

Extraction of Subcellular Location Assertions and Models from Structured and Unstructured Sources

Robert F. Murphy
 Ray and Stephanie Lane Professor of Computational Biology
 Molecular Biosensors and Imaging Center, Departments of Biological Sciences, Biomedical Engineering and Machine Learning and

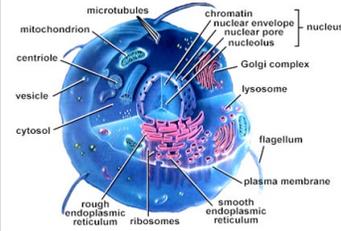


Center for BioImage Informatics
From Image to Knowledge

RAY AND STEPHANIE LANE
Center for Computational Biology
Carnegie Mellon

Central questions

- How many distinct locations within cells can proteins be found in? What are they?



Carnegie Mellon

Automated Interpretation

- Traditional analysis of fluorescence microscope images has occurred by visual inspection
- Our goal over the past twelve years has been to automate interpretation with the ultimate goal of fully automated learning of protein location from images

Carnegie Mellon

Approach

Combine fluorescence microscopy with pattern recognition techniques to automatically determine protein patterns

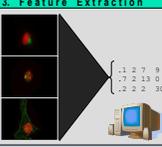
1. Image Acquisition



2. Image Processing

- Segmentation
- Denoising
- Deconvolution
- Signal unmixing

3. Feature Extraction



4. Feature Selection

Remove features

$$\begin{Bmatrix} -1 & 2 & 7 & 9 \\ -7 & 2 & 13 & 0 \\ -2 & 2 & 2 & 30 \end{Bmatrix} \rightarrow \begin{Bmatrix} -1 & 7 & 9 \\ -7 & 2 & 13 & 0 \end{Bmatrix}$$

or

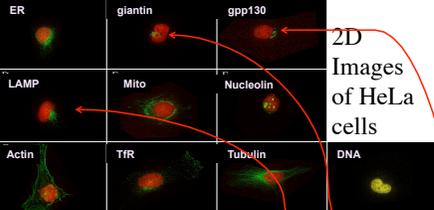
Combine features

$$\begin{Bmatrix} -1 & 2 & 7 & 9 \\ -7 & 2 & 13 & 0 \\ -2 & 2 & 2 & 30 \end{Bmatrix} \rightarrow \begin{Bmatrix} -6 & -3 & -7 \\ -2 & -8 & -4 \\ -5 & -7 & -1 \end{Bmatrix}$$

5. Classification



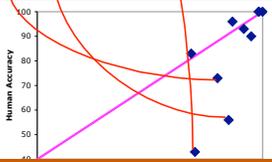
Carnegie Mellon



2D Images of HeLa cells

Murphy et al 2000; Boland & Murphy 2001; Murphy et al 2003; Huang & Murphy 2004

Subcellular Pattern Classification: Computer vs. Human



Even better results using multiresolution methods by Kovacevic group
Even better results for 3D images

Carnegie Mellon

yeastgfp.ucsf.edu
YEAST GFP FUSION LOCALIZATION DATABASE

[Advanced Query](#) < Go > [info](#)
[Quick Search](#) [faq](#)
[help](#)

Welcome to yeastgfp.ucsf.edu

The database of our global analysis of protein localization studies in the budding yeast, *S. cerevisiae*.

- quick case-insensitive searches of the database may be performed on yeastorf names (yaf001c) or gene names (TFC3)
- separate multiple orfs/genes with a space (e.g. yaf001c zvf1 bu02 etc.)
- more advanced searching and downloading can be done in Advanced Query
- GFP-tagged strains can be obtained from Invitrogen
- TAP-tagged strains can be obtained from Open Biosystems.
- more details available in [info](#) [faq](#) [help](#)

This web site supports Huh, et al., Nature 425, 686-691 (2003) pdf
 The visualization data presented here is published in Shalem-Balshani, et al., Nature 425, 737-741 (2003) pdf
 Detailed collection construction methods can be found in Hopson et al., Comp Funct Genom 6, 2-16 (2005). pdf

This research is the work of the laboratories of Erin O'Shea and Jonathan Weissman at the University of California San Francisco. Please direct comments, concerns, and questions to erin.oshea@gmail.com

© Copyright 2001 - 2006 University of California Regents. All rights reserved.

