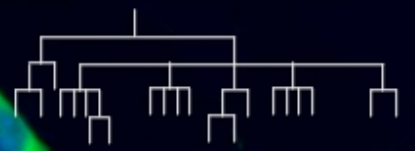
Principles of Bioimage Informatics: Focus on machine learning of cell patterns

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



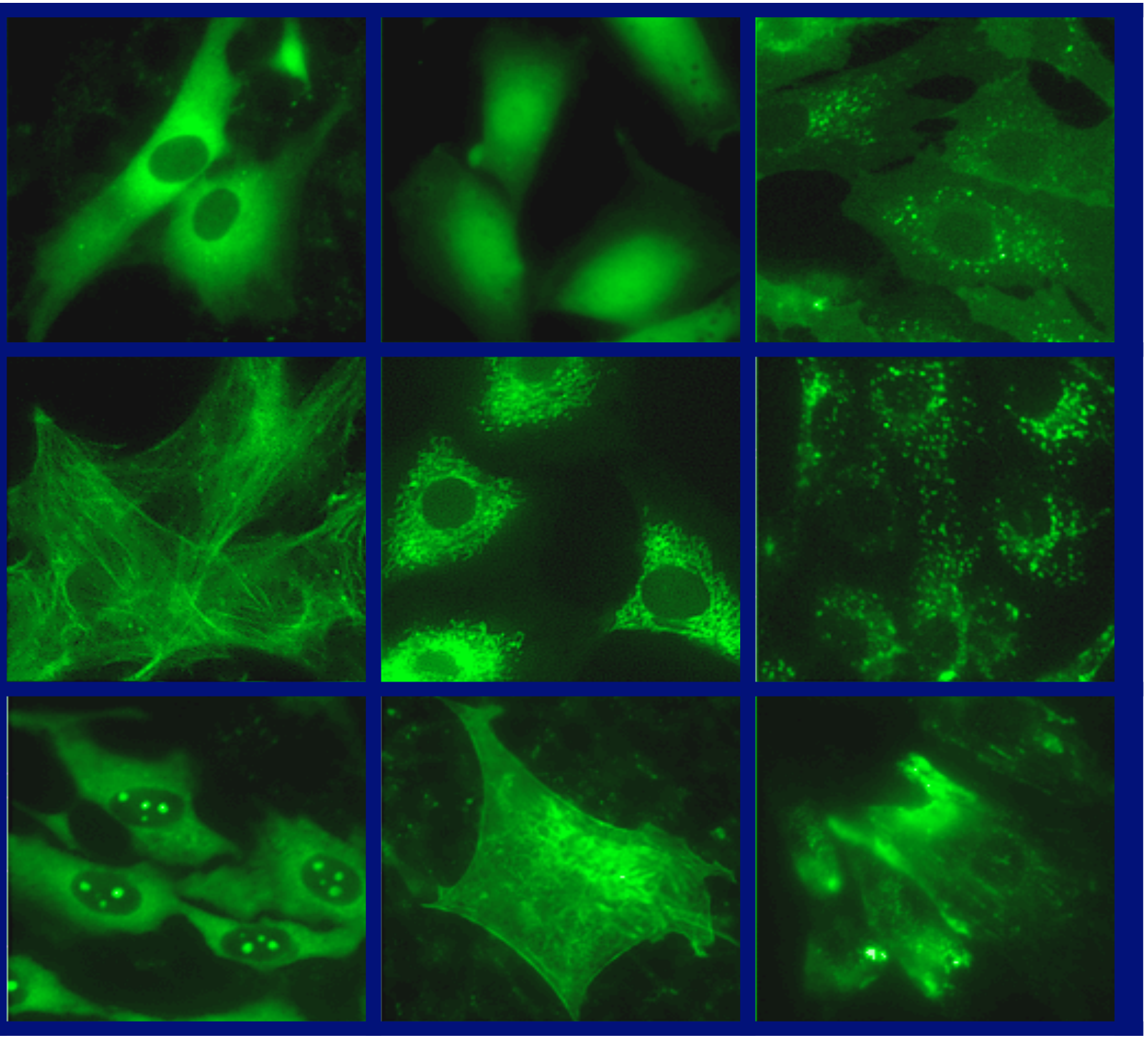
Center for Bioimage Informatics

from image to knowledge

RAY AND STEPHANIE LANE

Center for Computational Biology

Carnegie Mellon



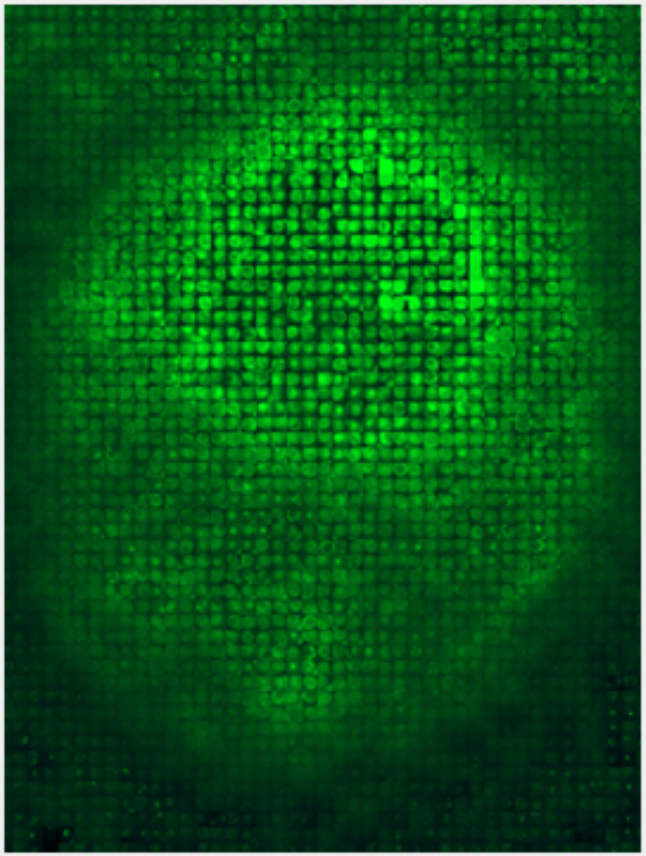
Images from Randtag project: Murphy, Jarvik, Berget

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>

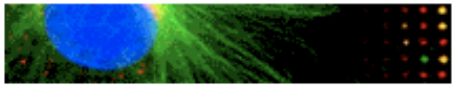
http://www.dkfz.de/LIFEdb/(rze40cvcqowayfu44rtthp55)/LIFEdb.aspx

lifedb

Database for Localization, Interaction, Functional assays and Expression of Proteins

home@mga
home@dkfz
home@smp-cell.org
LIFEdb description
credits

LIFEdb



Simple Query | ID Query | Localization Query | Assay Query | Complex Query | Table | Help

Configure output:

- [x] Gene Name
- [x] Chromosomal Location
- [ ] Insert size (bp)
- [ ] Orf size (aa)
- [ ] IEP
- [ ] Mol. wt.
- [+] identifiers

- [x] Localization
- [ ] S-Phase data
- [ ] Electronic Northen
- [ ] Pred. localization
- [ ] Motifs and Domains
- [ ] Best BlastP hit(s) swissprot

Enter identifier, keyword, subcellular compartment or chromosomal location

Search DB | Reset

Carnegie Mellon

http://www.proteinatlas.org/

human protein atlas

hpr | the project | protein atlas | dictionary | disclaimer | submission of antibodies | help

HUMAN PROTEIN ATLAS

The human protein atlas shows expression and localization of proteins in a large variety of normal human tissues, cancer cells and cell lines with the aid of immunohistochemistry (IHC) images and immunofluorescence (IF) confocal microscopy images.

Enter search: search

Advanced search

Select a chromosome:

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

OTHER

Or a protein class:

Enzymes | GPCRs excl olfactory receptors | Kinases | Peptidases | Transcription factors | Transporters | More...

Filter search to show genes with tissue profiles

Version: **5.0** Atlas updated: **2009-06-16** (release history)
Atlas content: **8832** antibodies and **7,334,244** images.

2009-06-16
A new version (5.0) has been released including antibodies targeting protein products from 1/3 of the human protein-coding genes. Tutorials on how to use the Human Protein Atlas have also been added.

The Human Protein Atlas now contains subcellular localization data (IF) for 3541 genes.

See release history for further details.

Knut och Alice Wallenbergs Stiftelse

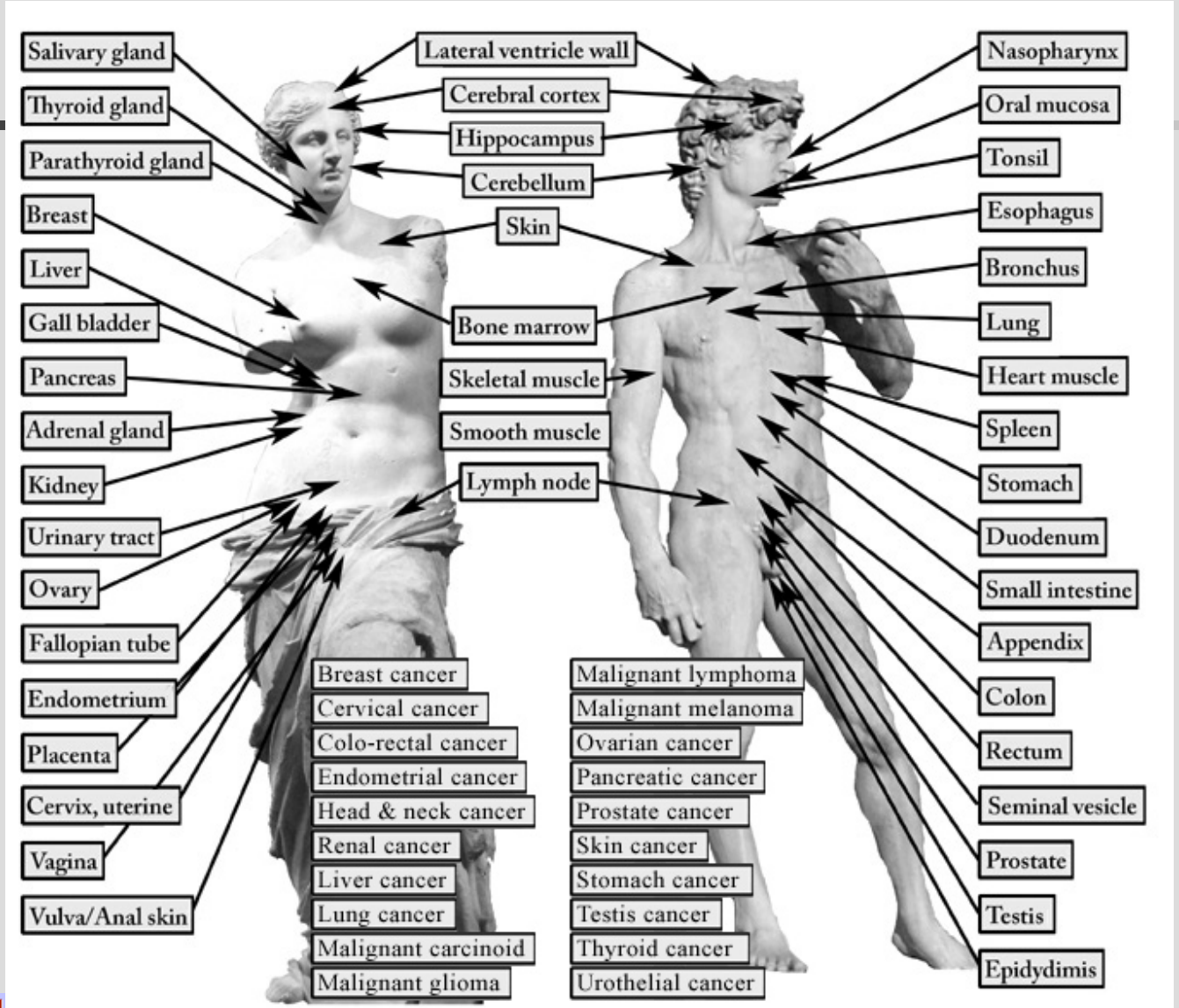
The HPR project is funded by the Knut & Alice Wallenberg foundation. The atlas is part of the HUPO Human Antibody Initiative (HAI).

KTH | UPPSALA

Send questions, comments or suggestions to: contact@hpr.se. | FAQ / Help

Carnegie Mellon

Human Protein Atlas



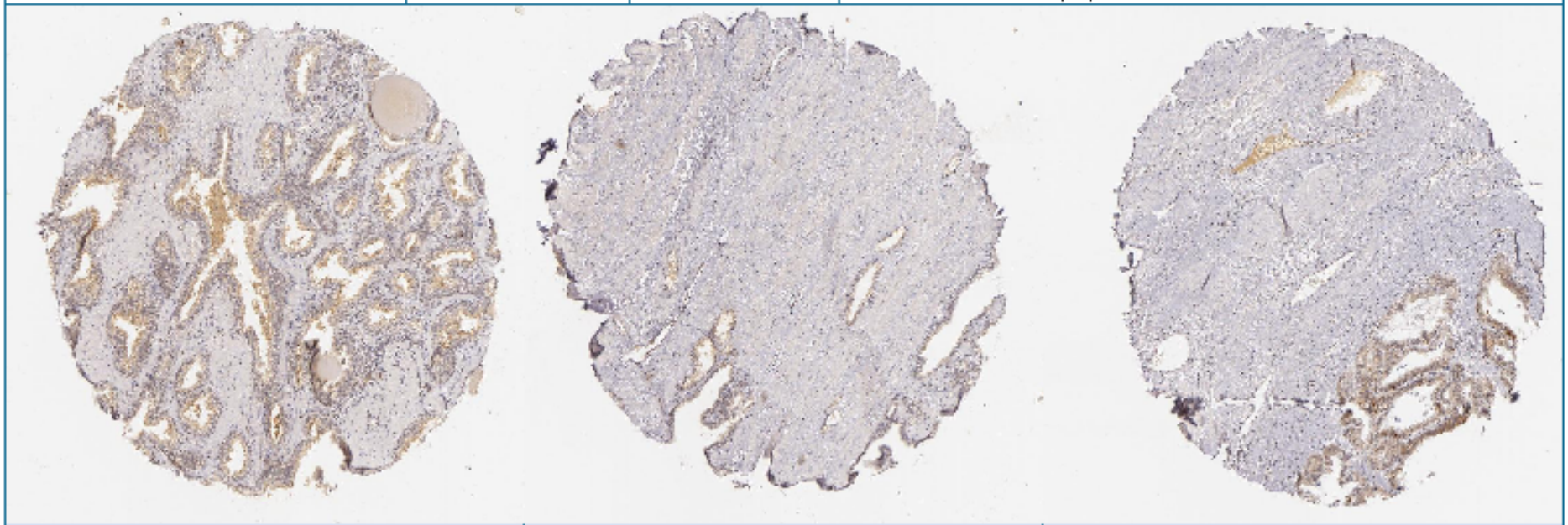
Courtesy www.proteinatlas.org (Uhlén, Pontén et al)

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

Courtesy www.proteinatlas.org (Uhlén, Pontén et al)

http://www.proteinatlas.org/cell_if_unit.php?antibody_id=1141&mainannotation_id=200000025

human protein atlas

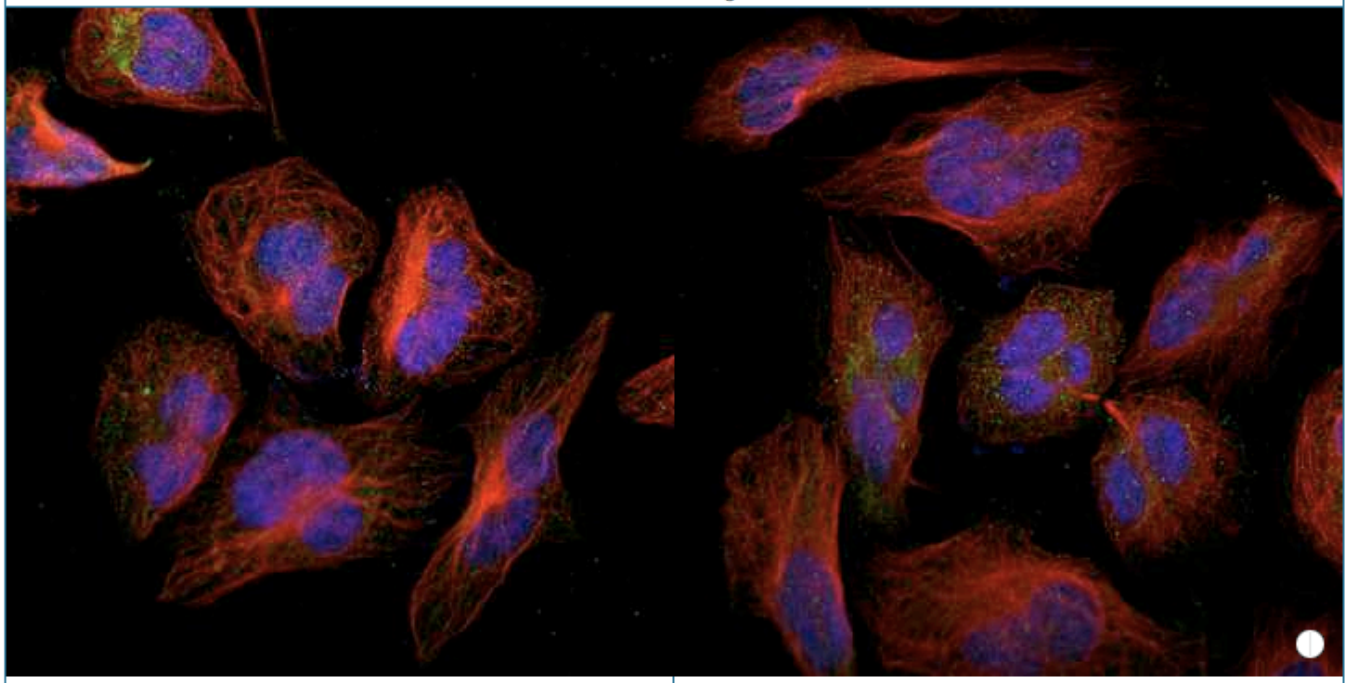
hpr | the project | protein atlas | dictionary | disclaimer | submission of antibodies | help

U-2 OS - [A4GALT]

Annotation		
Intensity	**Localization**	**Characteristics**
negative	undetermined	-

Validation

7 - No staining.



