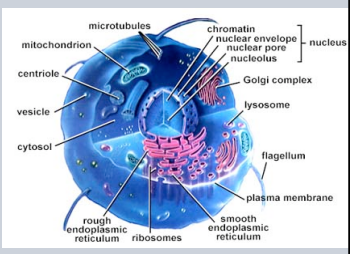
Extraction of Subcellular Location Assertions and Models from Structured and Unstructured Sources

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
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

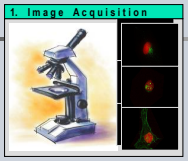
- Traditional analysis of fluorescence microscope images has occurred by visual inspection
- Our goal over the past twelve years has to been to automate interpretation with the ultimate goal of fully automated learning of protein location from images

Carnegie Mellon

Approach

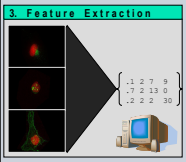
Combine fluorescence microscopy with pattern recognition techniques to automatically determine protein patterns

1. Image Acquisition

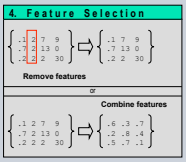


2. Image Processing
- Segmentation
- Denoising
- Deconvolution
- Signal unmixing

3. Feature Extraction

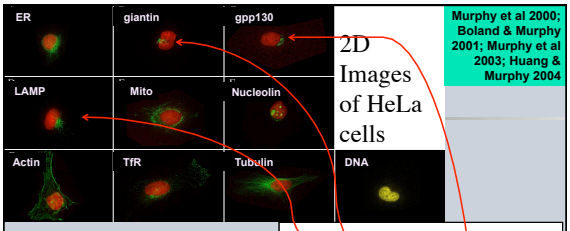


4. Feature Selection



5. Classification

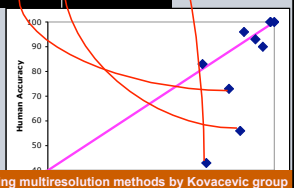
Carnegie Mellon



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

2D Images of HeLa cells

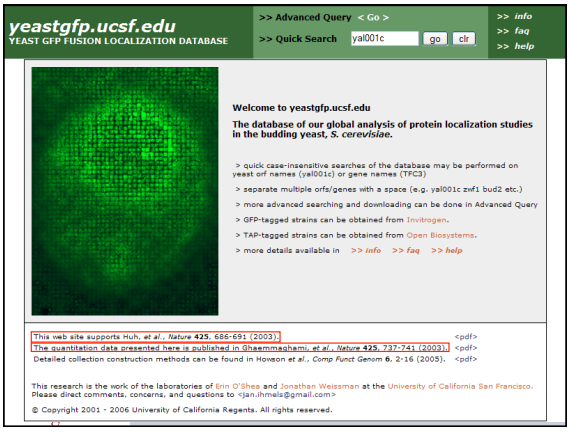
Subcellular Pattern Classification: Computer vs. Human



Even better results using multiresolution methods by Kovacevic group

Even better results for 3D images

Carnegie Mellon



Annotations of Yeast GFP Fusion Localization Database

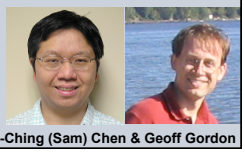
- Contains images of 4156 proteins (out of 6234 ORFs in all 16 yeast chromosomes).
- GFP tagged immediately before the stop codon of each ORF to minimize perturbation of protein expression.
- Annotations were done manually by two scorers and co-localization experiments were done for some cases using 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



Shann-Ching (Sam) Chen & Geoff Gordon

Carnegie Mellon

Example of Potentially Incorrect Label

ORF Name
YGR130C

UCSF Location
punctate_composite

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

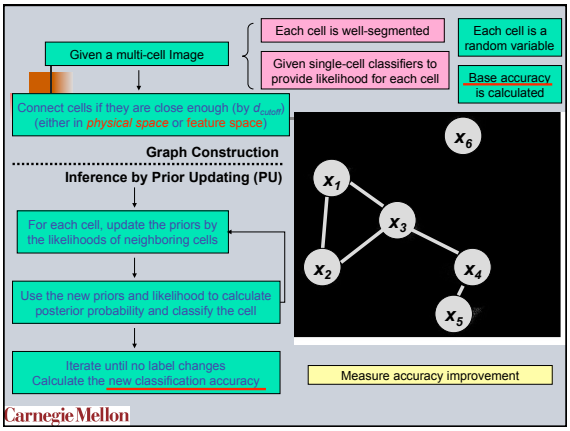
DNA GFP Segmentation

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.

