Automated Proteome-Wide Determination and Modeling of Subcellular Location for Systems Biology

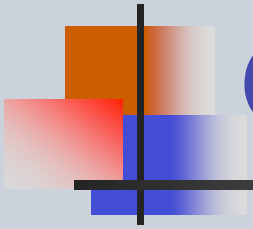
Robert F. Murphy

Ray and Stephanie Lane Professor of Computational Biology

Departments of Biological Sciences, Biomedical Engineering and Machine Learning and

RAY AND STEPHANIE LANE
Center for Computational Biology

Carnegie Mellon

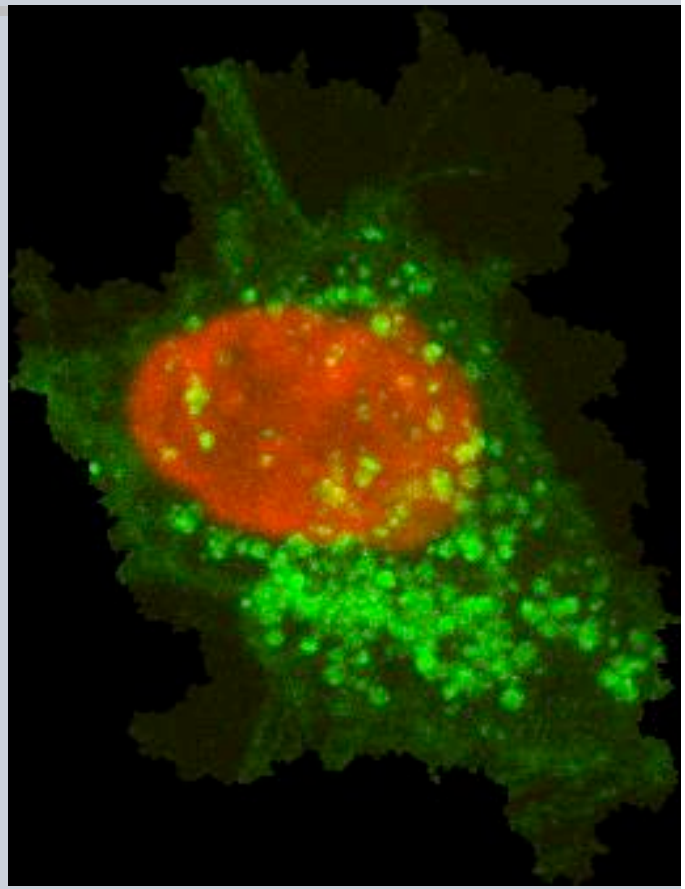


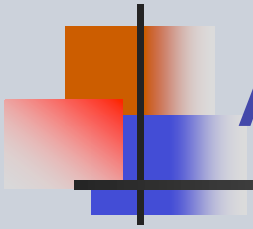
Open questions

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

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





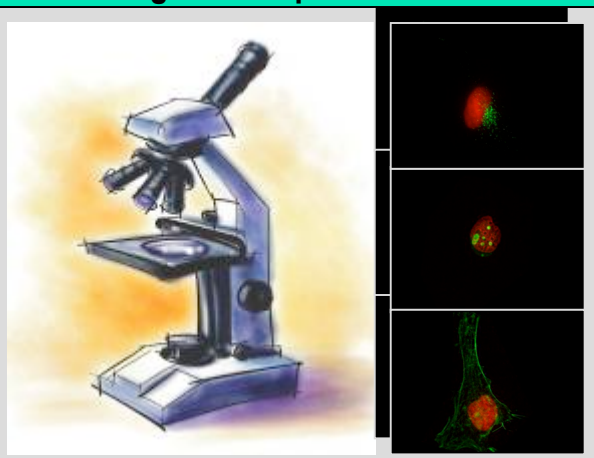
Automated Interpretation

- 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

Approach

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

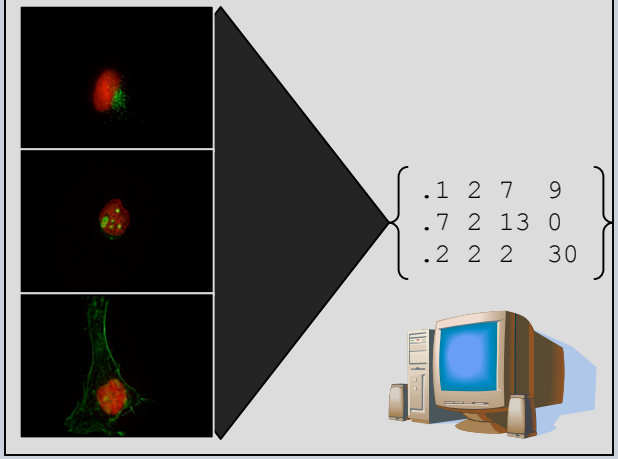
1. Image Acquisition



2. Image Processing

- Segmentation
- Denoising
- Deconvolution
- Signal unmixing

3. Feature Extraction



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

4. Feature Selection

$$\begin{Bmatrix} .1 & 2 & 7 & 9 \\ .7 & 2 & 13 & 0 \\ .2 & 2 & 2 & 30 \end{Bmatrix} \Rightarrow \begin{Bmatrix} .1 & 7 & 9 \\ .7 & 13 & 0 \\ .2 & 2 & 30 \end{Bmatrix}$$

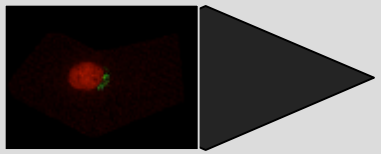
Remove features

or

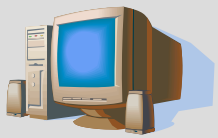
Combine features

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

5. Classification

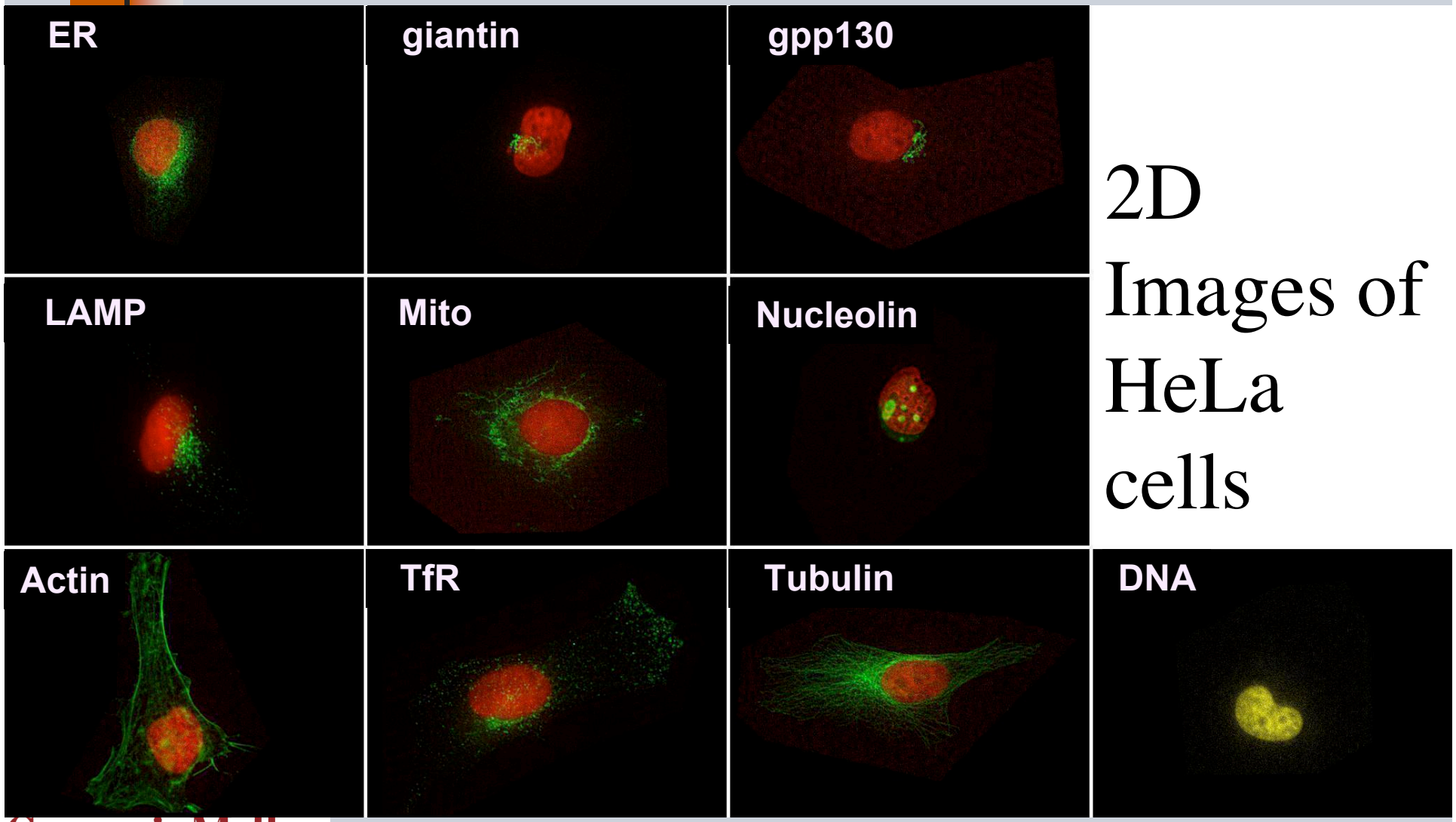


Golgi pattern!



Carnegie Mellon

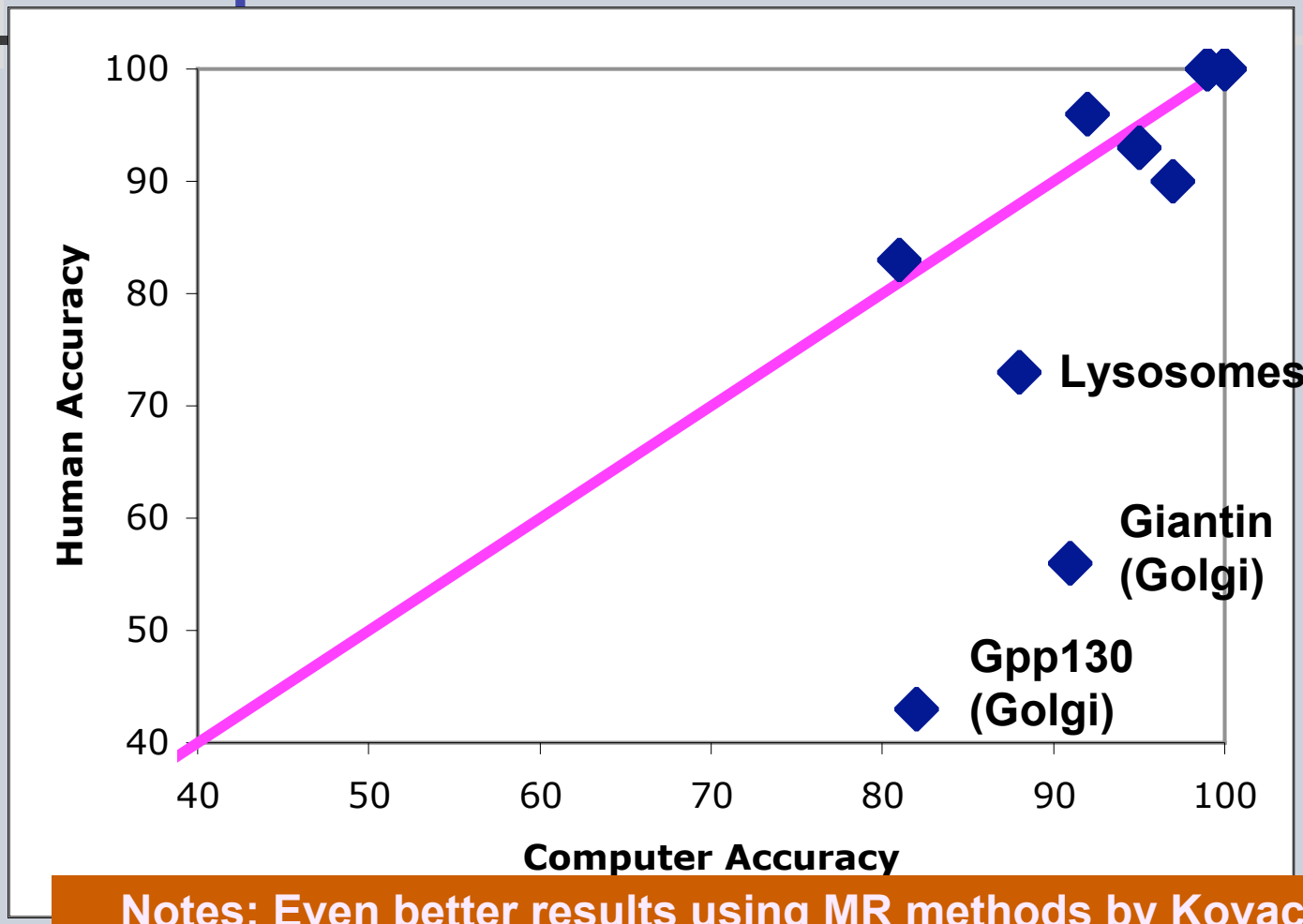
Initial goal: Learn to recognize all major subcellular patterns