Protein targeted by HPA antibody is shown in green, nucleus in blue, micro-tubules in red and endoplasmatic reticulum (ER) in yellow.
Image Usage Policy

Navigation

Home
Search result
HPA001141
- Expression profiles - IHC
  - Normal tissues
  - Cancer tissues
  - Cells
- Expression profiles - IF
  - Cells
  - **U-2 OS**
- Gene/Protein info
- Antigen/Antibody info

Validation Summary

IH | IF | PA | WB

Search

search

help for this page

Carnegie Mellon

BMC Cell Biology

BioMed Central

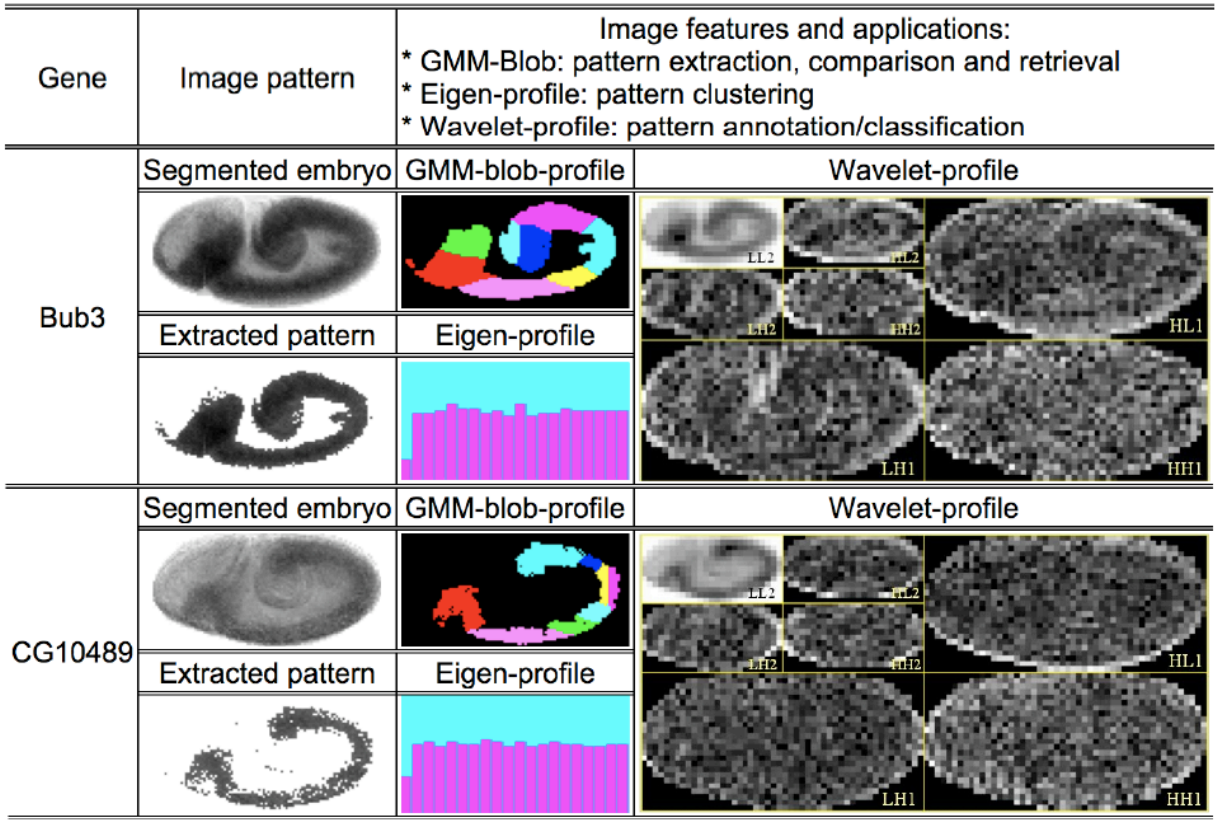
Research

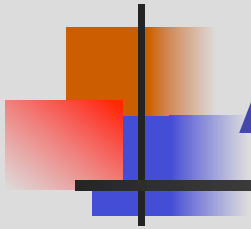
Open Access

Automatic image analysis for gene expression patterns of fly embryos

Hanchuan Peng*[1], Fuhui Long[1], Jie Zhou[2], Garmay Leung[3], Michael B Eisen[3,4] and Eugene W Myers[1]

Address: [1]Janelia Farm Research Campus, Howard Hughes Medical Institute, Ashburn, VA 20147, USA, [2]Department of Computer Science, Northern Illinois University, DeKalb, IL 60115, USA, [3]Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA and [4]Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

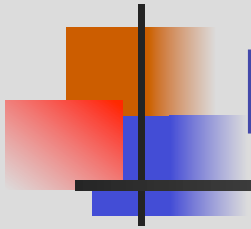




Automated Interpretation

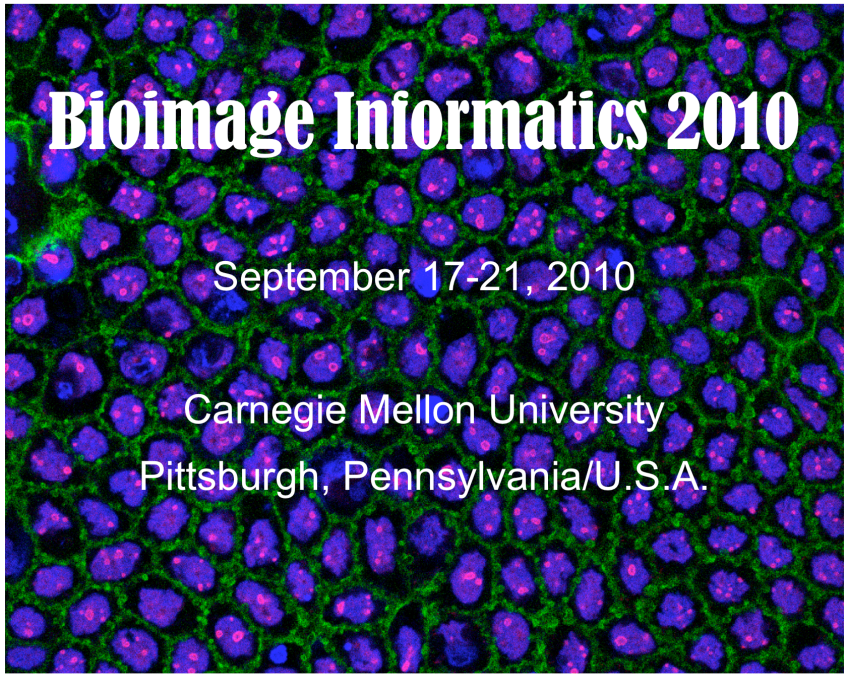
- Traditional analysis of microscope images has occurred by visual inspection
- Our group's goal over the past fourteen years has to automating interpretation with the ultimate goal of fully automated learning of patterns from images
- Much recent work from other groups

Major approaches within Bioimage Informatics



- Quantitation approaches
  - Expression and colocalization
  - Tracking of cells and structures
- Atlas approaches
  - Registration and construction
  - Change over time
  - Changes in mutants
- Pattern approaches
  - Focus of this introduction

Hold the date!



Organizers:

Gaudenz Danuser, Harvard Medical School

Michael Hawrylycz, Allen Institute for Brain Science

Robert F. Murphy, Carnegie Mellon University

Hosted by

Center for Bioimage Informatics
from image to knowledge

RAY AND STEPHANIE LANE
Center for Computational Biology
Carnegie Mellon

Local Organizers:

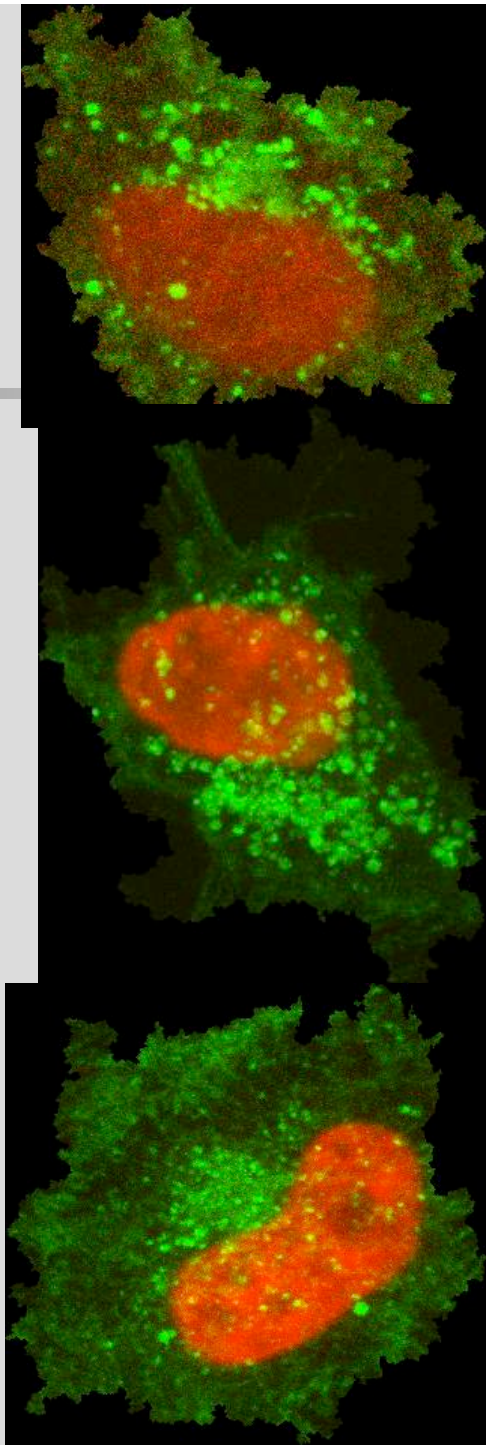
Jelena Kovačević, Gustavo Rohde, Ge Yang

http://www.cbi.cmu.edu

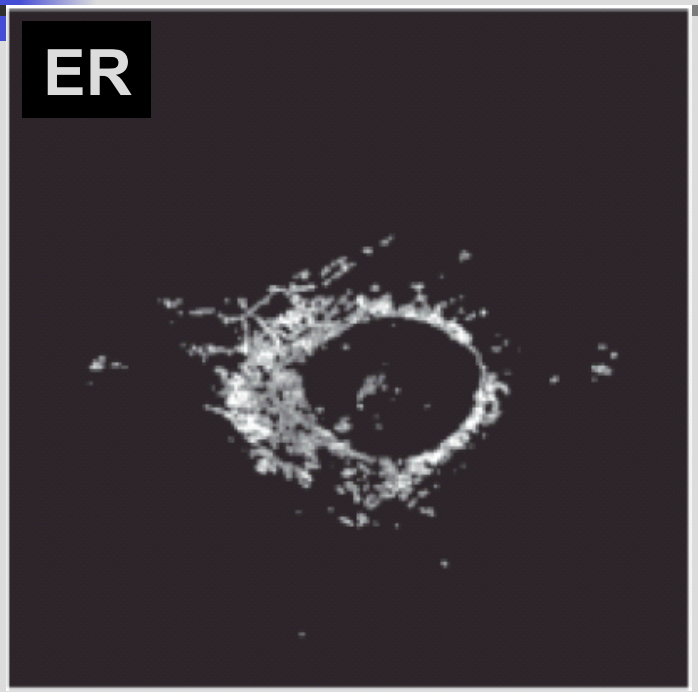
Carnegie Mellon

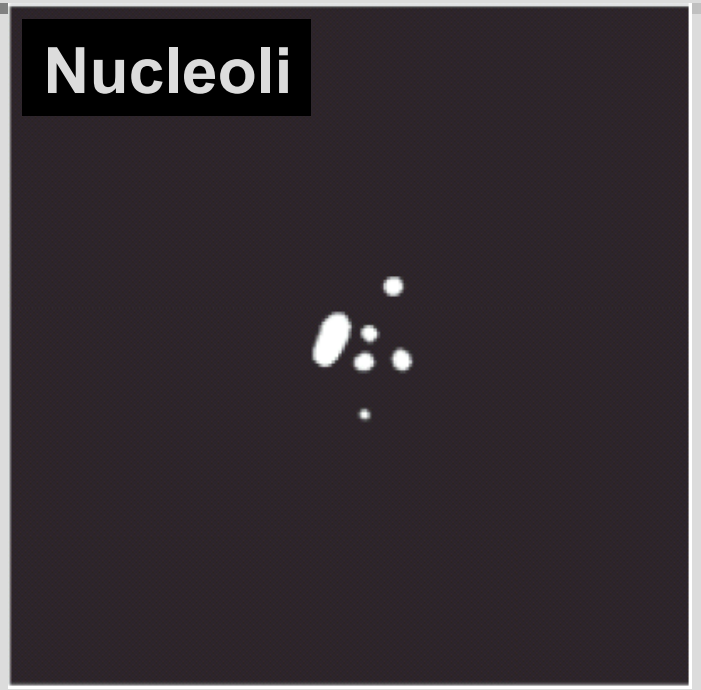
The Challenge

- Comparison of cell images pixel-by-pixel or region-by-region matching does not work for cell patterns because different cells have different **shapes, sizes, orientations**
- Organelles/structures within cells are **not found in fixed locations**
- ***Instead, describe each image numerically and operate on the descriptors ("SLF" - Subcellular Location Features)***