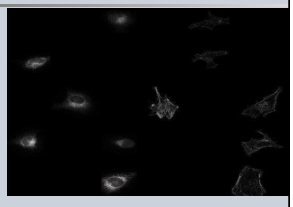
Carnegie Mellon



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

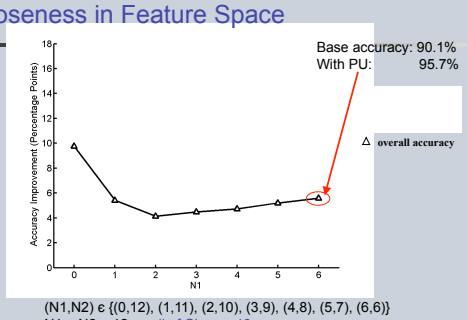


(N1,N2) ∈ {(0,12), (1,11), (2,10), (3,9), (4,8), (5,7), (6,6)}
N1 + N2 = 12 # of Class = 10

Carnegie Mellon

Results

Closeness in Feature Space

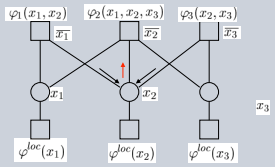


(N1,N2) ϵ {(0,12), (1,11), (2,10), (3,9), (4,8), (5,7), (6,6)}
N1 + N2 = 12 # of Class = 10 *(Chen and Murphy, 2006)*

Carnegie Mellon

Belief Propagation in Factor Graph

1.Messages from variable to factor

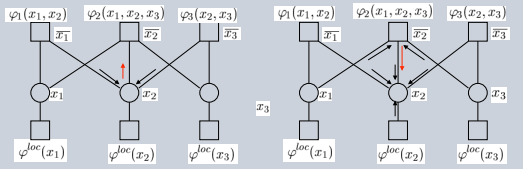


$$m_{j\to i}(x_j) = \varphi^{loc}(x_j) \prod_{l=1, l\neq i}^{k} m_{l\to j}(x_j)$$

Carnegie Mellon

Belief Propagation in Factor Graph

1.Messages from variable to factor 2.Messages from factor to variable



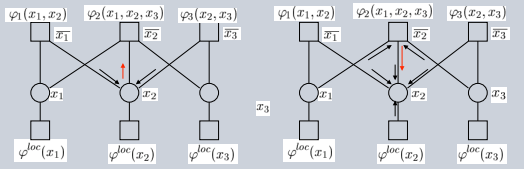
$$m_{j\to i}(x_j) = \varphi^{loc}(x_j) \prod_{l=1}^{k} m_{l\to j}(x_j) \quad m_{\vec{j}\to i}(x_i) = \sum_{\sim\{x_i\}} \varphi_j(x_1,\ldots,x_k) \prod_{l=1, l\neq i}^{k} m_{l\to \vec{j}}(x_l)$$

When converge $belief(x_j) = \varphi^{loc}(x_j) \prod_{l=1}^{\kappa} m_{\vec{l}\to j}(x_j)$

Carnegie Mellon

Posterior Probabilities can be calculated by

1. (Naïve) Exact Inference
2. (Loopy) Belief Propagation



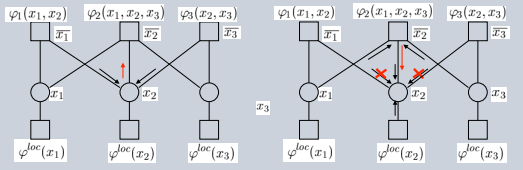
$$m_{j\to i}(x_j) = \varphi^{loc}(x_j) \prod_{l=1}^{k} m_{l\to j}(x_j) \quad m_{\vec{j}\to i}(x_i) = \sum_{\sim\{x_i\}} \varphi_j(x_1,\ldots,x_k) \prod_{l=1, l\neq i}^{k} m_{l\to \vec{j}}(x_l)$$

