Image Analysis of Subcellular Patterns for High Throughput Screening and Systems Biology

Robert F. Murphy
Departments of Biological Sciences, Biomedical Engineering, and Machine Learning, and

Contents

- Introduction to subcellular pattern analysis and recommendations regarding image acquisition for subsequent automated analysis
- methods for automated segmentation of multi-cell images into single cell regions
- types of features used to describe subcellular patterns and methods for extraction of these features (especially morphological, texture and wavelet features)
- statistical and machine learning methods for comparison, classification and clustering of patterns
- publicly available image database systems

Introduction to Protein Subcellular Location

Eukaryotic cells have many parts

Protein localization

- The sequence of each protein determines where it is localized in cells
- Subsequences (“motifs”) within a protein’s sequence are responsible for targeting it to one (or more) locations (structures/organelles)

Open questions

- How many distinct locations can proteins be found in? What are they?
- How many distinct motifs direct proteins to those locations? What are they?
Proteomics

- The set of proteins expressed in a given cell type or tissue is called its proteome.
- Proteomics projects:
  - sequence
  - structure
  - activity
  - partners
  - location

Location information in protein databases: Traditional approach

- conduct experiments of various types
  - Cell fractionation
  - Electron microscopy
  - Fluorescence microscopy
- describe the results in unstructured text (first in journal articles and then in summaries in databases)
  - "Protein X is located primarily in protrusions from the early endosomal membrane but is also found in the plasma membrane."

Location information in protein databases: Ontology approach

- Systematic analysis and comparison of these descriptions were made difficult by both the unstructured nature of the text and the variation in terminology used from one laboratory to another.
- To address this problem, a restricted vocabulary for cellular components was created by the Gene Ontology consortium.

Restricted Vocabulary Approaches
Use of GO terms
- Databases such as SwissProt use manual curation to assign GO terms to proteins based on reading of relevant literature
- A major problem is consistency of application of terms

Comparison of GO terms for two proteins

<table>
<thead>
<tr>
<th>GoliB1</th>
<th>GPP130</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integral to membrane;</td>
<td>Integral to membrane;</td>
</tr>
<tr>
<td>Golgi membrane;</td>
<td>Golgi cis-face;</td>
</tr>
<tr>
<td>Golgi stack;</td>
<td>Golgi lumen;</td>
</tr>
<tr>
<td></td>
<td>endocytic transport vesicle</td>
</tr>
</tbody>
</table>

Source: SwissProt

Words are not enough
- We learned that Giantin and GPP130 are both Golgi proteins, but do we know:
  - What part (i.e., cis, medial, trans) of the Golgi complex they each are found in?
  - If they have the same subcellular distribution?
  - If they also are found in other compartments?

Conclusion
- Current knowledge of subcellular locations of proteins is not sufficiently detailed or systematic
- Systematic description of subcellular locations should be created using a data-driven approach rather than a knowledge-capture approach

Determining protein location
- The primary method used to determine the subcellular location of a protein is to “tag” it with a fluorescent probe and then image its distribution within cells using fluorescence microscopy

Tagging proteins for fluorescence microscopy
- Immunofluorescence
  - “primary” antibody against the target.
  - “secondary” antibody against the “primary” and conjugated with a fluorescent probe
  - Fixed-cells only
- Gene/cDNA-tagging
  - merge DNA coding for a naturally fluorescent protein (or vital probe binding sequence) with coding sequence of a protein of interest
  - Live-cell possible
Tagging proteins for fluorescence microscopy

- **GFP-tagging**
  - Can create fusion between GFP and a cDNA, in which case all regulatory sequences that control expression of the corresponding protein is lost.
  - Can create fusion between GFP and the genomic sequence of a gene, in which case regulatory sequences preserved.
  - Example: CD-tagging

Principles of CD-Tagging (Jarvik & Berget) (CD = Central Dogma)

