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

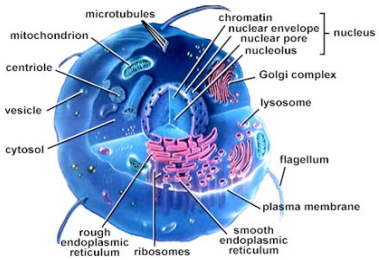
Robert F. Murphy
Departments of Biological Sciences, Biomedical Engineering, and Machine Learning, and
Center for Bioimage Informatics
Carnegie Mellon

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

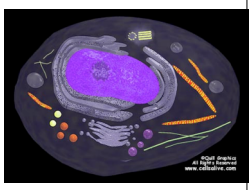
- The sequence of each protein determines where it is localized in cells
- Subsequences ("motifs") within a protein's sequence are responsible for targeting it to one (or more) locations (structures/organelles)

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**



Carnegie Mellon

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"

Carnegie Mellon

Location information in protein databases: Ontology approach

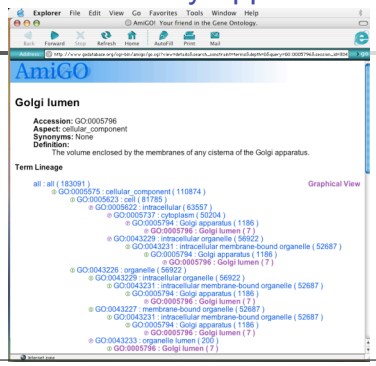
- Systematic analysis and comparison of these descriptions were made difficult by both the unstructured nature of the text and the variation in terminology used from one laboratory to another
- To address this problem, a **restricted** vocabulary for cellular components was created by the **Gene Ontology**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

Carnegie Mellon

Restricted Vocabulary Approaches

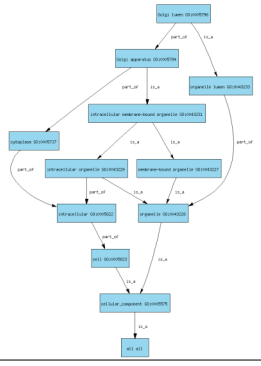
Carnegie Mellon

Restricted Vocabulary Approaches



Carnegie Mellon

Restricted Vocabulary Approaches



Carnegie Mellon

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

GolgB1	GPP130
Integral to membrane; Golgi membrane; Golgi stack;	Integral to membrane; Golgi cis-face; Golgi lumen; endocytotic transport vesicle

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

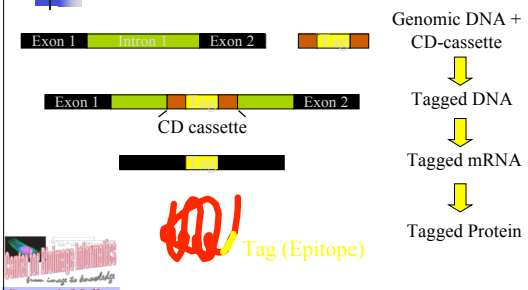
- Immunofluorescence
  - "primary" antibody against the target,
  - "secondary" antibody against the "primary" and conjugated with a fluorescent probe
  - Fixed-cells only
- Gene/cDNA-tagging
  - merge DNA coding for a naturally fluorescent protein (or vital probe binding sequence) with coding sequence 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

Carnegie Mellon

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



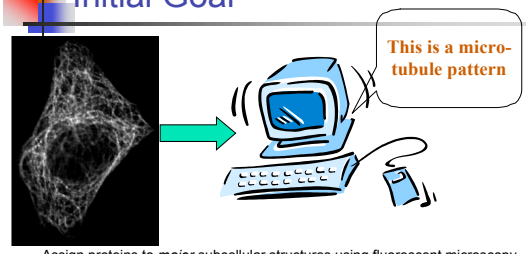
Carnegie Mellon

Automated Interpretation

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

Carnegie Mellon

Initial Goal

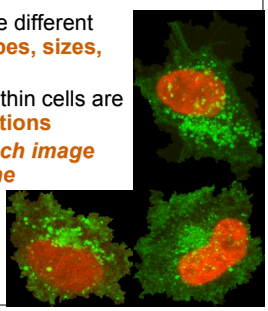


Assign proteins to *major* subcellular structures using fluorescent microscopy

Carnegie Mellon

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***



Carnegie Mellon

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

Carnegie Mellon

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

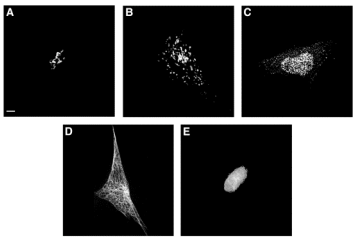
- We have collected datasets of fluorescence microscope images depicting the subcellular location patterns of a number of proteins in three different cell lines
- Available at http://murphylab.web.cmu.edu

