

Location Proteomics: Image Informatics for Systems Biology

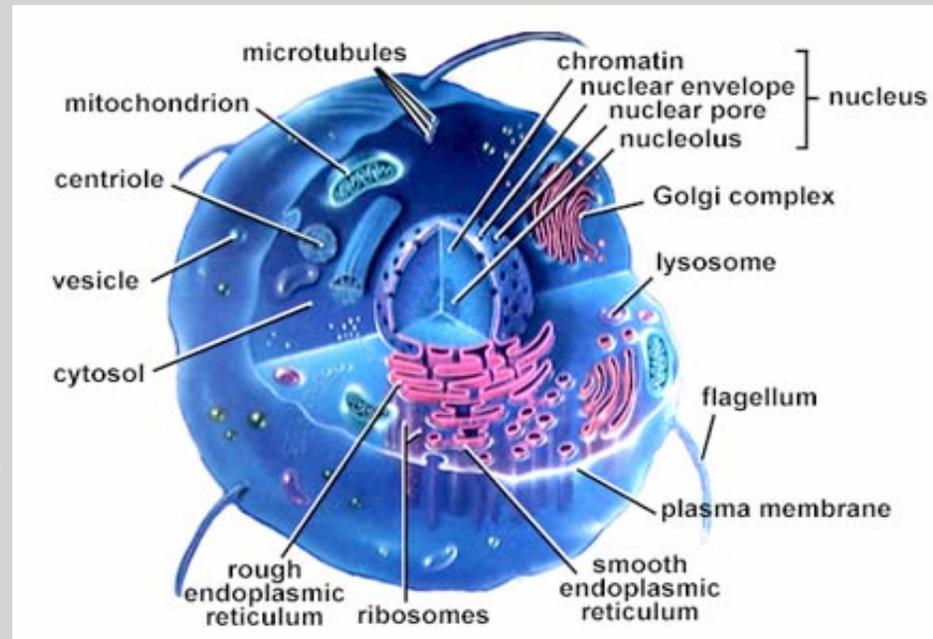
Robert F. Murphy

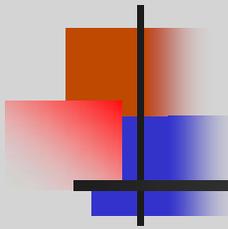
Departments of Biological Sciences, Biomedical Engineering and Machine Learning and



Systems Biology and Location Proteomics

- All systems biology must be data driven
- Key to progress
 - identification of aspect that needs to be analyzed “ome-wide”
 - development of assays and automated analysis approaches
- Systems biology needs systematic information on high-resolution subcellular location
 - Eventually, for every expressed protein for all cell types under all conditions
- Providing this information is the goal of Location Proteomics





Automated Interpretation

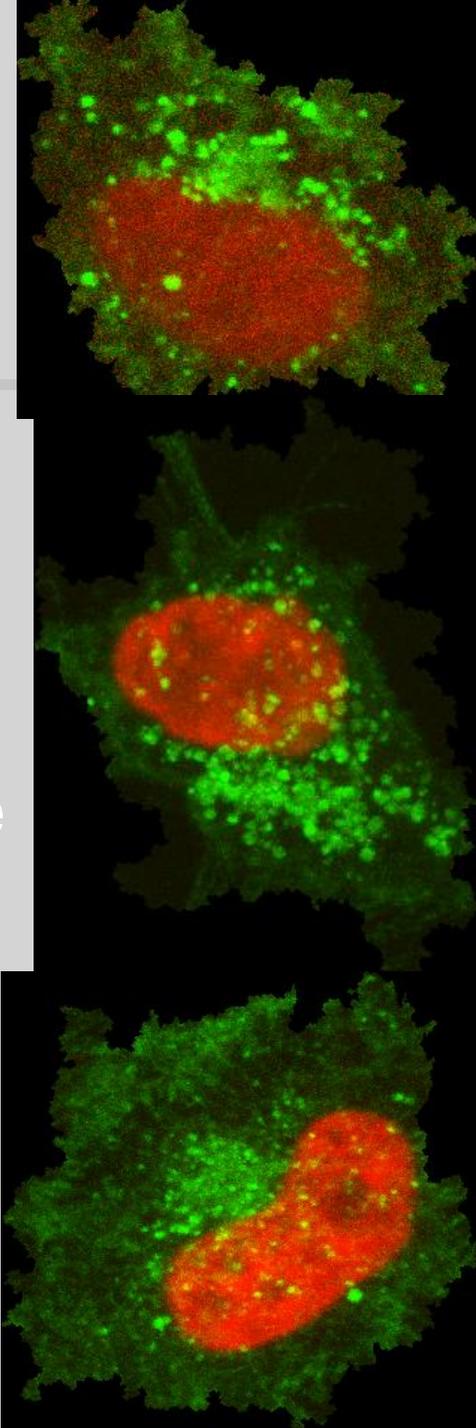
- Traditional analysis of fluorescence microscope images has occurred by visual inspection
- Our goal over the past ten years has to been automate the interpretation, to yield better
 - Objectivity
 - Sensitivity
 - Reproducibility

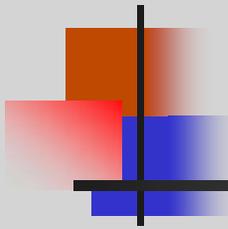
A. Supervised Learning of High-Resolution Subcellular Location Patterns



The Challenge

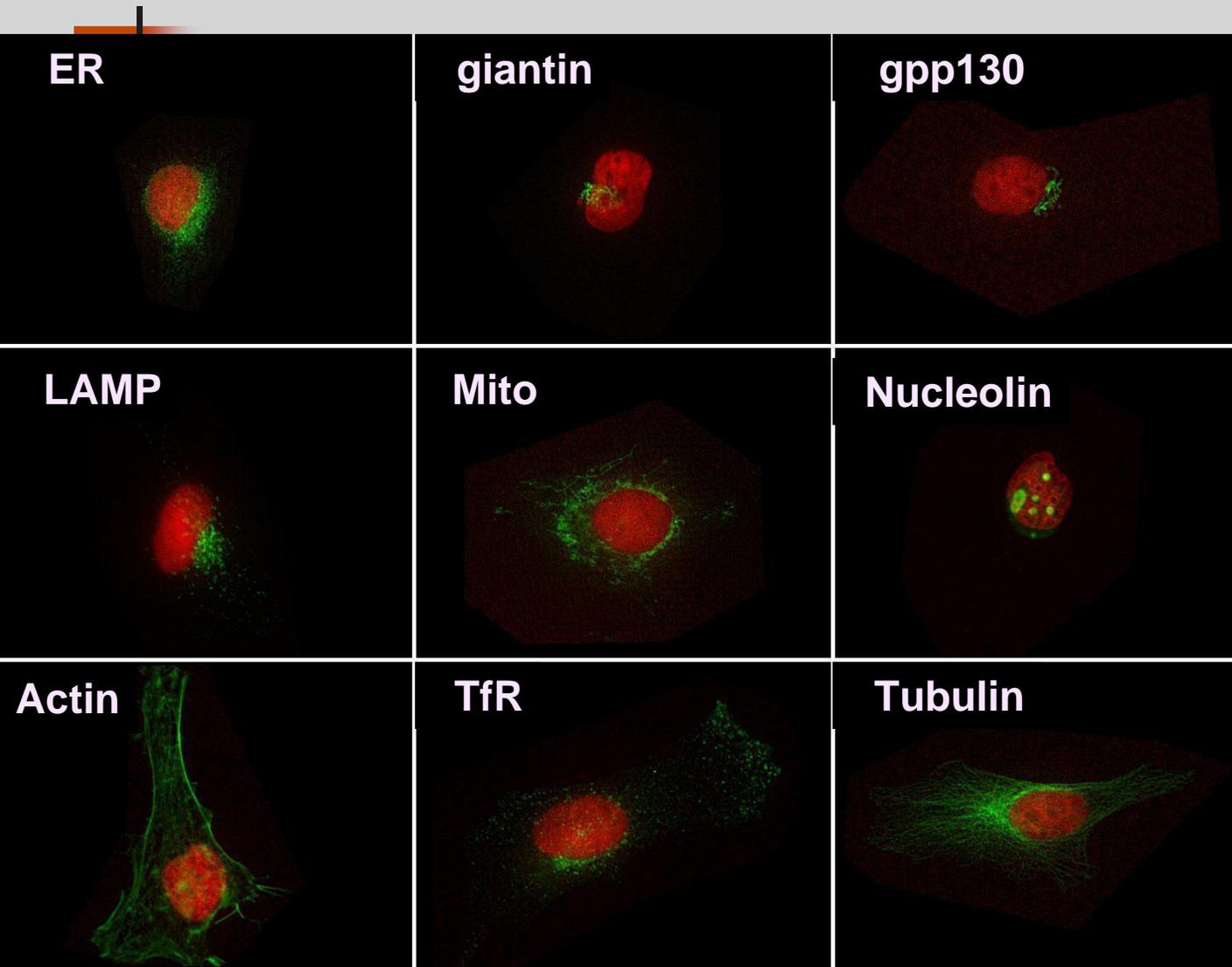
- Direct comparison of cell patterns to known examples does not work because different cells have different **shapes, sizes, orientations**
- Organelles/structures within cells are **not found in fixed locations**
- ***Instead, we describe each image numerically and compare the descriptors***