2D Images of HeLa cells

Classification Results: Computer vs. Human

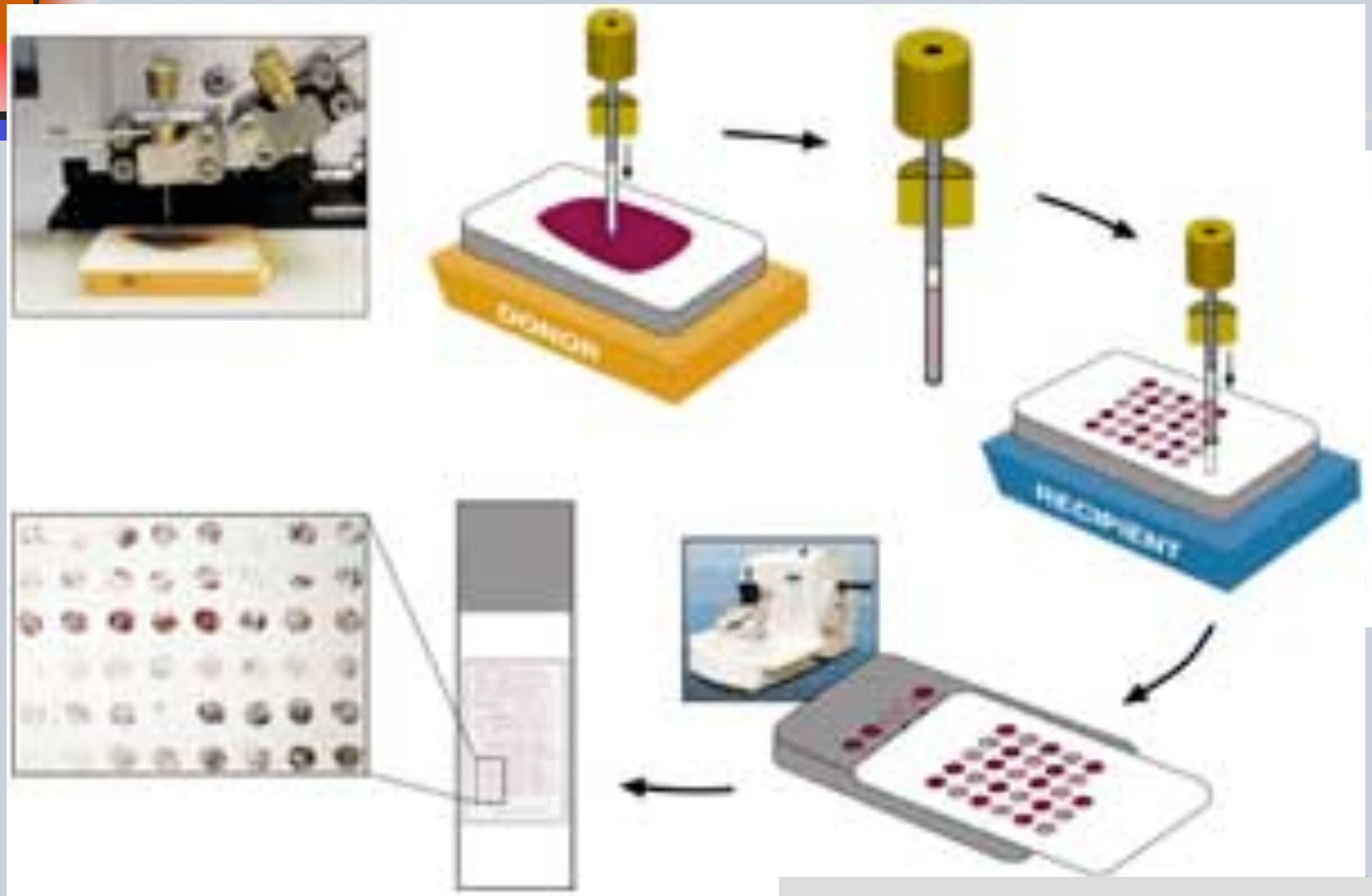
Murphy et al 2000; Boland & Murphy 2001; Murphy et al 2003; Huang & Murphy 2004



Notes: Even better results using MR methods by Kovacevic group

Even better results for 3D images

Tissue Microarrays



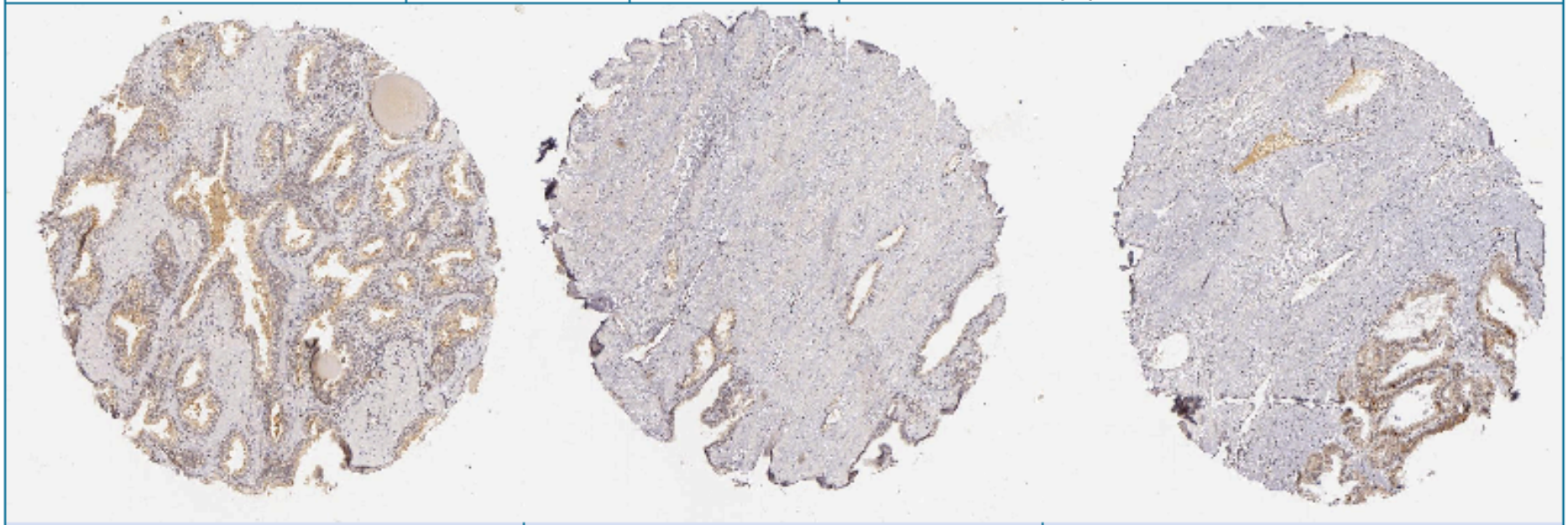
Courtesy www.microarraystation.com

Human Protein Atlas

hpa | the project | protein atlas | dictionary | disclaimer | submission of antibodies

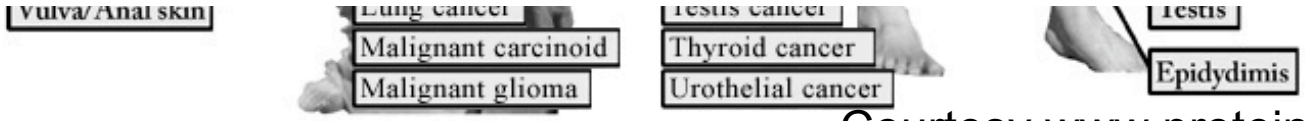
Prostate [CASP8]

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



Male, age 51	Male, age 64	Male, age 60

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



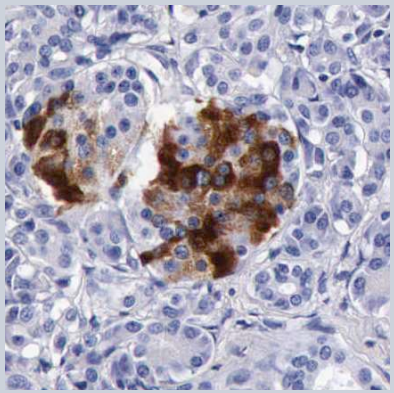
Carnegie Mellon

Courtesy www.proteinatlas.org

Test Dataset from Human Protein Atlas

- Selected set of 10 proteins from the Atlas that are similar to 2D HeLa dataset used to establish our methods (Nucleus, Nucleolar, 2 Golgi, ER, Endosome, Lysosome, Mitochondria, Actin Cytoskeleton, Tubulin Cytoskeleton)
- ~45 tissue types for each class (e.g. liver, muscle, skin)
- ~120 images per class
- Goal: Train classifier to recognize each subcellular pattern across all tissue types

Insulin in islet cells



Subcellular Pattern Classification over 45 tissues

Labels \ Prediction	MCM	DKC	GOL	TRI	HSP	TFR	LAM	SYN	TUB	ACT
MCM	**92.9**	0	7.1	0	0	0	0	0	0	0
DKC	0	**94.9**	0	0	0	2.6	0	2.6	0	0
GOL	0	4.9	**85.4**	0	0	7.3	0	2.4	0	0
TRI	0	0	0	**100**	0	0	0	0	0	0
HSP	0	0	0	0	**97.7**	0	2.3	0	0	0
TFR	0	2.6	2.6	0	0	**94.7**	0	0	0	0
LAM	0	0	0	0	0	0	**100**	0	0	0
SYN	0	0	3	0	3	0	0	**93.9**	0	0
TUB	0	0	0	2.6	2.6	0	0	0	**86.8**	7.9
ACT	0	0	0	0	0	0	0	0	18.6	**81.4**

Overall accuracy 92.8%

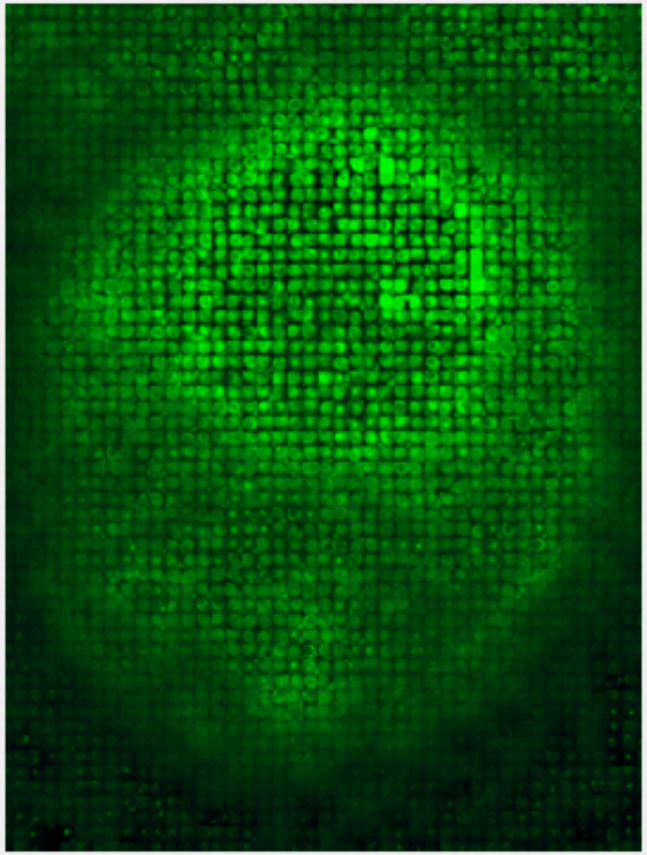
Accuracy for 60% of images with highest confidence: 97%

yeastgfp.ucsf.edu
YEAST GFP FUSION LOCALIZATION DATABASE

>> Advanced Query < Go >

>> Quick Search yal001c go clr

>> info
>> 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 yeast orf names (yal001c) or gene names (TFC3)

> separate multiple orfs/genes with a space (e.g. yal001c zwf1 bud2 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 quantitation data presented here is published in Ghaemmaghami, *et al.*, *Nature* **425**, 737-741 (2003). <pdf>

Detailed collection construction methods can be found in Howson *et al.*, *Comp Funct Genom* **6**, 2-16 (2005). <pdf>

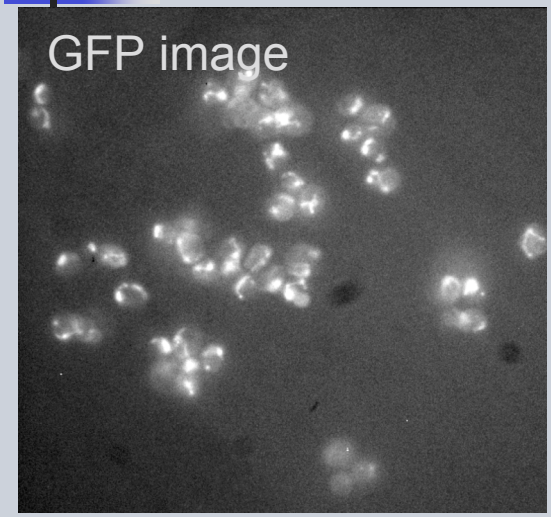
This research is the work of the laboratories of Erin O'Shea and Jonathan Weissman at the University of California San Francisco.
Please direct comments, concerns, and questions to <jan.ihmels@gmail.com>

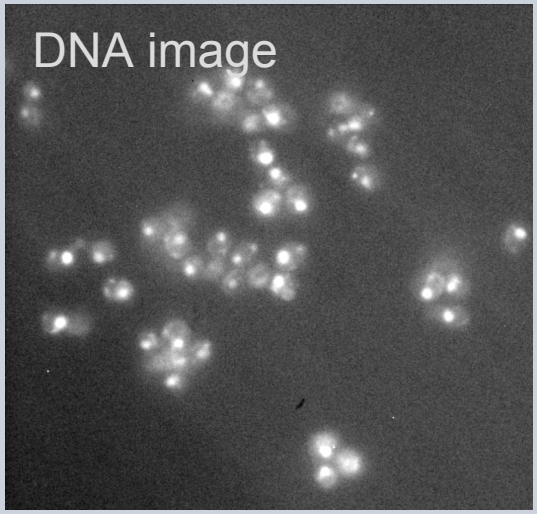
Annotations of Yeast GFP Fusion Localization Database

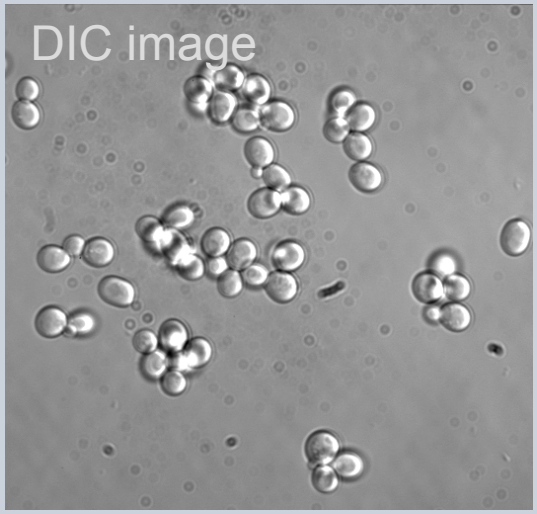
- Contains images of 4156 proteins (out of 6234 ORFs in all 16 yeast chromosomes).
- GFP tagged immediately before the stop codon of each ORF to minimize perturbation of protein expression.
- Annotations were done manually by two scorers and co-localization experiments were done for some cases using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