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

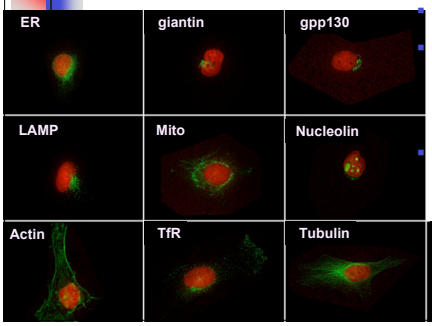
- Single color staining for specific protein
- Three 2D slices acquired and numerically deconvolved to yield one in focus 2D slice



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

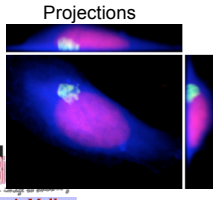
Microscope Datasets for Subcellular Location

- 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

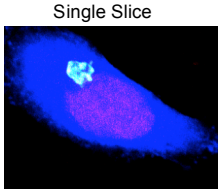
Example Image: 3D HeLa

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

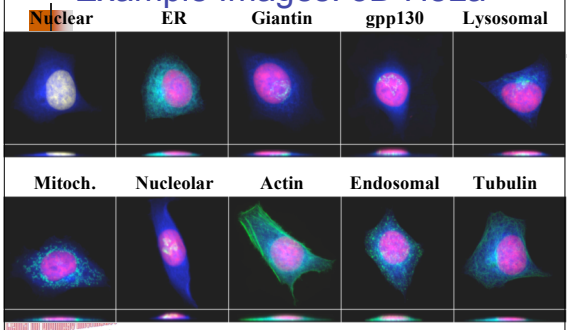
Projections



Single Slice



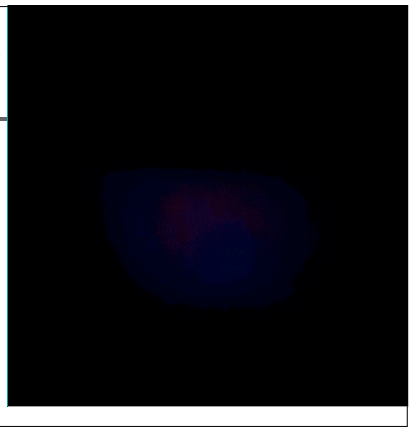
Example Images: 3D HeLa



Carnegie Mellon

3D HeLa

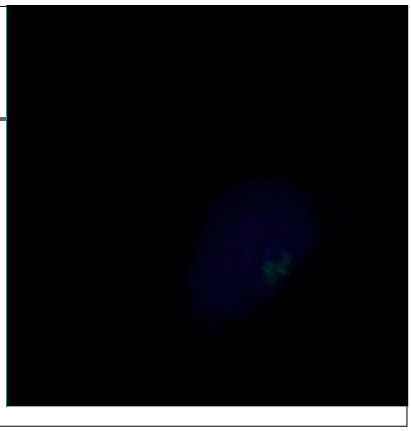
- 2D slices (from bottom to top) for cell labeled for **transferrin receptor** (primarily in endosomes)



Carnegie Mellon

3D HeLa

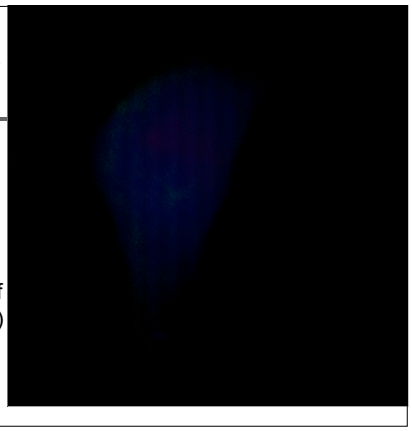
- 2D slices (from bottom to top) for cell labeled for **giantin** (primarily in Golgi)