- Genomic DNA + CD-cassette
- Exon 1
- Exon 2
- Tag
- Tagged DNA
- Tagged mRNA
- Tagged Protein
- Exon 1
- Intron 1
- Exon 2
- CD cassette
- Tag
- Tagged mRNA
- Tagged Protein

Automated Interpretation

- Traditional analysis of fluorescence microscope images has occurred by visual inspection.
- Our goal has been automate the interpretation, to yield better:
  - Objectivity
  - Sensitivity
  - Reproducibility

Initial Goal

- This is a microtubule pattern
- Assign proteins to major subcellular structures using fluorescent microscopy

The Challenge

- Problem is hard because different cells have different shapes, sizes, orientations.
- Organelles/structures within cells are not found in fixed locations.
- Therefore, describe each image numerically and use the descriptors.

Successful Classification and Clustering

- Murphy group has demonstrated classification of ten subcellular patterns in 2D and 3D images of HeLa cells with accuracy on single cells of 92% and 98% accuracy, respectively.
- Have also clustered 90 randomly-tagged proteins into 17 statistically-distinct patterns in 3T3 cells.
**Acquisition considerations**

- Ensure Nyquist Sampling at Rayleigh limit
- Maintain low cell density if single cell measurements desired
- Control acquisition variables
  - Select (initial) focal plane consistently
  - Select fields consistently (at least one full cell per field)
  - Maintain constant camera gain, exposure time, number of slices
  - Select interphase cells or ensure sampling of cell cycle

**Acquisition considerations (continued)**

- Collect sufficient images per condition
  - For classifier training or set comparison, more than number of features
  - For classification or clustering, base on confidence level desired
- Collect reference images if possible (DNA, membrane)

**Annotation considerations**

- Maintain adequate records of all experimental settings
- Organize images by cell type/probe/condition

**Preprocessing**

- Correction for/Removal of camera defects
- Background correction
- Autofluorescence correction
- Illumination correction
- Deconvolution

**Preprocessing (continued)**

- Registration
  - Not critical if only using DNA or membrane references
- Intensity scaling (constant scale or contrast stretched for each cell)
- Single cell segmentation
  - Manual, semi-automated, automated
- Region finding
  - Nucleus
  - Cytoplasmic annulus
  - Cell boundary

**Microscope Datasets for Subcellular Location**

- We have collected datasets of fluorescence microscope images depicting the subcellular location patterns of a number of proteins in three different cell lines
- Available at http://murphylab.web.cmu.edu
Microscope Datasets for Subcellular Location

- 2D Chinese hamster ovary cells
  - Widefield microscopy with numerical deconvolution (100x)
  - 5 different probes (classes)
  - 1 color
  - Pixel size = 0.23 µm x 0.23 µm
  - ~80 cell images per class

Example Images: 2D CHO

- Single color staining for specific protein
- Three 2D slices acquired and numerically deconvolved to yield one in focus 2D slice

Microscope Datasets for Subcellular Location

- 2D HeLa
  - Widefield microscopy with numerical deconvolution (100x)
  - 9 different antibodies plus a DNA stain
  - 2 colors per image
  - Pixel size = 0.23 µm x 0.23 µm
  - ~80 cell images per class

Example Images: 2D HeLa

- Red=DNA, Green=specific
- Three 2D slices acquired & numerically deconvolved to yield one in focus 2D slice
- Red and Green semi-automatically registered

Microscope Datasets for Subcellular Location

- 3D HeLa
  - Confocal Microscope (100x)
  - 9 different antibodies plus DNA stain and total protein stain
  - 3 colors per image
  - Voxel size = 0.049 µm x 0.049 µm x 0.2 µm
  - ~50 cell images per class

Example Image: 3D HeLa

- Red=DNA, Blue=Total Protein, Green=specific protein
- Acquired as stack of 2D slices by changing focal position

Example Images: 2D CHO
Example Images: 3D HeLa

- Nuclear
- ER
- Giantin
- gpp130
- Lysosomal

- Mitoch.
- Nucleolar
- Actin
- Endosomal
- Tubulin

3D HeLa

- 2D slices (from bottom to top) for cell labeled for transferrin receptor (primarily in endosomes)