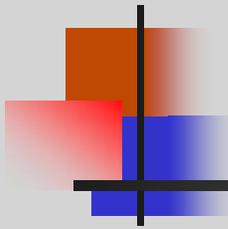


Feature-based, Supervised learning approach

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



2D
Images
of 10
Patterns
(HeLa)



Evaluating Classifiers

- Divide ~100 images for each class into **training** set and **test** set
- Use the **training** set to determine rules for the classes
- Use the **test** set to evaluate performance
- Repeat with different division into training and test
- Evaluate different sets of features chosen as most discriminative by feature selection methods
- Evaluate different classifiers

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

2D Classification Results

True Class	Output of the Classifier									
	DNA	ER	Gia	Gpp	Lam	Mit	Nuc	Act	TfR	Tub
DNA	99	1	0	0	0	0	0	0	0	0
ER	0	97	0	0	0	2	0	0	0	1
Gia	0	0	91	7	0	0	0	0	2	0
Gpp	0	0	14	82	0	0	2	0	1	0
Lam	0	0	1	0	88	1	0	0	10	0
Mit	0	3	0	0	0	92	0	0	3	3
Nuc	0	0	0	0	0	0	99	0	1	0
Act	0	0	0	0	0	0	0	100	0	0
TfR	0	1	0	0	12	2	0	1	81	2
Tub	1	2	0	0	0	1	0	0	1	95

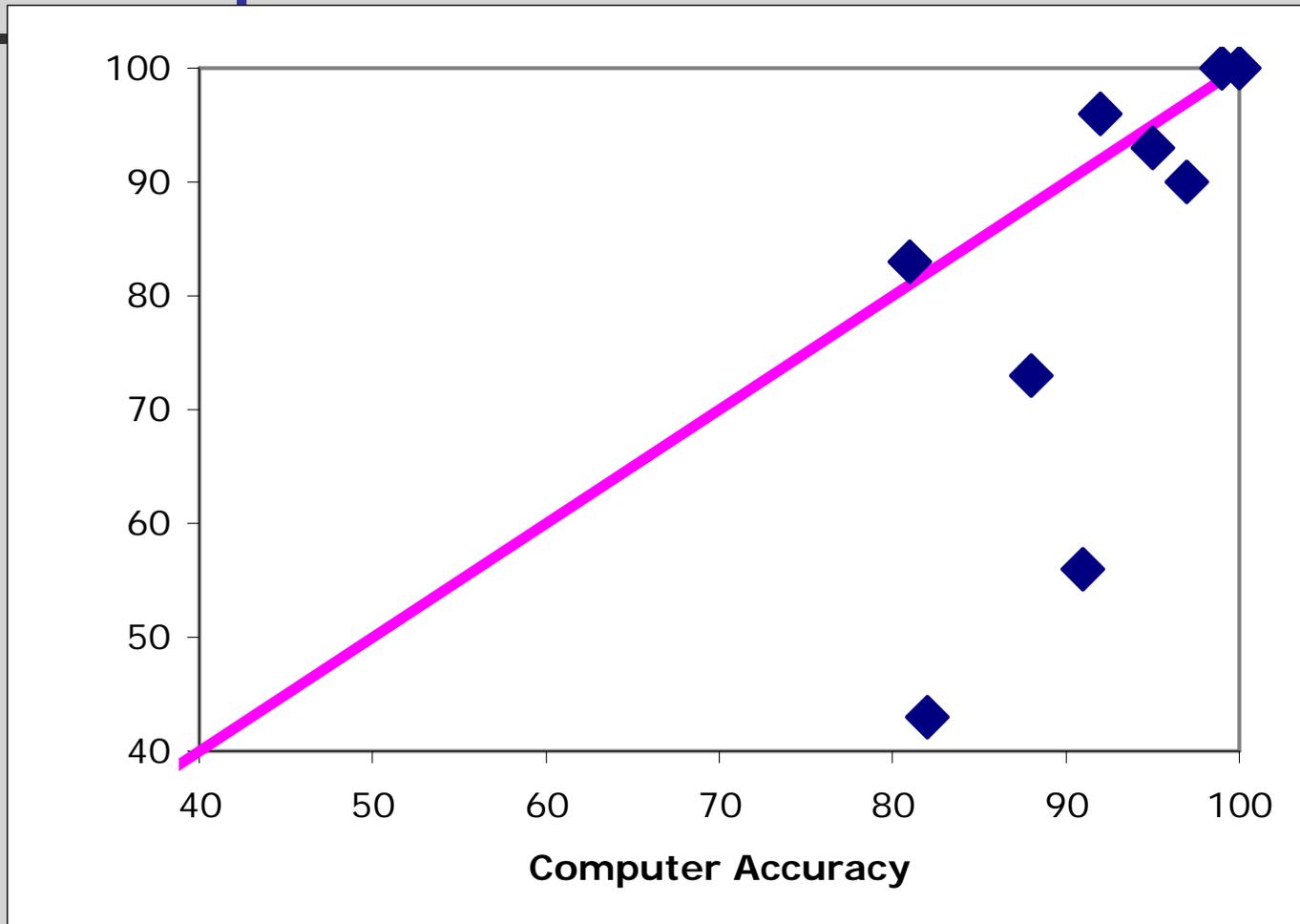
Overall accuracy = 92%

Human Classification Results

True Class	Output of the Classifier									
	DNA	ER	Gia	Gpp	Lam	Mit	Nuc	Act	TfR	Tub
DNA	100	0	0	0	0	0	0	0	0	0
ER	0	90	0	0	3	6	0	0	0	0
Gia	0	0	56	36	3	3	0	0	0	0
Gpp	0	0	54	33	0	0	0	0	3	0
Lam	0	0	6	0	73	0	0	0	20	0
Mit	0	3	0	0	0	96	0	0	0	3
Nuc	0	0	0	0	0	0	100	0	0	0
Act	0	0	0	0	0	0	0	100	0	0
TfR	0	13	0	0	3	0	0	0	83	0
Tub	0	3	0	0	0	0	0	3	0	93

Overall accuracy = 83%

Computer vs. Human



3D Classification Results

True Class	Output of the Classifier									
	DNA	ER	Gia	Gpp	Lam	Mit	Nuc	Act	TfR	Tub
DNA	98	2	0	0	0	0	0	0	0	0
ER	0	100	0	0	0	0	0	0	0	0
Gia	0	0	100	0	0	0	0	0	0	0
Gpp	0	0	0	96	4	0	0	0	0	0
Lam	0	0	0	4	95	0	0	0	0	2
Mit	0	0	2	0	0	96	0	2	0	0
Nuc	0	0	0	0	0	0	100	0	0	0
Act	0	0	0	0	0	0	0	100	0	0
TfR	0	0	0	0	2	0	0	0	96	2
Tub	0	2	0	0	0	0	0	0	0	98