Example of classification using Morphological Features

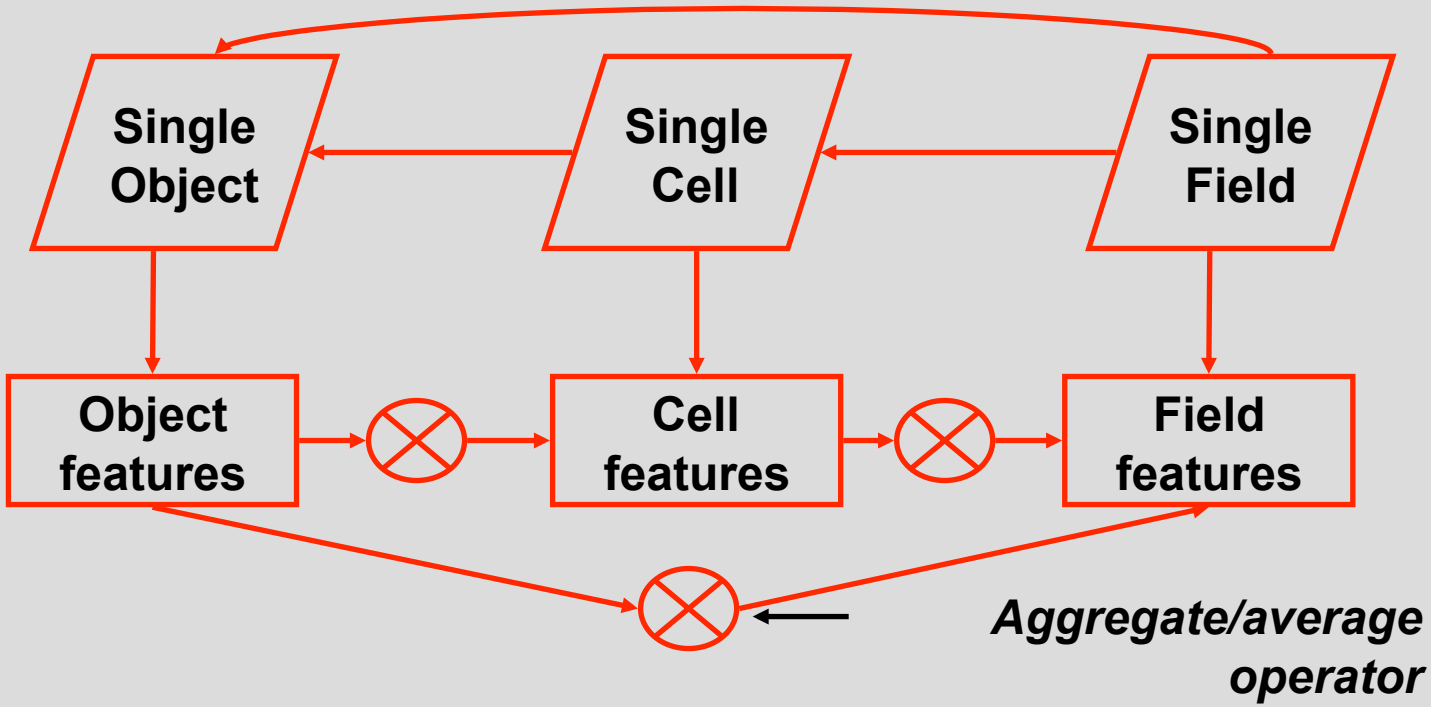




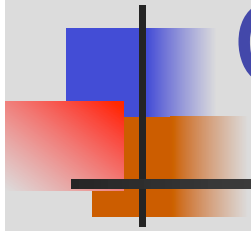
108	# of objects	6
83	Average size of objects	232
31	Average distance to COF	4

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

Feature levels and granularity

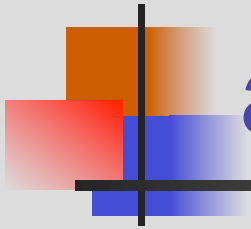


Granularity: 2D, 3D, 2Dt, 3Dt



Cell Segmentation

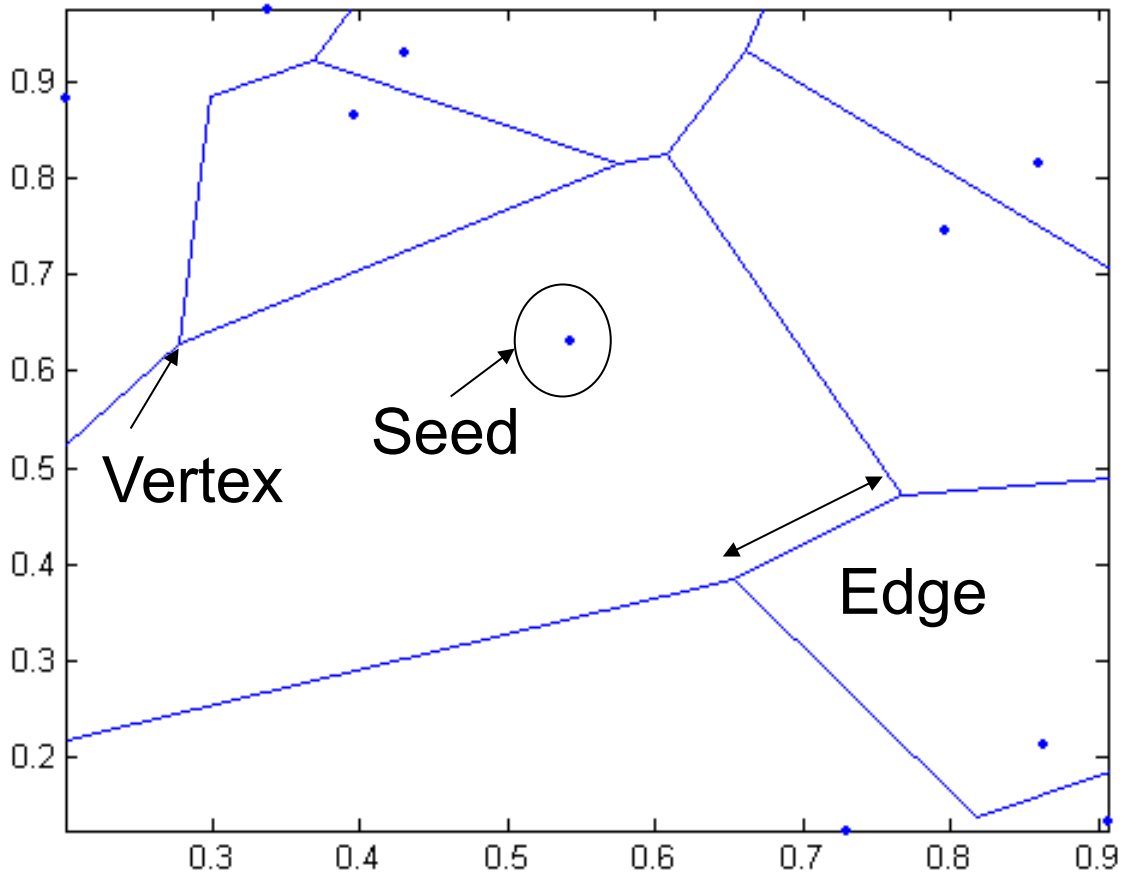
Single cell segmentation approaches



- Voronoi
- Watershed
- Seeded Watershed
- Active Contours and Level Set Methods
- Graphical Models
- Active Masks

Voronoi diagram

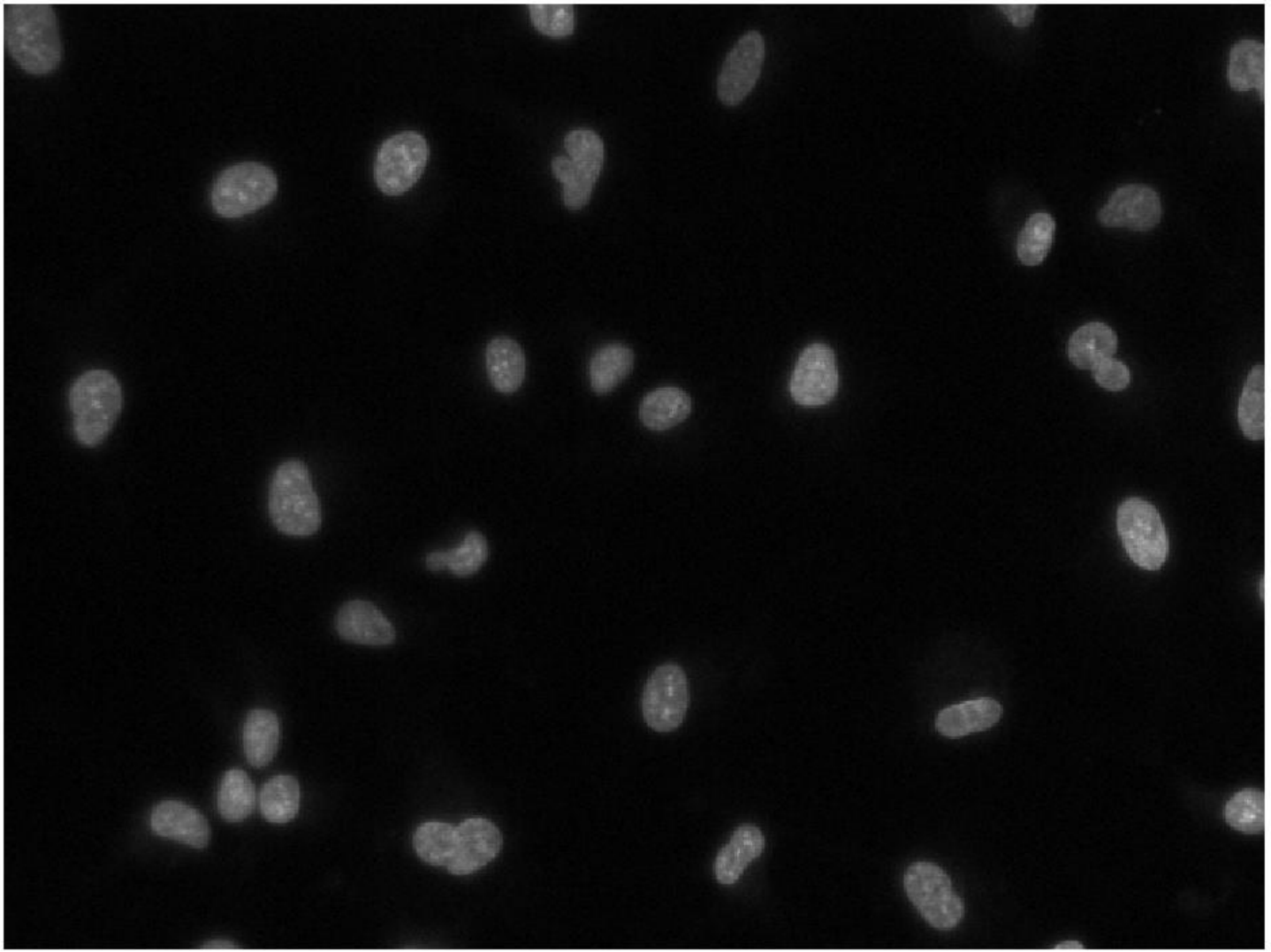
Given a set of seeds, draw vertices and edges such that each seed is enclosed in a single polygon where each edge is equidistant from the seeds on either side.



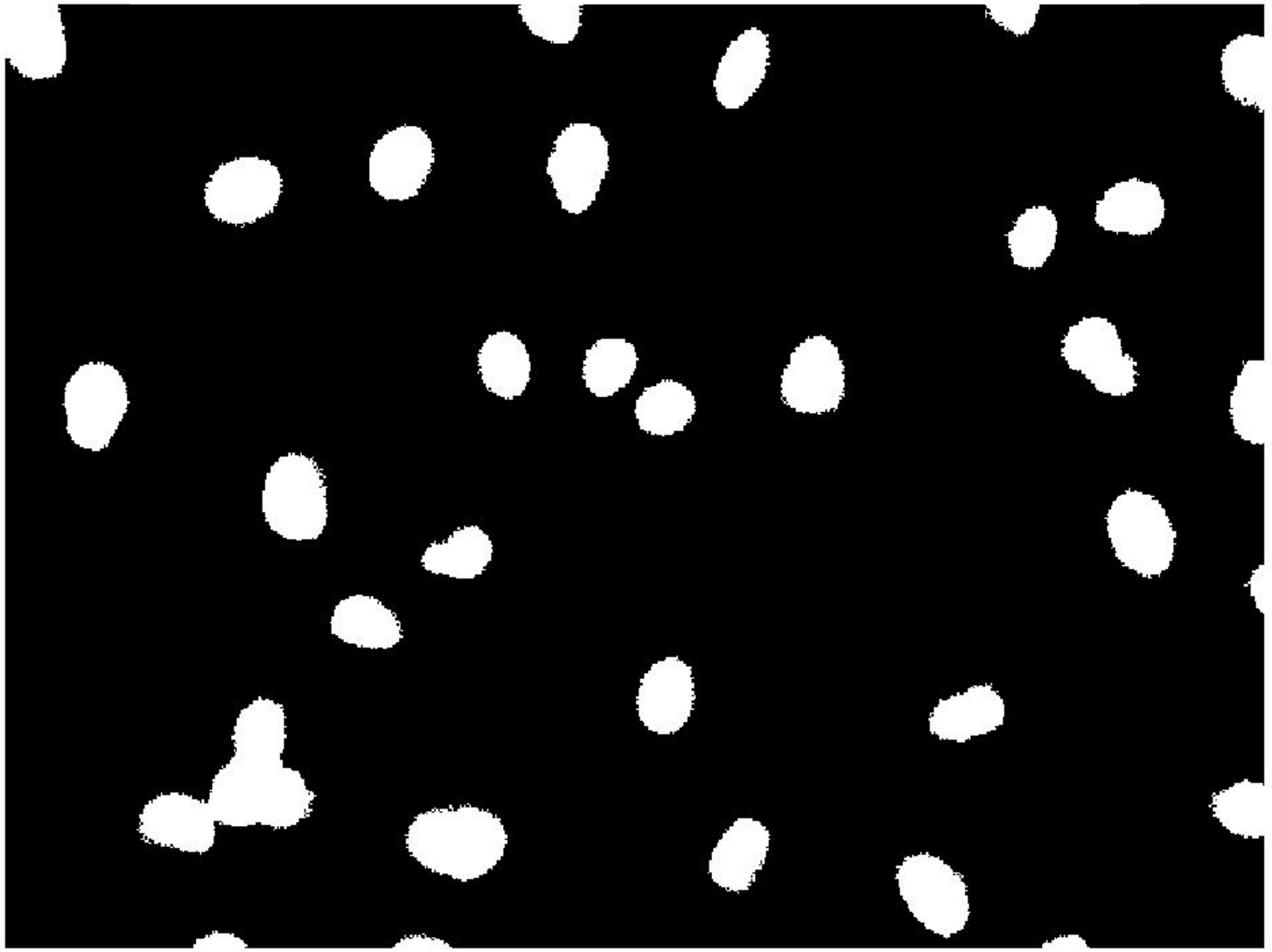
Voronoi Segmentation Process

- Threshold DNA image (downsample?)
- Find the objects in the image
- Find the centers of the objects
- Use as seeds to generate Voronoi diagram
- Create a mask for each region in the Voronoi diagram
- Remove regions whose object that does not have intensity/size/shape of nucleus