When converge $belief(x_j) = \varphi^{loc}(x_j) \prod_{l=1}^{\kappa} m_{\vec{l}\to j}(x_j)$

Carnegie Mellon

Posterior Probabilities can be calculated by

1. (Naïve) Exact Inference
2. (Loopy) Belief Propagation
3. Prior Updating (with Voting Potential)



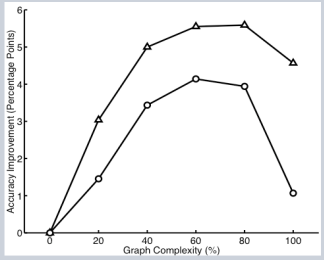
$$m_{j\to i}(x_j) = \varphi^{loc}(x_j) \prod_{l=1}^{k} m_{l\to j}(x_j) \quad m_{\vec{j}\to i}(x_i) = \sum_{\sim\{x_i\}} \varphi_j(x_1,\ldots,x_k) \prod_{l=1, l\neq i}^{k} m_{l\to \vec{j}}(x_l)$$

When converge $belief(x_j) = \varphi^{loc}(x_j) \prod_{l=1}^{\kappa} m_{\vec{l}\to j}(x_j)$

Carnegie Mellon

Results of Large Graphs

Base accuracy: **91.22%**

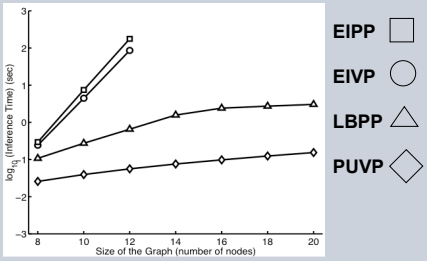


PUVP △
LBPP ○

(N1,N2) = (6,6) # of Class = 10 *(Chen, Gordon, Murphy, 2006)*

Carnegie Mellon

Inference Time vs. Graph Size

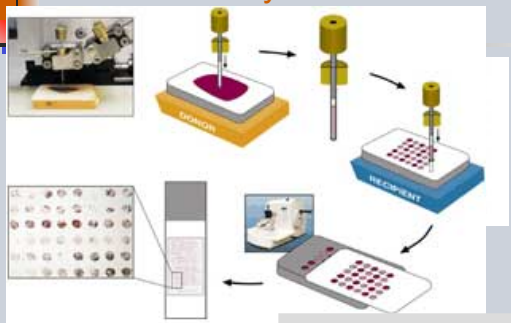


# of Class = 4

(Chen, Gordon, Murphy, 2006)