Annotations of Yeast GFP Fusion Localization Database

- Contains images of 4156 proteins (out of 6234 ORFs in all 16 yeast chromosomes).
- GFP tagged immediately before the stop codon of each ORF to minimize perturbation of protein expression.
- Annotations were done manually by two scorers and colocalization experiments were done for some cases using mRFP.
- Each protein is assigned one or more of 22 location categories.

Carnegie Mellon

Chen et al 2007

Classification of Yeast Subcellular Patterns

- Selected only those assigned to single unambiguous location class (21 classes)
- Trained classifier to recognize those classes
- 81% agreement with human classification**
- 94.5% agreement for high confidence assignments (without using colocalization!)**
- Examination of proteins for which methods disagree suggests machine classifier is correct in at least some cases



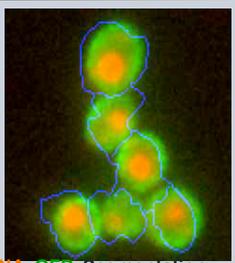

Carnegie Mellon Shann-Ching (Sam) Chen & Geoff Gordon

Example of Potentially Incorrect Label

ORF Name
YGR130C

UCSF Location
punctate_composite

Automated Prediction
cell_periphery (60.67%)
cytoplasm (30%)
ER (9.33%)



DNA GFP Segmentation

Carnegie Mellon

Graphical models for multi-cell images

- Cells with same location pattern are often close to each other.
- Considering *multiple cells* may improve the classification accuracy.
- Propose a *novel graphical model* to describe the relationship between cells such that the classification of a cell is influenced by other neighboring cells.

Carnegie Mellon

Given a multi-cell image

Connect cells if they are close enough (by d_{center}) (either in physical space or feature space)

Graph Construction

Inference by Prior Updating (PU)

For each cell, update the priors by the likelihoods of neighboring cells

Use the new priors and likelihood to calculate posterior probability and classify the cell

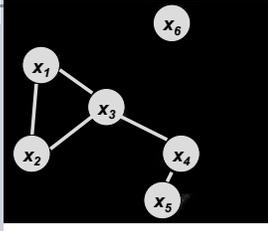
Iterate until no label changes
Calculate the new classification accuracy

