arbitrary node *i* and its neighbor v_k is 1/9). The probabilities *I(v_k,i)* are updated iteratively. The iterative updating is stopped when the probabilities of each node stay unchanged in two successive iterations.

8. Results

To classify individual cells, we set the size of the random forest to 500 trees and the number of random features used to generate the decision of each node to 8. The overall accuracy of the random forest was 89.24% with a large disparity between classes, from 76.5% accuracy for predicting tubulin to 100% accuracy for predicting nuclei (see Table 2).

To improve this accuracy, we tested various graphical model variations. The two parameters in the graphical model, λ and d_{cutoff}, were optimized through a grid search. λ was varied from 0 to 1.2 with an increment of 0.2, and d_{cutoff} was varied from 0 to 1200 pixels with an increment of 100 in the physical space and from 0 to 1.2 with an increment of 0.1 in the feature space. A 5-fold cross-validation was used to optimize the parameters. We observed that variations in λ had almost no effect on the accuracy (data not shown). During the optimization, the highest accuracies were reached for λ = 0 and d_{cutoff}=800 for the physical space. This gave an overall accuracy after updating labels of 99.57% (Table 2). By contrast, the graphical model tested in the feature space gave an overall accuracy of only 89.5%.

The graphical model in the physical space improved the accuracy by 10 percentage points over the random forest classifier (see Figure 2 for illustration). The approach does not seem to be as effective in the feature space, probably because the cells which are mislabeled by the random forest already have atypical features for their class. Linking these cells with cells which have similar features does not help in updating labels.

The graphical model built in the physical space mimics the situation of a tissue composed of different cell types which are organized in such a way that similar cells are close to each other to fulfill their function. Several variations can be tested to build the graph, for example by linking only cells which have the same cell type in a tissue or considering only touching cells.

The approach described here can easily be extended to the study of 3D tissues by building the graph using distance in three-dimensional physical space without changing anything else.

8. Conclusion

This work illustrates a new application of graphical models for the determination of subcellular protein patterns in real tissue from fluorescence confocal microscopy images. It

shows a significant improvement over a random forest classifier allowing an 99.57% accuracy in distinguishing 9 major subcellular patterns. This work also provides a complete collection of 3D fluorescent images of artificial tissue with 3 channels, DNA, actin and immunostained protein which represent the major subcellular compartments. The dataset and code used in this paper are available from http://murphylab.web.cmu.edu/software.

Acknowledgments

The authors would like to thank Naomi Gunawardena and James Fitzpatrick for their help in cell culture and image acquisition, respectively. The work described here was supported in part by NIH grants GM075205 (R.F.M.) and U54 RR022241 (Alan Waggoner).

References

1. Boland MV, Markey MK, Murphy RF. Automated Recognition of Patterns Characteristic of Subcellular Structures in Fluorescence Microscopy Images. Cytometry. 1998; 33:366–375. [PubMed: 9822349]
2. Velliste, M.; Murphy, RF. Automated Determination of Protein Subcellular Locations from 3D Fluorescence Microscope Images. Proc 2002 IEEE Intl Symp Biomed Imaging (ISBI 2002); 2002. p. 867-870.
3. Glory E, Murphy RF. Automated Subcellular Location Determination and High Throughput Microscopy. Developmental Cell. 2007; 12:7–16. [PubMed: 17199037]
4. Uhlen M, Bjorling E, Agaton C, Szigyarto CAK, Amini B, Andersen E, Andersson AC, Angelidou P, Asplund A, Cerjan D, Ekstrom M, Elobeid A, Eriksson C. A Human Protein Atlas for Normal and Cancer Tissues based on Antibody Proteomics. Mol Cell Proteomics. 2005; 4:1920–1932. [PubMed: 16127175]
5. Newberg JY, Murphy RF. A Framework for the Automated Analysis of Subcellular Patterns in Human Protein Atlas Images. J Proteome Res. 2008; 7:2300–2308. [PubMed: 18435555]
6. Chen SC, Murphy RF. A Graphical Godel approach to automated classification of protein subcellular location patterns in multi-cell images. BMC Bioinformatics. 2006; 7:90–103. [PubMed: 16504075]
7. Chen SC, Gordon GJ, Murphy RF. Graphical Models for Structured Classification, with an Application to Interpreting Images of Protein Subcellular Location Patterns. J Machine Learning Res. 2008; 9:651–682.
8. Glory-Afshar E, Osuna-Highley E, Granger B, Murphy RF. A Framework for the Determination of Protein Subcellular Location Within Tissues. in preparation.
9. Nair P, Schaub BE, Huang K, Chen X, Murphy RF, Griffith JM, Geuze HJ, Rohrer J. Characterization of the TGN Exit Signal of the Human Mannose 6-Phosphate Uncovering Enzyme. J Cell Sci. 2005; 11:2949–2956. [PubMed: 15976452]
10. Breiman L. Random Forests. Machine Learning. 2001; 45:5–32.