3D HeLa

- 2D slices (from bottom to top) for cell labeled for giantin (primarily in Golgi)

3D HeLa

- 2D slices (from bottom to top) for cell labeled for tubulin (major constituent of microtubules)

Microscope Datasets for Subcellular Location

- 3D 3T3
  - Spinning Disk Confocal Microscope (60x)
  - GFP for a specific protein
  - Images collected for 90 different clones
  - 1 color
  - Voxel size = 0.11 µm x 0.11 µm x 0.5 µm
  - ~30 cell images per class
  - Also have some 2D time series images

Example Images: 3D 3T3

Thirty slices acquired by spinning disk confocal microscope
2D slices over time for cell labeled for Glut1 (membrane protein on surface and in vesicles)

Other subcellular location projects
- O’Shea group - Yeast
  - GFP-labeled cDNAs
  - GFP and DNA images with some additional markers
- Pepperkok group - human (MCF7 cells)
  - GFP-labeled cDNAs
  - GFP and DNA images
- Uhlen group (Protein Atlas) - human
  - Immunohistochemistry with monoclonal antibodies
  - DAB and hematoxylin images
  - Fixed tissues
- Schubert group (MELK technology)
  - Cycles of immunofluorescence, imaging and bleaching
  - Fixed tissues

Feature Extraction for Subcellular Pattern Analysis

Goal
- This is a microtubule pattern
- Assign proteins to major subcellular structures using fluorescent microscopy

The Challenge
- Problem is hard because different cells have different shapes, sizes, orientations
- Organelles/structures within cells are not found in fixed locations
- Therefore, describe each image numerically and use the descriptors

Feature-Based, Supervised Learning Approach
1. Create sets of images showing the location of many different proteins (each set defines one class of pattern)
2. Reduce each image to a set of numerical values (“features”) that are insensitive to position and rotation of the cell
3. Use statistical classification methods to “learn” how to distinguish each class using the features
### Subcellular Location Features (SLF)
- Combinations of features of different types that describe different aspects of patterns in fluorescence microscope images have been created.
- Motivated in part by descriptions used by biologists (e.g., punctate, perinuclear).
- To ensure that the specific features used for a given experiment can be identified, they are referred to as Subcellular Location Features (SLF) and defined in sets (e.g., SLF1).

### Feature levels and granularity
- **Object features**
- **Cell features**
- **Field features**
- Aggregate/average operator

Granularity: 2D, 3D, 2Dt, 3Dt

### Thresholding
- First type of feature is morphological.
- Morphological features require some method for defining objects.
- Most common approach is global thresholding.
- Methods exist for automatically choosing a global threshold (e.g., Riddler-Calvard method).

### Ridler-Calvard Method
- Find threshold that is equidistant from the average intensity of pixels below and above it.

Blue line shows histogram of intensities, green lines show average to left and right of red line, red line shows midpoint between them or the RC threshold.

Ridler-Calvard Illustration

![Ridler-Calvard Method](image)
Otsu Method

- Find threshold to minimize the variances of the pixels below and above it

Adaptive Thresholding

- Various approaches available
- Basic principle is use automated methods over small regions and then interpolate to form a smooth surface

Suitability of Automated Thresholding for Classification

- For the task of subcellular pattern analysis, automated thresholding methods perform quite well in most cases, especially for patterns with well-separated objects
- They do not work well for images with very low signal-noise ratio
- Can tolerate poor behavior on a fraction of images for a given pattern while still achieving good classification accuracies

Object finding

- After choice of threshold, define objects as sets of touching pixels that are above threshold