Each cell is well-segmented

Each cell is a random variable

Given single-cell classifiers to provide likelihood for each cell

Base accuracy is calculated

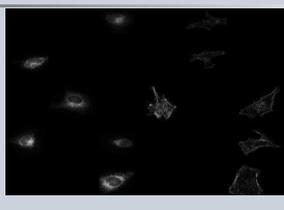


Measure accuracy improvement

Carnegie Mellon

Evaluating PU

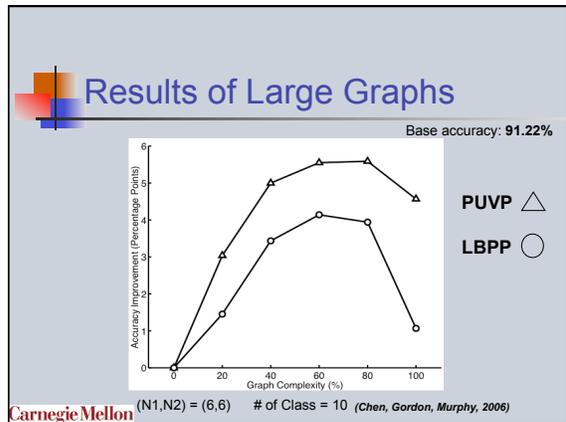
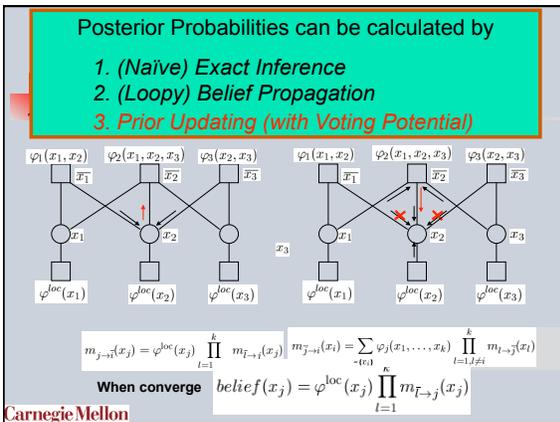
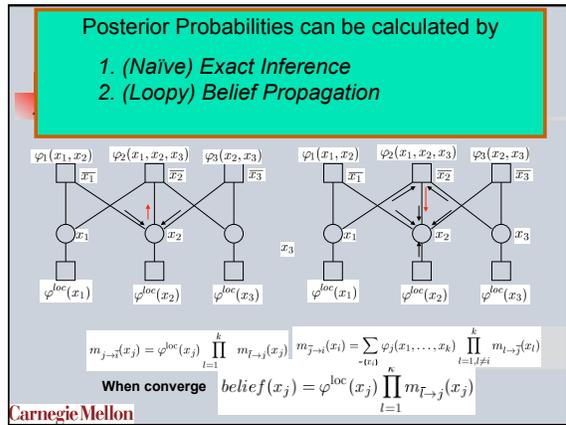
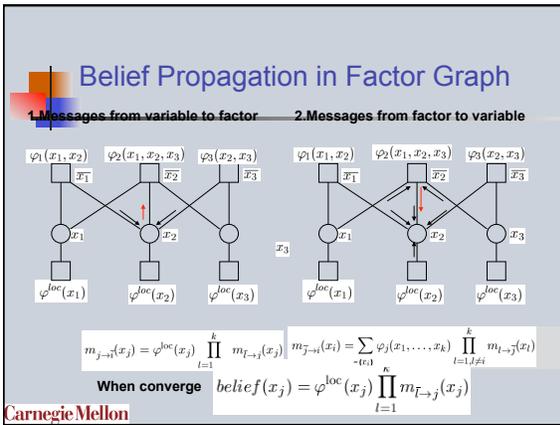
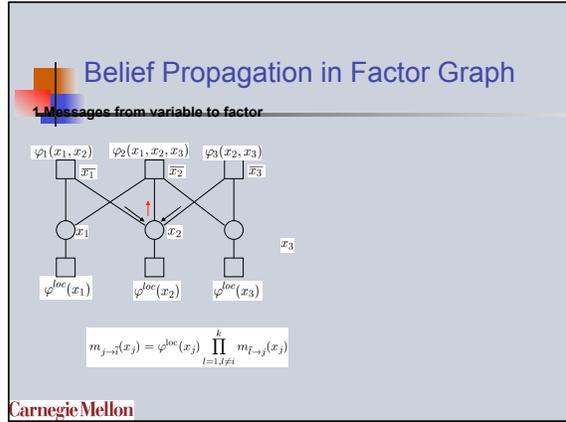
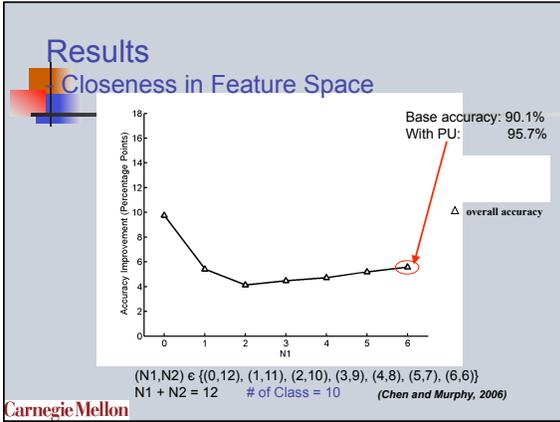
- Use the single-cell images in 10 class 2D HeLa data set to create synthetic multi-cell images
- Each cell is well-segmented
- Single-cell classifiers are trained
- Simulate fields containing only **two location patterns** in various proportions of cells

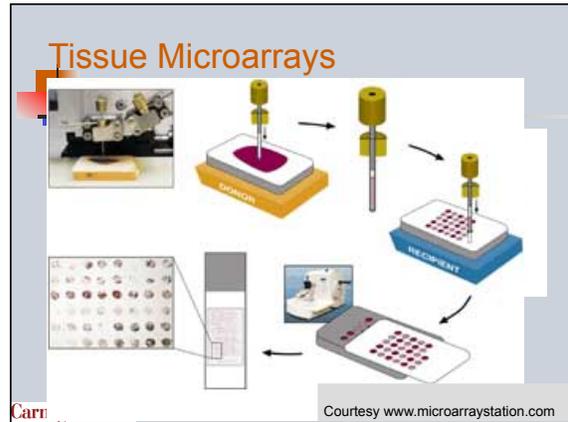
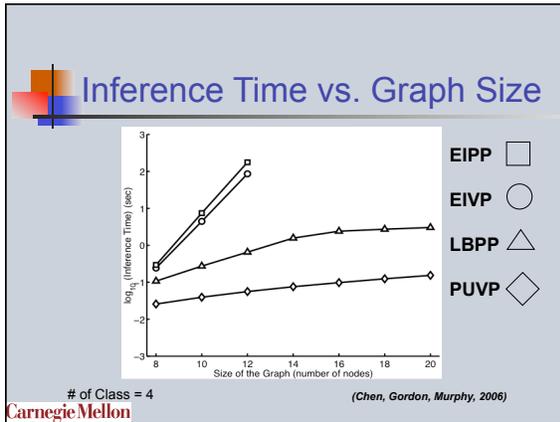