Chen et al 2006

Cell Image Segmentation







DNA potential, a function of one pixel p_i
The likelihood of a pixel to be foreground/background

→ **Generate Mask**

boundary potential, a function of two neighboring pixels p_i *and* p_j
The likelihood that there is a cell boundary between p_i and p_j

Chen et al 2007

Classification of Yeast Subcellular Patterns

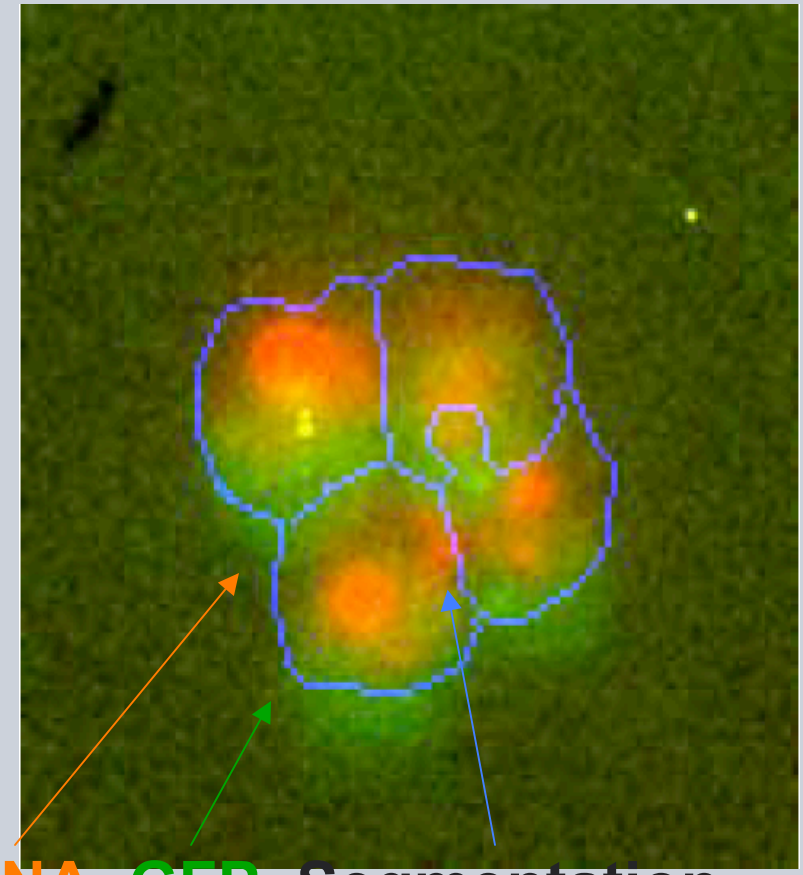
- Selected only those assigned to single unambiguous location class (21 classes)
- Trained classifier to recognize those classes
- **81% agreement with human classification**
- **94.5% agreement for high confidence assignments (without using colocalization!)**
- Examination of proteins for which methods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

Example of Potentially Incorrect Label

ORF Name
YAL009W

UCSF Location
nucleus

Automated Prediction
vacuole (52%)
cytoplasm (44%)
Mitochondrion (4%)

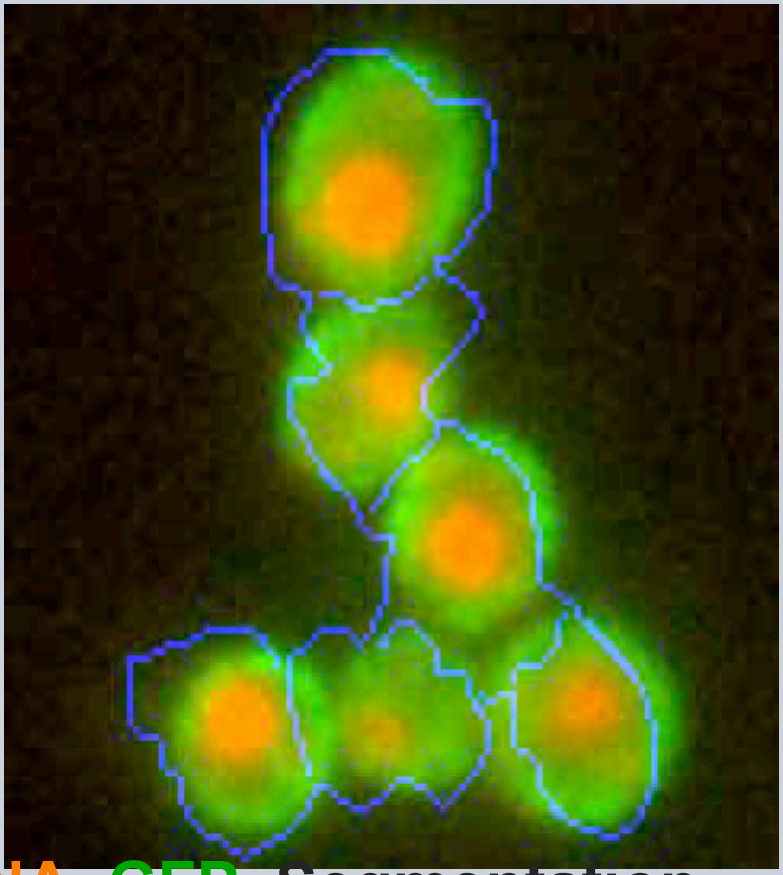


Example of Potentially Incorrect Label

ORF Name
YGR130C

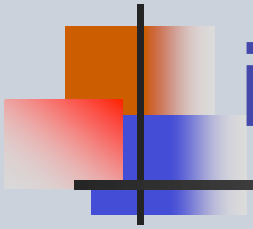
UCSF Location
punctate_composite

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



DNA GFP Segmentation

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.

Given a multi-cell Image

Each cell is well-segmented

Given single-cell classifiers to provide likelihood for each cell

Each cell is a random variable

Base accuracy is calculated

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

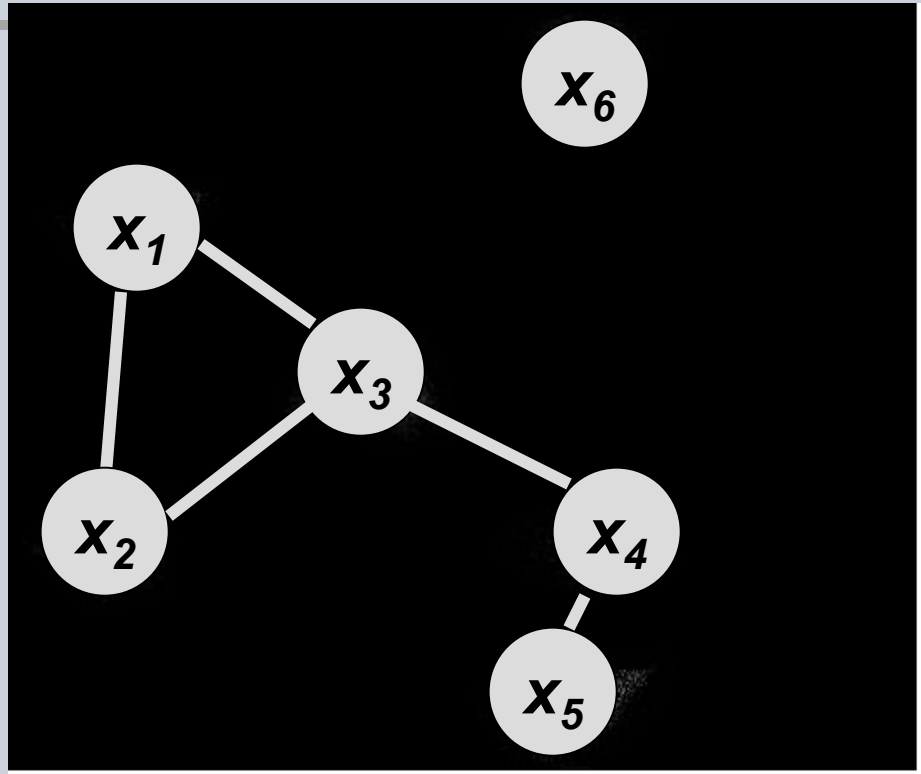
Graph Construction

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

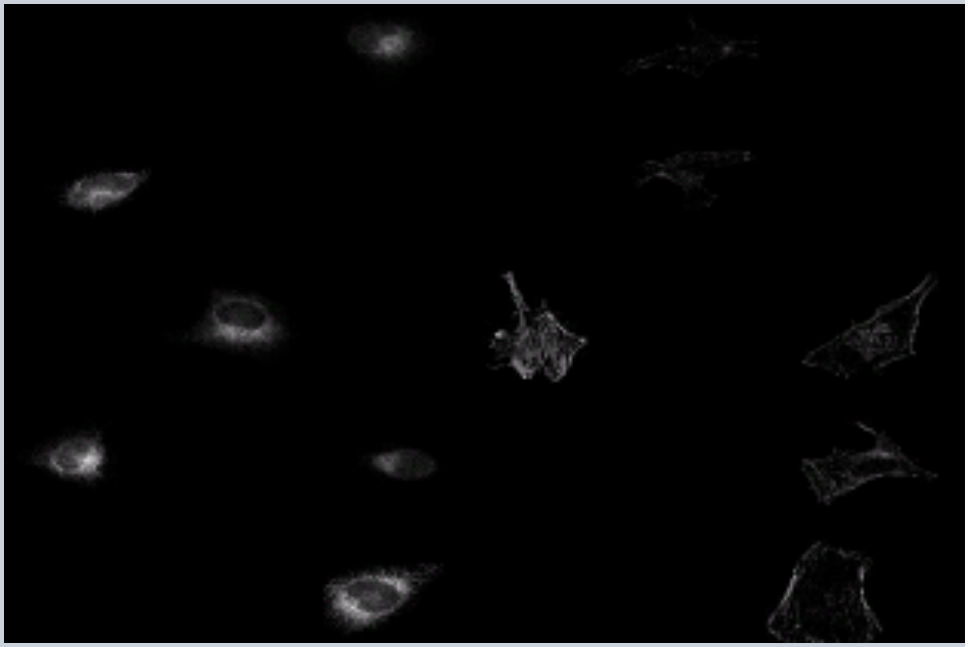
Iterate until no label changes
Calculate the new classification accuracy



Measure accuracy improvement

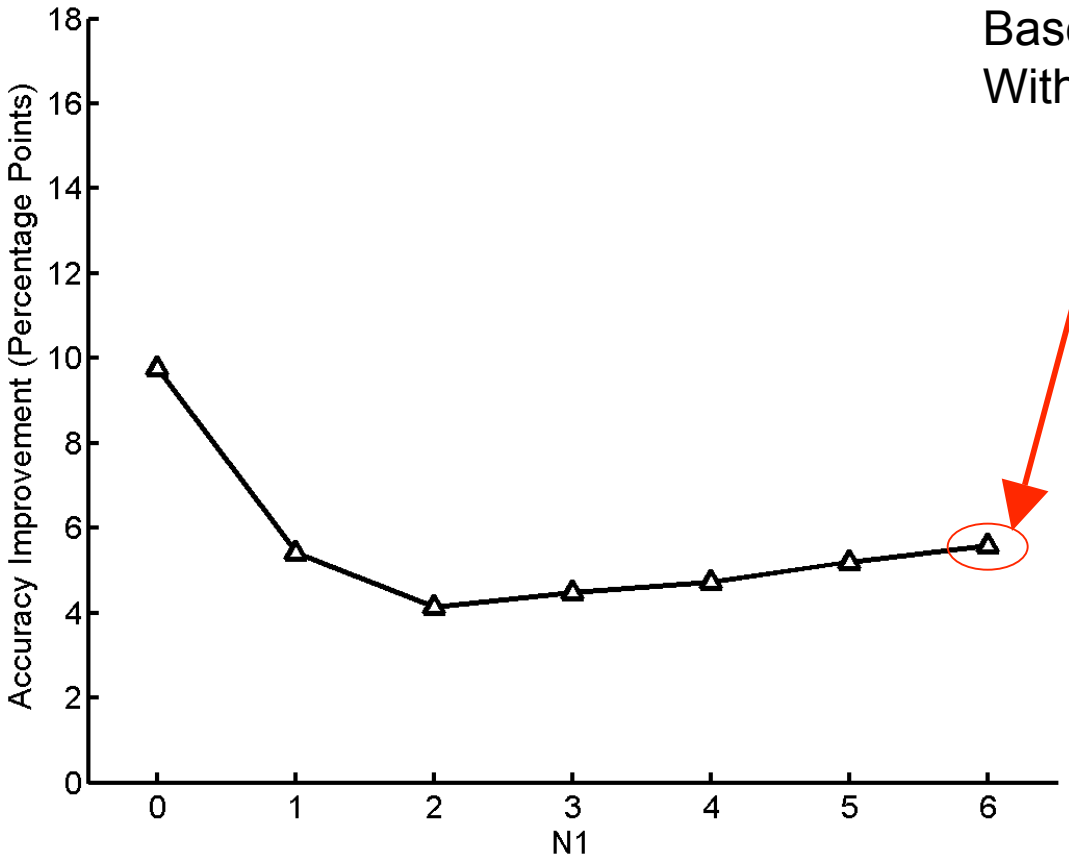
Evaluating PU

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

Results
- Closeness in Feature Space



Base accuracy: 90.1%
With PU: 95.7%

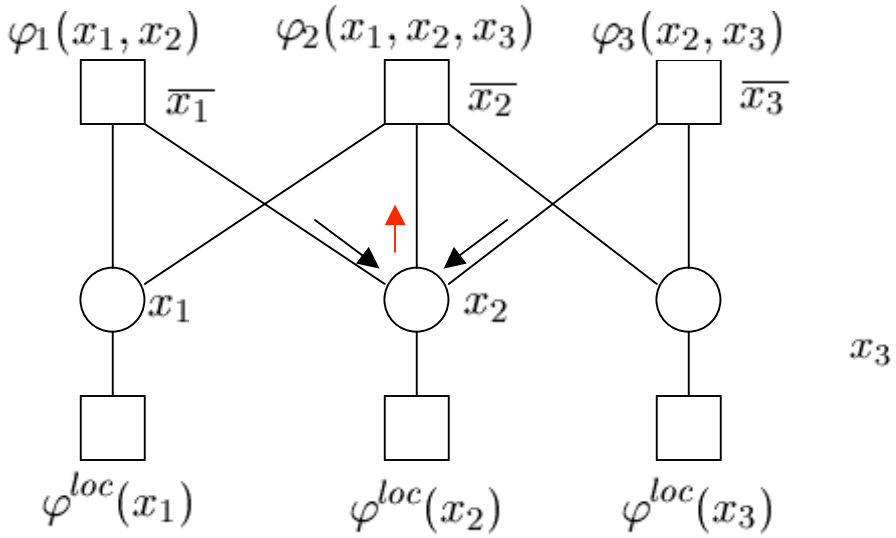
△ overall accuracy

(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)