Carnegie Mellon

3D HeLa

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



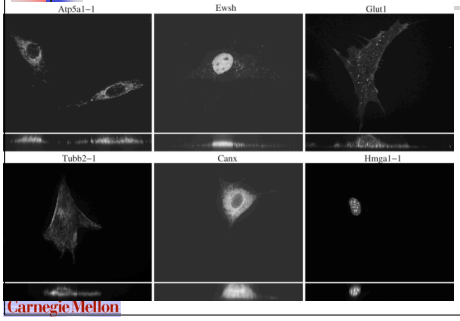
Carnegie Mellon

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

Carnegie Mellon

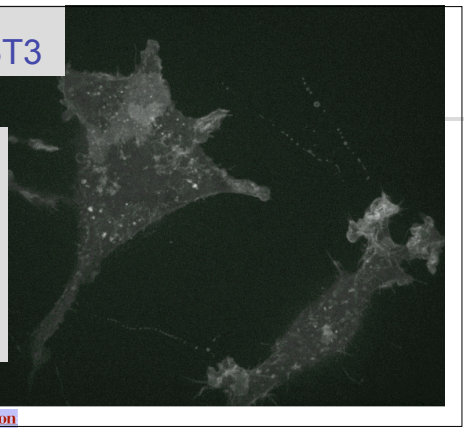
Example Images: 3D 3T3



Thirty slices acquired by spinning disk confocal microscope

Carnegie Mellon

2Dt 3T3



2D slices over time for cell labeled for **Glut1** (membrane protein on surface and in vesicles)

Carnegie Mellon

Other subcellular location projects

- O'Shea group - Yeast
  - GFP-tagged cDNAs
  - GFP and DNA images with some additional markers
- Pepperkok group - human (MCF7 cells)
  - GFP-tagged cDNAs
  - GFP and DNA images
- Uhlen group (Protein Atlas) - human
  - Immunohistochemistry with monospecific antibodies
  - DAB and hematoxylin images
  - Fixed tissues
- Schubert group (MELK technology)
  - Cycles of immunofluorescence, imaging and bleaching
  - Fixed tissues

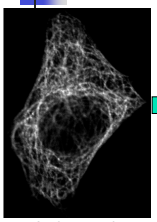
Carnegie Mellon

Feature Extraction for Subcellular Pattern Analysis

Center for Bioimage Informatics

Carnegie Mellon

Goal

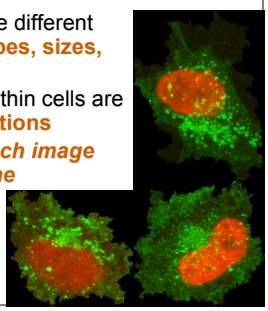


Assign proteins to *major* subcellular structures using fluorescent microscopy

Carnegie Mellon

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***



Carnegie Mellon

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

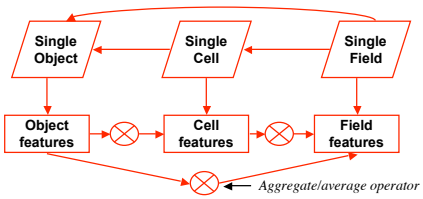
Carnegie Mellon

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 **S**ubcellular **L**ocation **F**eatures (**SLF**) and defined in sets (e.g., SLF1)

Carnegie Mellon

Feature levels and granularity



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

Carnegie Mellon

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)

Carnegie Mellon

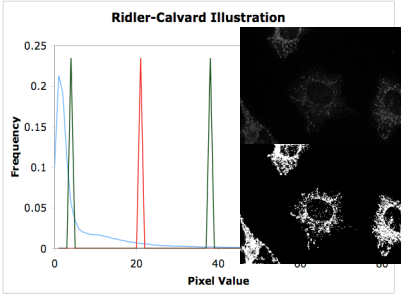
Ridler-Calvard Method

- Find threshold that is equidistant from the average intensity of pixels below and above it
- Ridler, T.W. and Calvard, S. (1978) Picture thresholding using an iterative selection method. *IEEE Transactions on Systems, Man, and Cybernetics* 8:630-632.

Carnegie Mellon

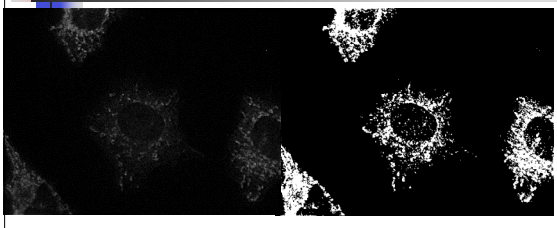
Ridler-Calvard Method

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



Carnegie Mellon

Ridler-Calvard Method



original thresholded

Carnegie Mellon

Otsu Method

- Find threshold to minimize the variances of the pixels below and above it
- Otsu, N., (1979) A Threshold Selection Method from Gray-Level Histograms, *IEEE Transactions on Systems, Man, and Cybernetics*, 9:62-66.