Overall accuracy = 98%

B. Unsupervised Learning to Identify High-Resolution Protein Patterns



Location Proteomics

- **Tag** many proteins
 - We have used **CD-tagging** (developed by **Jonathan Jarvik** and **Peter Berget**): Infect population of cells with a retrovirus carrying DNA sequence that will “tag” in a random gene



Isolate separate **clones**, each of which produces express one tagged protein

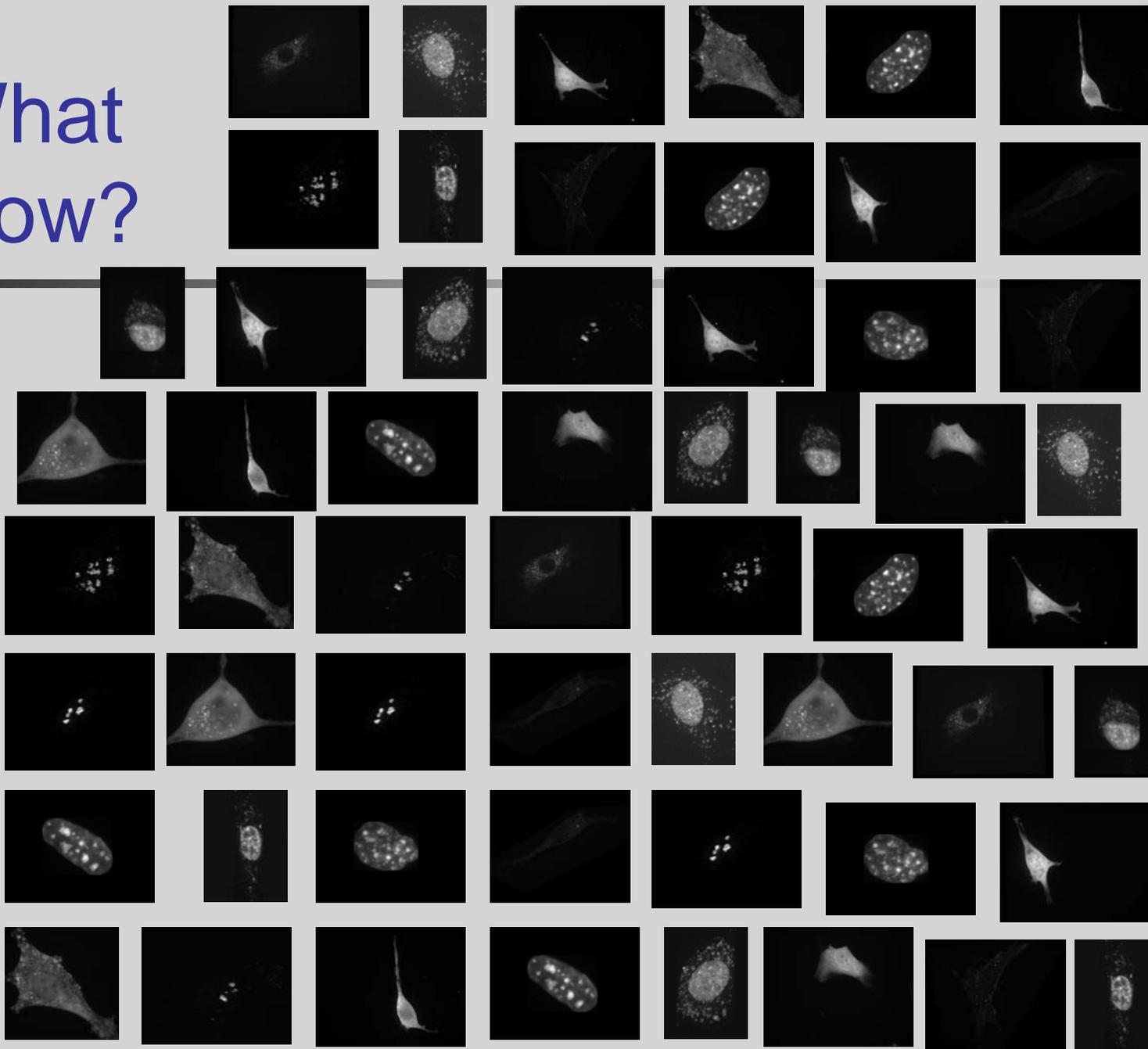
Use RT-PCR to **identify tagged gene** in each clone

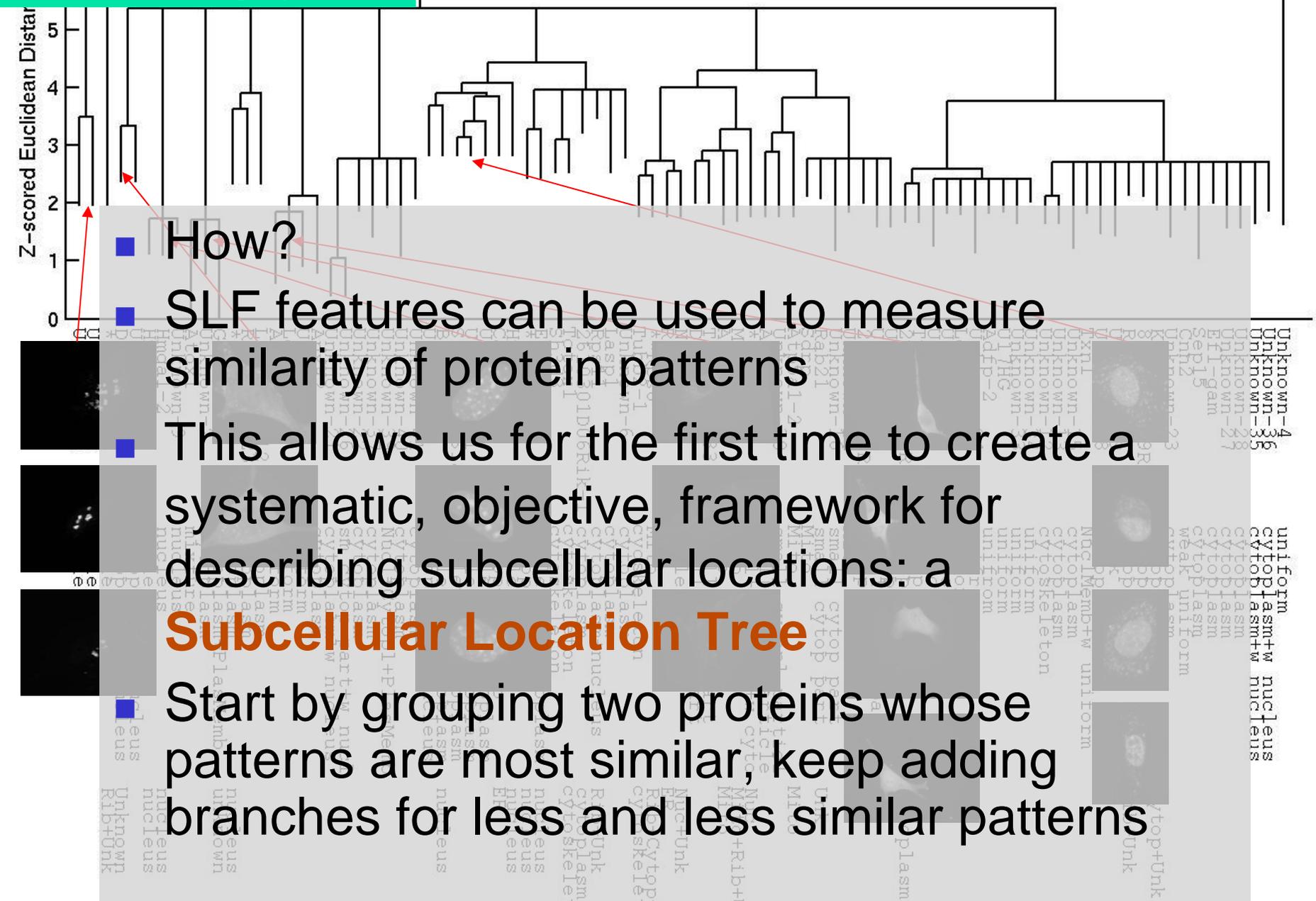
- Collect **many live cell images** for each clone using spinning disk confocal fluorescence microscopy

Jarvik
et al
2002

What Now?