$(N1, N2) \in \{(0,12), (1,11), (2,10), (3,9), (4,8), (5,7), (6,6)\}$

$N1 + N2 = 12$ # of Class = 10

Carnegie Mellon





Human Protein Atlas

Salivary gland | Lateral ventricle wall | Nasopharynx

Prostate [CASP8]

Cell Type	Intensity	Quantity	Localization
Glandular cells	weak	~75%	cytoplasmic and/or membranous

Male, age 51 | Male, age 64 | Male, age 80

Brown color indicates presence of protein. Blue color shows cell nuclei. Image Usage Policy

[LIVER CARCINOMA](#)
[MALIGNANT GLIOMA](#)
[MELANOMA](#)
[MELANOCYTOBLASTOMA](#)
[MELANOCYTOBLASTOMA](#)
[MELANOCYTOBLASTOMA](#)

Courtesy www.proteinatlas.org

Test Dataset from Human Protein Atlas

- Selected 16 proteins from the Atlas
- Two each from all major organelles (class)
- ~45 tissue types for each class (e.g. liver, skin)
- Goal: Train classifier to recognize each subcellular pattern across all tissue types

Justin Newberg

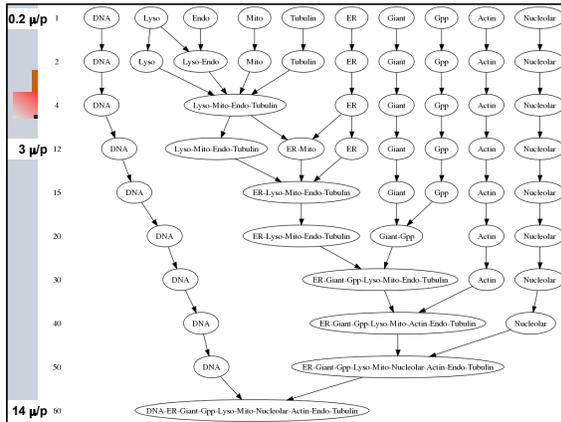
Subcellular Pattern Classification over 45 tissues

	ER	Cyto	Endo	Golgi	Lyso	Mito	Nucleolus	Nucleus
ER (53)	100	0	0	0	0	0	0	0
Cyto (21)	4.8	76.2	0	0	14.3	4.8	0	0
Endo (2)	0	0	100	0	0	0	0	0
Golgi (88)	1.1	0	0	98.9	0	0	0	0
Lyso (52)	0	1.9	0	0	96.2	0	1.9	0
Mito (64)	0	0	0	0	0	98.4	1.6	0
Nucleolus (94)	0	0	0	2.1	2.1	1.1	94.7	0
Nucleus (78)	0	0	0	0	0	0	0	100

Accuracy for 50% of images with highest confidence: 97%

Image resolution and pattern discrimination

- What effect does image resolution have on our ability to discriminate subcellular patterns?
- Start from high-resolution images of HeLa cells and downsample
- Determine how accuracy decreases
- Determine which patterns can still be determined (merge patterns to achieve original accuracy)



Huang et al 2002; Huang et al 2007

PSLID: Protein Subcellular Location Image Database

- First public domain software for automated analysis of subcellular patterns in images from large scale microscopy/high content screening experiments
- Publicly accessible image database at <http://pslid.cbi.cmu.edu>
 - Version 3 released February 2, 2007
 - 2D and 3D images (single cell regions defined)
 - Two cell types, HeLa and 3T3
 - Over 120,000 images/3000 unique fields/14,000 cells
 - 111 classes; 55 known proteins; 11 targeting mutants of one protein
 - Programmatic search via URL
 - Adding yeast and tissue images
 - Version 4 to be released May 9, 2008

CarnegieMellon Protein Subcellular Location Image Database

Supervised vs. Unsupervised Learning

- This work demonstrates the feasibility of using classification methods to assign all proteins to known major classes
- Similar approach being taken in location prediction from sequence
- Do we know all locations? Are assignments to major classes enough?
- Need approach to discover classes