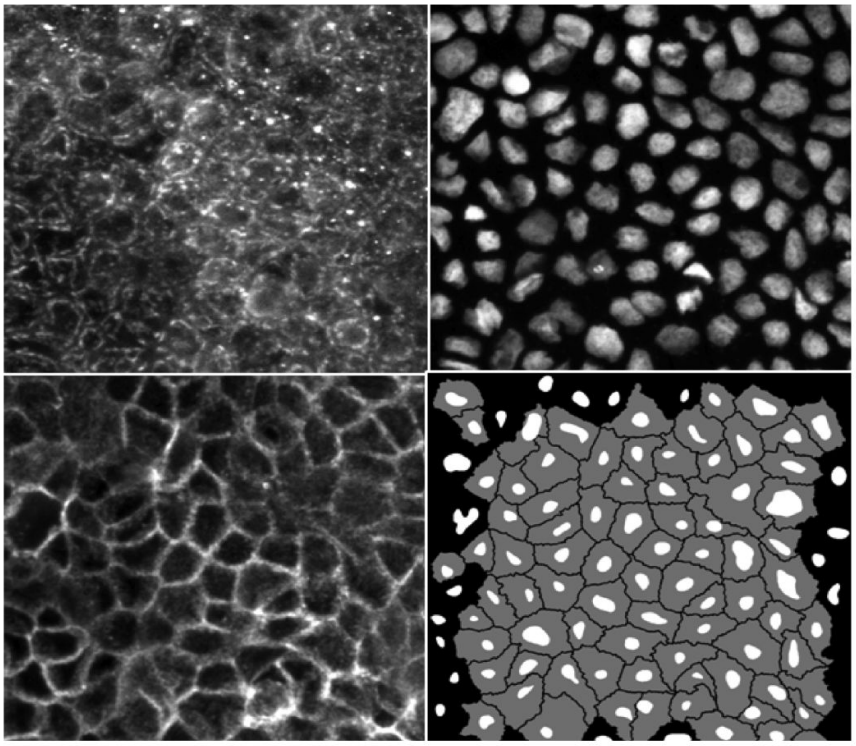

Figure 1.
The top left panel shows the central slices projection of the protein channel depicting a Golgi pattern. The top right and bottom left panels represent respectively the projected DNA and actin channels. The bottom right panel shows the seeds (segmented nuclei) in white and the regions delineated by the watershed in gray. The regions touching the image borders were discarded from subsequent analysis since their patterns are likely incomplete.

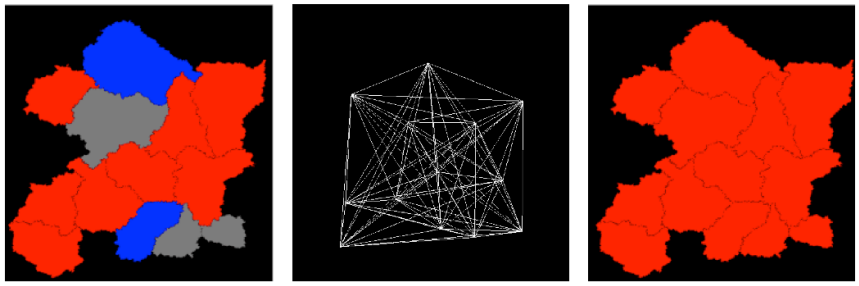

Figure 2.
Illustration of graphical model updating for an image stained for ER. The first panel shows the output of the random forest where the red regions represent cells correctly assigned the ER pattern, while the blue and gray regions represent cells misclassified as lysosomes and tubulin, respectively. The propagation of the class probabilities along the edges of the graph (middle panel) modifies the labels to give the correct labeling (right panel).

Table 1

Description of dataset composed of 9 subcellular protein patterns.

Target	Antibody or Stain	Cells/Images
Mitochondia	mouse anti-mitochondrial inner membrane	669/17
Endosomes	mouse anti-CD71	446/21
Lysosomes	mouse anti-LAMP2	514/16
Endoplasmic Reticulum	mouse anti-ERp57	808/18
Golgi Apparatus	mouse anti-Golgi 58K protein	793/21
Nucleoli	mouse anti-nucleolin	505/16
Tubulin	mouse anti-ß Tubulin I + II	635/11
DNA	DRAQ5	669/17
Actin	FITC-Phalloidin	669/17

Table 2

Confusion matrices of classes assigned by the computer (columns) versus the true pattern (rows). The first number corresponds to the accuracy of the random forest classifier (overall accuracy=89.24%) and the second number represents the accuracy of the graphical model (overall accuracy=99.57%).

	Nucleus	ER	Golgi	Lysosomes	Mito.	Nucleoli	Actin	Endosomes	Tubulin
Nucleus	**100/100**	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
ER	0/0	**90.34/99.1**	0.91/0	0.27/0	2.73/0.18	0.27/0	0.09/0	5.01/0.73	0.36/0
Golgi	0/0	1.21/0.23	**83.28/99.77**	2.19/0	2.65/0	0.08/0	.38/0	7.87/0.23	2.34/0
Lysosome	0/0	.61/0.21	4.27/0.31	**87.69/99.08**	2.54/0	0.20/0	0.20/0	2.75/0.10	1.73/0.51
Mitocho.	0/0	0.88/0	1.77/0	2.74/0	**87.62/99.20**	0.27/0	0.27/0	5.39/0.27	1.06/0.53
Nucleoli	0.21/0	0.32/0	0.85/0	0.32/0	0.53/0	**94.03/100**	0.21/0	1.60/0	1.92/0
Actin	0/0	1.41/0.19	0.56/0	0/0	0.19/0	0/0	**96.43/99.81**	0.66/0	0.75/0
Endosomes	0/0	2.43/0	4.35/0	0.64/0	3.57/0.07	0.21/0	0.93/0	**87.23/99.93**	0.64/0
Tubulin	0.18/0	3.91/0	6.41/0	1.60/0	0.53/0	1.78/0.71	3.20/0	5.87/0	**76.51/99.29**

Proc IEEE Int Symp Biomed Imaging. Author manuscript; available in PMC 2011 May 27.