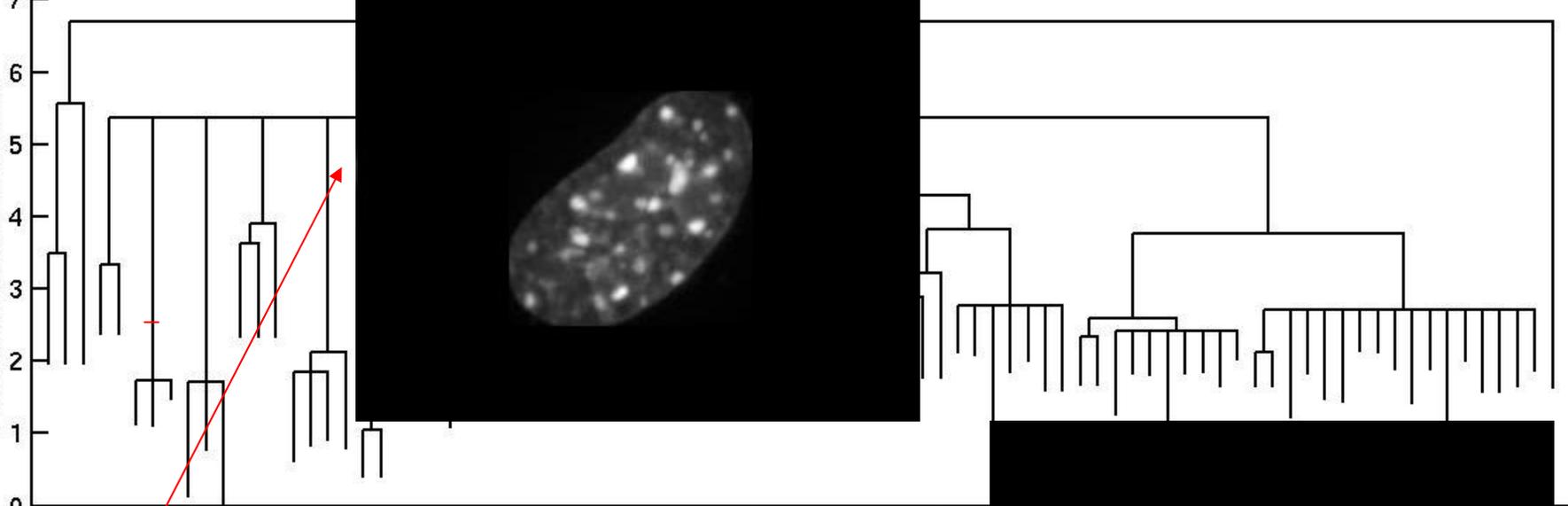
Group
~90
tagged
clones
by
pattern



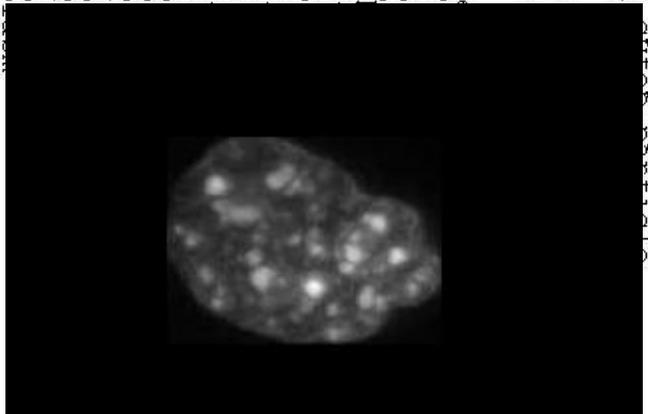
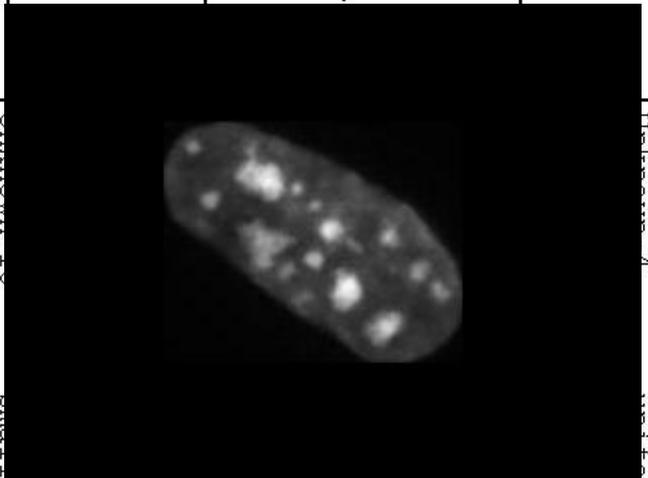
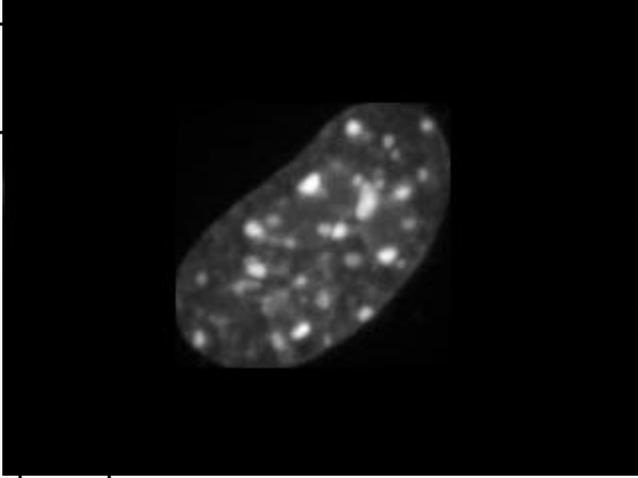


■ How?
■ SLF features can be used to measure similarity of protein patterns
■ This allows us for the first time to create a systematic, objective, framework for describing subcellular locations: a **Subcellular Location Tree**
■ Start by grouping two proteins whose patterns are most similar, keep adding branches for less and less similar patterns

Z-scored Euclidean Distance



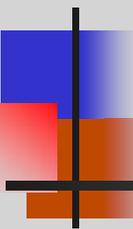
- Unknwn-25
- Unknwn-32
- Unknwn-33
- Unknwn-35
- Unknwn-36
- Unknwn-37
- Unknwn-38
- Unknwn-39
- Unknwn-40
- Unknwn-41
- Unknwn-42
- Unknwn-43
- Unknwn-44
- Unknwn-45
- Unknwn-46
- Unknwn-47
- Unknwn-48
- Unknwn-49
- Unknwn-50
- Unknwn-51
- Unknwn-52
- Unknwn-53
- Unknwn-54
- Unknwn-55
- Unknwn-56
- Unknwn-57
- Unknwn-58
- Unknwn-59
- Unknwn-60
- Unknwn-61
- Unknwn-62
- Unknwn-63
- Unknwn-64
- Unknwn-65
- Unknwn-66
- Unknwn-67
- Unknwn-68
- Unknwn-69
- Unknwn-70
- Unknwn-71
- Unknwn-72
- Unknwn-73
- Unknwn-74
- Unknwn-75
- Unknwn-76
- Unknwn-77
- Unknwn-78
- Unknwn-79
- Unknwn-80
- Unknwn-81
- Unknwn-82
- Unknwn-83
- Unknwn-84
- Unknwn-85
- Unknwn-86
- Unknwn-87
- Unknwn-88
- Unknwn-89
- Unknwn-90
- Unknwn-91
- Unknwn-92
- Unknwn-93
- Unknwn-94
- Unknwn-95
- Unknwn-96
- Unknwn-97
- Unknwn-98
- Unknwn-99
- Unknwn-100