CarnegieMellon

Location Proteomics

Group ~90 tagged clones by pattern

CarnegieMellon

Chen et al 2003; Chen and Murphy 2005

Z-scored Euclidean Dist

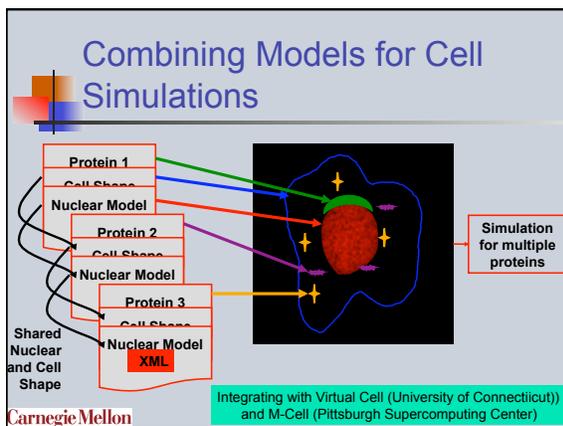
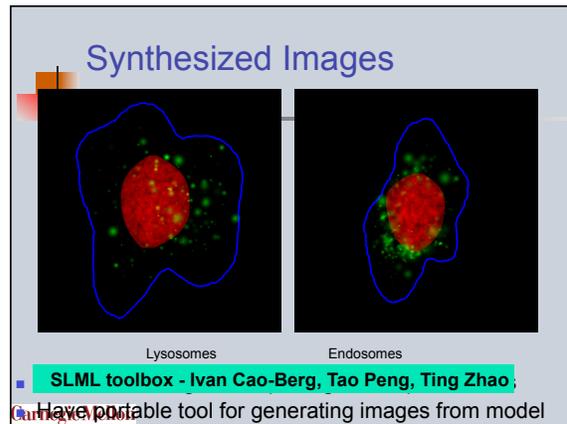
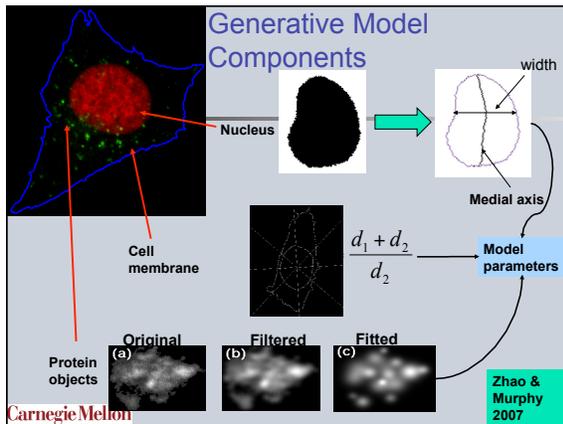
- How?
- Features can be used to measure similarity of protein patterns
- Build **Subcellular Location Tree**
- Have multiple images per protein
- Sample repeatedly from available images, build cluster tree for each subsample, and form consensus tree

Xiang Chen

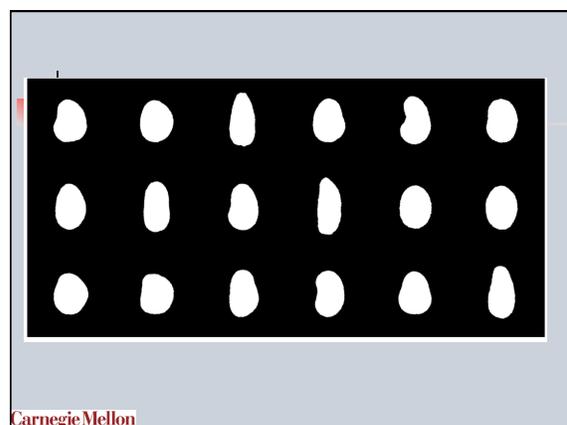
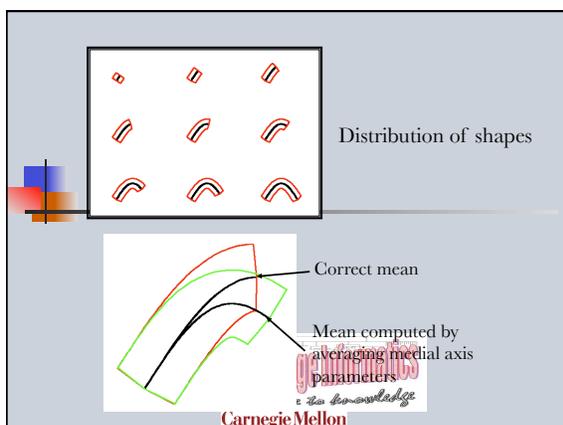
Subcellular Location Families and Generative Models