Carnegie Mellon

Adaptive Thresholding

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

Carnegie Mellon

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

Carnegie Mellon

Object finding

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

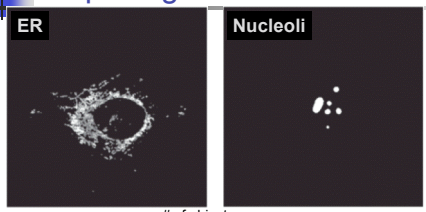
Carnegie Mellon

2D Features Morphological Features

SLF No.	Description
SLF1.1	The number of fluorescent objects in the image
SLF1.2	The Euler number of the image
SLF1.3	The average number of above-threshold pixels per object
SLF1.4	The variance of the number of above-threshold pixels per object
SLF1.5	The ratio of the size of the largest object to the smallest
SLF1.6	The average object distance to the cellular center of fluorescence(COF)
SLF1.7	The variance of object distances from the COF
SLF1.8	The ratio of the largest to the smallest object to COF distance

Carnegie Mellon

2D Features Morphological Features



108	# of objects	6
83	Average size of objects	232
31	Average distance to COF	4

Any of these features could be used to distinguish these two classes

Carnegie Mellon

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

DNA features (objects relative to DNA reference)

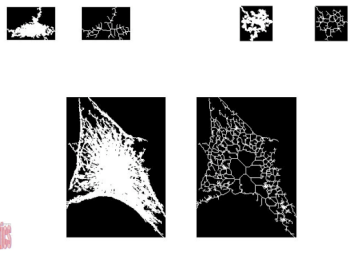
SLF No.	Description
SLF2.17	The average object distance from the COF of the DNA image
SLF2.18	The variance of object distances from the DNA COF
SLF2.19	The ratio of the largest to the smallest object to DNA COF distance
SLF2.20	The distance between the protein COF and the DNA COF
SLF2.21	The ratio of the area occupied by protein to that occupied by DNA
SLF2.22	The fraction of the protein fluorescence that co-localizes with DNA

2D Features Skeleton Features

Skeleton features

SLF No.	Description
SLF7.80	The average length of the morphological skeleton of objects
SLF7.81	The ratio of object skeleton length to the area of the convex hull of the skeleton, averaged over all objects
SLF7.82	The fraction of object pixels contained within the skeleton
SLF7.83	The fraction of object fluorescence contained within the skeleton
SLF7.84	The ratio of the number of branch points in the skeleton to the length of skeleton

Illustration – Skeleton



2D Features Edge Features

Edge features

SLF No.	Description
SLF1.9	The fraction of the non-zero pixels that are along an edge
SLF1.10	Measure of edge gradient intensity homogeneity
SLF1.11	Measure of edge direction homogeneity 1
SLF1.12	Measure of edge direction homogeneity 2
SLF1.13	Measure of edge direction difference

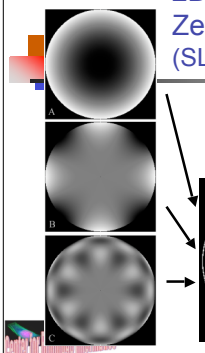
2D Features Hull Features

Convex hull (geometrical) features

SLF1.14	The fraction of the convex hull area occupied by protein fluorescence
SLF1.15	The roundness of the convex hull
SLF1.16	The eccentricity of the convex hull

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

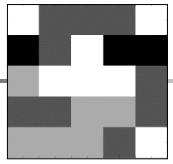
Carnegie Mellon

2D Features
Haralick Texture Features
(SLF7.66-7.78)

- 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

Carnegie Mellon

Example image with 4 gray levels



4	2	2	2	4
1	2	4	1	1
3	4	4	4	2
2	2	3	3	2
3	3	3	2	4

Co-occurrence Matrices

	1	2	3	4
1	0	2	1	3
2	2	4	4	4
3	1	4	2	2
4	2	3	2	2

	1	2	3	4
1	2	1	0	1
2	1	6	3	4
3	0	3	6	2
4	1	4	2	4

	1	2	3	4
1	0	1	0	3
2	1	4	3	3
3	0	3	4	1
4	3	3	1	2

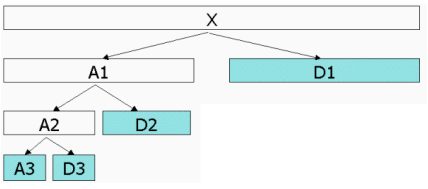
	1	2	3	4
1	0	3	0	1
2	3	0	4	4
3	0	4	0	3
4	1	4	3	2