Punctate Nuclear Proteins

- Unknwn-25
- Unknwn-32
- Unknwn-33
- Unknwn-35
- Unknwn-36
- Unknwn-37
- Unknwn-38
- Unknwn-39
- Unknwn-40
- Unknwn-41
- Unknwn-42
- Unknwn-43
- Unknwn-44
- Unknwn-45
- Unknwn-46
- Unknwn-47
- Unknwn-48
- Unknwn-49
- Unknwn-50
- Unknwn-51
- Unknwn-52
- Unknwn-53
- Unknwn-54
- Unknwn-55
- Unknwn-56
- Unknwn-57
- Unknwn-58
- Unknwn-59
- Unknwn-60
- Unknwn-61
- Unknwn-62
- Unknwn-63
- Unknwn-64
- Unknwn-65
- Unknwn-66
- Unknwn-67
- Unknwn-68
- Unknwn-69
- Unknwn-70
- Unknwn-71
- Unknwn-72
- Unknwn-73
- Unknwn-74
- Unknwn-75
- Unknwn-76
- Unknwn-77
- Unknwn-78
- Unknwn-79
- Unknwn-80
- Unknwn-81
- Unknwn-82
- Unknwn-83
- Unknwn-84
- Unknwn-85
- Unknwn-86
- Unknwn-87
- Unknwn-88
- Unknwn-89
- Unknwn-90
- Unknwn-91
- Unknwn-92
- Unknwn-93
- Unknwn-94
- Unknwn-95
- Unknwn-96
- Unknwn-97
- Unknwn-98
- Unknwn-99
- Unknwn-100

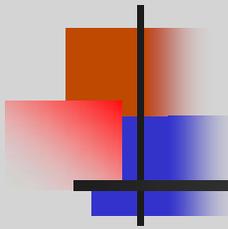
C. The Protein Subcellular Location Image Database (PSLID)



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



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

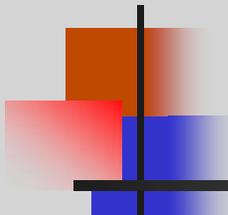
QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

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

The Subcellular Location Image Finder (SLIF)





Objectives of SLIF

- 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*.

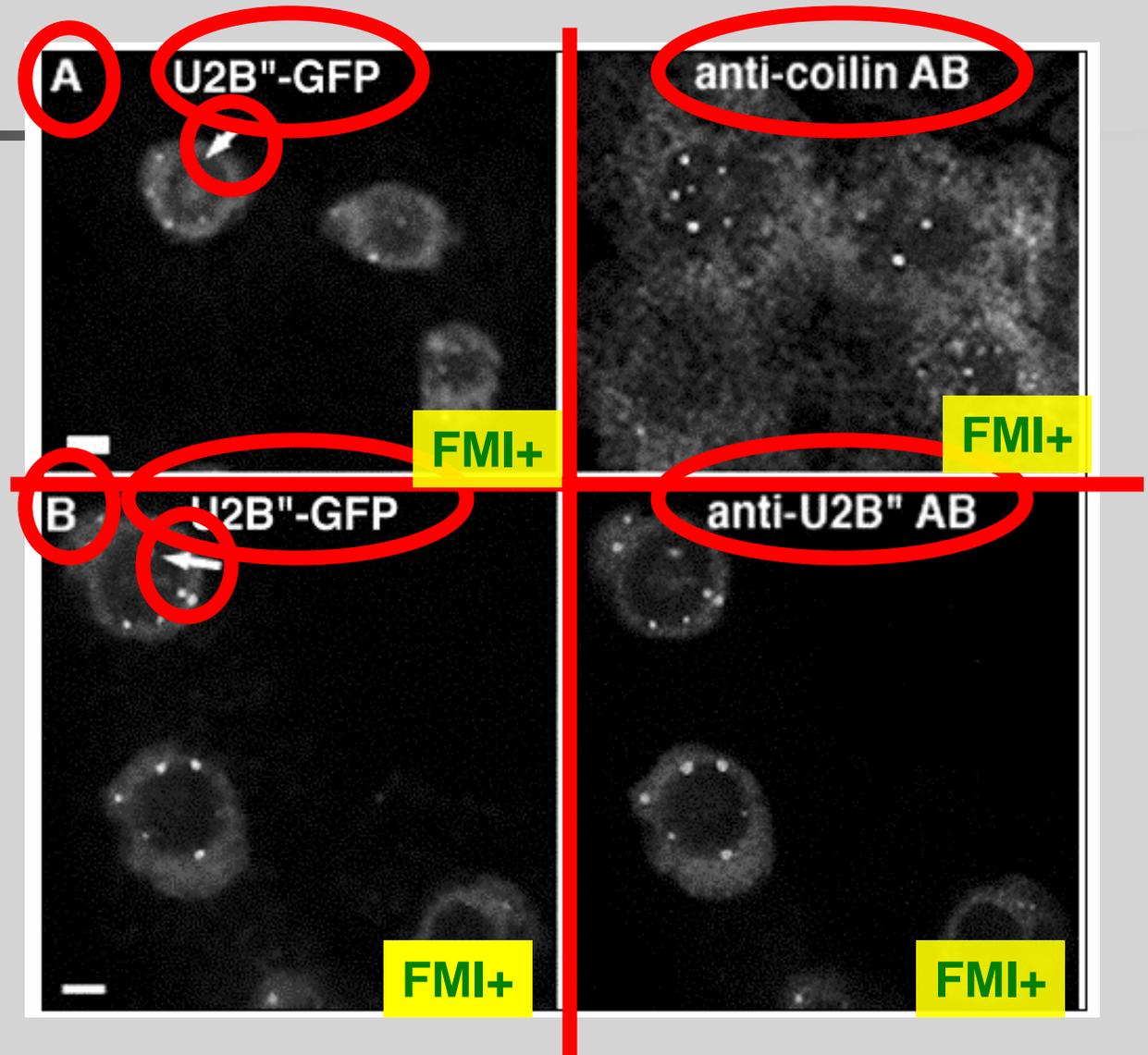
Overview: Image processing in SLIF

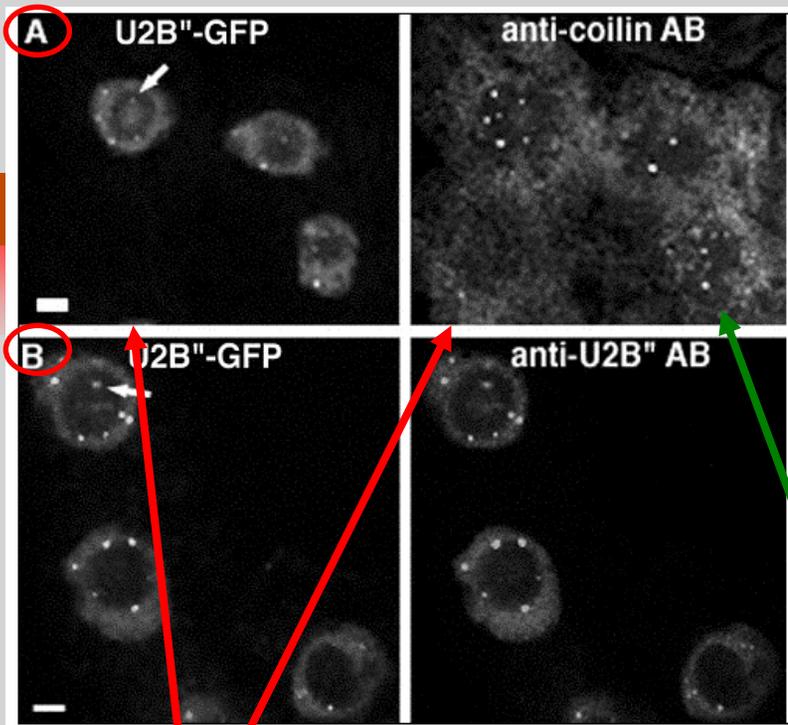
Segment

into
"panels"

Detect & remove
annotations

Classify
panels



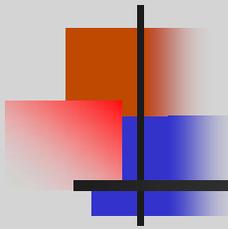


Overview: Text Processing in SLIF

- 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 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 μ m.