Belief Propagation in Factor Graph

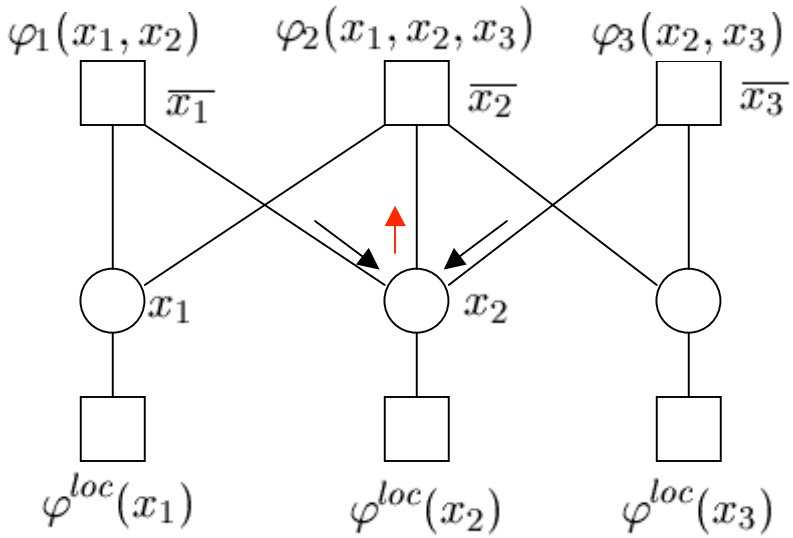
1.Messages from variable to factor



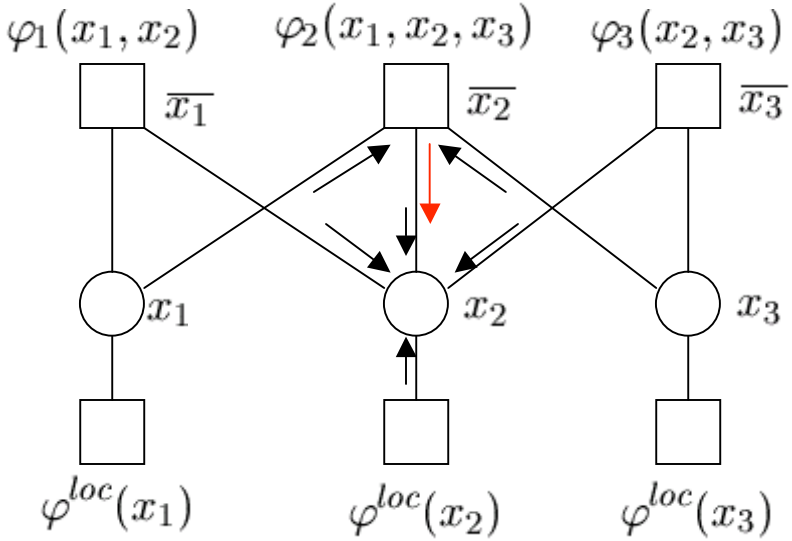
$$m_{j\to\bar{i}}(x_j) = \varphi^{\mathrm{loc}}(x_j) \prod_{l=1, l\neq i}^{k} m_{\bar{l}\to j}(x_j)$$

Belief Propagation in Factor Graph

1.Messages from variable to factor



2.Messages from factor to variable

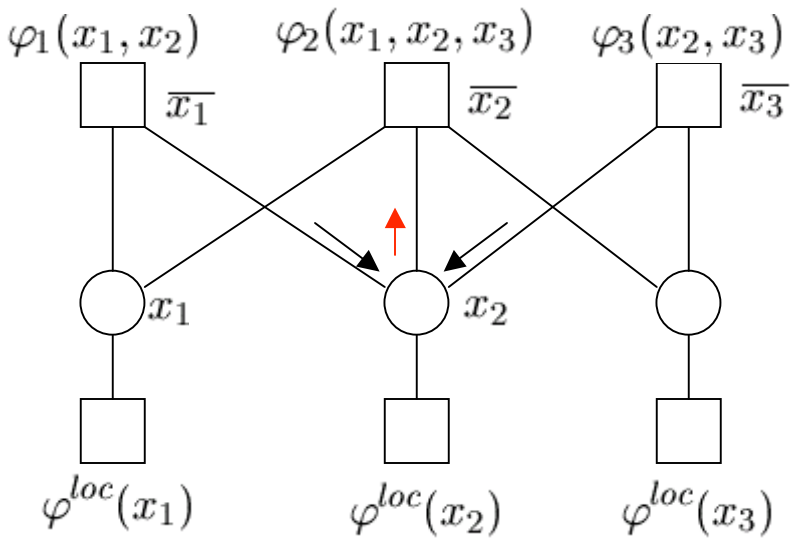


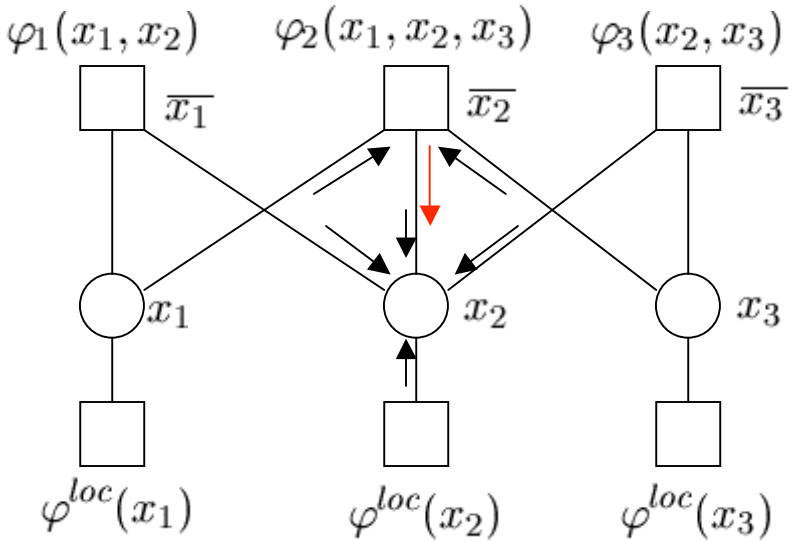
$$m_{j\to\bar{i}}(x_j) = \varphi^{\mathrm{loc}}(x_j) \prod_{l=1}^{k} m_{\bar{l}\to j}(x_j) \qquad m_{\bar{j}\to i}(x_i) = \sum_{\sim\{x_i\}} \varphi_j(x_1, \ldots, x_k) \prod_{l=1, l\neq i}^{k} m_{l\to\bar{j}}(x_l)$$

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

Posterior Probabilities can be calculated by

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



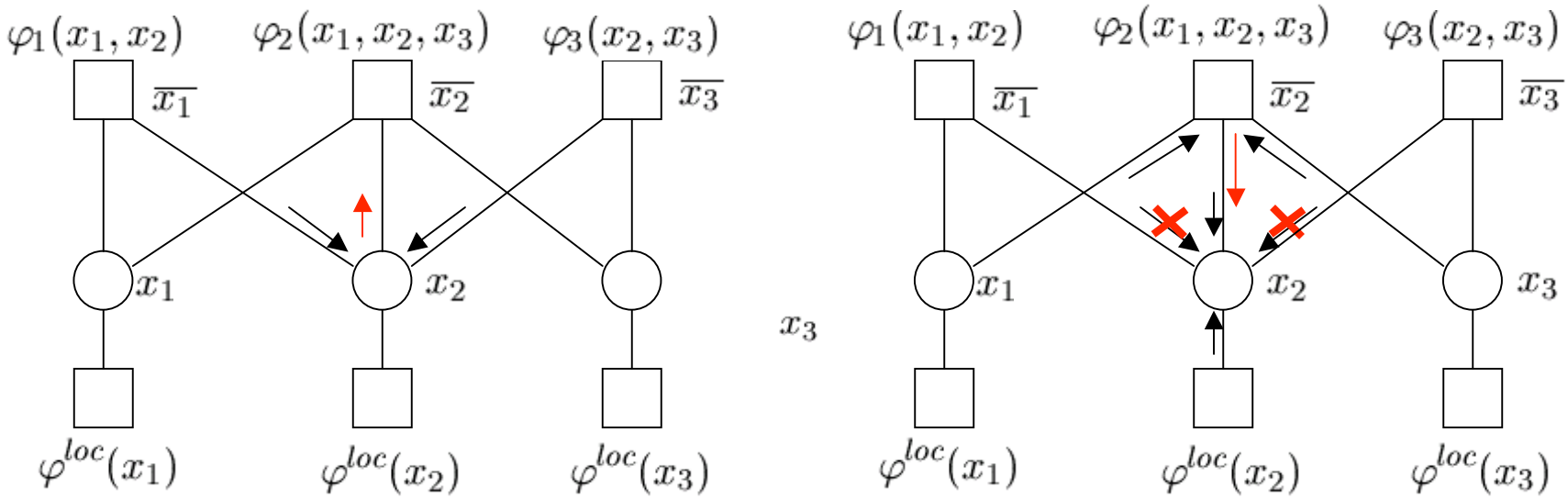


$$m_{j\to\bar{i}}(x_j)=\varphi^{\mathrm{loc}}(x_j)\prod_{l=1}^{k} m_{\bar{l}\to j}(x_j) \qquad m_{\bar{j}\to i}(x_i)=\sum_{\sim\{x_i\}}\varphi_j(x_1,\ldots,x_k)\prod_{l=1,l\neq i}^{k} m_{l\to\bar{j}}(x_l)$$

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

Posterior Probabilities can be calculated by

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



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

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

Inference Methods

EIPP Exact Inference with Potts Potential
LBPP Loopy Belief Propagation with Potts Potential
EIVP **Exact Inference with Voting Potential**
PUVP **Prior Updating with Voting Potential**

Results in small graphs (Considering Closeness in Feature Space)

Base accuracy: **88.29%**

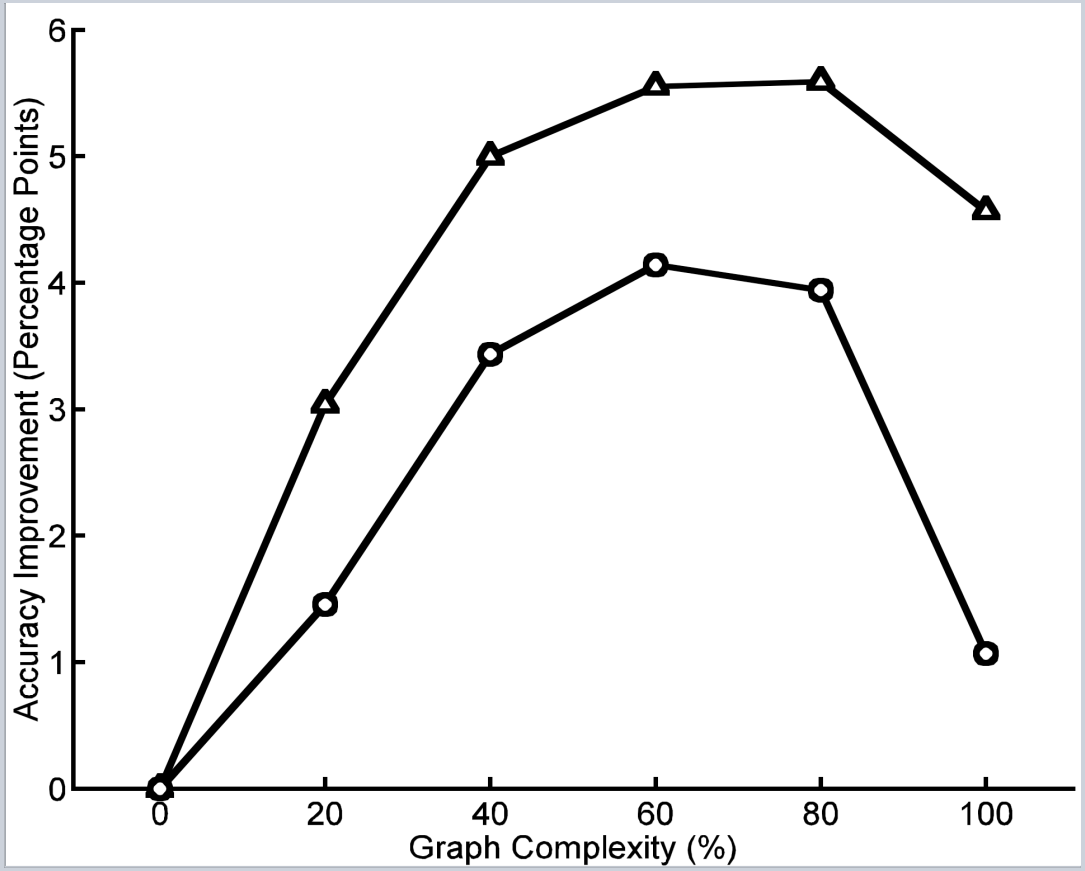
Accuracy Improvement	**EIPP**	**LBPP**	**EIVP**	**PUVP**
	1.58%	1.58%	2.84%	3.04%