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

Carnegie Mellon

Wavelet Transformation - 1D

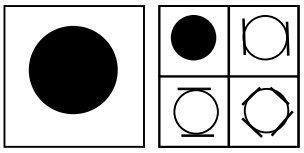


A: approximation (low frequency)
D: detail (high frequency)
X=A3+D3+D2+D1

Carnegie Mellon

2D Wavelets - intuition

- Apply some filter to detect edges (horizontal; vertical; diagonal)

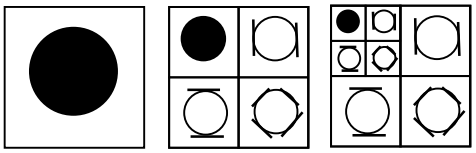


After Christos Faloutsos

Carnegie Mellon

2D Wavelets - intuition

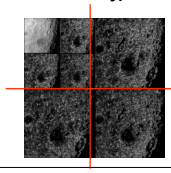
- Recurse



Slide courtesy of Christos Faloutsos

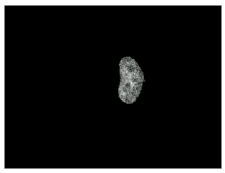
2D Wavelets - intuition

- Many wavelet basis functions (filters):
  - Haar
  - Daubechies (-4, -6, -20)
- http://www331.jpl.nasa.gov/public/wave.html

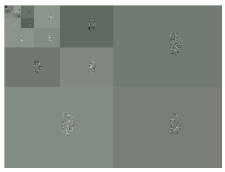


Slide courtesy of Christos Faloutsos

Daubechies D4 decomposition



Original image

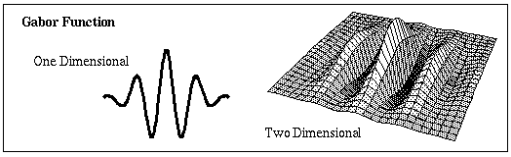


Wavelet Transformation

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-occurence 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

Index	Feature Description
SOF1.1	Number of pixels in object
SOF1.2	Distance between object Center of Fluorescence (COF) and DNA COF
SOF1.3	Fraction of object pixels overlapping with DNA
SOF1.4	A measure of eccentricity of the object
SOF1.5	Euler number of the object
SOF1.6	A measure of roundness of the object
SOF1.7	The length of the object's skeleton
SOF1.8	The ratio of skeleton length to the area of the convex hull of the skeleton
SOF1.9	The fraction of object pixels contained within the skeleton
SOF1.10	The fraction of object fluorescence contained within the skeleton
SOF1.11	The ratio of the number of branch points in skeleton to length of skeleton

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)

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: N_{level} by N_{level} matrix, Element 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

Carnegie Mellon

Image at t0

4	2	2	2	4
1	2	4	1	1
3	4	4	4	2
2	2	3	3	2
3	3	3	2	4

Image at t1

4	2	2	2	4
1	2	4	1	1
3	4	4	4	2
2	2	3	3	2
3	3	3	2	4

Temporal co-occurrence matrix (for image that does not change)

	1	2	3	4
1	3	0	0	0
2	0	9	0	0
3	0	0	6	0
4	0	0	0	7

Carnegie Mellon

Image at t0

4	2	2	2	4
1	2	4	1	1
3	4	4	4	2
2	2	3	3	2
3	3	3	2	4

Image at t1

2	1	4	4	3
1	4	2	3	3
2	3	3	2	2
4	4	2	2	3
2	4	2	1	4

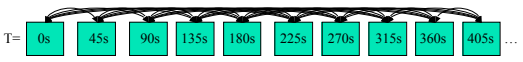
Temporal co-occurrence matrix (for image that changes)