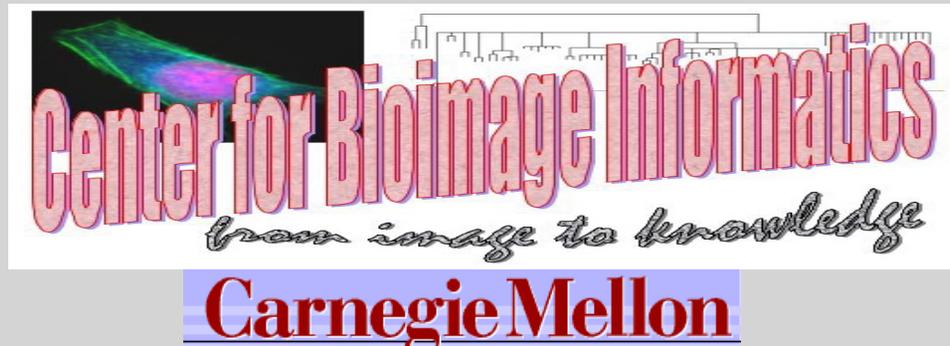
QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.



SLIF programmatic interface

- <http://slif.cbi.cmu.edu/search.jsp?arguments>
 - protein=<protein name>
 - level=figure OR level=panel
 - type=FMI
 - pixel_size_lo=<lower bound>
 - pixel_size_hi=<upper bound>
 - location=<subcellular location>

E. Preliminary Automated Analysis of Images from the Human Protein Atlas



Human Protein Atlas

http://proteinallas.org/tissue_profile.php?antibody_id=1949


[about the project](#)
[about protein atlas](#)
[tissue dictionary](#)
[disclaimer](#)
[submission of antibodies](#)

CASK tissue profiles. Validation score: N/A

Gene data

Description: Peripheral plasma membrane protein CASK (EC 2.7.1.-) (hCASK) (Calcium/calmodulin-dependent serine protein kinase) (Lin-2 homolog).
Source: O14936 (Uniprot)
Chromosome: X:p11.4
EnsEMBL ID: ENSG00000147044

Splice variant	Protein Ensembl ID	Transcript Ensembl ID	No of aa	Mw	Signal Peptide	TM Region(s)
Splice variant 1:	ENSP00000354641	ENST00000361962	921	105 kDa	No	No
Splice variant 2:	ENSP00000322727	ENST00000318588	921	105 kDa	No	No
Splice variant 3:	ENSP00000347218	ENST00000355101	897	102 kDa	No	No

Normal Tissues

Tissue	Expression	Cell Type
Adrenal gland	●	cortical cells
	●	medullar cells
Appendix	●	glandular cells
	●	lymphoid tissue
Bone marrow	○	bone marrow poetic cells
Breast	●	glandular cells
Bronchus	●	surface epithelial cells
Cerebellum	●	cells in granular layer
	●	cells in molecular layer
	●	purkinje cells
Cerebral cortex	●	neuronal cells
	●	non-neuronal cells
Cervix, uterine	●	glandular cells
Lung	○	alveolar cells
	○	macrophages
Lymph node	●	follicle cells (cortex)
	●	non-follicle cells (paracortex)
Nasopharynx	●	surface epithelial cells
Oral mucosa	○	surface epithelial cells
Ovary	●	follicle cells
	●	ovarian stromal cells
Pancreas	●	exocrine pancreas
	●	islet cells
Parathyroid gland	●	glandular cells
Placenta	●	decidual cells
	●	trophoblastic cells

Navigation

- Home
- Search result
- CAB001949**
 - Tissue profiles**
 - Antibody info

Search

1	▬▬▬▬▬▬	14	▬▬▬▬
2	▬▬▬▬▬▬	15	▬▬▬▬
3	▬▬▬▬▬▬	16	▬▬▬▬
4	▬▬▬▬▬▬	17	▬▬▬▬
5	▬▬▬▬▬▬	18	▬▬▬▬
6	▬▬▬▬▬▬	19	▬▬▬▬
7	▬▬▬▬▬▬	20	▬▬▬▬
8	▬▬▬▬▬▬	21	▬▬▬▬
9	▬▬▬▬▬▬	22	▬▬▬▬
10	▬▬▬▬▬▬	X	▬▬▬▬
11	▬▬▬▬▬▬	Y	▬▬▬▬
12	▬▬▬▬▬▬		
13	▬▬▬▬▬▬	OTHER	▬▬▬▬

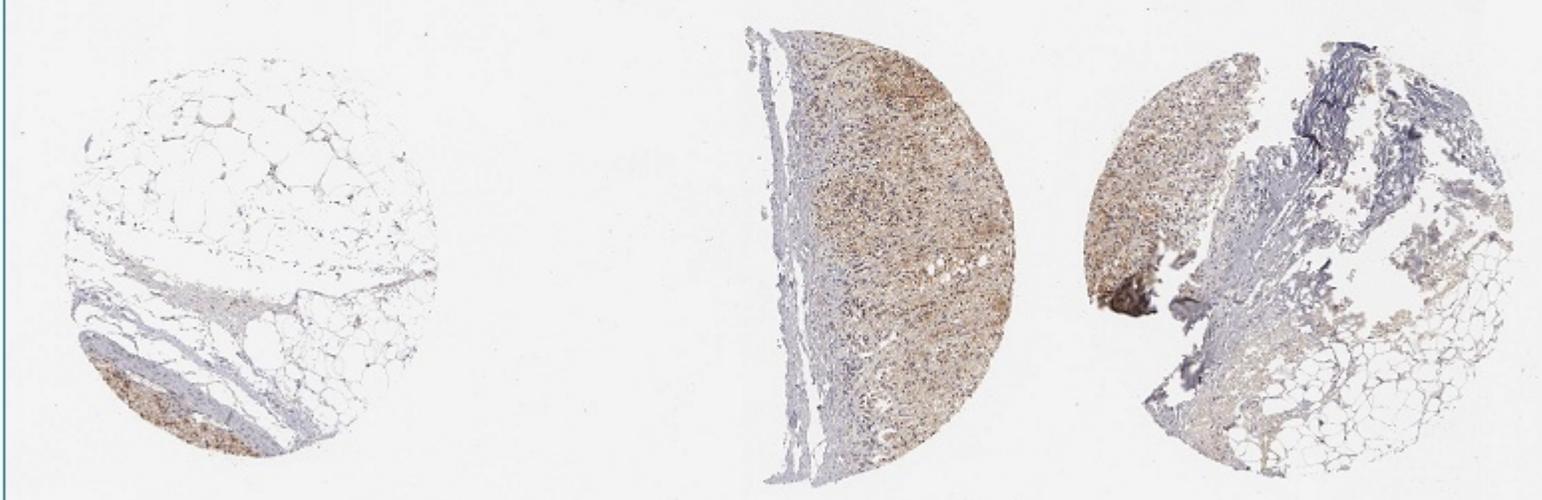
Human Protein Atlas

http://proteinatlas.org/normal_unit.php?antibody_id=1949&mainannotation_id=202310

hpa about the project about protein atlas tissue dictionary disclaimer submission of antibodies

Adrenal gland [CASK]