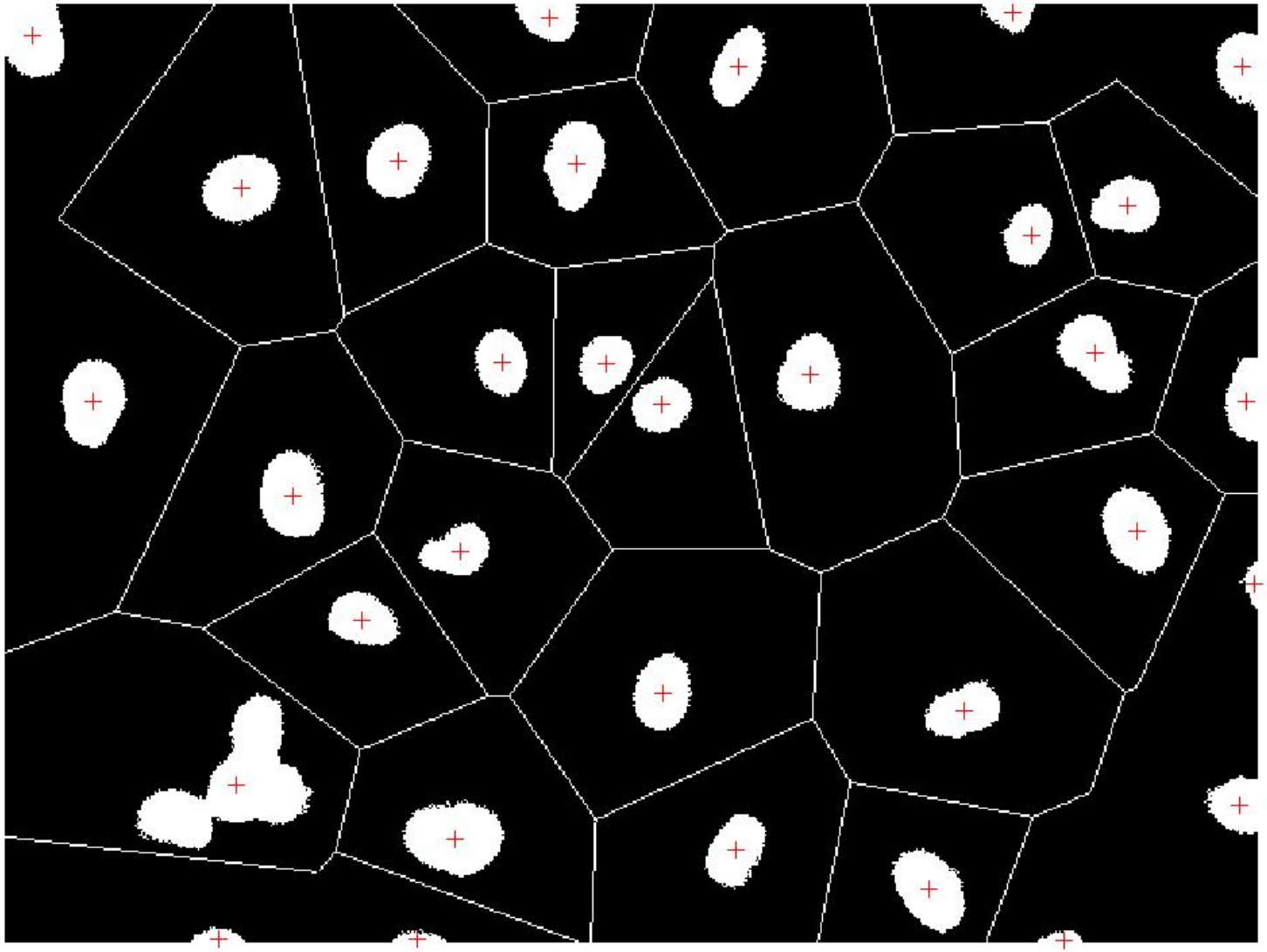
Original DNA image



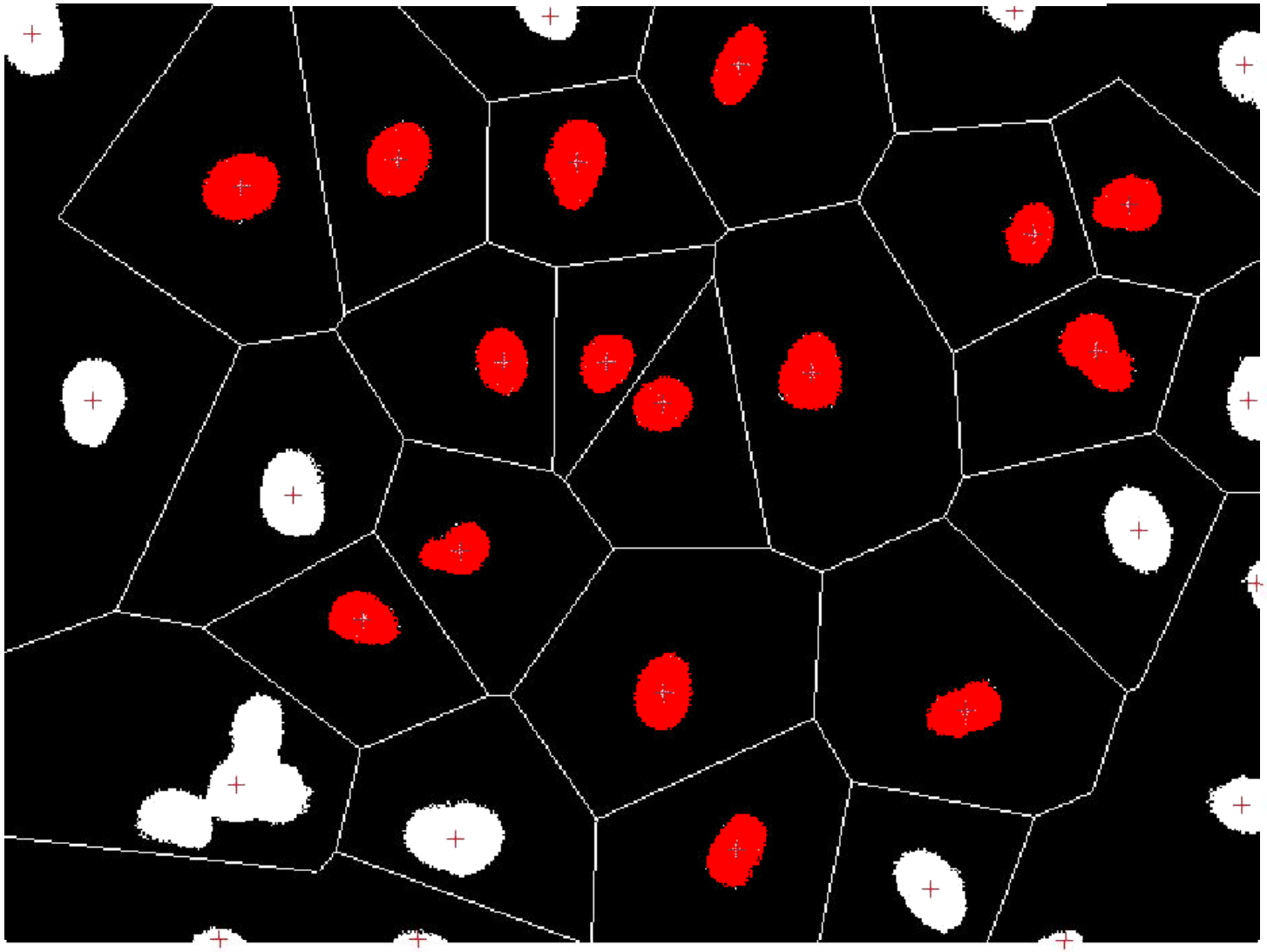
After thresholding and removing small objects



After triangulation

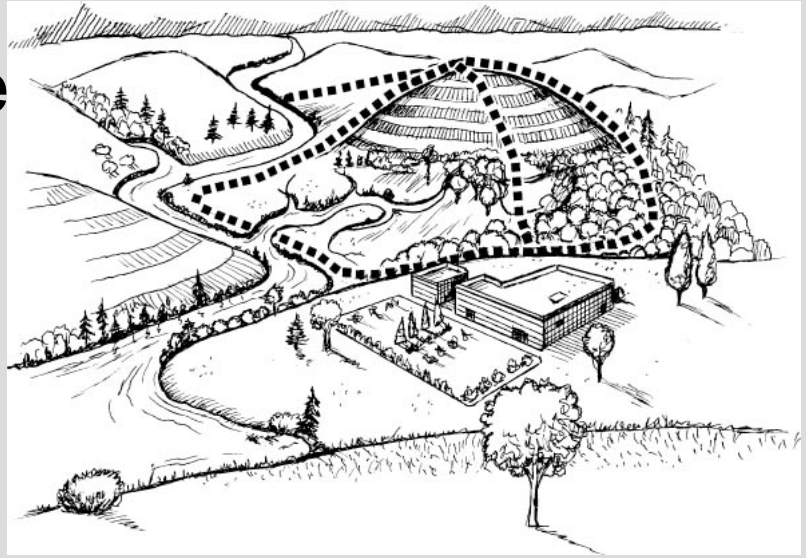


After removing edge cells and filtering



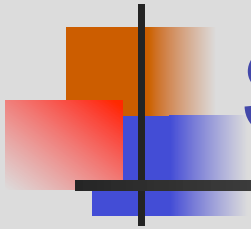
Watershed Segmentation

- Intensity of an image ~ elevation in a landscape
  - Flood from minima
  - Prevent merging of “catchment basins”
  - Watershed borders built at contacts between basins



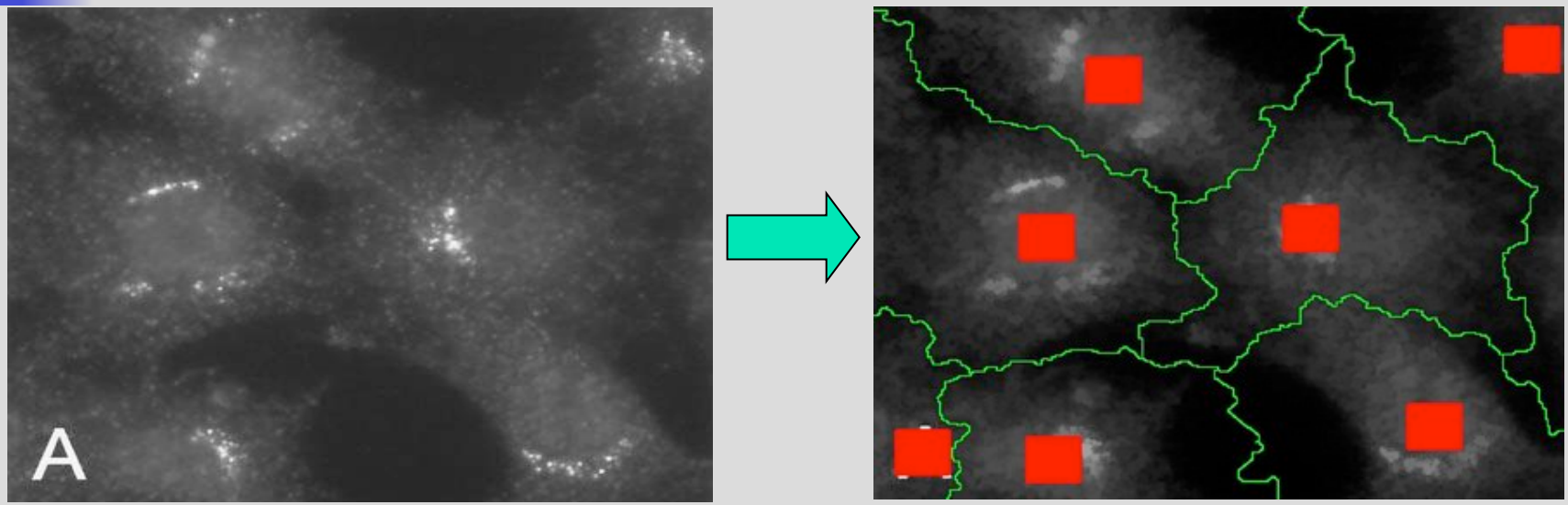
http://www.ctic.purdue.edu/KYW/glossary/whatisaws.html

Seeded Watershed Segmentation



- Drawback is that the number of regions may not correspond to the number of cells
- *Seeded* watershed allows water to rise only from predefined sources (seeds)
- If parallel DNA image available, can use same approach to generate these seeds as for Voronoi segmentation
- Can use seeds from DNA image but use total protein or membrane protein image for watershed segmentation

Seeded Watershed Segmentation

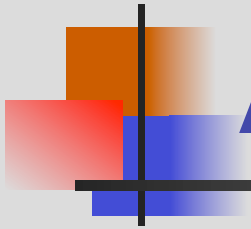


Original image

Seeds and boundary

Applied directly to protein image (no DNA image)

Note non-linear boundaries



Active masks

- Srinivasa, Fickus, Guo, Linstedt, Kovacevic, IEEE Trans. Image Proc. 2009
- Marriage of multiple methods
  - Active contours
  - Multiresolution
  - Multiscale
  - Region growing

AM segmentation: MS loop

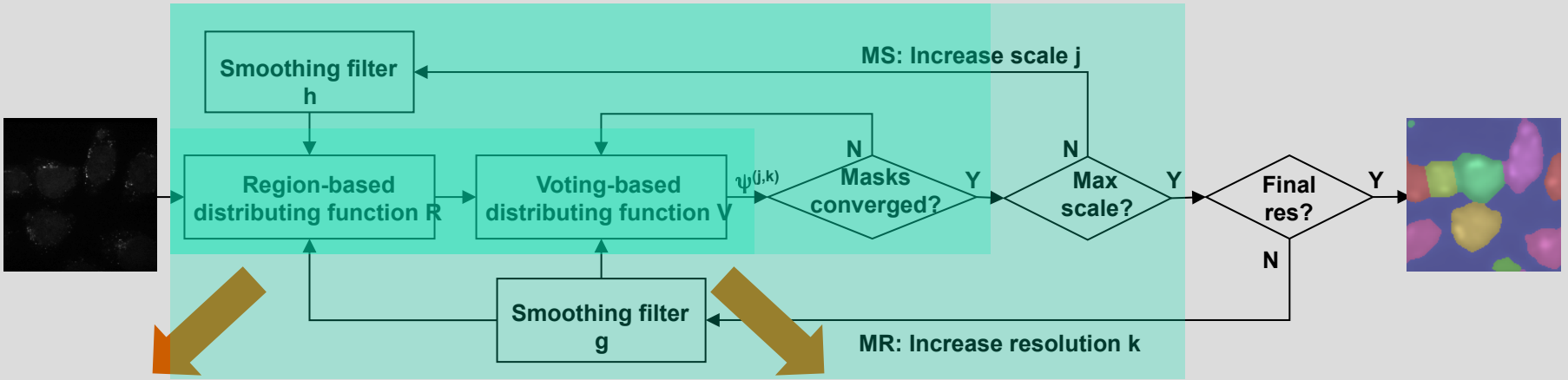
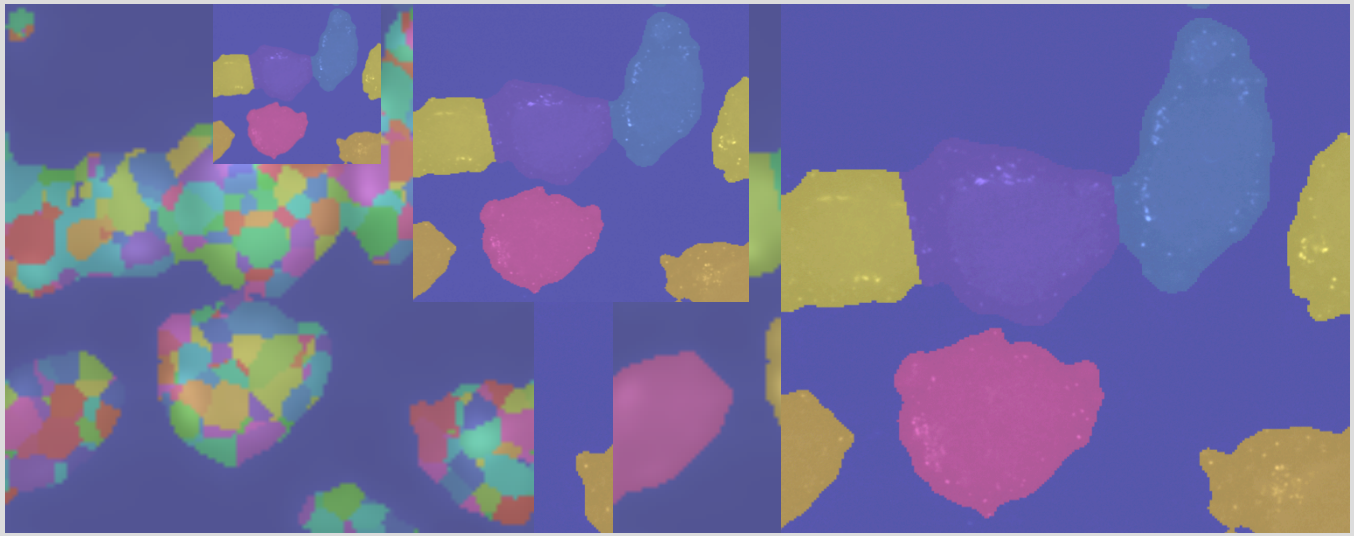
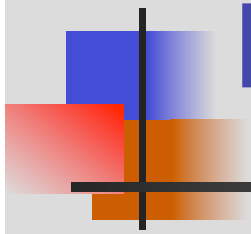