Carnegie Mellon

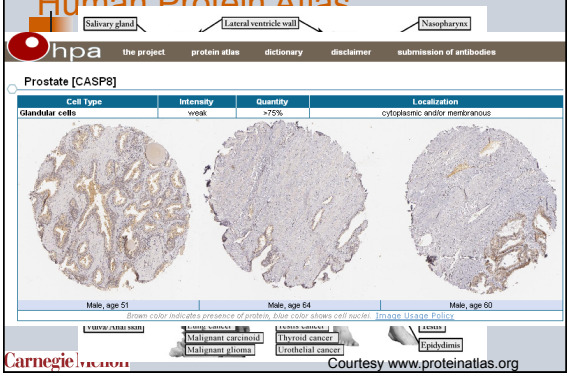
Tissue Microarrays



Courtesy www.microarraystation.com

Carnegie Mellon

Human Protein Atlas

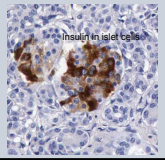


Courtesy www.proteinatlas.org

Carnegie Mellon

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

Carnegie Mellon

Subcellular Pattern Classification over 45 tissues

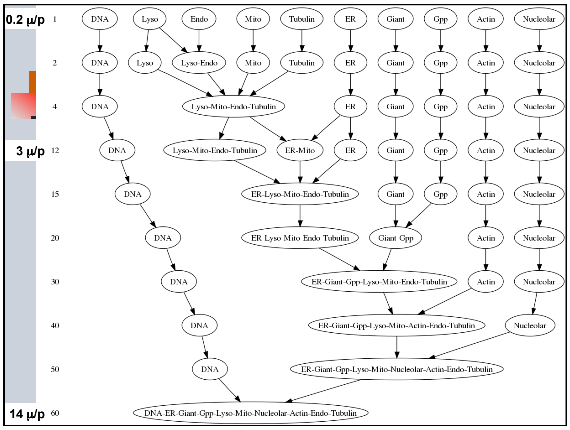
	Prediction							
	ER	Cyto	Endo	Golgi	Lyso	Mito	Nucleolus	Nucleus
ER (53)	**100**	0	0	0	0	0	0	0
Cyto (21)	4.8	**76.2**	0	0	14.3	4.8	0	0
Endo (2)	0	0	**100**	0	0	0	0	0
Golgi (88)	1.1	0	0	**98.9**	0	0	0	0
Lyso (52)	0	1.9	0	0	**96.2**	0	1.9	0
Mito (64)	0	0	0	0	0	**98.4**	1.6	0
Nucleolus (94)	0	0	0	2.1	2.1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