- Rather than using words (e.g., GO terms) to describe location patterns, can make entries in protein databases that give its Subcellular Location Family - a specific node in a Subcellular Location Tree
- Provides necessary resolution that is difficult to obtain with words
- How do we communicate patterns: Use generative models learned from images to capture **pattern** and **variation** in pattern

CarnegieMellon



- ### Diffeomorphic analysis of nuclear shape (w/ Gustavo Rohde)
- Use of medial axis model assumes parameters lie in Euclidean space
 - Actual shape space is non-Euclidean
 - Can use distance between shapes instead
- Carnegie Mellon



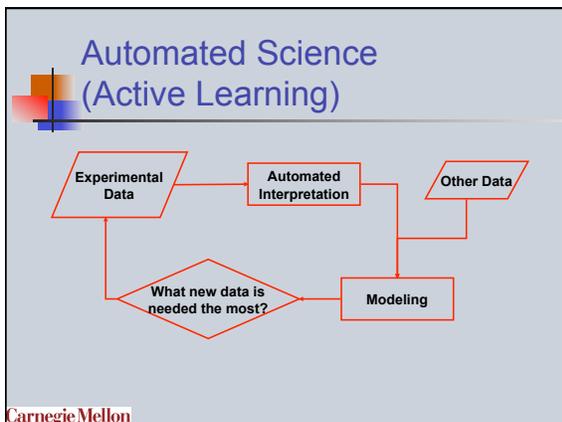
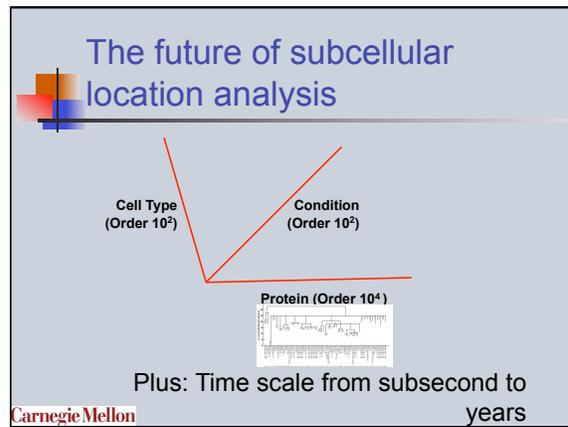
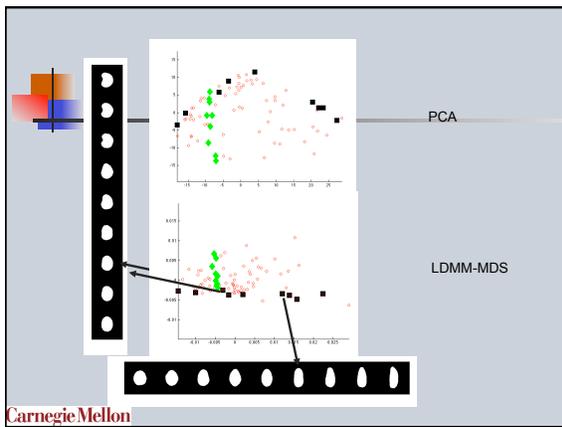
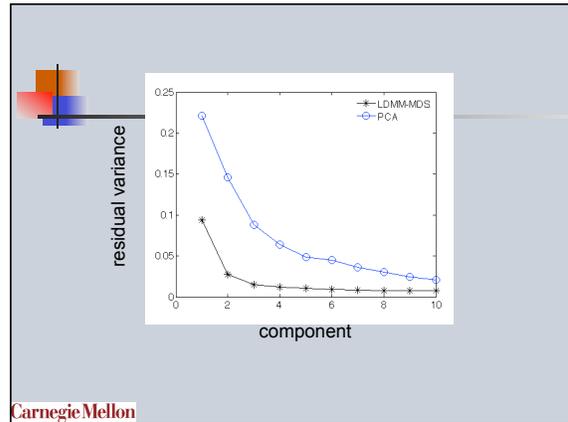
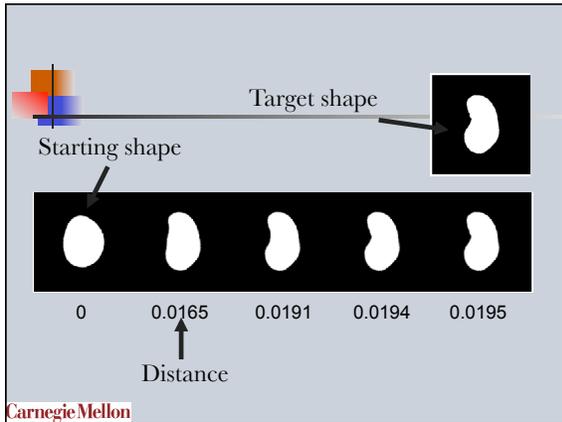