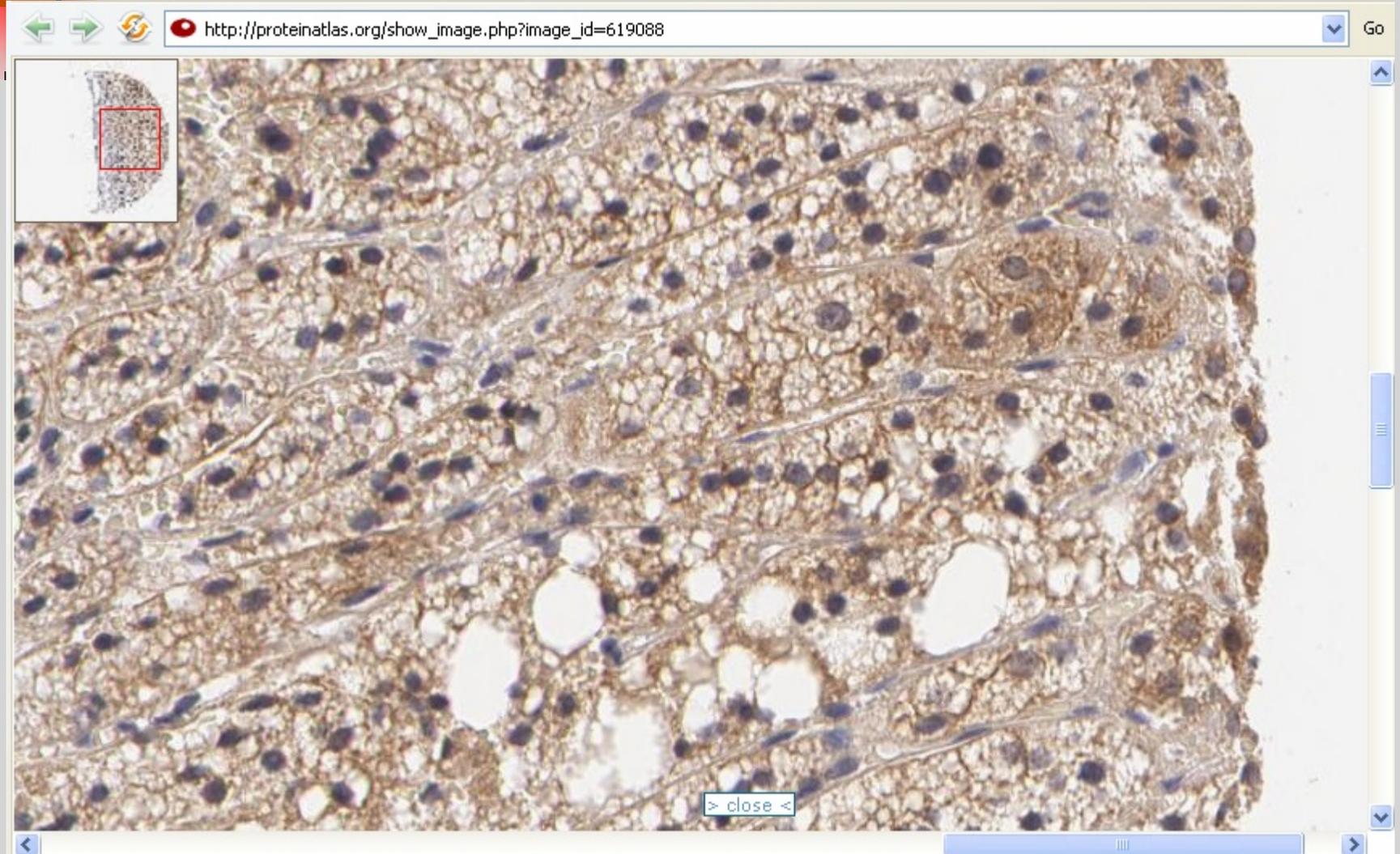
Cell Type	Intensity	Quantity	Localization
Cortical cells	moderate	>75%	cytoplasmic and/or membranous
Medullar cells			-- cell type not present --

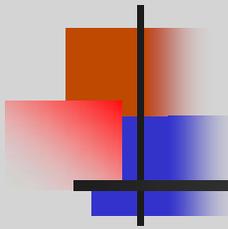


Male, age 35 Female, age 44 Female, age 71

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

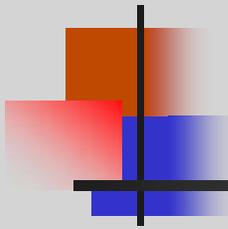
Human Protein Atlas





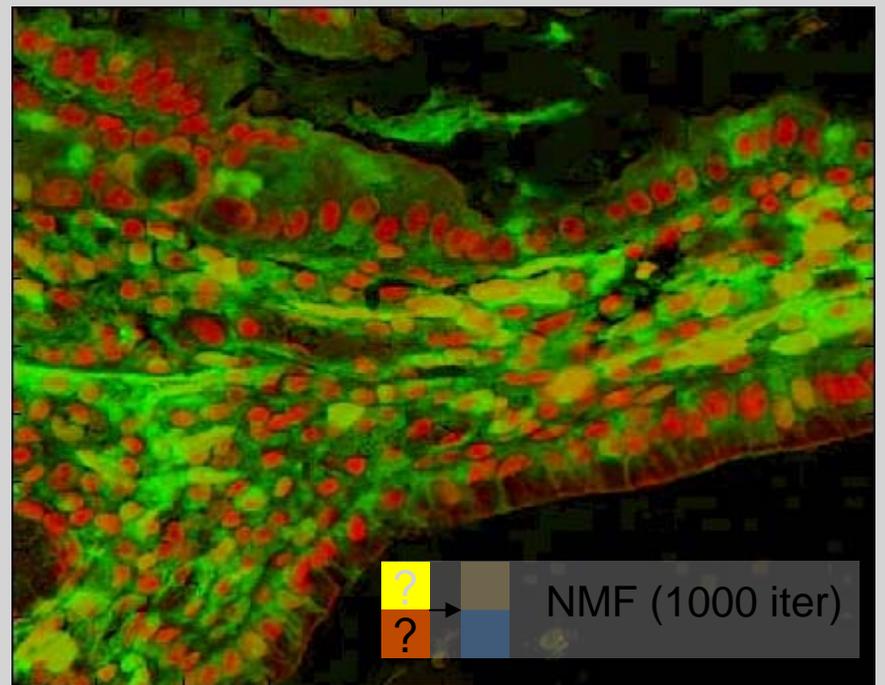
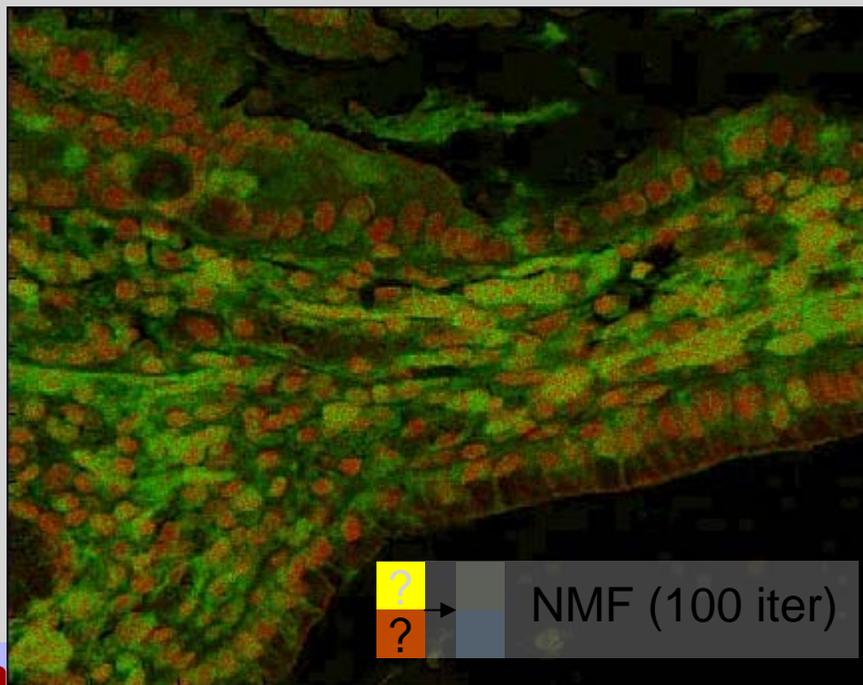
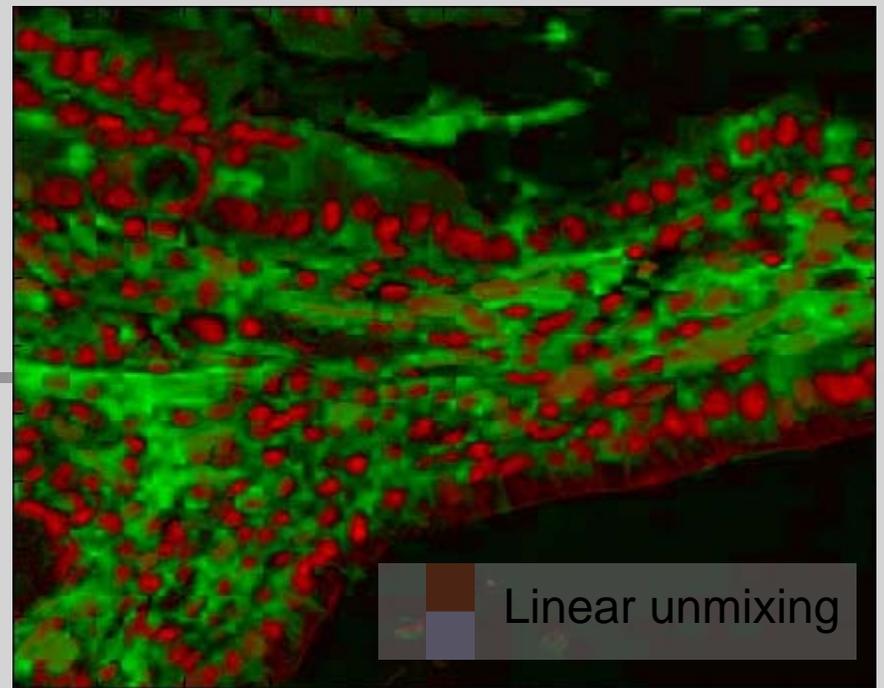
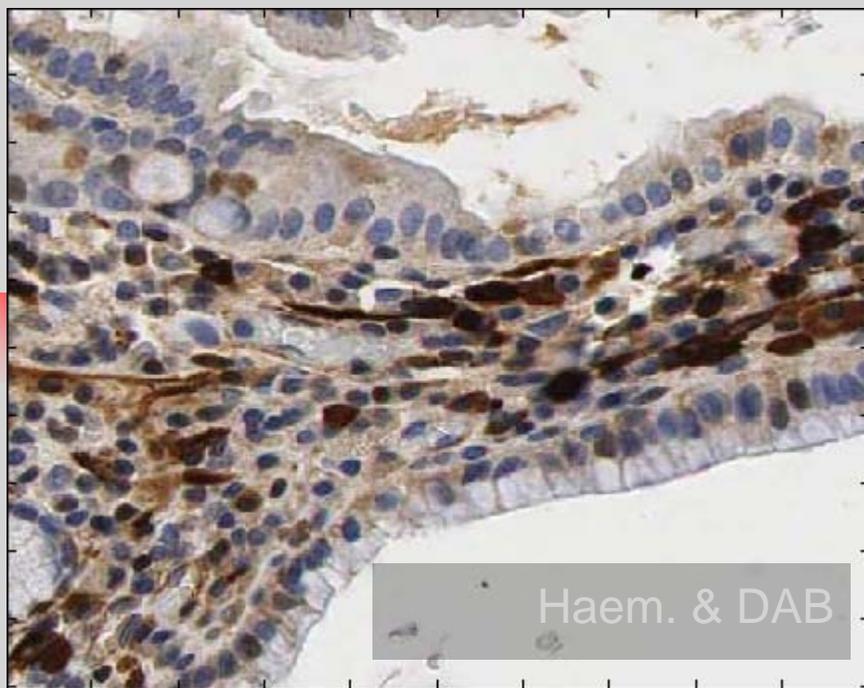
Issues