(N1,N2) = (4,4) # of Class = 5

(Chen, Gordon, Murphy, 2006)

Results of Large Graphs

Base accuracy: **91.22%**



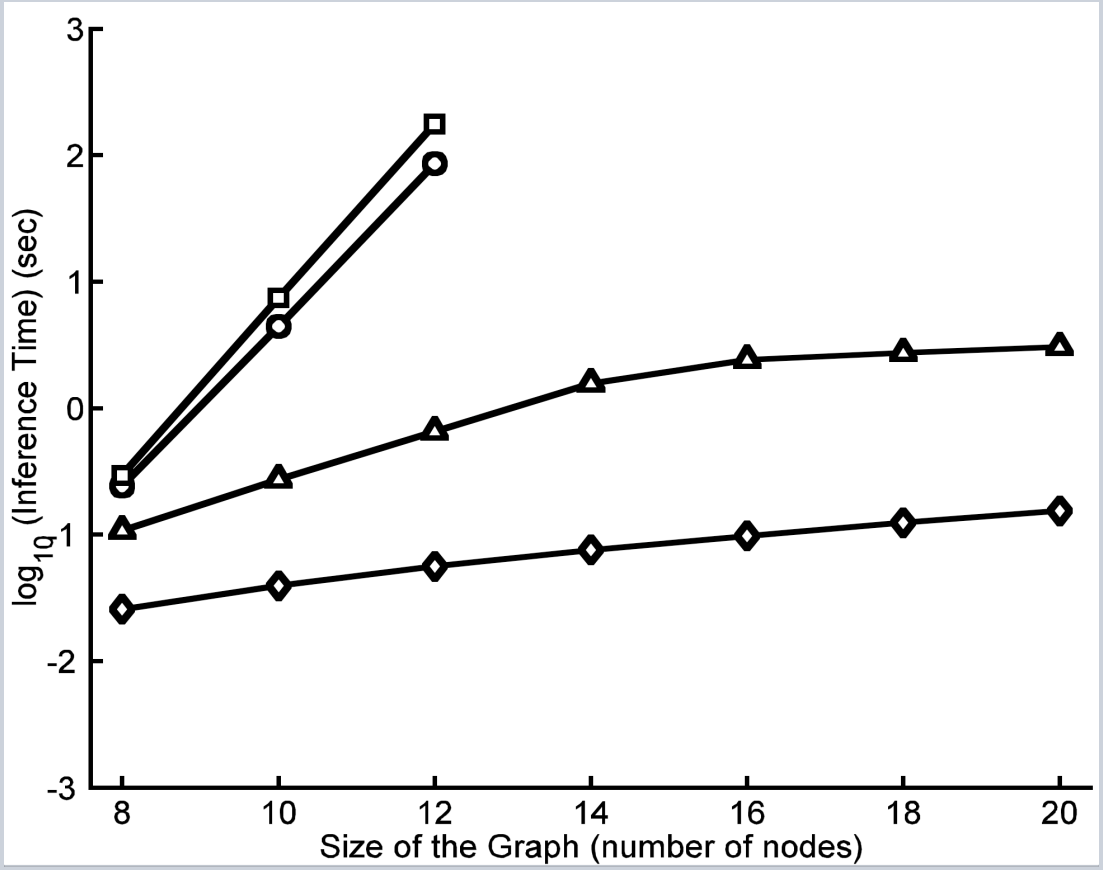
PUVP △

LBPP ○

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

Carnegie Mellon

Inference Time vs. Graph Size



EIPP

EIVP

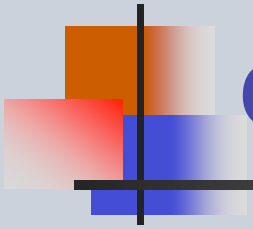
LBPP

PUVP

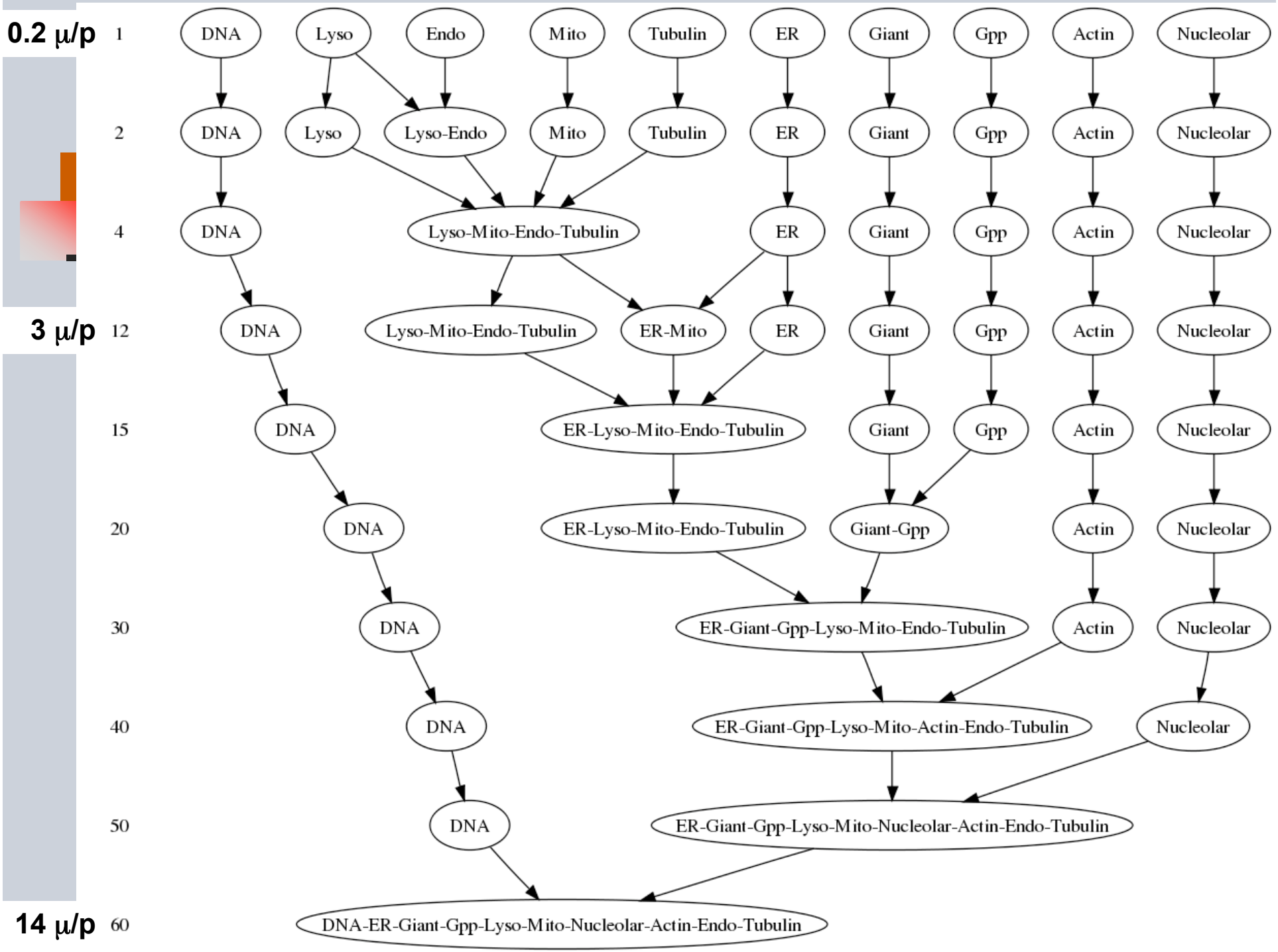
# of Class = 4

(Chen, Gordon, Murphy, 2006)

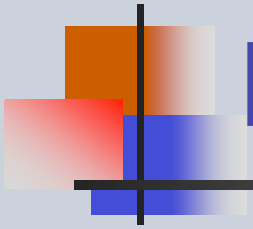
Image resolution and pattern discrimination



- 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)


0.2 µ/p
3 µ/p
14 µ/p
1
2
4
12
15
20
30
40
50
60
DNA
Lyso
Endo
Mito
Tubulin
ER
Giant
Gpp
Actin
Nucleolar
Lyso-Endo
Lyso-Mito-Endo-Tubulin
ER-Mito
ER-Lyso-Mito-Endo-Tubulin
Giant-Gpp
ER-Giant-Gpp-Lyso-Mito-Endo-Tubulin
ER-Giant-Gpp-Lyso-Mito-Actin-Endo-Tubulin
ER-Giant-Gpp-Lyso-Mito-Nucleolar-Actin-Endo-Tubulin
DNA-ER-Giant-Gpp-Lyso-Mito-Nucleolar-Actin-Endo-Tubulin

Supervised vs. Unsupervised Learning



- This work demonstrated the feasibility of using classification methods to assign all proteins to known major classes
- Do we know all locations? Are assignments to major classes enough?
- Need approach to discover classes

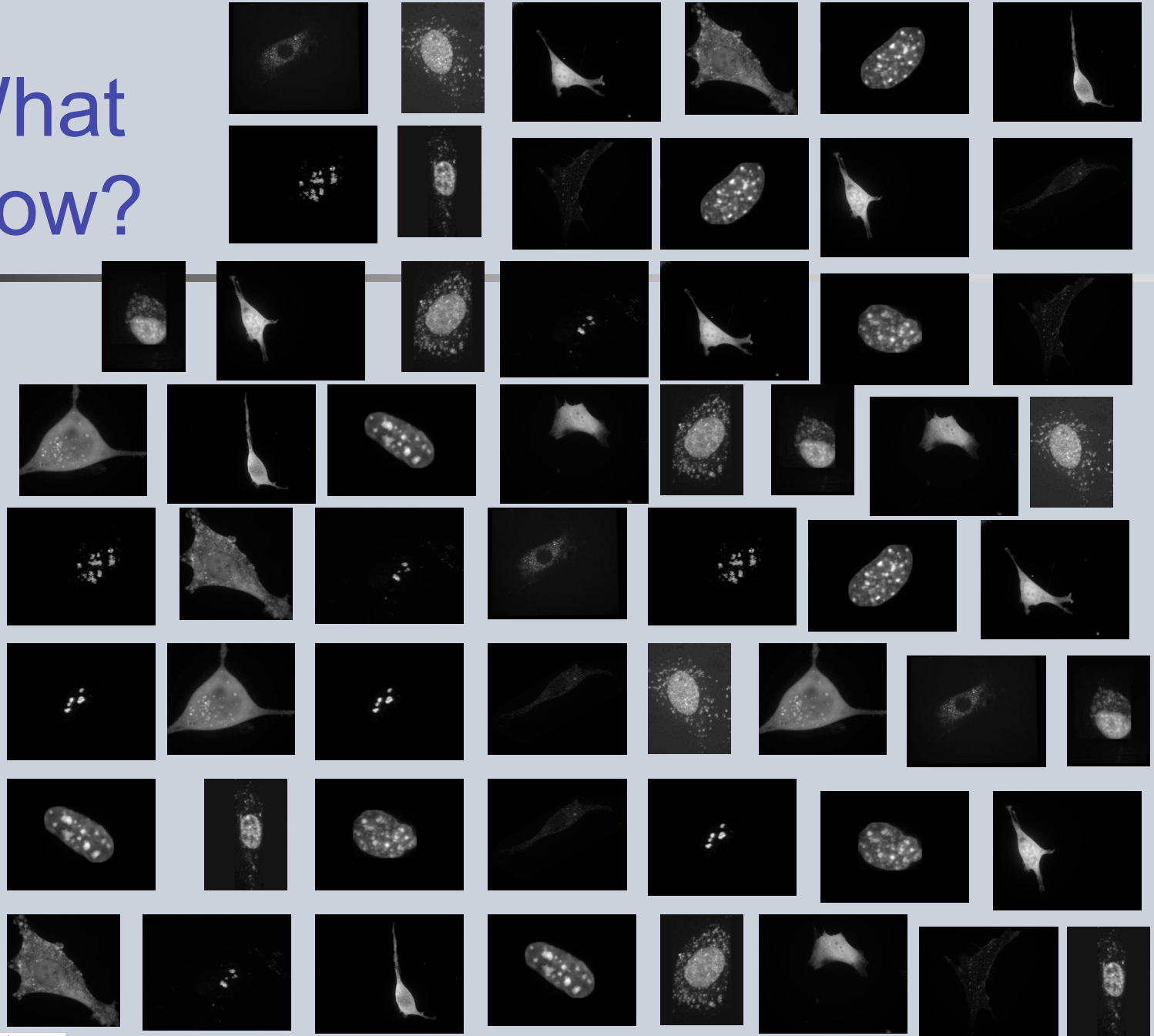
Location Proteomics

- **Tag** many proteins (many methods available; we use **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 in each cell
- 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