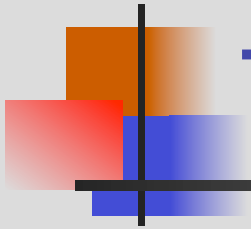
Image based **Geometry based**





Feature Extraction

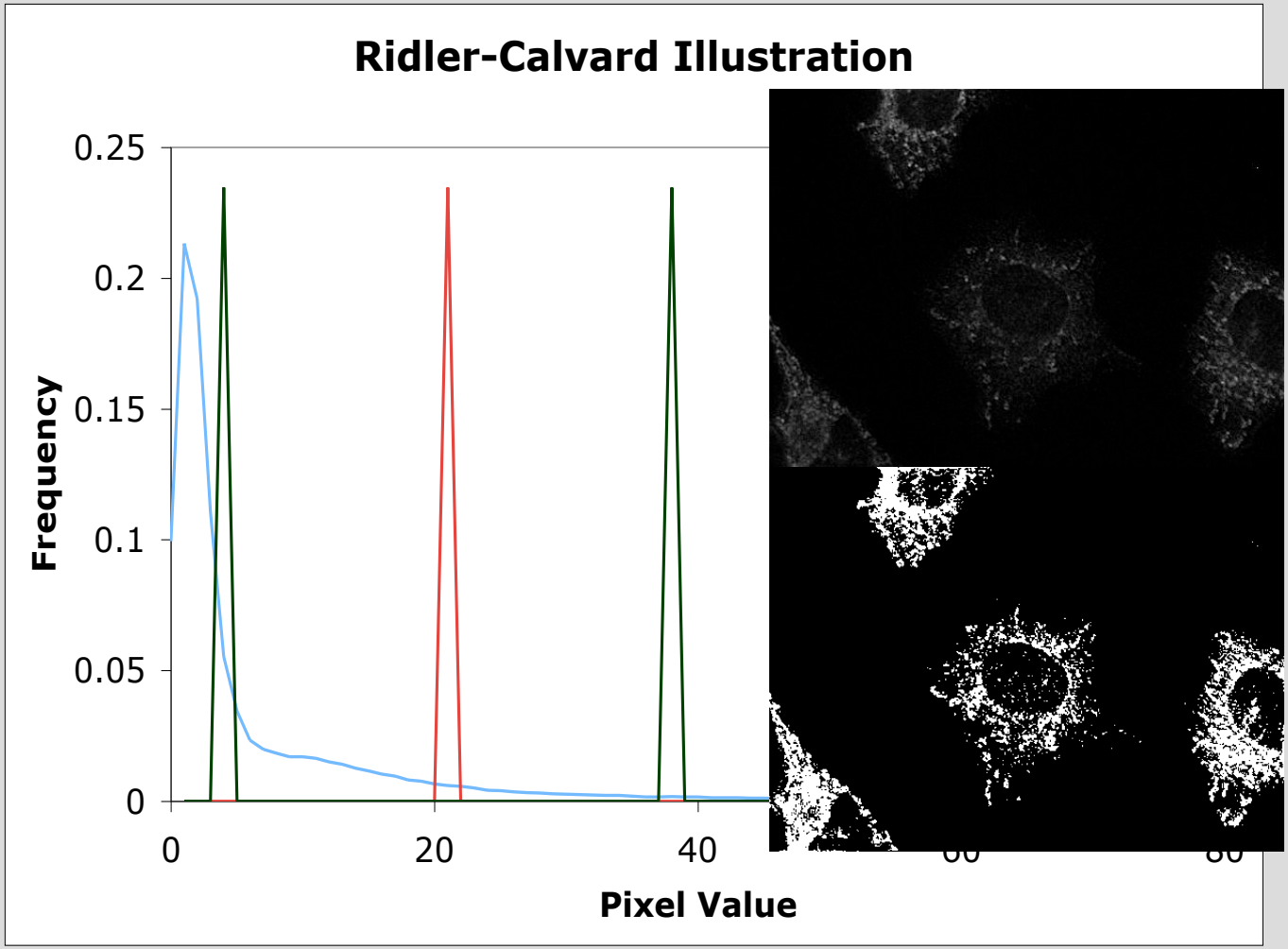
Morphological Features - Thresholding

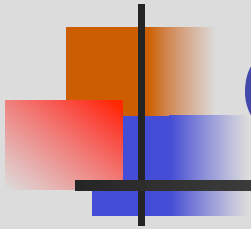


- Morphological features require some method for defining objects
- Most common approach is global thresholding
- Methods exist for automatically choosing a global threshold (e.g., Riddler-Calvard method)
  - Find threshold that is equidistant from the average intensity of pixels below and above it

Ridler-Calvard Method

Blue line shows histogram of intensities, green lines show average to left and right of red line, red line shows midpoint between them or the RC threshold



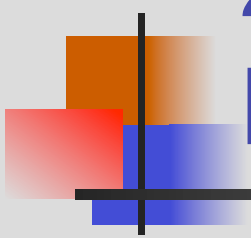


Object finding

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

2D Features
Morphological Features

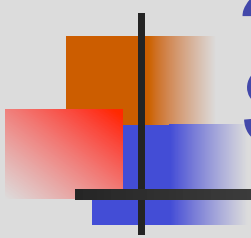
Description
The number of fluorescent objects in the image
The Euler number of the image
The average number of above-threshold pixels per object
The variance of the number of above-threshold pixels per object
The ratio of the size of the largest object to the smallest
The average object distance to the cellular center of fluorescence(COF)
The variance of object distances from the COF
The ratio of the largest to the smallest object to COF distance



2D Features
DNA Features

DNA features (objects relative to DNA reference)

Description
The average object distance from the COF of the DNA image
The variance of object distances from the DNA COF
The ratio of the largest to the smallest object to DNA COF distance
The distance between the protein COF and the DNA COF
The ratio of the area occupied by protein to that occupied by DNA
The fraction of the protein fluorescence that co-localizes with DNA

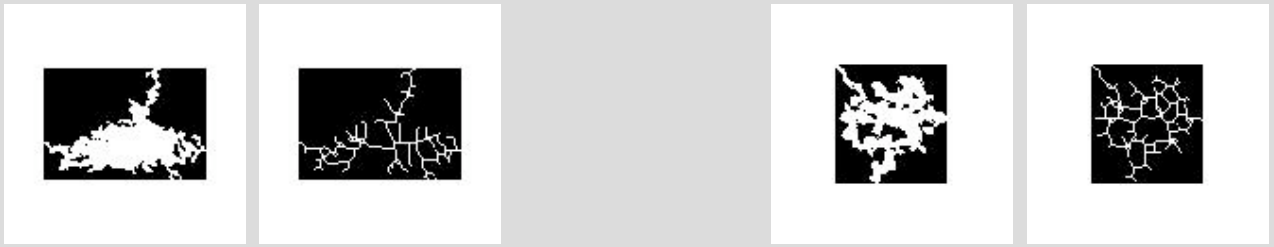


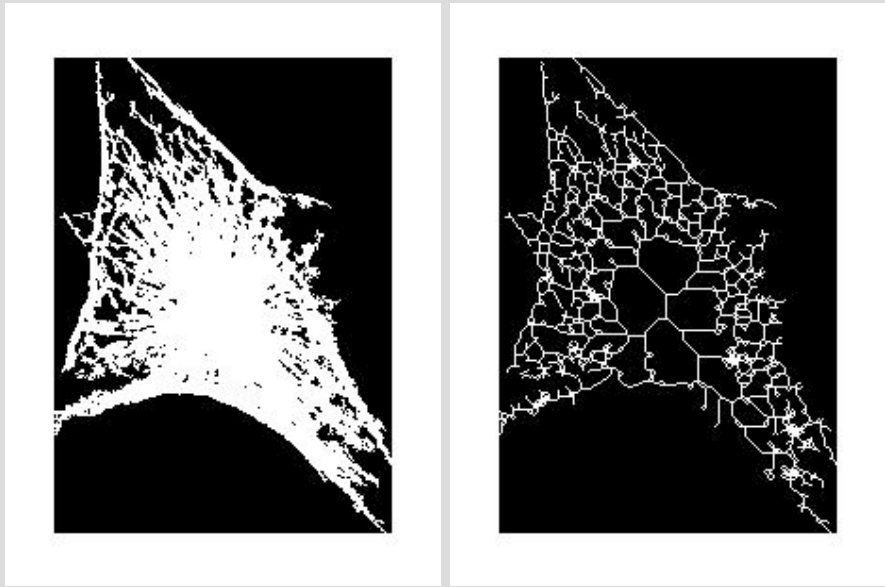
2D Features
Skeleton Features

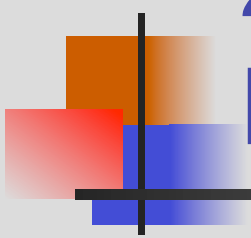
Skeleton features (object shape)

Description
The average length of the morphological skeleton of objects
The ratio of object skeleton length to the area of the convex hull of the skeleton, averaged over all objects
The fraction of object pixels contained within the skeleton
The fraction of object fluorescence contained within the skeleton
The ratio of the number of branch points in the skeleton to the length of skeleton

Illustration – Skeleton







2D Features
Edge Features

Edge features

Description
The fraction of the non-zero pixels that are along an edge
Measure of edge gradient intensity homogeneity
Measure of edge direction homogeneity 1
Measure of edge direction homogeneity 2
Measure of edge direction difference

2D Features
Haralick Texture Features

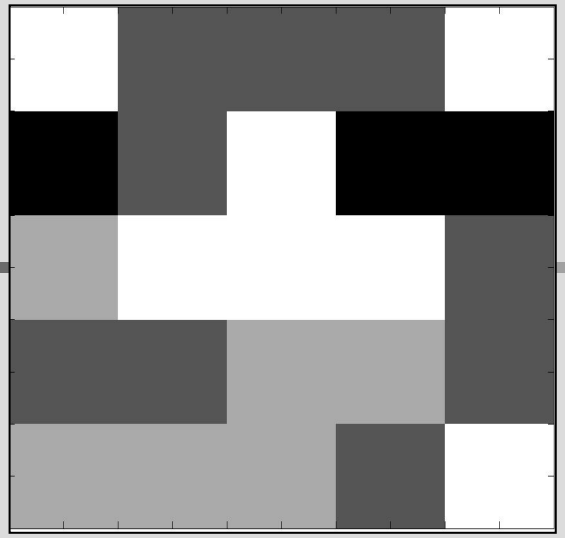
- Correlations of adjacent pixels in gray level images
- Start by calculating gray level co-occurrence matrix P:

  N by N matrix, N=number of gray level.

  Element P(i,j) is the probability of a pixel with value i being adjacent to a pixel with value j
- Four directions in which a pixel can be adjacent

Co-occurrence Matrices

Example image with 4 gray levels



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

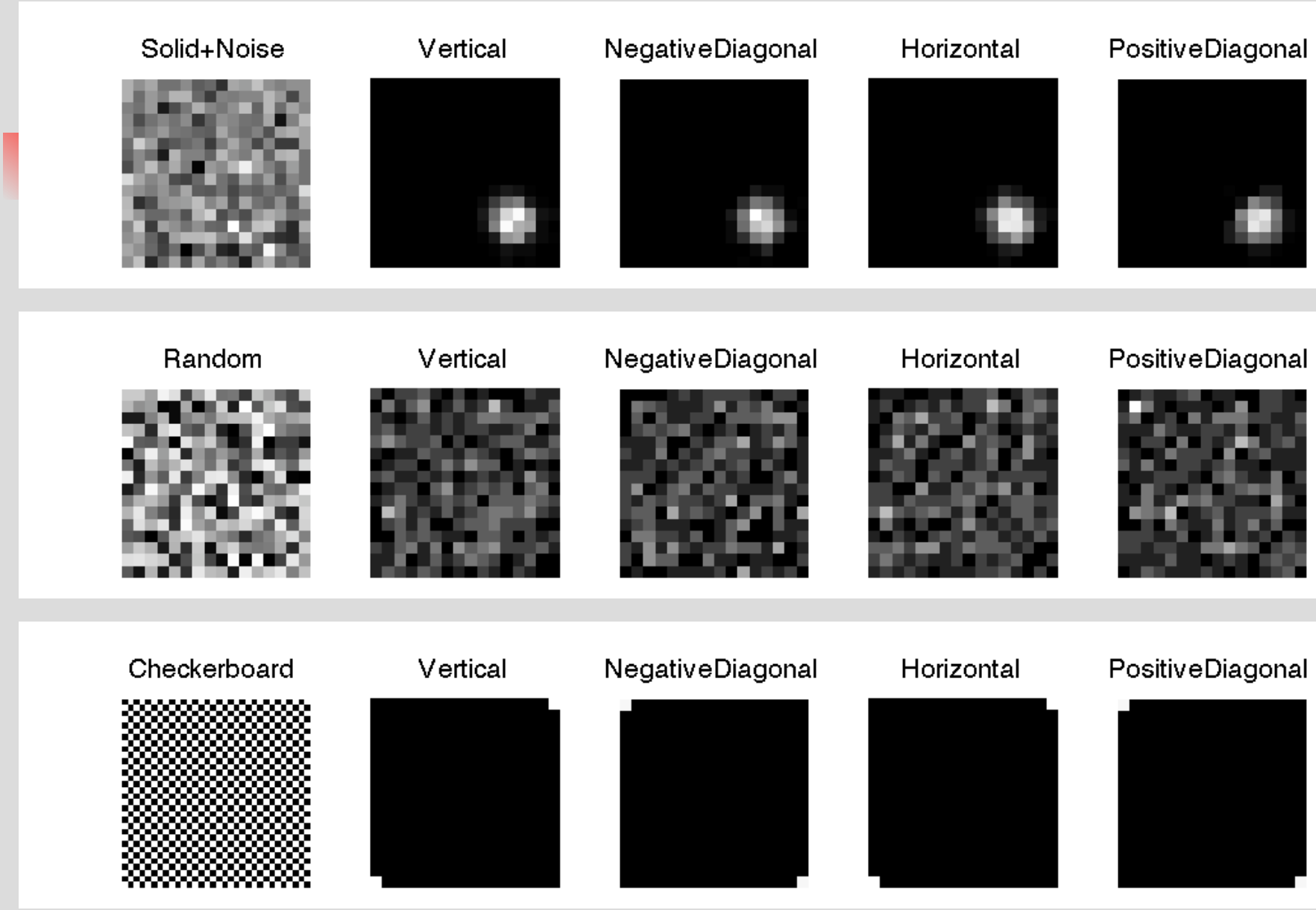
	0	1	2	3
0	0	2	1	2
1	2	4	4	4
2	1	4	2	2
3	2	4	2	2

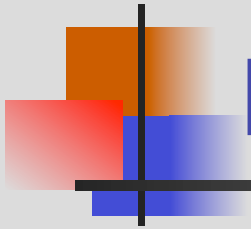
	0	1	2	3
0	2	1	0	1
1	1	6	3	4
2	0	3	6	2
3	1	4	2	4

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

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

Example Gray Level Co-occurrence Matrices for Various Textures



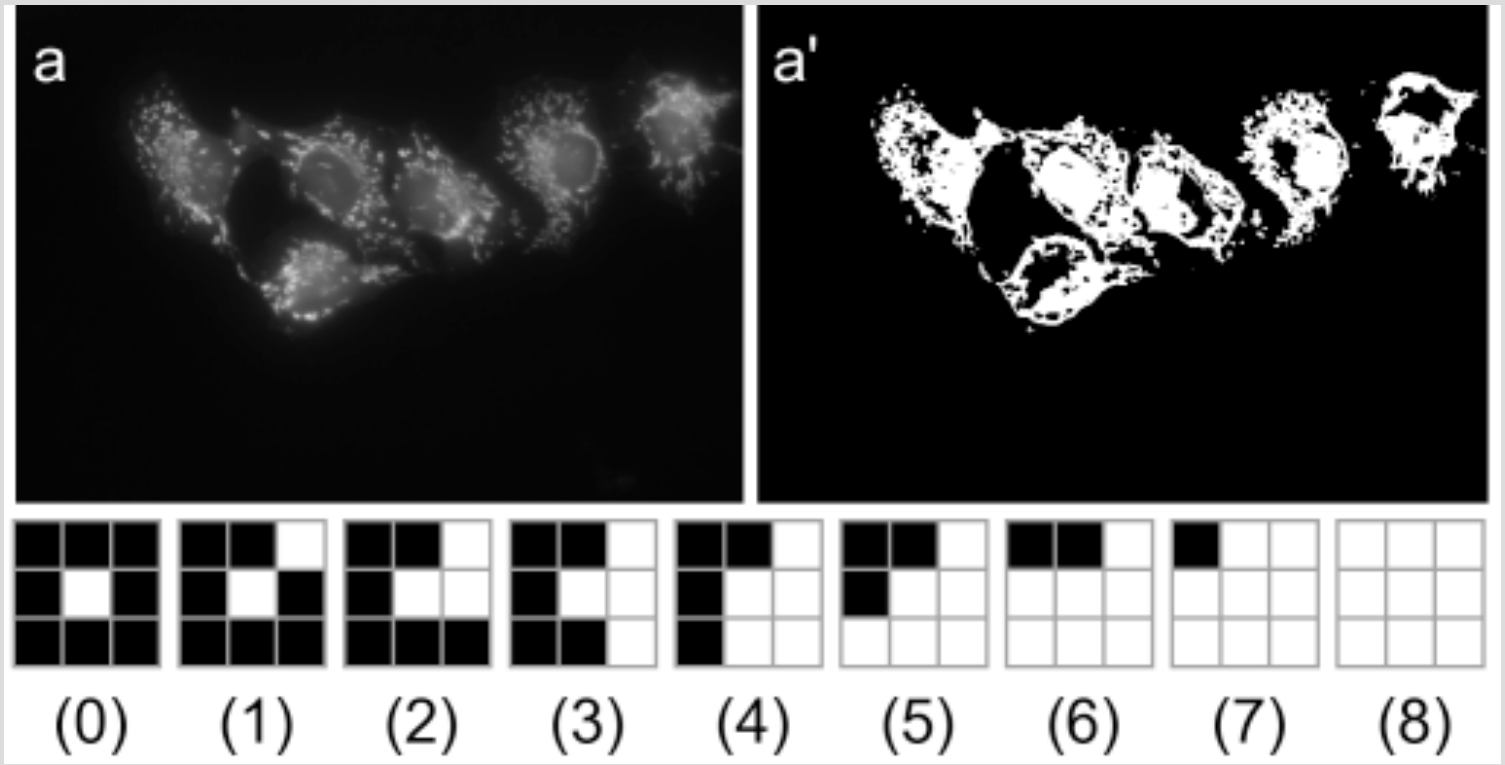


Pixel Resolution and Gray Levels

- Texture features are influenced by the number of gray levels and pixel resolution of the image
- Optimization for each image dataset required
- Alternatively, features can be calculated for many resolutions