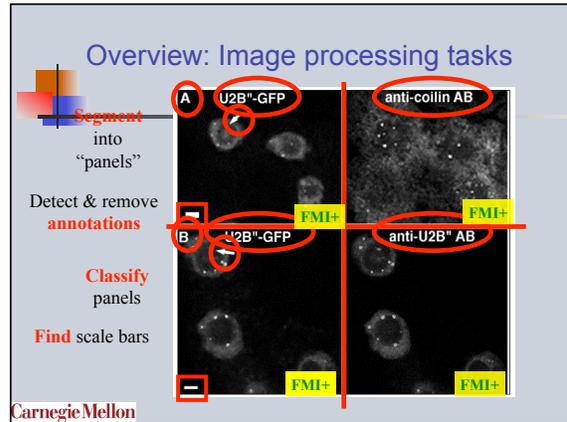
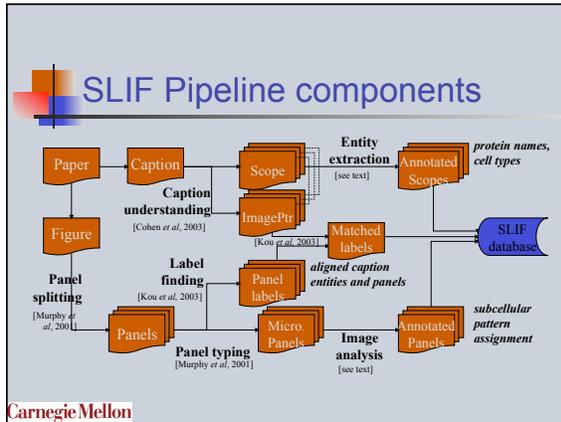
Image Content-based Retrieval and Interpretation of Micrographs from On-line Journal Articles

The Subcellular Location Image Finder

Center for Bioimage Informatics
Carnegie Mellon University

William Cohen Carnegie Mellon Eric Xing

Carnegie Mellon



Overview: Text processing tasks

- 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-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 are 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. Movement of Coiled Bodies Vol. 10, July 1999 2299

CarnegieMellon

Subcellular Location Image Finder

WELCOME TO

SLIF

Subcellular Location Image Finder

SLIF (Subcellular Location Image Finder) automatically extracts information about protein subcellular locations from figure-caption pairs in biological literature. SLIF separates figures into panels and decides which panels contain fluorescence microscope images (FMI). It applies image processing methods to analyze the FMI and extract a quantitative description of the localization patterns they display. The associated captions are also processed to identify which portions of the caption refer to which panels and to identify the names of proteins contained in the captions. The results of this analysis are stored in the SLIF database.

Our long-term goal is to develop a large library of annotated and analyzed fluorescence microscope images, in order to support data-mining.

PNAS, version 3.0

The current version of the database contains records for 15180 papers from volumes 94-99 of the Proceedings of the National Academy of Sciences (USA), generously made available by the Academy for demonstration purposes.

BioMed Central, version 1.0

Due for release 22 January 2007.

Pubmed Central, version 1.0

The database will be expanded shortly to include all open access articles in Pubmed Central, including BMC papers but not PNAS papers (approximately 45,000 as of 31 December 2007).

A service of the Robert F. Murphy laboratory
Departments of Biological Sciences, Biomedical Engineering, and Machine Learning
and Center for Biomagnetic Informatics
Carnegie Mellon University, Pittsburgh, Pennsylvania, U.S.A.