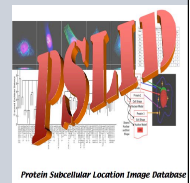
Carnegie Mellon



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



Carnegie Mellon

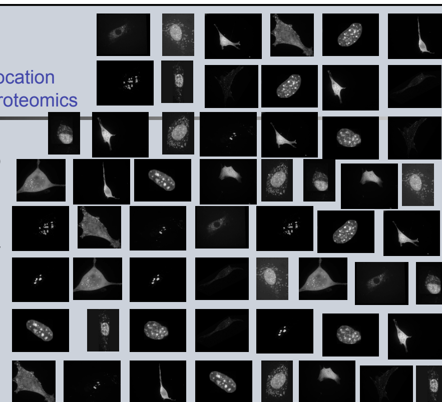
Supervised vs. Unsupervised Learning

- 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

Carnegie Mellon

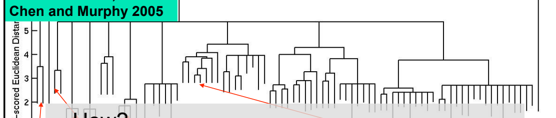
Location Proteomics

Group ~90 tagged clones by pattern



Carnegie Mellon

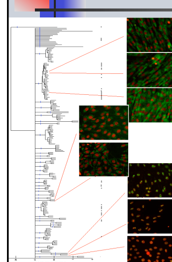
Chen et al 2003; Chen and Murphy 2005



- 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

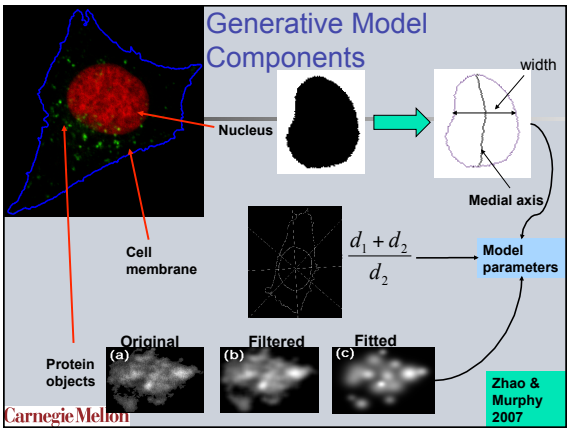
Xiang Chen

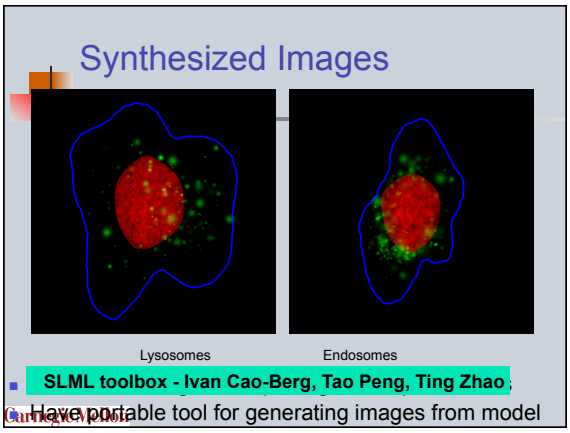
Subcellular Location Families and Generative Models

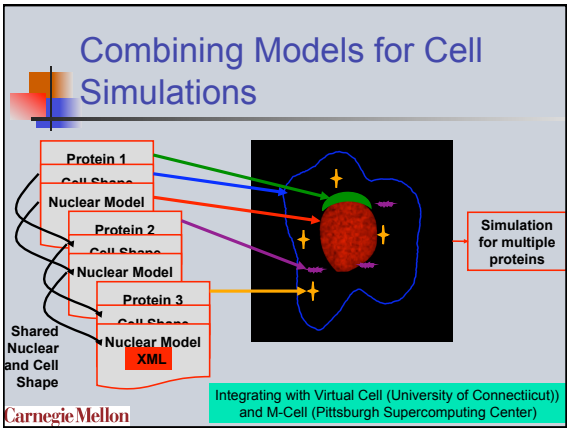


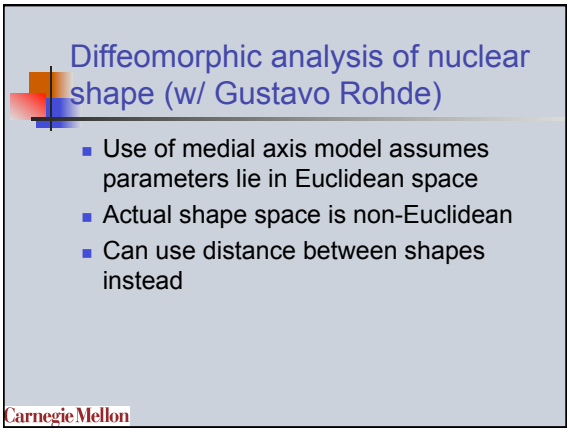
- Rather than using words (e.g., GO terms) to describe location patterns, can make entries in protein databases that give its Subcellular Location Family - a specific node in a Subcellular Location Tree
- Provides necessary resolution that is difficult to obtain with words
- How do we communicate patterns: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