Threshold Adjacency Statistics

For a specified range of gray level values, count how many neighbors each above threshold pixel has



N A Hamilton,[3] R S Pantelic, K Hanson, and R D Teasdale
BMC Bioinformatics. 2007; 8: 110

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**

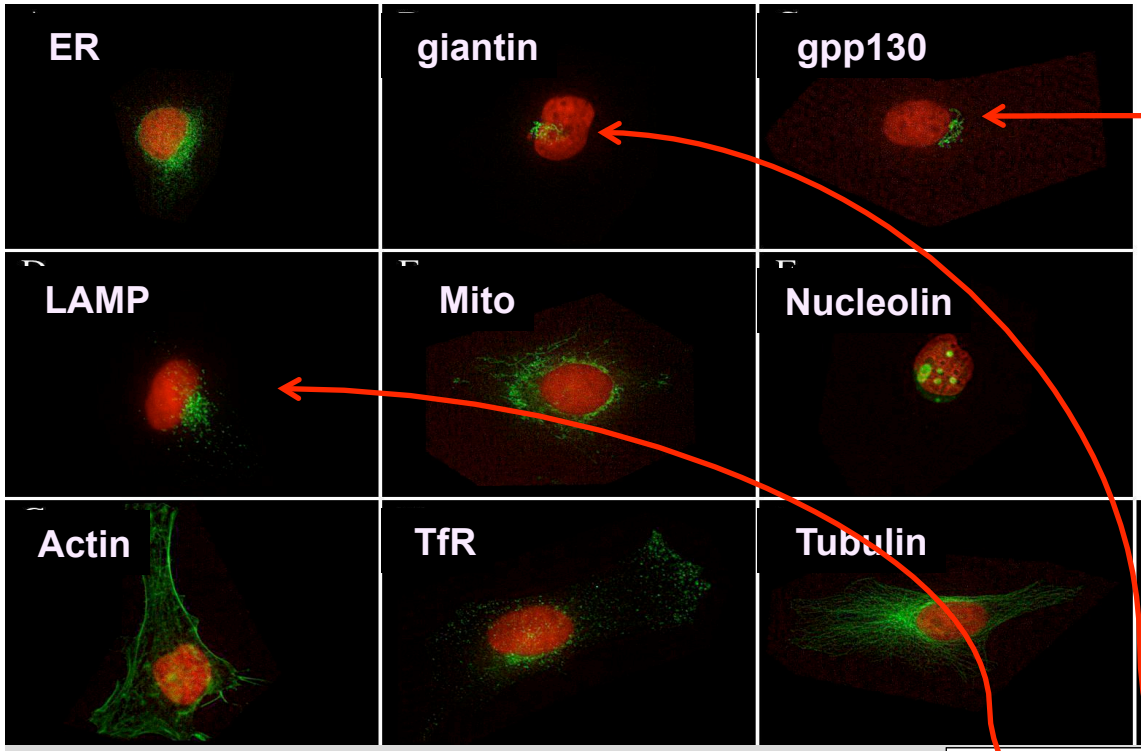
Overall accuracy = 92%

Murphy et al 2003

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**

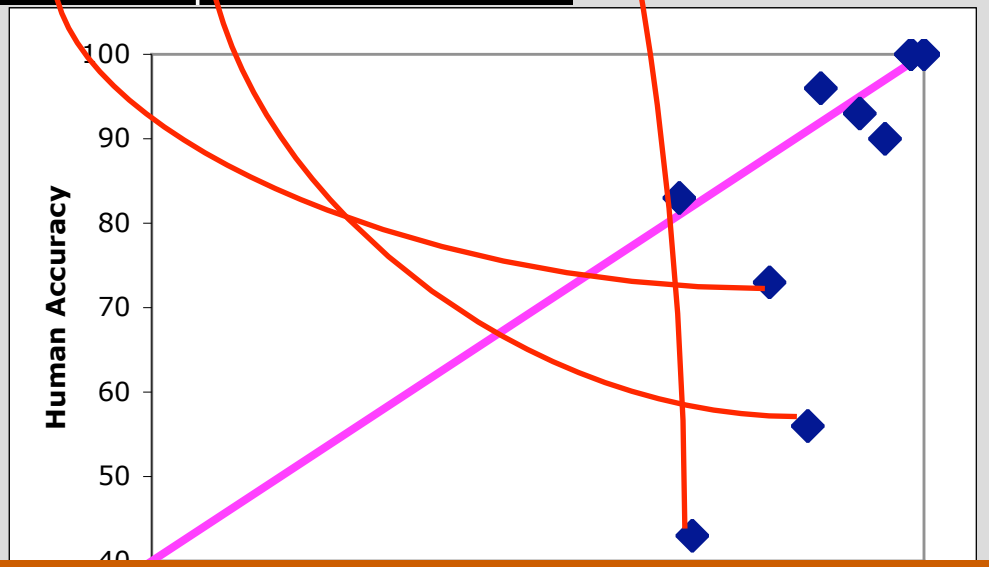
Overall accuracy = 83%



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



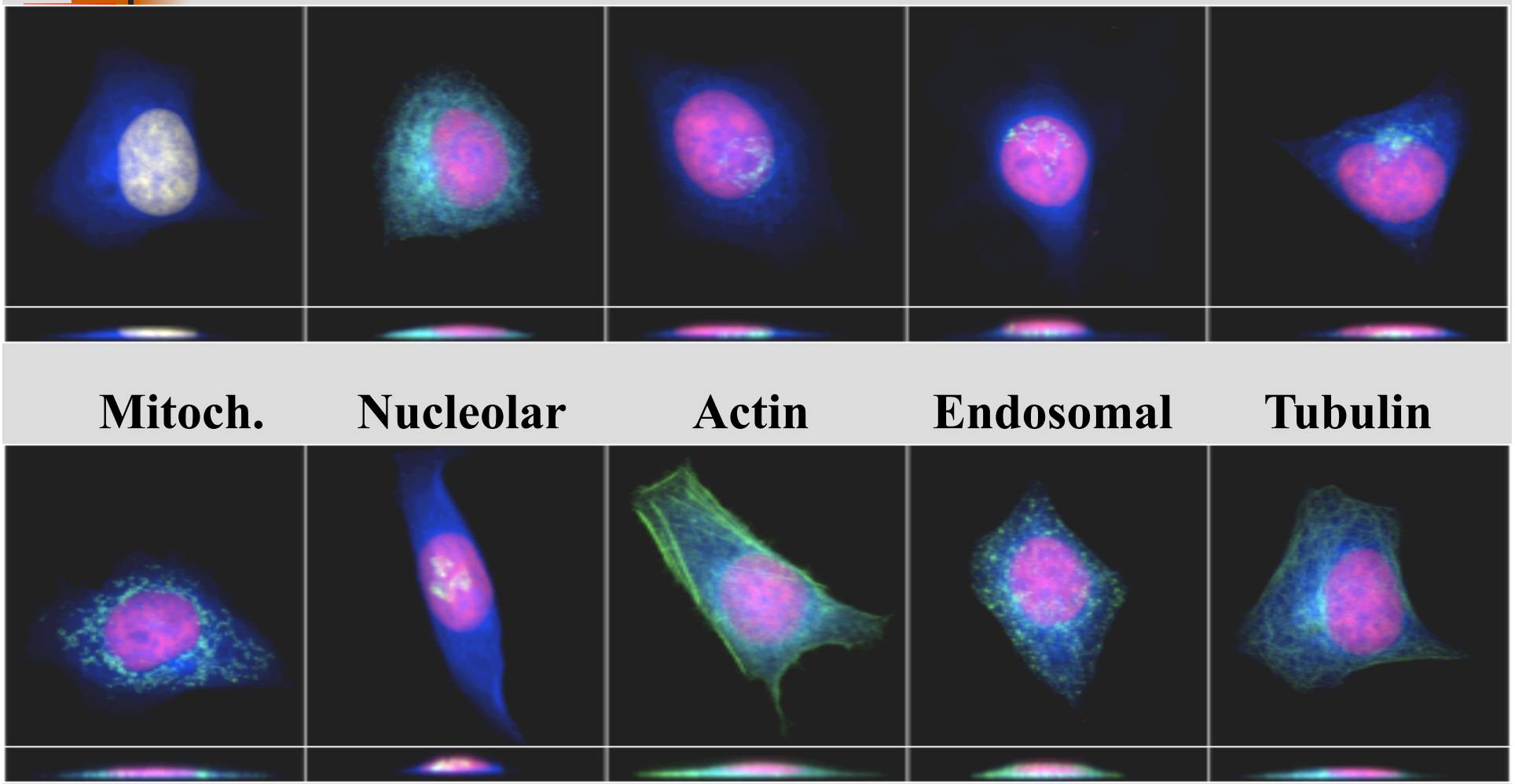
Even better results using multiresolution methods by Kovacevic group

Even better results for 3D images

Carnegie Mellon

Velliste & Murphy 2002

3D HeLa cell images



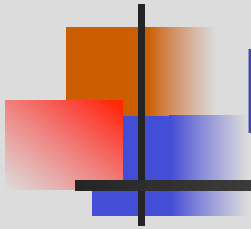
Images collected using facilities at the Center for Biologic Imaging courtesy of Simon Watkins

Velliste & Murphy 2002; Chen & Murphy 2004

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**

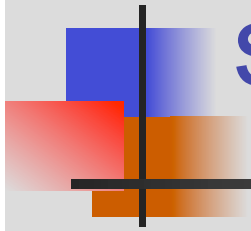
Overall accuracy = 98%



Results for large collections

Collection	Number of classes	Overall accuracy	Recall for high precision	Precision for that recall
Yeast GFP	20	81%	80%	95%
HPA tissue	8	81%	50%	97%

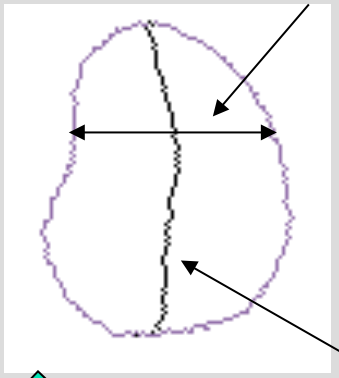
Cell, nuclear and organelle shape modeling



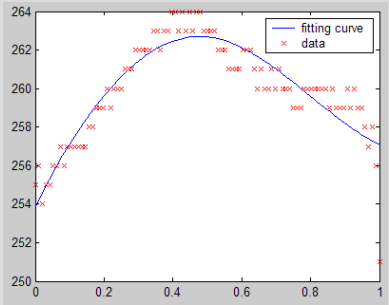
Carnegie Mellon

Nuclear Shape - Medial Axis Model

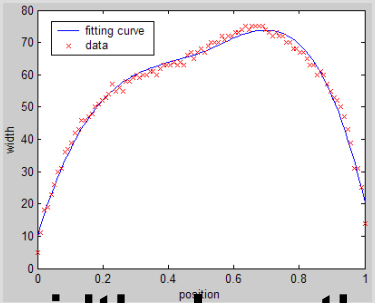
Rotate



Represented by two curves



the medial axis



width along the medial axis

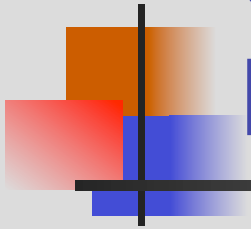
Carnegie Mellon

Synthetic Nuclear Shapes

Diffeomorphic analysis of nuclear shape

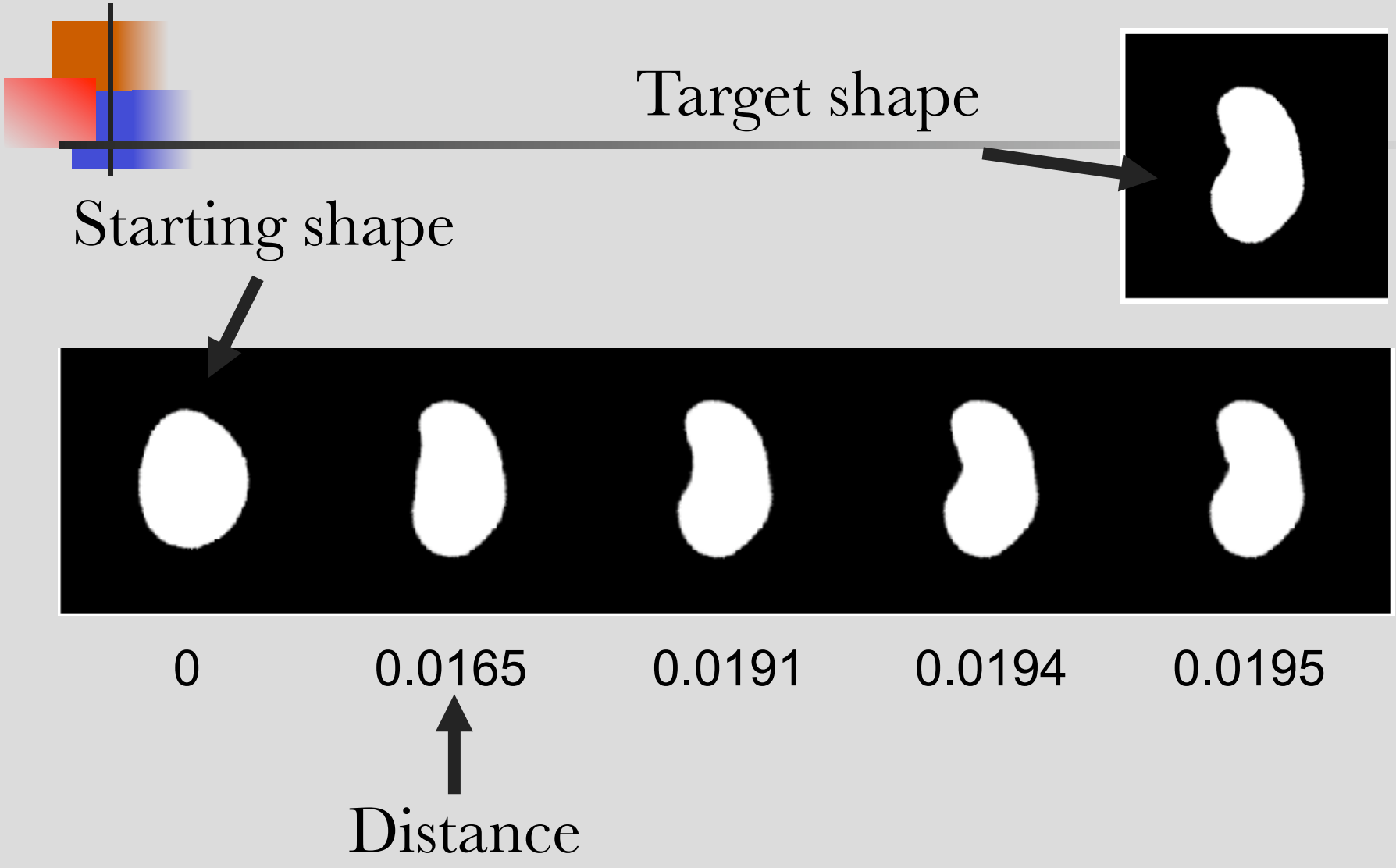
- Can use distance between shapes to characterize shape space instead of parameters of model
- First application of morphing methods to nuclei: S. Yang, D. Köhler, K. Teller, T, Cremer, P. Le Baccon, E. Heard, R. Eils, and K. Rohr, LNCS 4190, pp. 907–914, 2006