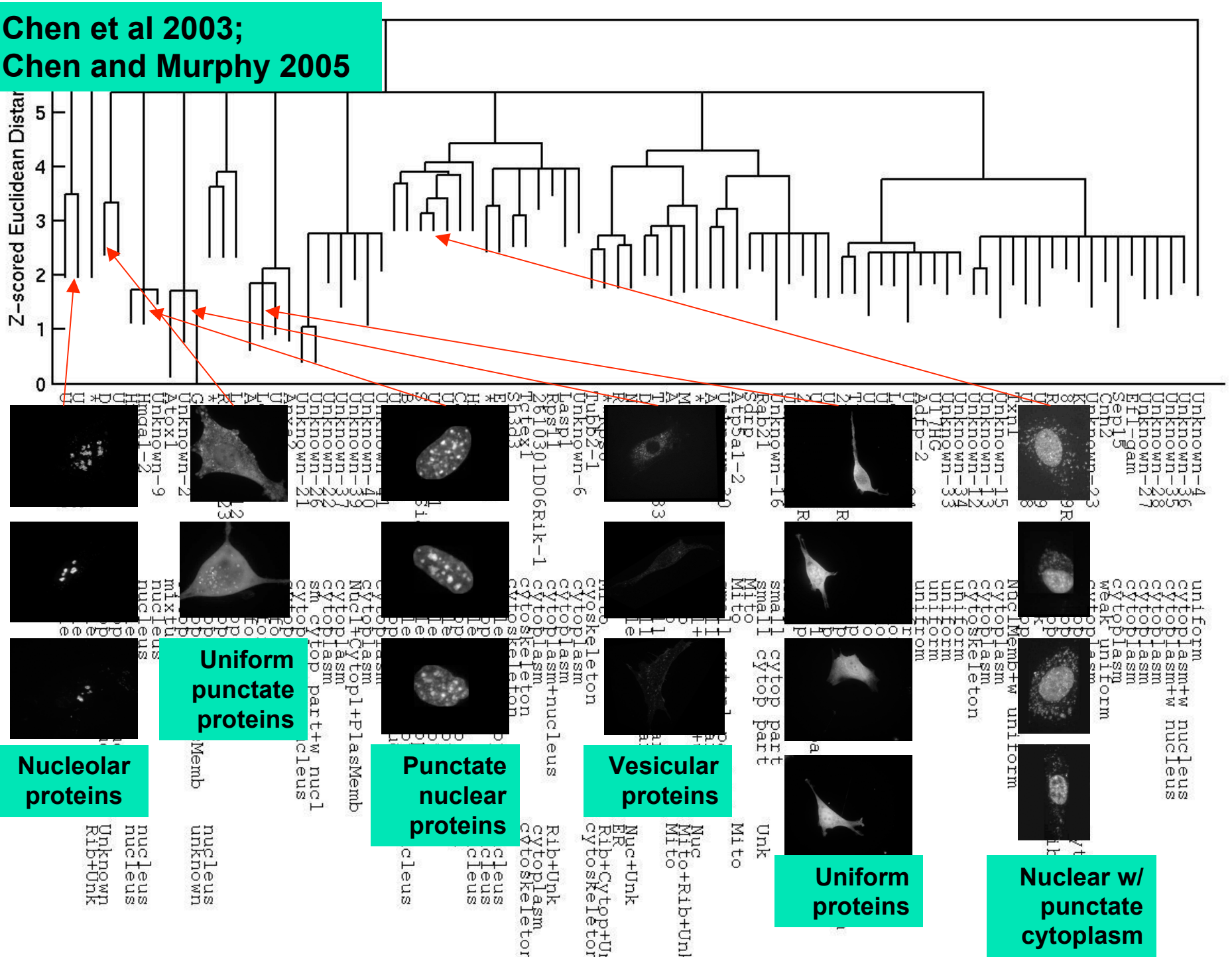
What Now?

Group ~90 tagged clones by pattern



Carnegie Mellon

Chen et al 2003; Chen and Murphy 2005



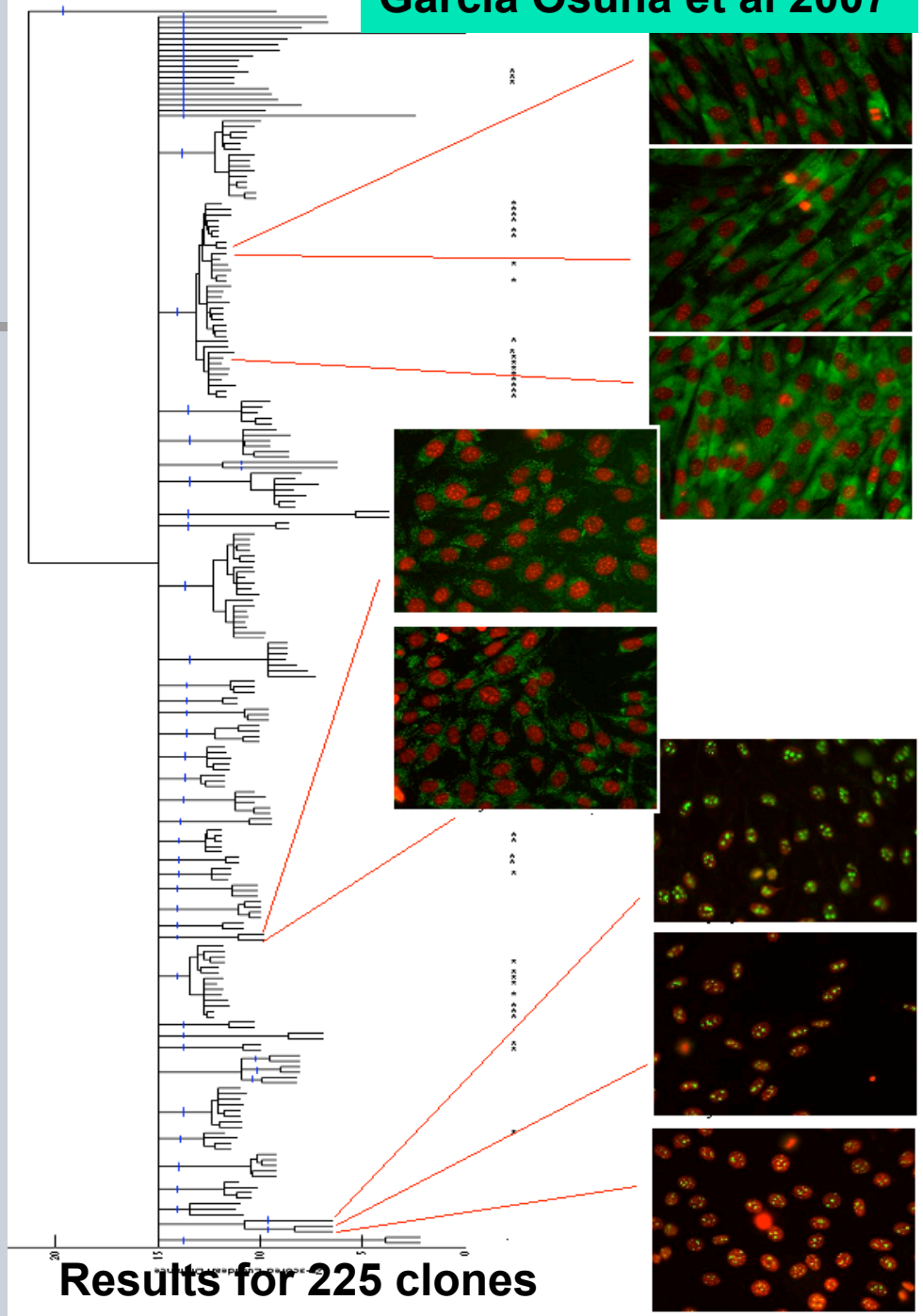
CD-tagging project

- Running ~100 clones/wk
- Automated imaging

Elvira Garcia Osuna

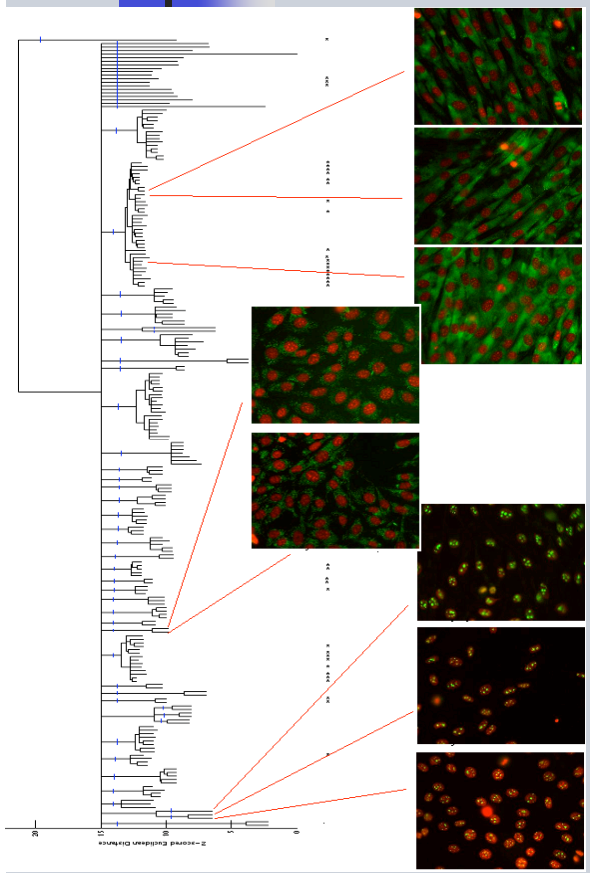
Carnegie Mellon

Garcia Osuna et al 2007



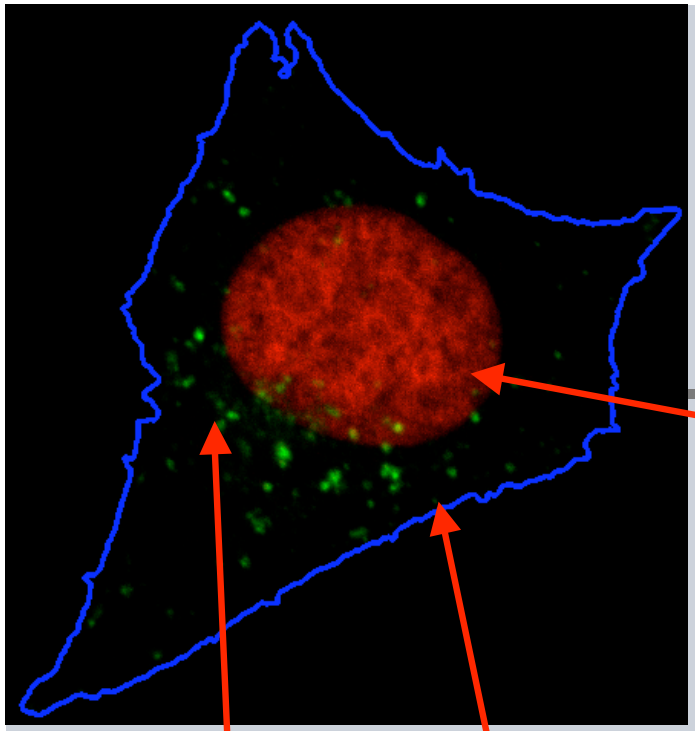
Results for 225 clones

Subcellular Location Families and Generative Models

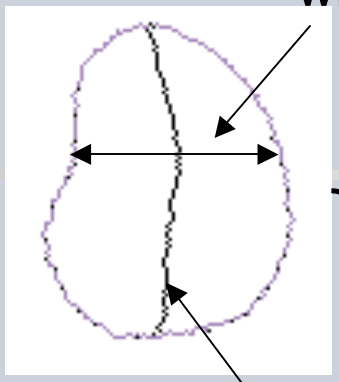


- Rather than using words (e.g., GO terms) to describe location patterns, can make entries in protein databases that give its Subcellular Location Family - a specific node in a Subcellular Location Tree
- Provides necessary resolution that is difficult to obtain with words
- How do we communicate patterns: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