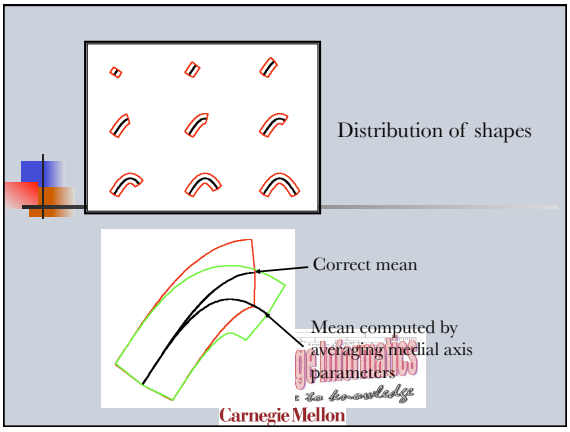
Carnegie Mellon

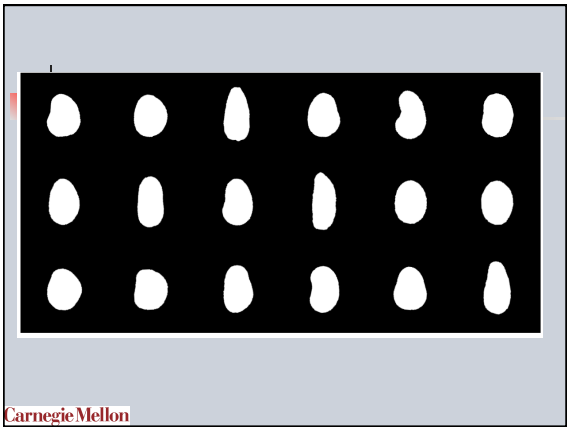

Generative Model Components
Nucleus
width
Medial axis
Cell membrane
$\frac{d_1 + d_2}{d_2}$
Model parameters
Original
Filtered
Fitted
Protein objects
(a)
(b)
(c)
Zhao & Murphy 2007
Carnegie Mellon

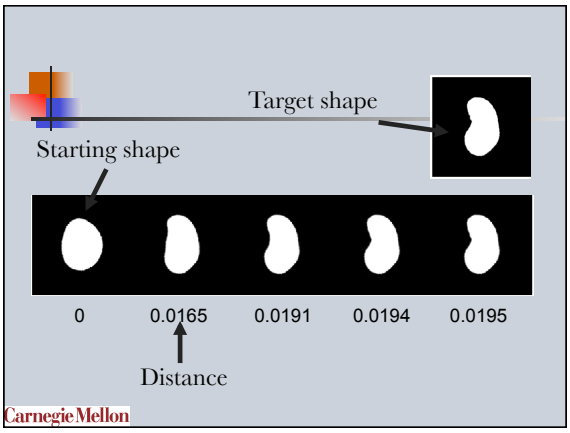

Synthesized Images
Lysosomes
Endosomes
SLML toolbox - Ivan Cao-Berg, Tao Peng, Ting Zhao
Have portable tool for generating images from model
Carnegie Mellon

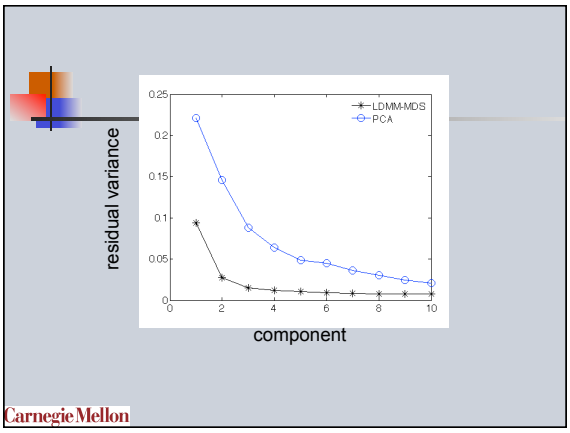

Combining Models for Cell Simulations
Protein 1
Nuclear Model
Protein 2
Nuclear Model
Protein 3
Nuclear Model
XML
Shared Nuclear and Cell Shape
Simulation for multiple proteins
Integrating with Virtual Cell (University of Connectiicut)) and M-Cell (Pittsburgh Supercomputing Center)
Carnegie Mellon

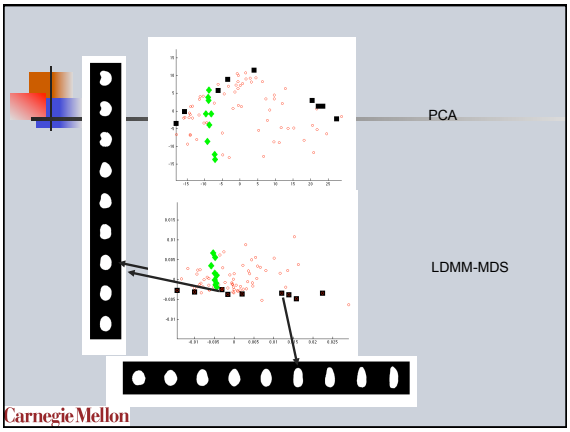

Diffeomorphic analysis of nuclear shape (w/ Gustavo Rohde)
Use of medial axis model assumes parameters lie in Euclidean space
Actual shape space is non-Euclidean
Can use distance between shapes instead
Carnegie Mellon