	1	2	3	4
1	1	0	2	0
2	2	1	1	5
3	0	5	0	1
4	0	3	3	1

Carnegie Mellon

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*5*2=130 features

Carnegie Mellon

Subcellular Location Databases

Carnegie Mellon

PSLID: Protein Subcellular Location Image Database

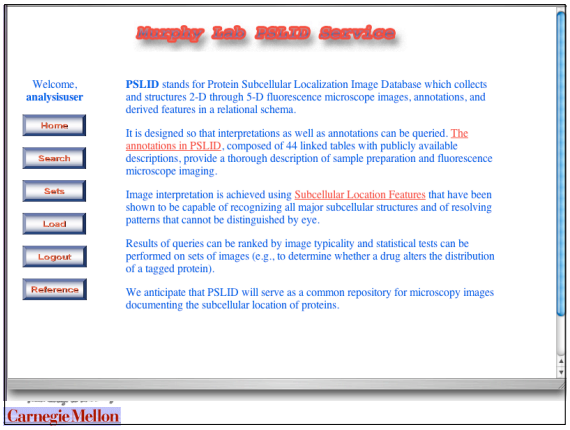
- A publicly accessible image database at http://murphylab.web.cmu.edu/services/PSLID
- A downloadable open source database system for creating local databases
  - Focused on subcellular pattern analysis
  - Subcellular Location Features integrated into database
  - Integrated comparison, classification,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

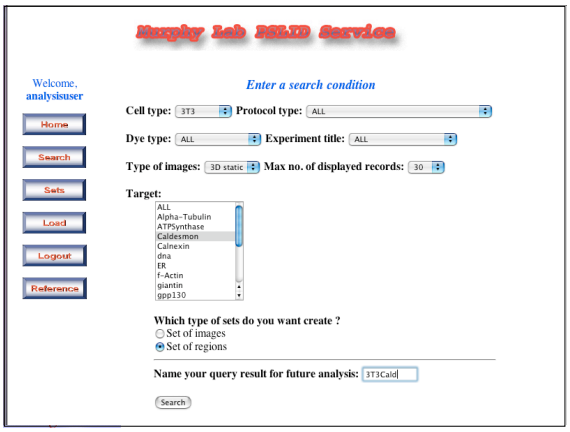
Carnegie Mellon

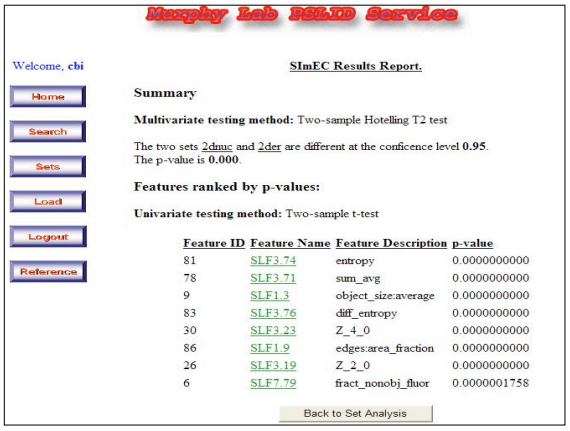
PSLID contents

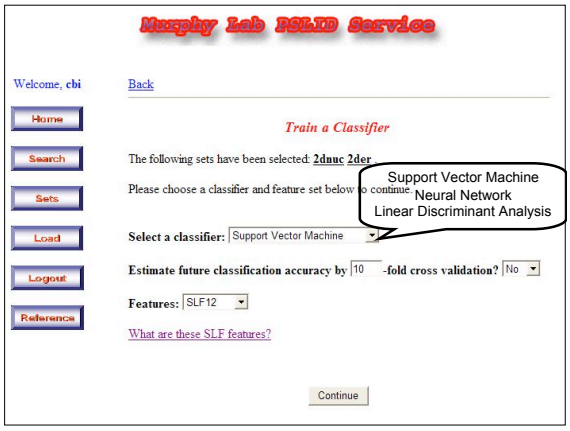
- ~1000 2D images of 10 patterns in HeLa cells
- ~1500 3D images of 23 patterns in HeLa cells
- ~2500 3D images of 90 patterns in 3T3 cells
- ~1000 4D images of 32 patterns in 3T3 cells
- More being added

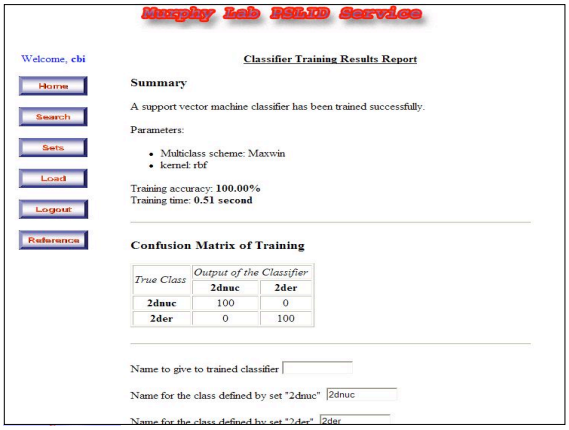
Carnegie Mellon











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*.