2D Features

Morphological Features

<table>
<thead>
<tr>
<th>SLF No.</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLF1.1</td>
<td>The number of fluorescent objects in the image</td>
</tr>
<tr>
<td>SLF1.2</td>
<td>The average size of objects</td>
</tr>
<tr>
<td>SLF1.3</td>
<td>The variance of the number of above-threshold pixels per object</td>
</tr>
<tr>
<td>SLF1.4</td>
<td>The ratio of the size of the largest object to the smallest</td>
</tr>
<tr>
<td>SLF1.5</td>
<td>The average object distance to the cellular center of fluorescence(COF)</td>
</tr>
<tr>
<td>SLF1.6</td>
<td>The variance of object distances from the COF</td>
</tr>
<tr>
<td>SLF1.7</td>
<td>The ratio of the largest to the smallest object to COF distance</td>
</tr>
</tbody>
</table>

2D Features

Morphological Features

- Any of these features could be used to distinguish these two classes:
- ER
- Nucleoli
Suitability of Morphological Features for Classification

- Images for some subcellular patterns, such as those for cytoskeletal proteins, are not well-segmented by automated thresholding.
- When combined with non-morphological features, classifiers can learn to "ignore" morphological features for those classes.

2D Features

DNA Features

<table>
<thead>
<tr>
<th>SLF No.</th>
<th>Description</th>
<th>Description</th>
<th>Description</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLF2.17</td>
<td>The average object distance from the COF of the DNA image</td>
<td>SLF2.18</td>
<td>The variance of object distances from the DNA COF</td>
<td>SLF2.19</td>
</tr>
<tr>
<td>SLF2.20</td>
<td>The distance between the protein COF and the DNA COF</td>
<td>SLF2.21</td>
<td>The ratio of the area occupied by protein to that occupied by DNA</td>
<td>SLF2.22</td>
</tr>
</tbody>
</table>

Skeleton Features

<table>
<thead>
<tr>
<th>SLF No.</th>
<th>Description</th>
<th>Description</th>
<th>Description</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLF7.80</td>
<td>The average length of the morphological skeleton of objects</td>
<td>SLF7.81</td>
<td>The ratio of object skeleton length to the area of the convex hull of the skeleton, averaged over all objects</td>
<td>SLF7.82</td>
</tr>
<tr>
<td>SLF7.83</td>
<td>The fraction of object fluorescence contained within the skeleton</td>
<td>SLF7.84</td>
<td>The ratio of the number of branch points in the skeleton to the length of the skeleton</td>
<td></td>
</tr>
</tbody>
</table>

Edge Features

<table>
<thead>
<tr>
<th>SLF No.</th>
<th>Description</th>
<th>Description</th>
<th>Description</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLF1.10</td>
<td>Measure of edge gradient intensity homogeneity</td>
<td>SLF1.11</td>
<td>Measure of edge direction homogeneity 1</td>
<td>SLF1.12</td>
</tr>
<tr>
<td>SLF1.13</td>
<td>Measure of edge direction difference</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Convex Hull Features

<table>
<thead>
<tr>
<th>SLF No.</th>
<th>Description</th>
<th>Description</th>
<th>Description</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLF1.14</td>
<td>The fraction of the convex hull area occupied by protein fluorescence</td>
<td>SLF1.15</td>
<td>The roundness of the convex hull</td>
<td>SLF1.16</td>
</tr>
</tbody>
</table>
2D Features
Zernike Moment Features
(SLF 3.17-3.65)
- Shape similarity of protein image to Zernike polynomials $Z(n,l)$
- 49 polynomials and 49 features

Left: Zernike polynomials
A: $Z(2,0)$
B: $Z(4,4)$
C: $Z(10,6)$

Right: lamp2 image

• Shape similarity of protein image to Zernike polynomials $Z(n,l)$
• 49 polynomials and 49 features

2D Features
Haralick Texture Features
(SLF 7.66-7.78)
- Correlations of adjacent pixels in gray level images
- Start by calculating co-occurrence matrix $P$:
  - $N$ by $N$ matrix, $N=$number of gray level.
  - Element $P(i,j)$ is the probability of a pixel with value $i$ being adjacent to a pixel with value $j$
- Four directions in which a pixel can be adjacent
- Each direction considered separately and then features averaged across all directions