Murphy Lab SLIF Service

Searched for: **Captions with the words tubulin.**

Results 1 - 10 of 1029

Caption	Figure	Panel	Protein	Cell	Location	μm/pixel
Fluorescently labeled MTs in PKC cells. (A) Fluorescent MTs in a living PKC cell. The nucleus excludes MTs and tubulin monomers, creating a region in the center of the cell with little or no background fluorescence. In the cell shown here, the centrosome is in the center of this darkened region, making it easier to see single MTs. (Bar = 3 μm.) (B) Diagram depicting the geometrical relationship among the ventral cortex, the centrosome, MTs, and the nucleus as seen in side view; the components are not necessarily drawn to scale. The dashed lines show the approximate region of focus. Click to view paper			PKC	N/A	TIR	9.2

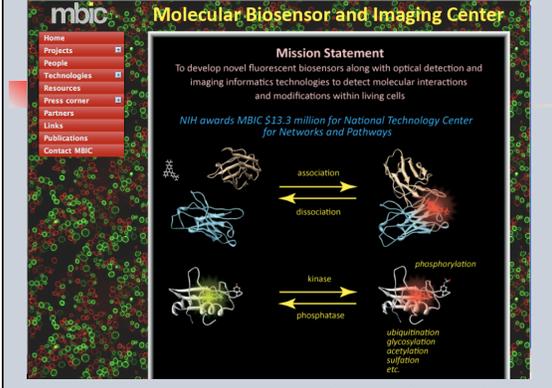
Conclusions

- Computers better than people at recognizing complex subcellular patterns
- Colocalization not necessary and not appropriate for assigning proteins to subcellular locations
- Generative models can be built directly from data to summarize results and make predictions
- Active machine learning methods needed/ appropriate to exploring complex multivariate spaces

CarnegieMellon

Acknowledgments

- Past and Present Students and Postdocs
 - Michael Boland (Hopkins), Mia Markey (UT Austin), Gregory Porreca (Harvard), Meel Velliste (U Pitt), Kai Huang, Xiang Chen (Yale), Yanhua Hu, Juchang Hua, Ting Zhao (HHMI Janelia Farm), Shann-Ching Chen (Scripps), Elvira Garcia Osuna (CMU), Justin Newberg, Estelle Glory, Tao Peng, Luis Coelho
- Funding
 - NSF, NIH, Commonwealth of Pennsylvania
- Collaborators/Consultants
 - David Casasent, Simon Watkins, **Jon Jarvik**, **Peter Bergert**, Jack Rohrer, Tom Mitchell, Christos Faloutsos, **Jelena Kovacevic**, **William Cohen**, **Geoff Gordon**, B. S. Manjunath, Ambuj Singh, Les Loew, Ion Moraru, Jim Schaff, Paul Campagnola, **Gustavo Rohde**
- Slides/Data
 - Jelena Kovacevic, Les Loew, Badri Roysam
- Centers
 - Molecular Biosensors and Imaging Center - TCNP (Waggoner)
 - National Center for Integrative Biomedical Informatics - NCBI (Athey)

Molecular Biosensor and Imaging Center

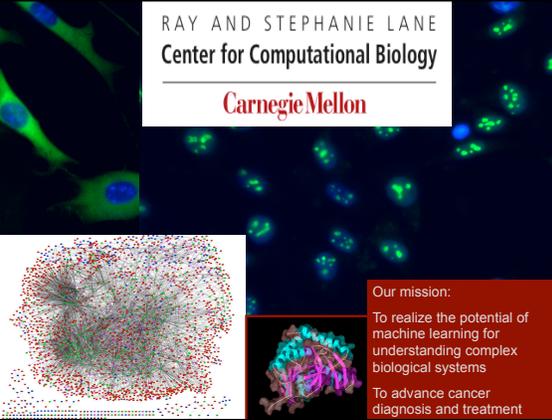
Mission Statement
To develop novel fluorescent biosensors along with optical detection and imaging informatics technologies to detect molecular interactions and modifications within living cells

NIH awards MBIC \$13.3 million for National Technology Center for Networks and Pathways

The diagram illustrates molecular interactions and modifications:

- Association and dissociation of protein complexes.
- Phosphorylation and dephosphorylation (catalyzed by kinase and phosphatase).
- Other modifications: ubiquitination, glycosylation, acetylation, sulfation, etc.

Carnegie Mellon Alan Waggoner (CMU) and Simon Watkins (Pitt)



RAY AND STEPHANIE LANE
Center for Computational Biology
Carnegie Mellon

Our mission:
To realize the potential of machine learning for understanding complex biological systems
To advance cancer diagnosis and treatment