LDDMM – Large deformation diffeomorphic metric mapping: Miller and colleagues

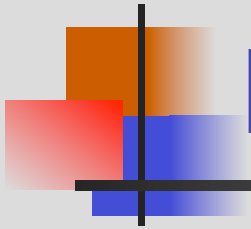


- Goal: Find a function g(x,t) which smoothly transforms an image I_n into an image I_m as t goes from 0 to T
- Choose g(x,t) to minimize sum of
  - Total deformation in g from 0 to T
  - Distance between I_m and $I_n(g(x,T))$

Mapping two shapes to each other


Target shape
Starting shape
0
0.0165
0.0191
0.0194
0.0195
Distance

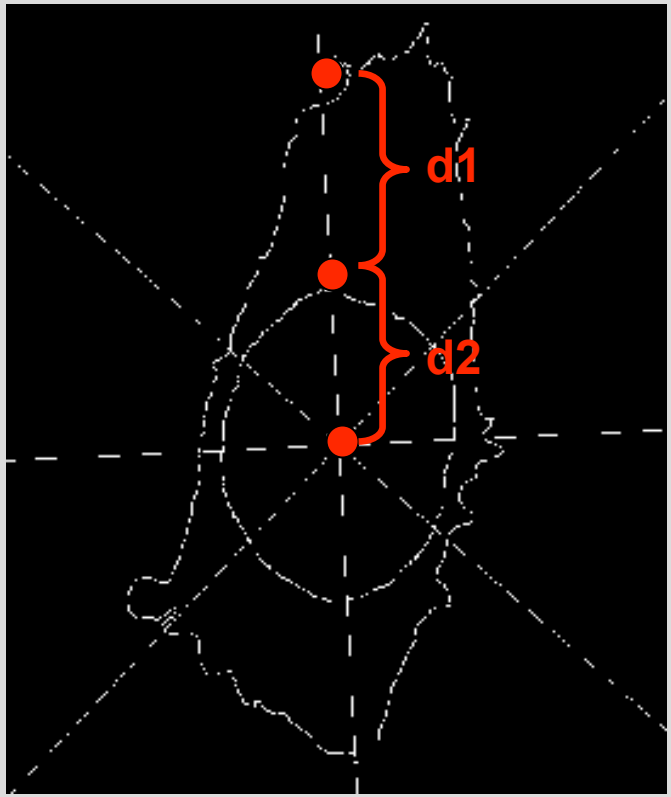
Diffeomorphic analysis of cell shape



Eigenshapes

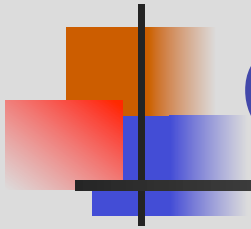
- Orient objects to align (e.g., major axis)
- Find fixed number of coordinates of sampled/interpolated points on outline of object (nucleus, cell)
- Represent variation in these coordinates using principal components

Conditional Cell Shape Description: Distance Ratio



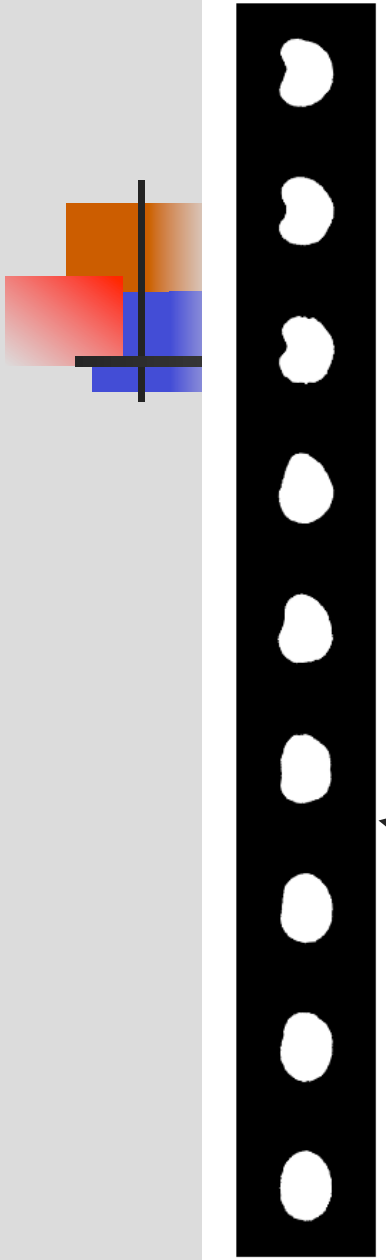
$$r = \frac{d_1 + d_2}{d_2}$$

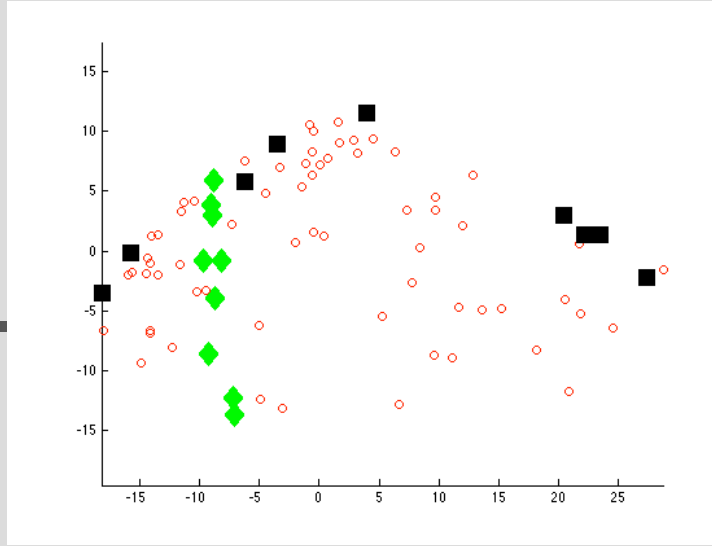
Represent single shape as vector of ratios for *n* angles and represent variation using PCA



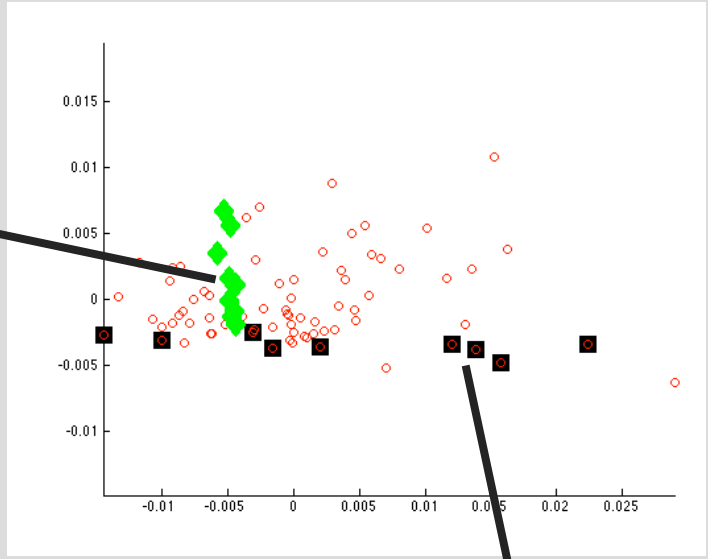
Characterizing shape space

- Find deformation fields from each image to every other image
- Calculate distance between each pair of images as total deformation required between them
- Use multidimensional scaling (MDS) to find variables (principal components) that compactly represent variation



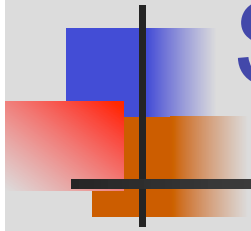


First 2 components from MDS directly on perimeter coordinates



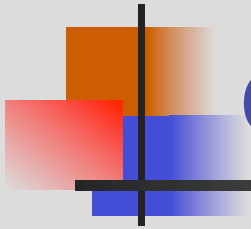
First 2 components from MDS on distance matrix from LDDMM

Generative Models for Subcellular Location Patterns



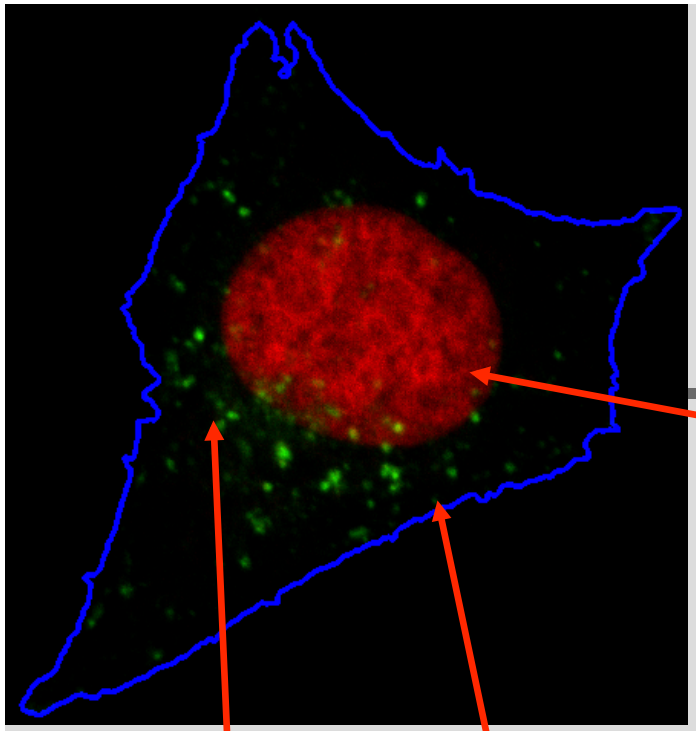
Carnegie Mellon

Generative models for communicating patterns



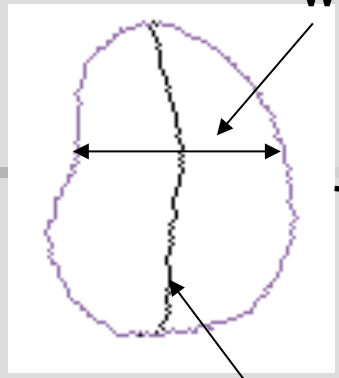
- How do we communicate results learned about subcellular patterns?
- Proposal: Use generative models learned from images to capture pattern and *variation* in pattern

Generative Model Components



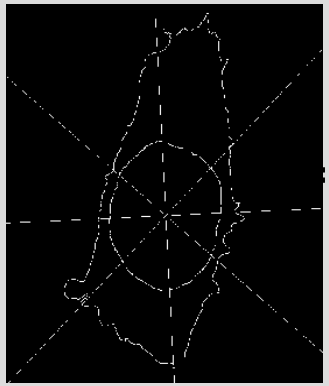
Nucleus

width



Medial axis

Cell membrane

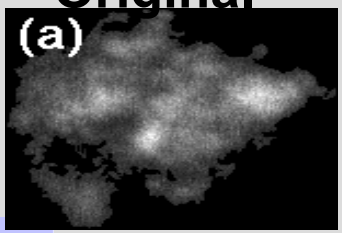


$$\frac{d_1 + d_2}{d_2}$$

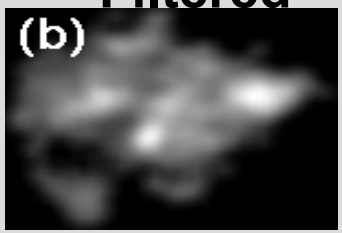
Model parameters

Protein objects

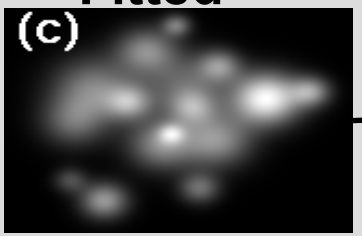
Original



Filtered

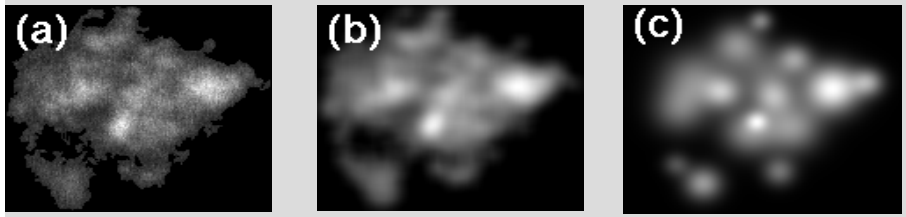


Fitted

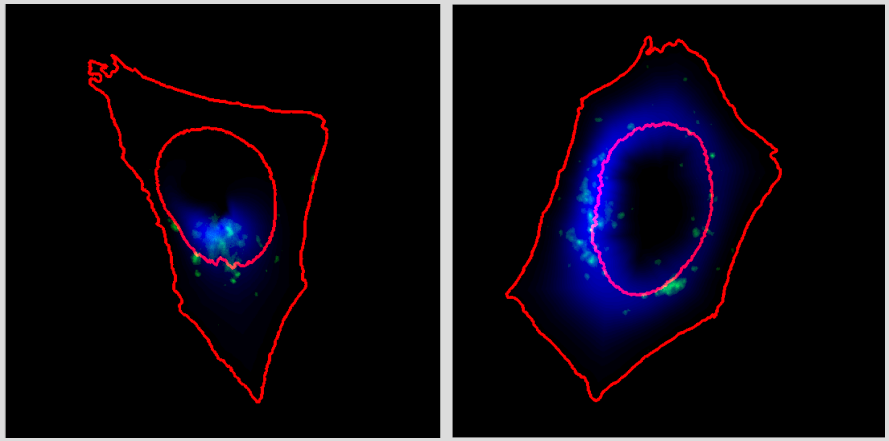


Zhao & Murphy 2007

Models for protein-containing objects

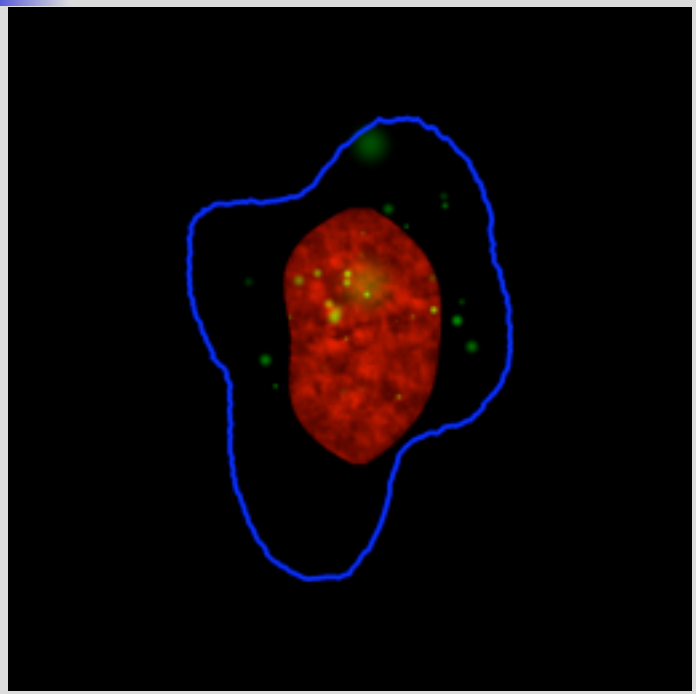


r: normalized distance, a: angle to major axis

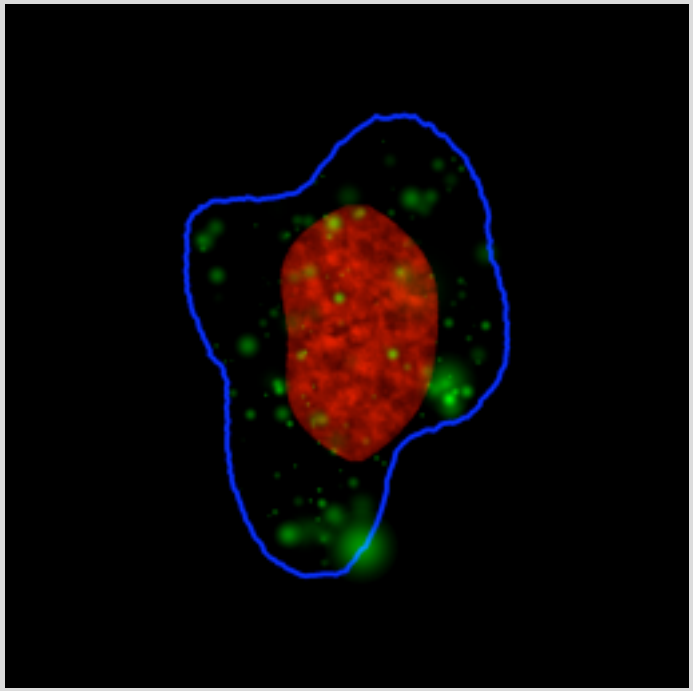


- Mixture of Gaussian objects
- Learn distributions for number of objects and object size
- Learn probability density function for objects relative to nucleus and cell