Pixel Resolution and Gray Levels
- Texture features are influenced by the number of gray levels and pixel resolution of the image
- Optimization for each image dataset required
- Alternatively, features can be calculated for many resolutions

Wavelet Transformation - 1D
- A: approximation (low frequency)
- D: detail (high frequency)
- $X=A3+D3+D2+D1$

2D Wavelets - intuition
- Apply some filter to detect edges (horizontal; vertical; diagonal)
2D Wavelets - intuition

- Recurse

Many wavelet basis functions (filters):
- Haar
- Daubechies (-4, -6, -20)

Daubechies D4 decomposition

Original image
Wavelet Transformation

2D Features
Wavelet Feature Calculation

- Preprocessing
  - Background subtraction and thresholding
  - Translation and rotation
- Wavelet transformation
  - The Daubechies 4 wavelet
  - 10 level decomposition
  - Use the average energy of the three high-frequency components at each level as features

Gabor Function

Can extend the function to generate Gabor filters by rotating and dilating

2D Features
Gabor Feature Calculation

- Preprocessing same as Wavelet
- 30 Gabor filters were generated using five different scales and six different orientations
- Convolve an input image with a Gabor filter
- Take the mean and standard deviation of the convolved image
- 60 Gabor texture features
3D Features
Morphological (SLF-9)
- 28 features, 14 from protein objects and 14 from their relationship to corresponding DNA images
  - Based on number of objects, object size, object distance to COF
  - Corresponding DNA image required

SLF-14
- 14 SLF-9 features that do not require DNA images
- 2 Edge features
  - Ratio of above threshold pixel along an edge
  - Ratio of fluorescence along an edge
- 26 3D Haralick texture features
  - Gray level co-occurrence matrix for 13 directions
  - Calculate 13 Haralick statistics for each direction
  - Average each statistic over 13 directions and use mean and range as separate features: result is 26 features

SLF-17
- A feature subset with 7 features selected from SLF-14 at 256 gray levels and 0.4 micron pixel resolution
  - 1 morphological feature
  - 1 edge feature
  - 5 texture features

Object level features (SOF)
- Subset of SLFs calculated on single objects

Field level features (SLF21)
- Subset of SLFs that do not require segmentation into single cells
  - Average object features
  - Texture features (on whole field)
  - Edge features (on whole field)

Field level features (SLF21)
2Dt or 3Dt Features
Temporal Texture Features
- Haralick texture features describe the correlation in intensity of pixels that are next to each other in space.
  - These have been valuable for classifying static patterns.
- Temporal texture features describe the correlation in intensity of pixels in the same position in images next to each other over time.
Temporal Textures based on Co-occurrence Matrix

- Temporal co-occurrence matrix $P$: an $N_{level} \times N_{level}$ matrix, where $P[i, j]$ is the probability that a pixel with value $i$ has value $j$ in the next image (time point).
- Thirteen statistics calculated on $P$ are used as features.

Implementation of Temporal Texture Features

- Compare image pairs with different time interval, compute 13 temporal texture features for each pair.
- Use the average and variance of features in each kind of time interval, yields $13 \times 5 \times 2 = 130$ features.

PSLID: Protein Subcellular Location Image Database

- A publicly accessible image database at http://murphylab.web.cmu.edu/services/PSLID
- A downloadable open source database system for creating local databases
  - Focused on subcellular pattern analysis
  - Subcellular Location Features integrated into database
  - Integrated comparison, classification, clustering tools
  - Designed for high-throughput microscopy
  - Interface to OME in the works
  - Large ITR project with UCSB for distributed system

Subcellular Location Databases

- PSLID: Protein Subcellular Location Image Database
  - A publicly accessible image database at http://murphylab.web.cmu.edu/services/PSLID
  - A downloadable open source database system for creating local databases
    - Focused on subcellular pattern analysis
    - Subcellular Location Features integrated into database
    - Integrated comparison, classification, clustering tools
    - Designed for high-throughput microscopy
    - Interface to OME in the works
    - Large ITR project with UCSB for distributed system
PSLID contents

- ~1000 2D images of 10 patterns in HeLa cells
- ~1500 3D images of 23 patterns in HeLa cells
- ~2500 3D images of 90 patterns in 3T3 cells
- ~1000 4D images of 32 patterns in 3T3 cells
- More being added
SLIF: Subcellular Location Image Finder

- Extract structured assertions from unstructured Internet sources.
- Develop text and image processing methods to identify specific data that supports relevant assertions.
- Apply data mining methods to assertion knowledge bases to develop new hypotheses, form consensus conclusions, and distinguish differing conditions.