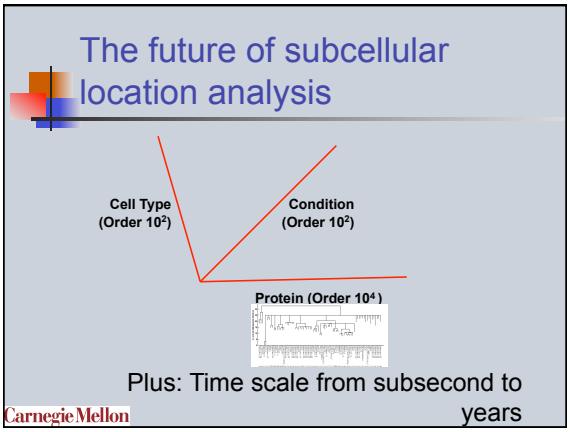

Distribution of shapes
Correct mean
Mean computed by averaging medial axis parameters
Carnegie Mellon

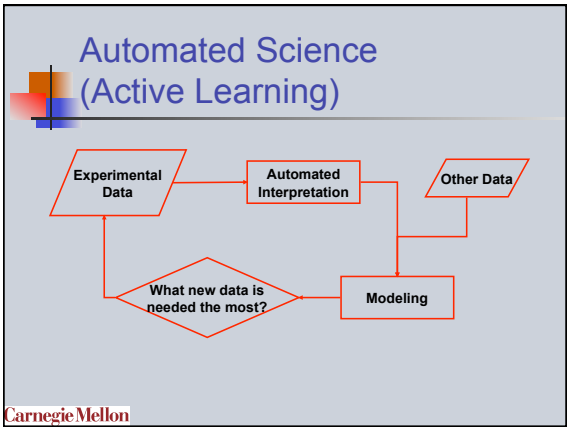

Carnegie Mellon


Target shape
Starting shape
0
0.0165
0.0191
0.0194
0.0195
Distance
Carnegie Mellon


residual variance
LDMM-MDS
PCA
component
Carnegie Mellon


PCA
LDMM-MDS
Carnegie Mellon


The future of subcellular location analysis
Cell Type (Order 10^2)
Condition (Order 10^2)
Protein (Order 10^4)
Plus: Time scale from subsecond to years
Carnegie Mellon


Automated Science (Active Learning)
Experimental Data
Automated Interpretation
Other Data
What new data is needed the most?
Modeling
Carnegie Mellon

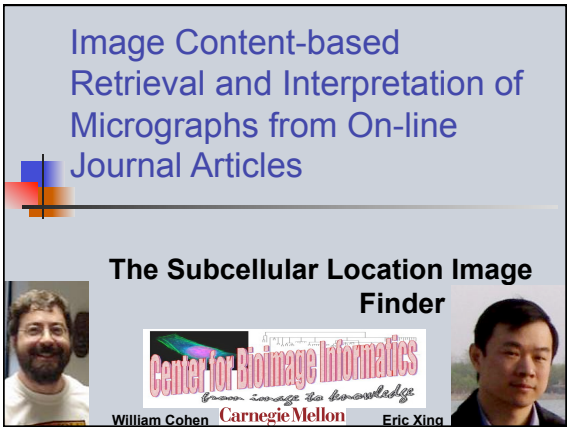
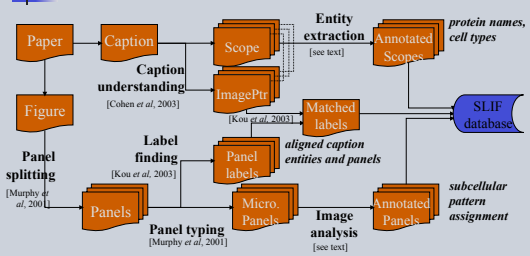


Image Content-based Retrieval and Interpretation of Micrographs from On-line Journal Articles
The Subcellular Location Image Finder
Center for Bioimage Informatics
William Cohen
Carnegie Mellon
Eric Xing