Generative Model Components



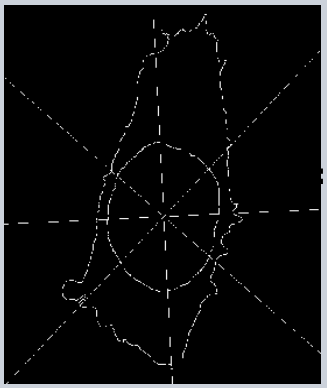
Nucleus



width

Medial axis

Cell membrane

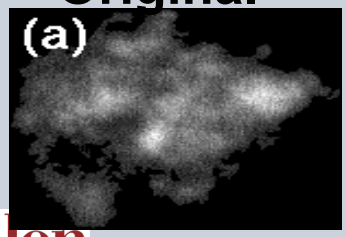


$$\frac{d_1 + d_2}{d_2}$$

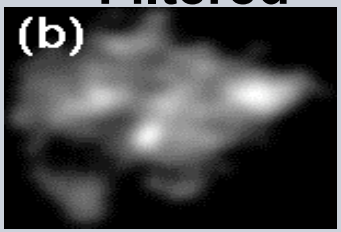
Model parameters

Protein objects

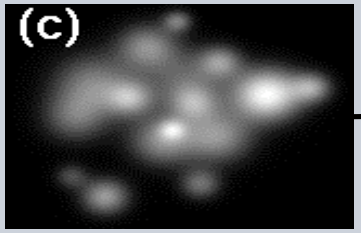
Original



Filtered



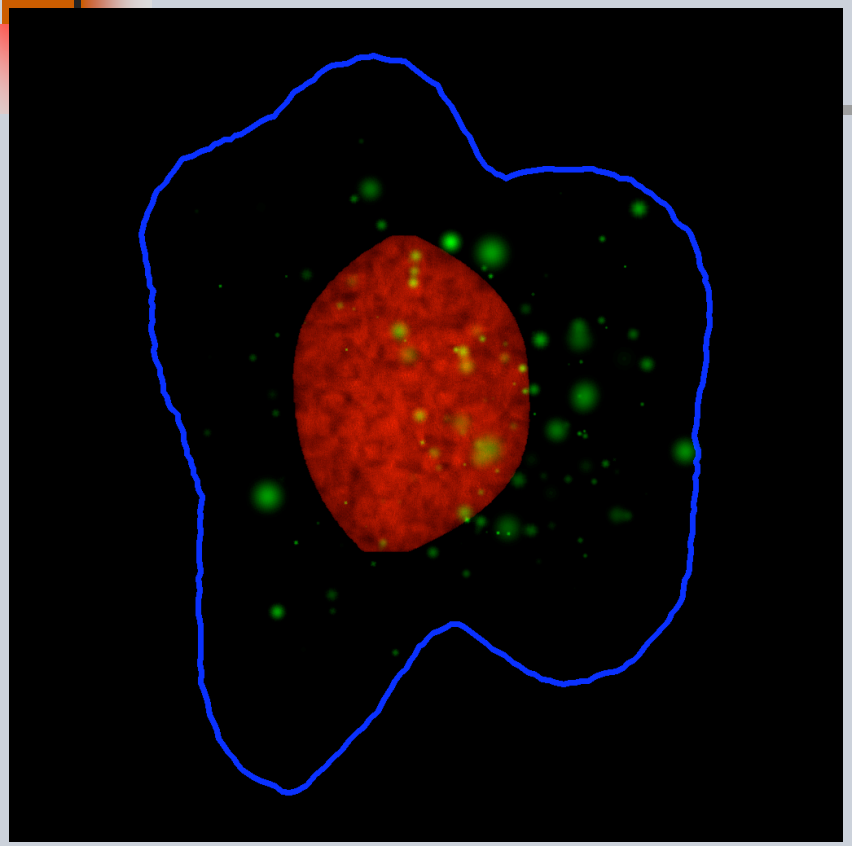
Fitted



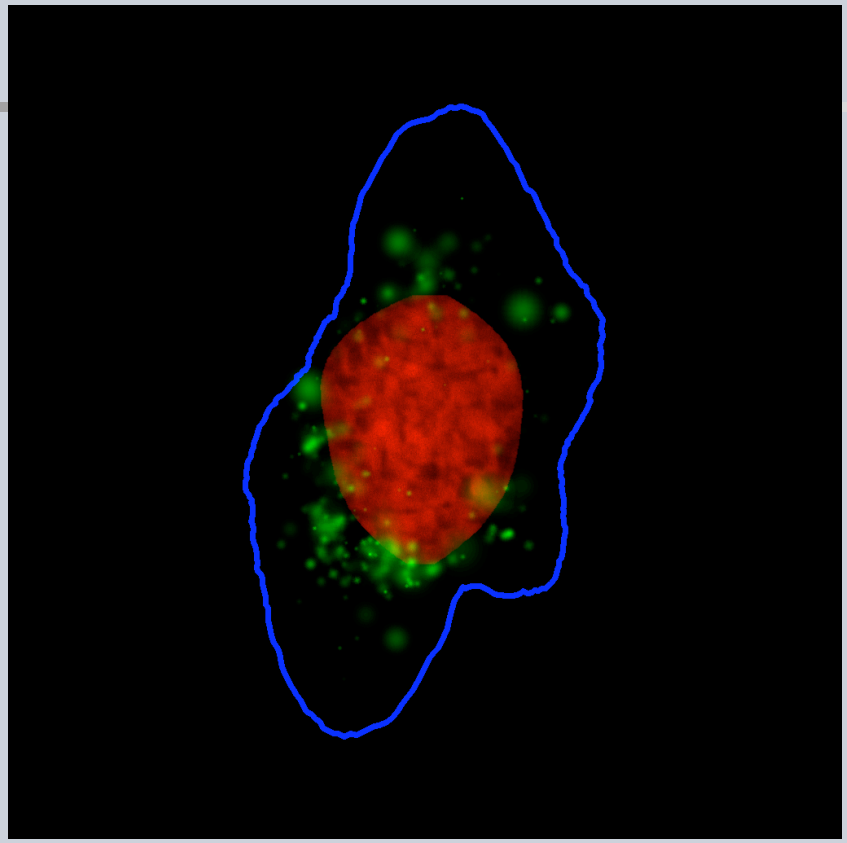
Zhao & Murphy 2007

Carnegie Mellon

Synthesized Images



Lysosomes



Endosomes

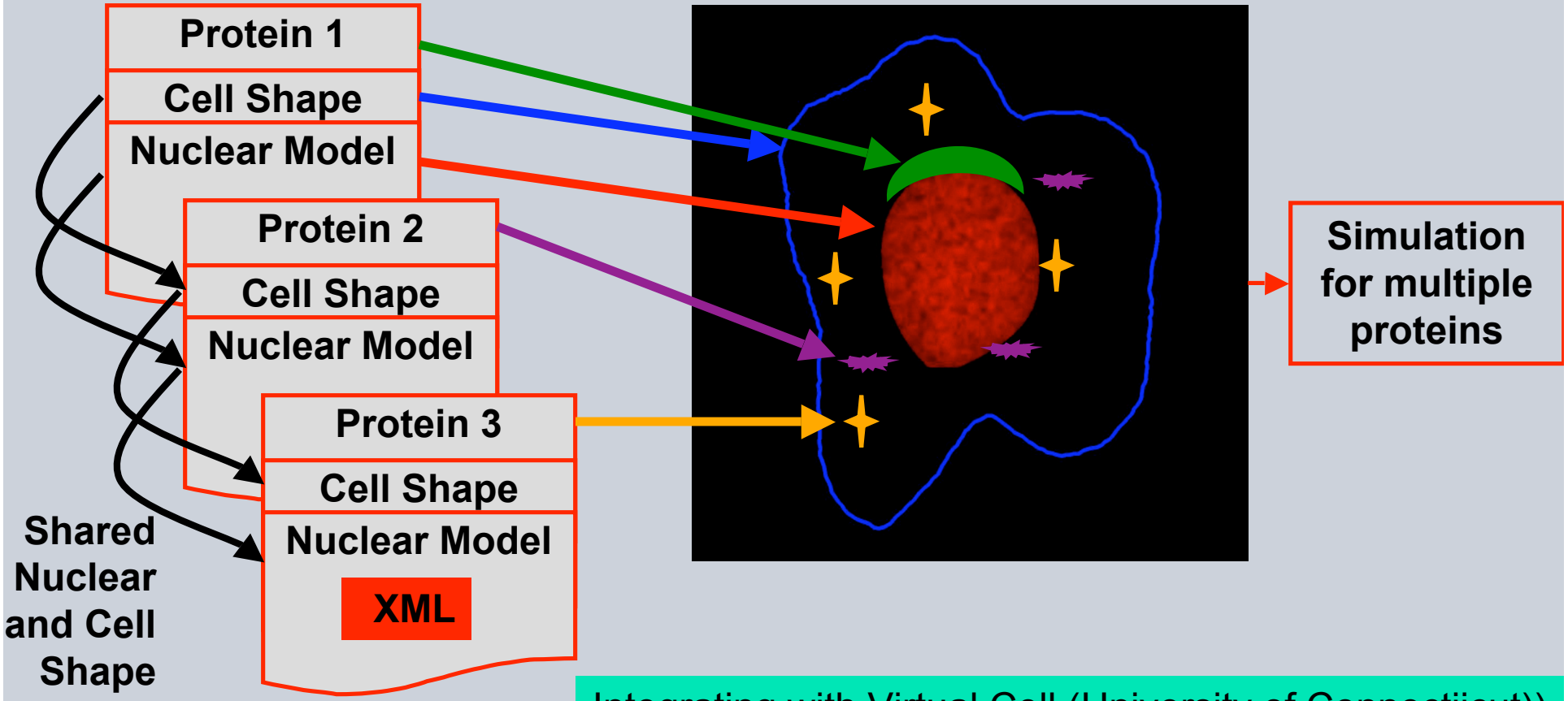
- Have XML design for capturing model parameters
- Have portable tool for generating images from model

Evaluation of synthesized images

Classification of synthesized images by a classifier trained on real images. Classification based on features that made 94% of real images distinguishable

True Classification	*Output of Classifier*									
	DNA	ER	Actin	Gia	Gpp	Lyso.	Mit.	Nuc	Endo.	Tub.
DNA	**100**	0	0	0	0	0	0	0	0	0
Gia	0	0	0	**31**	**54**	13	0	1	1	0
Gpp	0	0	0	**24**	**62**	11	0	2	1	0
Lyso.	0	0	0	7	4	**50**	7	0	32	0
Mit.	0	0	0	0	0	2	**18**	0	80	0
Nuc.	1	0	0	4	15	0	0	**80**	0	0
Endo.	0	2	0	0	0	1	2	0	**91**	4

Combining Models for Cell Simulations

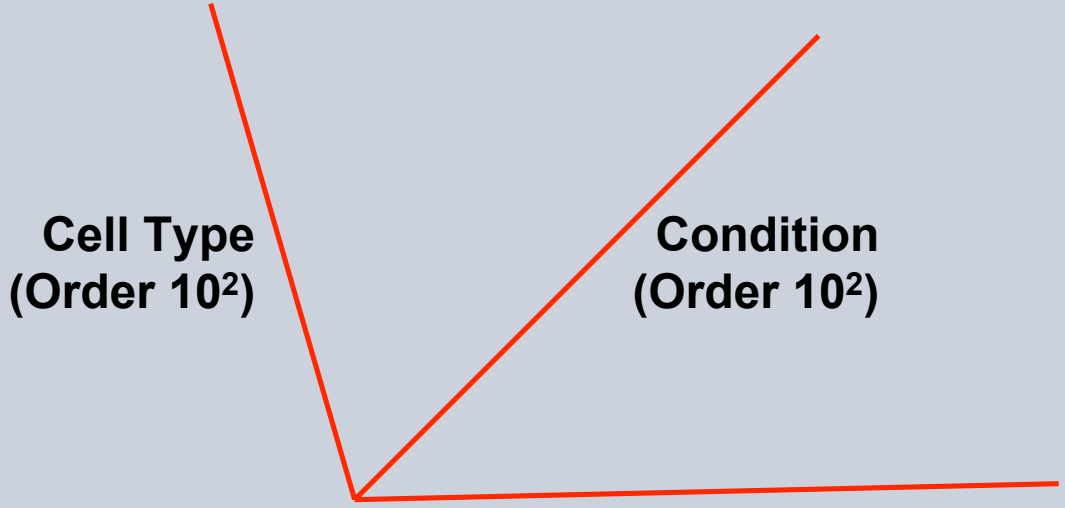


Integrating with Virtual Cell (University of Connectiicut)) and M-Cell (Pittsburgh Supercomputing Center)

PSLID: Protein Subcellular Location Image Database

- Version 4 to be released January 2008
  - Adding ~50,000 analyzed images (~1,000 clones, ~350,000 cells) from **3T3 cell random tagging project**
  - Adding ~7,500 analyzed images (~2,500 genes, ~40,000 cells) from **UCSF yeast GFP database**
  - Adding ~400,000 analyzed images (~3,000 proteins, 45 tissues) from **Human Protein Atlas**
  - Adding **generative models** to describe subcellular patterns consisting of discrete objects (e.g., lysosomes, endosomes, mitochondria)
  - Return **XML file with real images** that match a query
  - Return **XML file with generative model** for a pattern
  - Connecting to MBIC TCNP **fluorescent probes database**
  - Connecting to CCAM TCNP **Virtual Cell system**

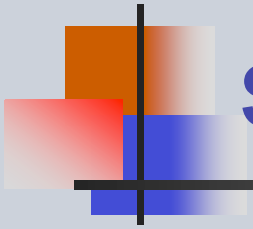
The future of subcellular location analysis



Plus: Time scale from subsecond to years

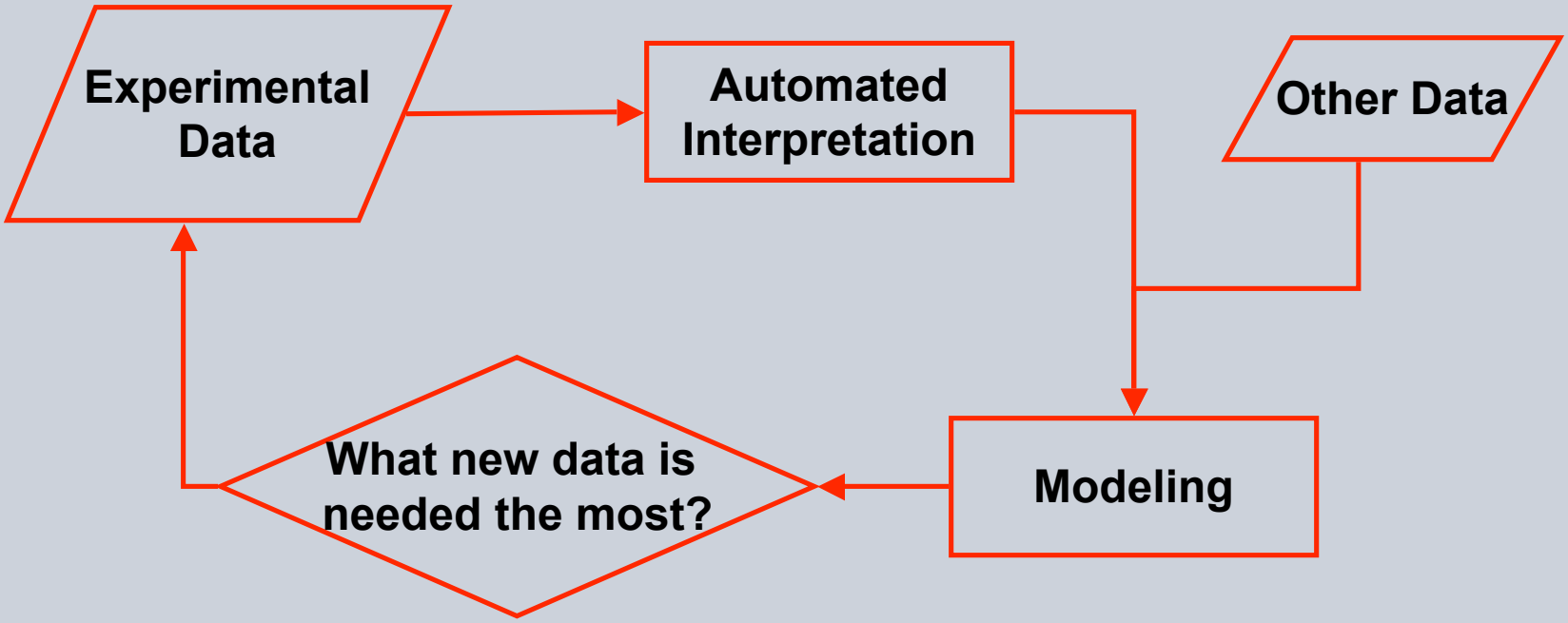
Carnegie Mellon

How do we really analyze subcellular location?