- ~700 antibodies on ~20 tissues
- Staining: DAB immunohistochemistry (for specific protein) and hematoxylin (for DNA)
- Color unmixing needed to extract separate images of protein and DNA distributions
- Images collected at low resolution (20x) - how well can subcellular patterns be distinguished?

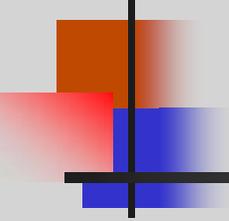


Initial Analysis

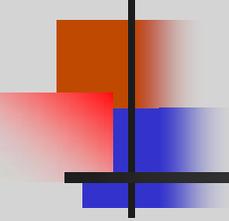
- Try some color unmixing methods
 - Linear unmixing
 - Non-negative matrix factorization
- Try training classifiers to recognize patterns for five test proteins



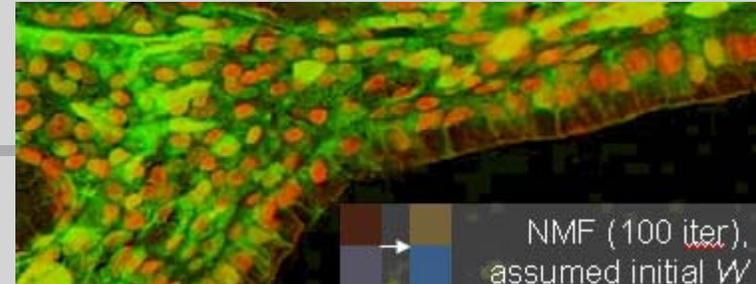
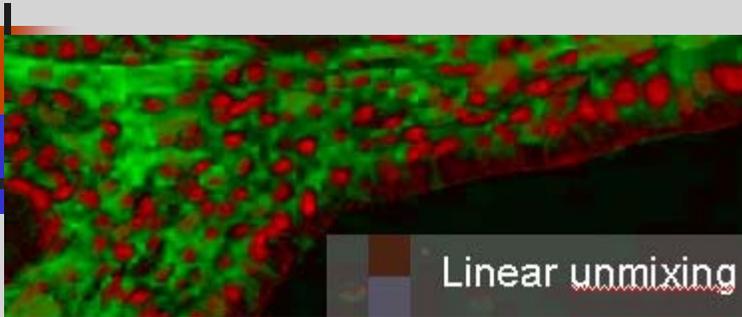
Dataset

- 
- 5 protein classes
 - Actin (cytoskeletal)
 - SNRP (nuclear)
 - TfR (lysosomal)
 - Thioredoxin (mitochondrial)
 - Tubulin (cytoskeletal)
 - Various (~10) tissue types for each class (ovary, liver, breast, skin, muscle, etc.)
 - 1 field for each tissue/protein pair

Dataset

- 
- Each field split into smaller 300x300 pixel images
 - Regions with less than 40% pixels above background are removed
 - A field gives 30-75 useable regions, usually 60

Classification Results –Across Tissues



Classified as

	Actin	SNRP	TfR	Thio	Tubul
Actin	36	9	8	13	34
SNRP	5	81	2	7	6
TfR	3	7	54	32	5
Thio	6	9	7	71	7
Tubul	15	6	13	8	57

Accuracy 61.5%

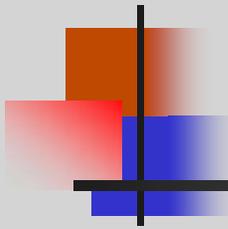
Classified as

	Actin	SNRP	TfR	Thio	Tubul
Actin	36	9	8	13	34
SNRP	5	81	2	7	6
TfR	3	7	54	32	5
Thio	6	9	7	71	7
Tubul	15	6	13	8	57

Accuracy 54.1%

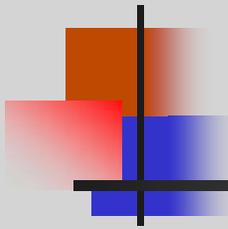
True class

True class



Summary

- Automated tools for analyzing subcellular patterns in fluorescence microscope images
- PSLID and SLIF databases for querying results of automated analysis
- Preliminary work on automated analysis of immunocytochemistry images in Human Protein Atlas



Future

- Build SLIF database for all of Pubmed Central and Biomed Central
- Additional datasets being added to PSLID
- SOAP interfaces to SLIF and PSLID in works
- Continue work on analysis of Human Protein Atlas
- Provide generative models for each location family