SLIF Pipeline components



Carnegie Mellon

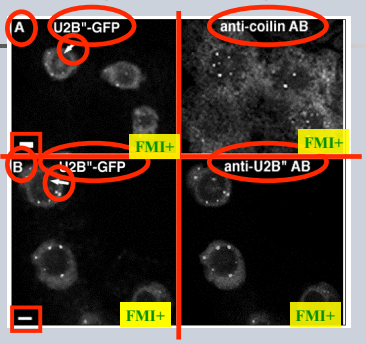
Overview: Image processing tasks

Segment into "panels"

Detect & remove **annotations**

Classify panels

Find scale bars



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.

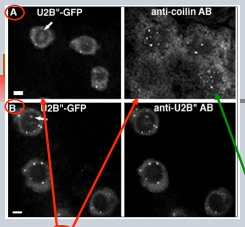
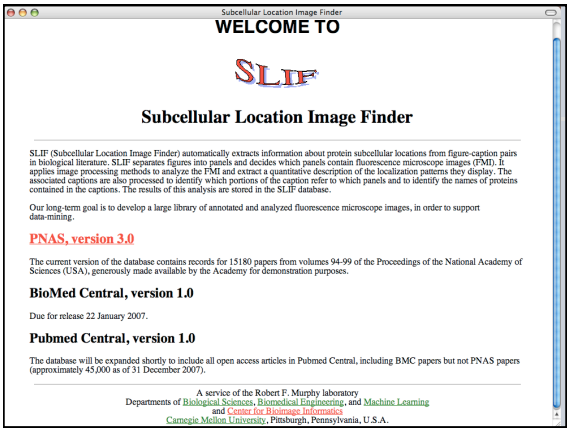
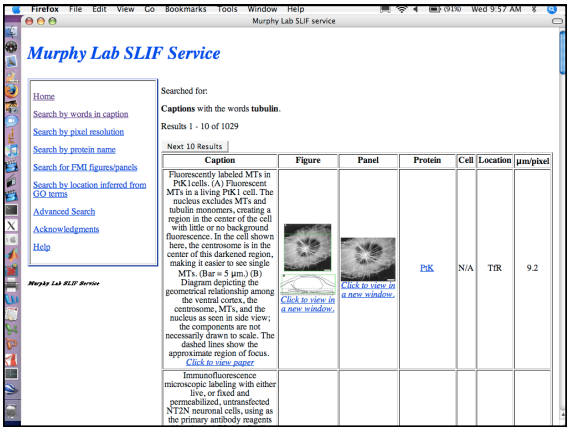


Figure 1. (A) Single confocal optical section of BY-2 cells expressing U2B 0-GFP, double labeled with GFP (left panel) and autoantibody against p80 coilin (right panel). Three nuclei are shown, and the bright GFP spots colocalize with bright foci of anti-coilin labeling. There is some labeling of the 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 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 m m. Movement of Coiled Bodies Vol. 10, July 1999 2299

Carnegie Mellon





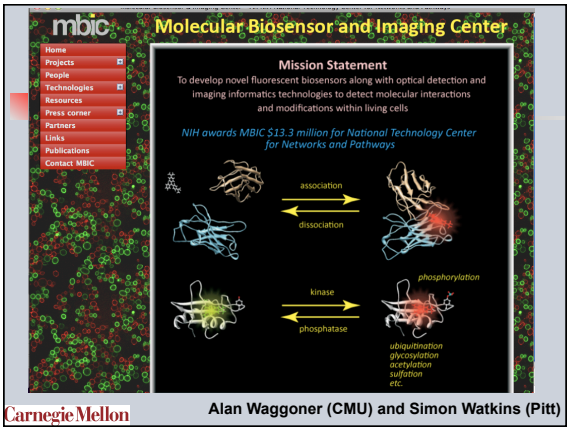
Conclusions

- 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

Carnegie Mellon

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 Berget**, 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 - NCBC (Athey)



Alan Waggoner (CMU) and Simon Watkins (Pitt)