Overview: Image processing in SLIF

- Segment into “panels”
- Detect & remove annotations
- Classify panels

Overview: Text Processing in SLIF

- Find entity names in text, and panel labels in text and the image.
- Match panels labels in text to panel labels on the image.
- Associate entity names to textual panel labels using scoping rules.

Figure 1.
(A) Single confocal optical section of BY-2 cells expressing U2B0-GFP, double labeled with GFP (left panel) and anti-p80 coilin (right panel). Three nuclei are shown, and the bright GFP spots colocalize with bright anti-p80 coilin labeling. There is some labeling of the cytoplasm by anti-p80 coilin. All coiled bodies that contain U2B0 also express the U2B0-GFP fusion. Bars, 5 µm.

(B) Single confocal optical section of BY-2 cells expressing U2B0-GFP, double labeled with GFP (left panel) and 4G3 antibody (right panel). Three nuclei are shown. Most coiled bodies are in the nucleoplasm, but occasionally are seen in the nucleolus (arrows). All coiled bodies that contain U2B0 also express the U2B0-GFP fusion. Bars, 5 µm.

SLIF components

Paper
Caption
Caption understanding
Image
Image processing
Entity extraction
Entity names, cell types
Label finding
Labels
Labels and Metadata
Metadata
Panel splitting
Panel analysis
Panel typing
Subcellular pattern assignment

Murphy Lab SLIF Service

SLIF supports the identification of protein localization, aiming to extract information about protein subcellular distribution from image datasets. The process involves the following steps:

1. Obtain the image dataset
2. Preprocess the images
3. Extract features
4. Classify the images
5. Assign subcellular localization

This approach utilizes deep learning techniques to analyze the images and determine the subcellular localization of proteins. The results are then used to support data mining and further research.
Conclusions

- Methods well worked out for classifying and learning protein patterns (3D images better than 2D images)
- Both better than visual examination
- Can be applied at field, cell or object level
- Image database integrated with interpretation tools (PSLID)
- Information extractor for online text and images (SLIF)

Review Articles


Acknowledgments

- Thanks to Michael Boland, Meel Velliste, Kai Huang, Xiang Chen, Juchang Hua, Yanhua Hu, Zhenzhen Kou, William Cohen, Jonathan Jarvik, and Peter Berget for contributions to the slides in this tutorial and/or the research they describe.
- The research was supported in part by research grant RPG-95-009-03-MCG from the American Cancer Society, by grant 98-295 from the Rockefeller Brothers Fund Charles E. Culpeper Biomedical Pilot Initiative, by NSF grants BIR-9217391, MCB-8800118, and BIR-9256343, by NIH grants R33 CA83219 and R01 GM068845 and by Commonwealth of Pennsylvania Tobacco Settlement Fund research grant 017393, and by graduate fellowships from the Merck Computational Biology and Chemistry Program at Carnegie Mellon University funded by the Merck Company Foundation.

First published system for recognizing subcellular location patterns - 2D CHO (5 patterns)


2D HeLa pattern classification (10 major patterns)


3D HeLa pattern classification (11 major patterns)

Improving features, feature selection, classification method


Classification of multi-cell images


Subcellular Location Trees - 3D 3T3 CD-tagged images


Subcellular Location Trees - Analysis of Location Mutants


PSLID - Protein Subcellular Location Image Database

SLIF - Subcellular Location Image Finder


[http://murphylab.web.cmu.edu/publications](http://murphylab.web.cmu.edu/publications)