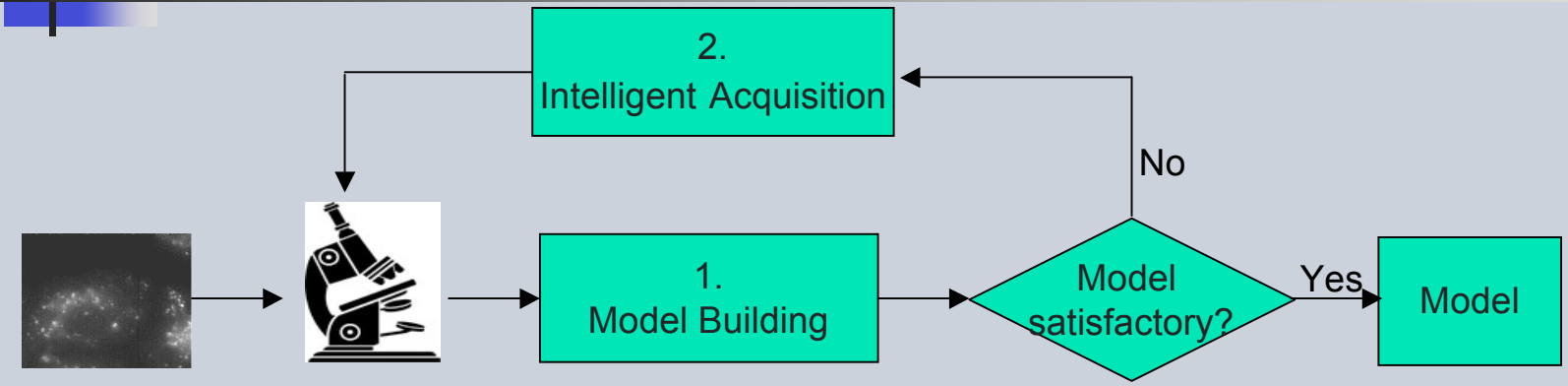


- Scope of problem argues for cooperation on grand scale
- Need intelligent (optimized) data collection: probabilistic methods to integrate available data, make predictions, suggest experiments and iterate

Automated Science (Active Learning)



Efficient Acquisition and Learning of Fluorescence Microscope Data Models - with Jelena Kovacevic



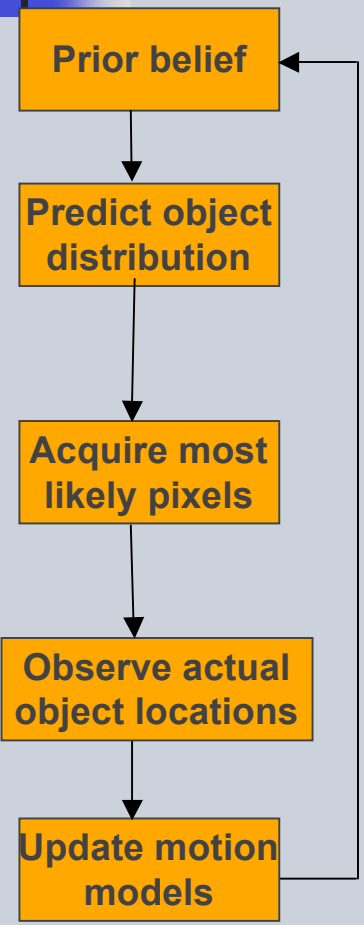
Develop a mathematical framework and algorithms
to build accurate models of fluorescence microscope data sets
as well as design intelligent acquisition systems based on those models

1. Use all the input from the microscope to model the data set

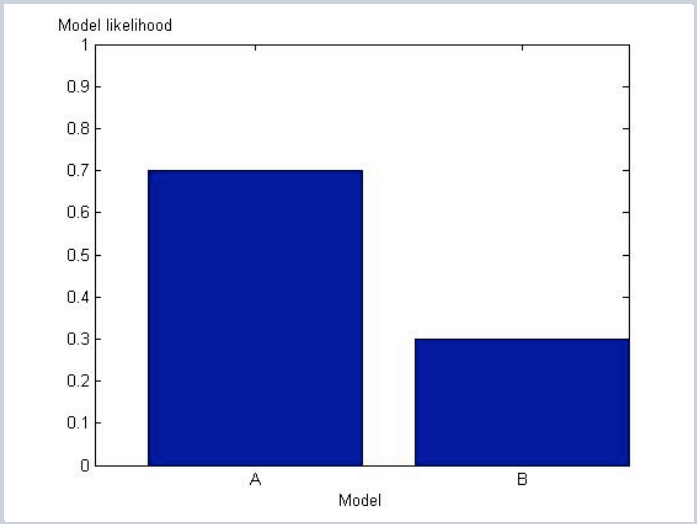
2. Choose acquisition requests that allow us to construct an accurate model in the shortest amount of time

Carnegie Mellon

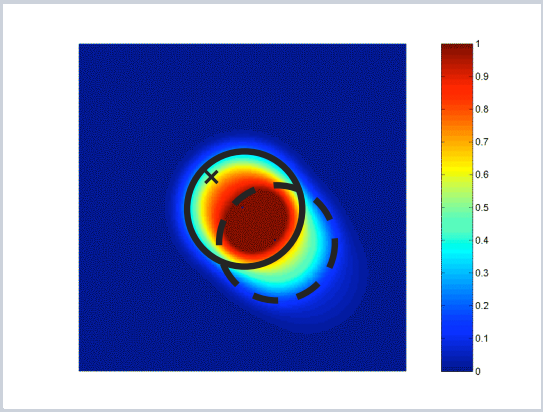
Intelligent Acquisition - Unknown Motion Model

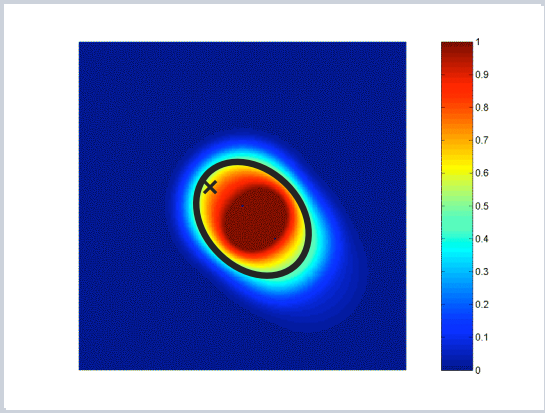


Model likelihood



Model

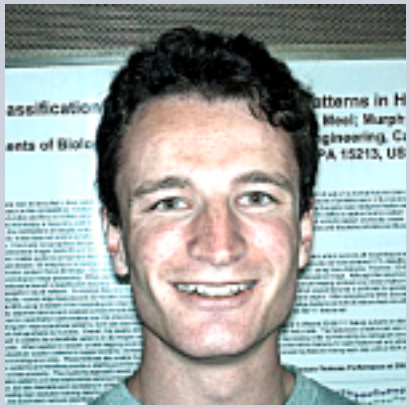




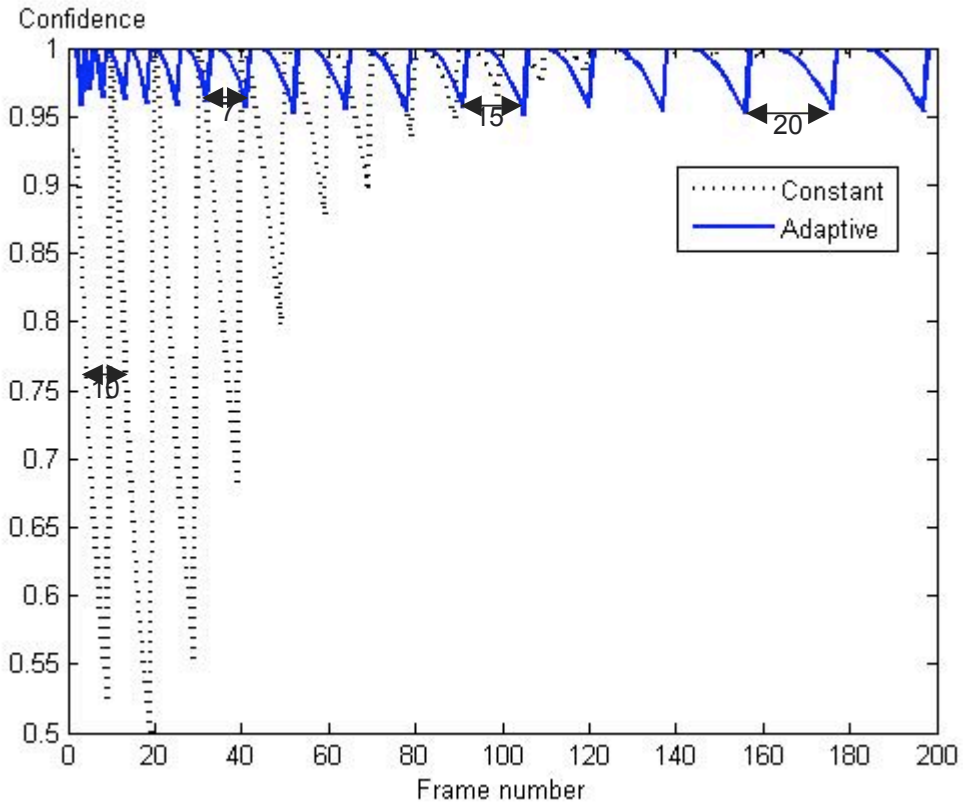
Carnegie Mellon

Intelligent Acquisition - Frame Rate

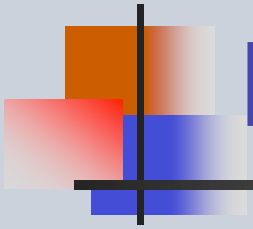
- Acquire frame if confidence in object's location falls below 95%
- We acquire less frequently when motion model is learned



Charles Jackson



Carnegie Mellon



More Challenges

- Models and conditional models for subcellular patterns
- Estimating model confidence in active learning of nested models

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)

Carnegie Mellon

Molecular Biosensor and Imaging Center

Welcome to the Molecular Biosensor and Imaging Center website.

Mission

To develop fluorescence detection technologies for biomedical research and NASA space exploration.

NIH Technology Center for Networks and Pathways

Carnegie Mellon

Alan Waggoner and Simon Watkins (Pitt)

NCIBI
NATIONAL CENTER FOR INTEGRATIVE BIOMEDICAL INFORMATICS

Search

home

log in · join

you are here: home

navigation

- Home
- About NCIBI
- Computational Technology
- Driving Biological Problems
- Resources and Software
- Education and Training
- Working With NCIBI
- Publications
- Sponsors and Collaborators
- Other NCBC Sites
- Events
- News

internal sites

- Collaboration Portal
- Wiki

National Center for Integrative Biomedical Informatics (NCIBI)

by plone — last modified 2005-09-29 09:29 AM

Mission

The mission of the NCIBI is to facilitate scientific exploration of complex disease processes on a much larger scale than is currently feasible.

The Center develops and interactively integrates analytical and modeling technologies to acquire or create context-appropriate molecular biology information from emerging experimental data, international genomic databases, and the published literature.

The NCIBI supports information access and data analysis workflow of collaborating biomedical researchers, enabling them to build computational and knowledge models of biological systems validated through focused work on specific diseases. The initial driving biological problems are prostate cancer progression, organ-specific complications of type 1 diabetes, genetic and metabolic heterogeneity of type 2 diabetes, and genetic susceptibility and phenotypic subclassification of bipolar depressive disease.

The Center also has outreach, training, and education programs.

Current NCIBI Collaborators

news

- NCIBI presents at the NIH New NCBC Kickoff in Bethesda 2005-12-23
- UM Press Release for NCIBI 2005-09-30
- R01 Collaboration Opportunity with NCIBI 2005-09-28

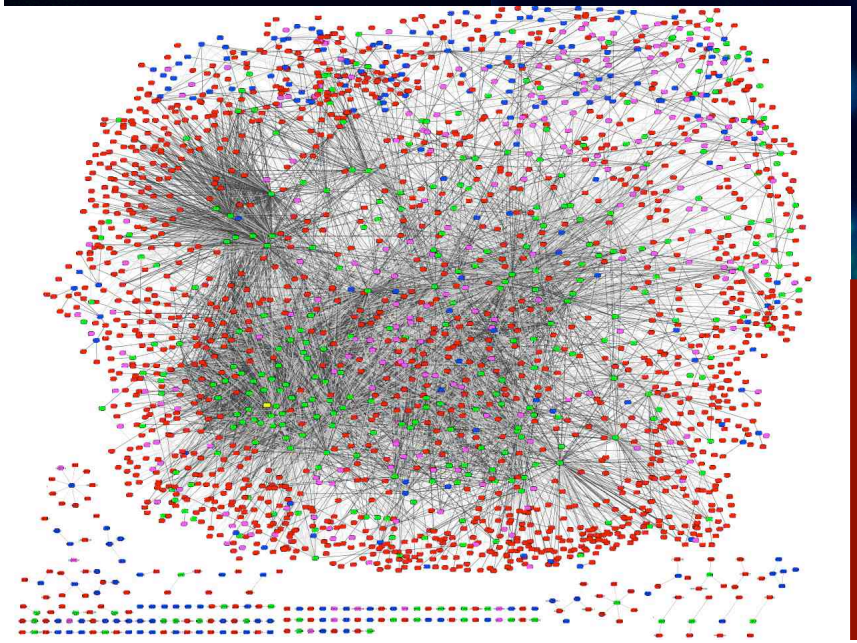
More news...

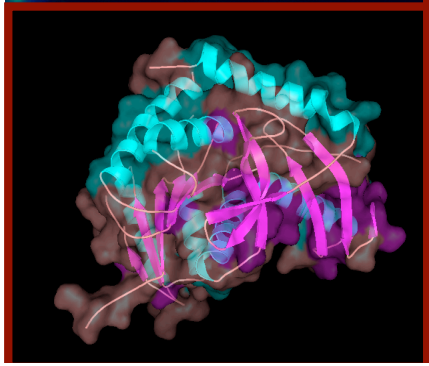
RAY AND STEPHANIE LANE

Center for Computational Biology

Carnegie Mellon

Recruiting postdocs and faculty!





Our mission:

To realize the potential of machine learning for understanding complex biological systems

To advance cancer diagnosis and treatment