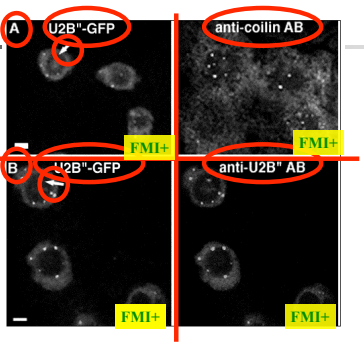
Carnegie Mellon

Overview: Image processing in SLIF

Segment into "panels"

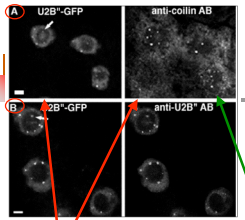
Detect & remove annotations

Classify panels



Carnegie Mellon

Overview: Text Processing in SLIF

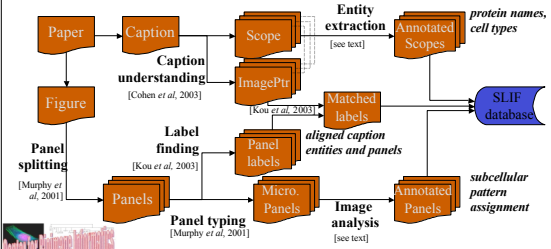


- Find *entity names* in text, and *panel labels* in text and the image.
- *Match* panels labels in text to panel labels on the image.
- *Associate* entity names to textual panel labels using *scoping* rules.

Figure 1. (A) Single confocal optical section of BY-2 cells expressing U2B 0-GFP, double labeled with GFP (left panel) and autoantibody against p80 coilin (right panel). Three nuclei are shown, and the bright GFP spots colocalize with bright foci of anti-coilin labeling. There is some labeling of the cytoplasm by anti-p80 coilin. (B) Single confocal optical section of BY-2 cells expressing U2B 0 -GFP, double labeled with GFP (left panel) and 4G3 antibody (right panel). Three nuclei are shown. Most coiled bodies in the nucleoplasm, but occasionally are seen in the nucleolus (arrows). All coiled bodies that contain U2B 0 also express the U2B 0-GFP fusion. Bars, 5 µm.

Carnegie Mellon

SLIF components



Carnegie Mellon

Carnegie Mellon

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)

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-95-099-03-MGO from the American Cancer Society, by grant 99-295 from the Rockefeller Brothers Fund Charles E. Culpeper Biomedical Pilot Initiative, by NSF grants BIR-9217091, MCB-8920118, 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.

http://murphylab.web.cmu.edu/publications

Review Articles

- Y. Hu and R. F. Murphy (2004). Automated Interpretation of Subcellular Patterns from Immunofluorescence Microscopy. *J. Immunol. Methods 290*:93-105.
- K. Huang and R. F. Murphy (2004). From Quantitative Microscopy to Automated Image Understanding. *J. Biomed. Optics 9*:893-912.
- R.F. Murphy (2005). Location Proteomics: A Systems Approach to Subcellular Location. *Biochem. Soc. Trans. 33*:535-538.
- R.F. Murphy (2005). Cytomics and Location Proteomics: Automated Interpretation of Subcellular Patterns in Fluorescence Microscope Images. *Cytometry 67A:1-3*.
- X. Chen, and R.F. Murphy (2006). Automated Interpretation of Protein Subcellular Location Patterns. *International Review of Cytology* 249:194-227.
- X. Chen, M. Velliste, and R.F. Murphy (2006). Automated Interpretation of Subcellular Patterns in Fluorescence Microscope Images for Location Proteomics. *Cytometry, in press*.

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

- M. V. Boland, M. K. Markey and R. F. Murphy (1997). Automated Classification of Cellular Protein Localization Patterns Obtained via Fluorescence Microscopy. *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 594-597.
- M. V. Boland, M. K. Markey and R. F. Murphy (1998). Automated Recognition of Patterns Characteristic of Subcellular Structures in Fluorescence Microscopy Images. *Cytometry 33*:366-375.

http://murphylab.web.cmu.edu/publications

http://murphylab.web.cmu.edu/publications

2D HeLa pattern classification (10 major patterns)

- R. F. Murphy, M. V. Boland and M. Velliste (2000). Towards a Systematics for Protein Subcellular Location: Quantitative Description of Protein Localization Patterns and Automated Analysis of Fluorescence Microscope Images. *Proc Int Conf Intell Syst Mol Biol 8*:251-259.
- M. V. Boland and R. F. Murphy (2001). A Neural Network Classifier Capable of Recognizing the Patterns of all Major Subcellular Structures in Fluorescence Microscope Images of HeLa Cells. *Bioinformatics 17*:1213-1223.

http://murphylab.web.cmu.edu/publications

3D HeLa pattern classification (11 major patterns)

- M. Velliste and R.F. Murphy (2002). Automated Determination of Protein Subcellular Locations from 3D Fluorescence Microscope Images. *Proceedings of the 2002 IEEE International Symposium on Biomedical Imaging (ISBI 2002)*, pp. 867-870.

http://murphylab.web.cmu.edu/publications

Improving features, feature selection, classification method

- R.F. Murphy, M. Velliste, and G. Porreca (2003). Robust Numerical Features for Description and Classification of Subcellular Location Patterns in Fluorescence Microscope Images. *J. VLSI Sig. Proc. 35:* 311-321.
- K. Huang, M. Velliste, and R. F. Murphy (2003). Feature reduction for improved recognition of subcellular location patterns in fluorescence microscope images. *Proc. SPIE 4962:* 307-318.

http://murphylab.web.cmu.edu/publications

Improving features, feature selection, classification method

- K. Huang and R.F. Murphy (2004). Boosting accuracy of automated classification of fluorescence microscope images for location proteomics. *BMC Bioinformatics 5:*78.
- X. Chen and R.F. Murphy (2004). Robust Classification of Subcellular Location Patterns in High Resolution 3D Fluorescence Microscope Images. *Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. *1632-1635.*

http://murphylab.web.cmu.edu/publications

Classification of multi-cell images

- K. Huang and R. F. Murphy (2004). Automated Classification of Subcellular Patterns in Multicell images without Segmentation into Single Cells. *Proceedings of the 2004 IEEE International Symposium on Biomedical Imaging (ISBI 2004),* pp. 1139-1142.
- S.-C. Chen, and R.F. Murphy (2006). A Graphical Model Approach to Automated Classification of Protein Subcellular Location Patterns in Multi-Cell Images. *BMC Bioinformatics* 7:90.

http://murphylab.web.cmu.edu/publications

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

- 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.
- X. Chen and R. F. Murphy (2005). Objective Clustering of Proteins Based on Subcellular Location Patterns. *Journal of Biomedicine and Biotechnology 2005*: 87-95.

http://murphylab.web.cmu.edu/publications

Subcellular Location Trees - Analysis of Location Mutants

- P. Nair, B.E. Schaub, K. Huang, X. Chen, R.F. Murphy, J.M. Griffith, H.J. Geuze, and J. Rohrer (2005). Characterization of the TGN Exit Signal of the human Mannose 6-Phosphate Uncovering Enzyme. *J. Cell Sci. 118:*2949-2956.

http://murphylab.web.cmu.edu/publications

PSLID - Protein Subcellular Location Image Database

- K. Huang, J. Lin, J.A. Gajnak, and R.F. Murphy (2002). Image Content-based Retrieval and Automated Interpretation of Fluorescence Microscope Images via the Protein Subcellular Location Image Database. *Proceedings of the 2002 IEEE International Symposium on Biomedical Imaging (ISBI 2002),* pp. 325-328.

http://murphylab.web.cmu.edu/publications

SLIF - Subcellular Location Image Finder

- R. F. Murphy, M. Velliste, J. Yao, and G. Porreca (2001). Searching Online Journals for Fluorescence Microscope Images Depicting Protein Subcellular Location Patterns. *Proceedings of the 2nd IEEE International Symposium on Bio-Informatics and Biomedical Engineering (BIBE 2001)*, pp. 119-128.
- R. F. Murphy, Z. Kou, J. Hua, M. Joffe, and W. W. Cohen (2004). Extracting and Structuring Subcellular Location Information from On-line Journal Articles: The Subcellular Location Image Finder. *Proceedings of the IASTED International Conference on Knowledge Sharing and Collaborative Engineering (KSCE 2004)*, pp. 109-114.