Synthesized Images

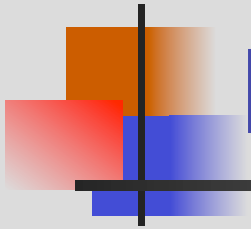


Lysosomes



Endosomes

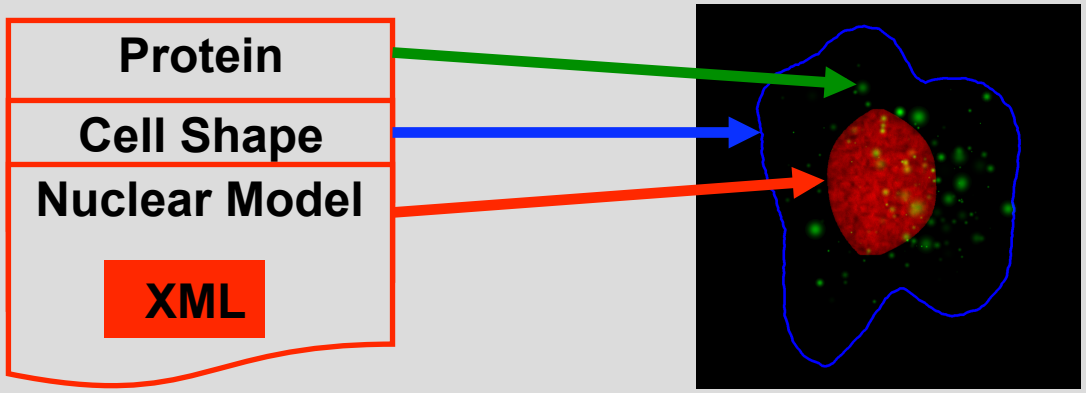
- **SLML toolbox - Ivan Cao-Berg, Tao Peng, Ting Zhao**
- Have portable tool for generating images from model



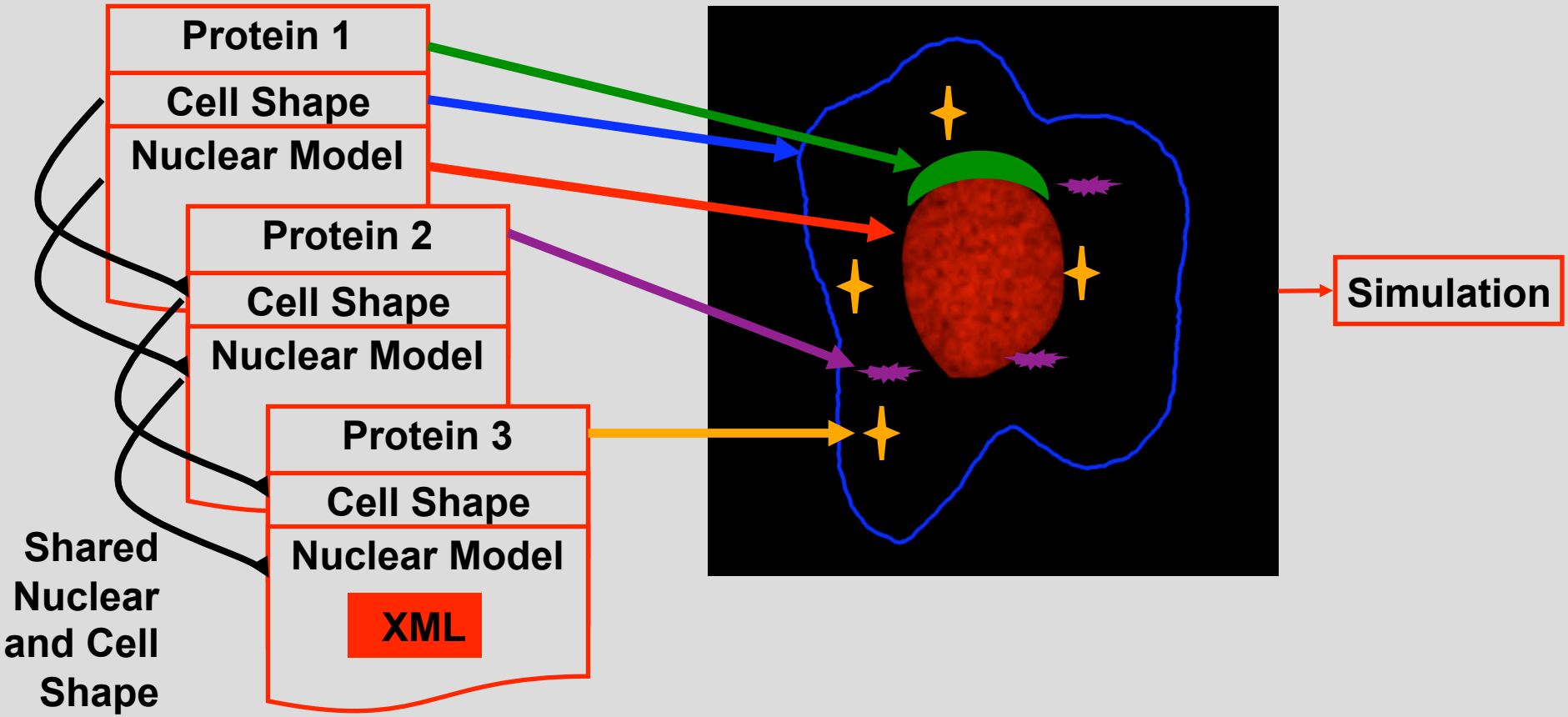
Model Distribution

- Generative models provide better way of distributing what is known about “subcellular location families” (or other imaging results, such as illustrating change due to drug addition)
- Have initial XML design for capturing the models for distribution
- Have portable tool for generating images from the model

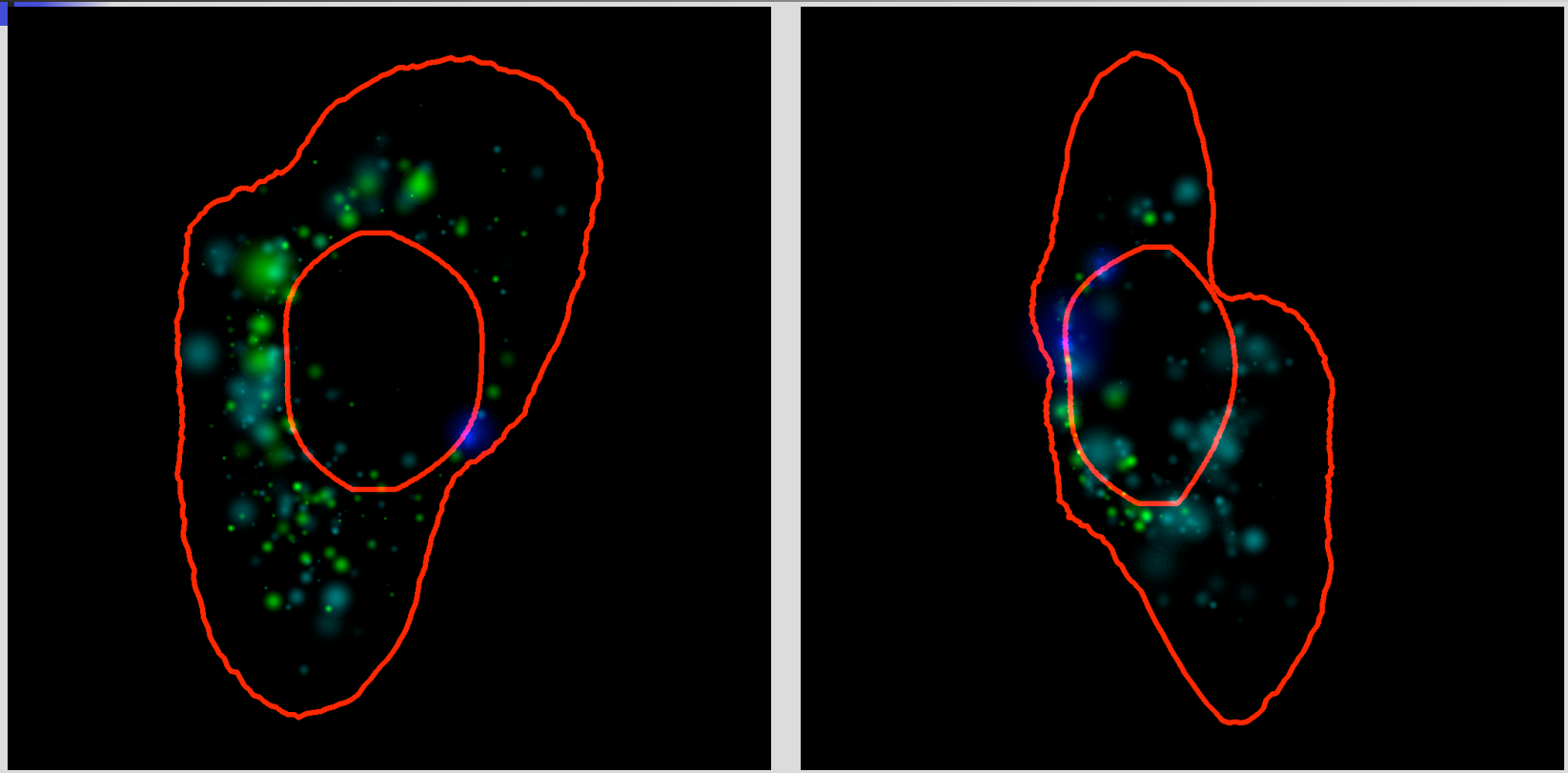
Generation Process



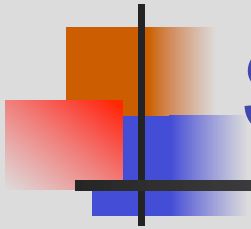
Combining Models for Cell Simulations



Example combination

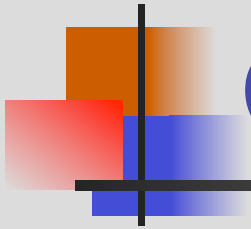


Red = nuclear membrane, plasma membrane
Blue = Golgi
Green = Lysosomes
Cyan = Endosomes



Some tools

- http://murphylab.web.cmu.edu/software
- http://www.openmicroscopy.org
- http://www.farsight-toolkit.org
- http://www.cbi-tmhs.org/Dcelliq
- http://icluster.imb.uq.edu.au
- http://www.cellprofiler.org

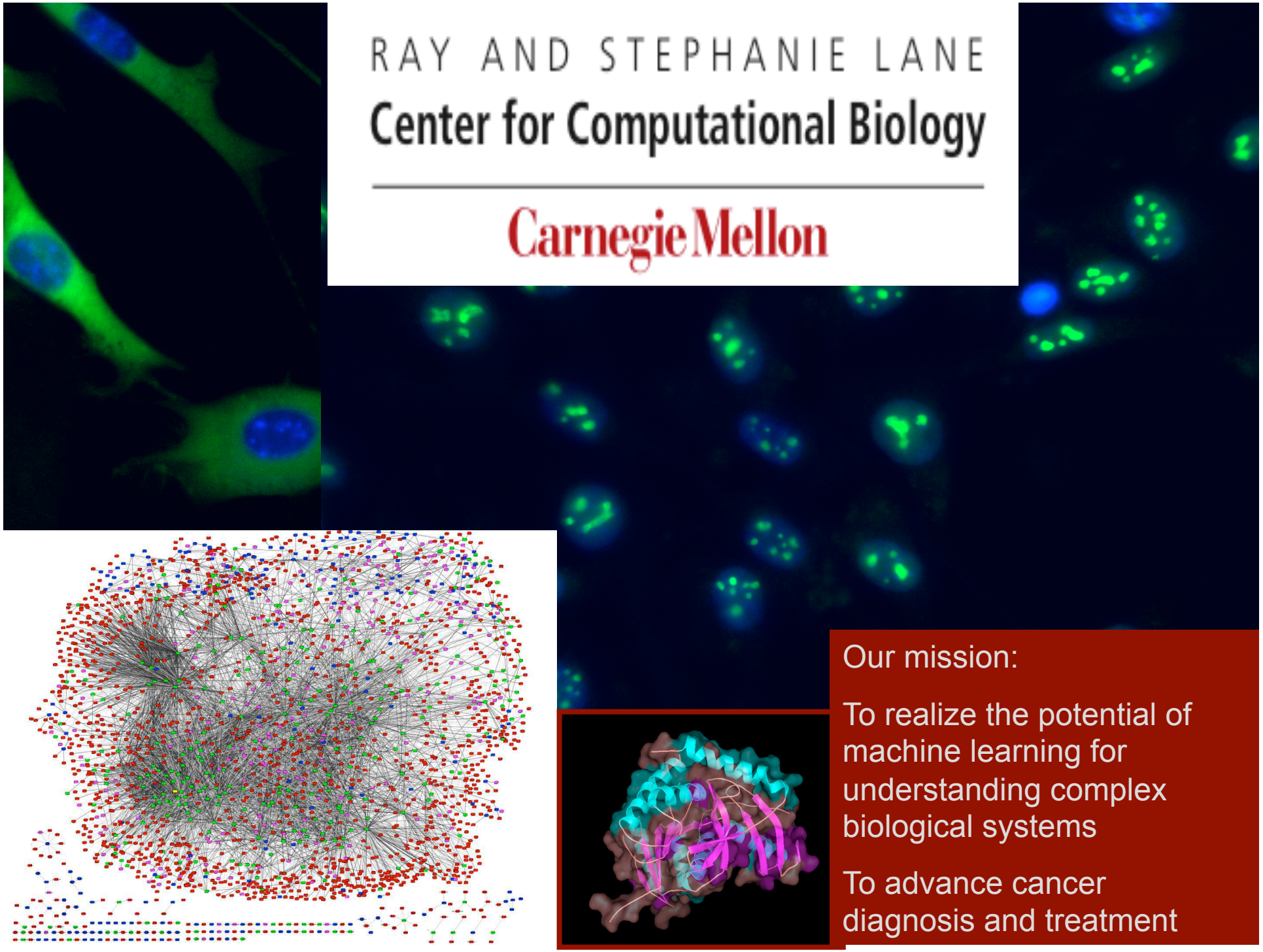


Conclusions

- Computers better than people at recognizing complex subcellular patterns
- Automated analysis of subcellular patterns in cells and tissues demonstrated – useful for potential biomarker discovery
- Generative models can be built directly from data to summarize results and make predictions – useful for cell simulations
- Many challenges remain!

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, Jieyue Li, Taraz Buck
- 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, Ghislain Bonamy, Sumit Chanda, Dan Rines**
- Centers
  - Molecular Biosensors and Imaging Center - TCNP (Waggoner)
  - National Center for Integrative Biomedical Informatics - NCBC (Athey)



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

Lane Fellows in Computational Biology

Recognize and support scientists of outstanding intellect dedicated to a career at the interface of computational and biological sciences so that they can pursue postdoctoral research in the rich computational environment at Carnegie Mellon

Candidates

- must be **nominated** by their thesis advisor or another faculty member from their Ph.D. granting institution **by October 15, 2009**
- must have received their doctoral degree after August 1, 2008 or be expected to receive their degree by August 1, 2010.

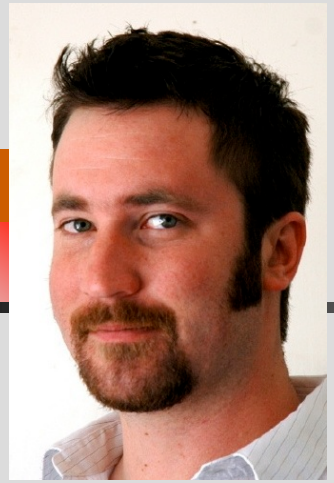
Fellows receive

- Stipend of **$56,000 per year** for up to three years
- Full fringe benefits including medical, dental, vision, life insurance
- Professional support allocation of **$6,000**

Carnegie Mellon

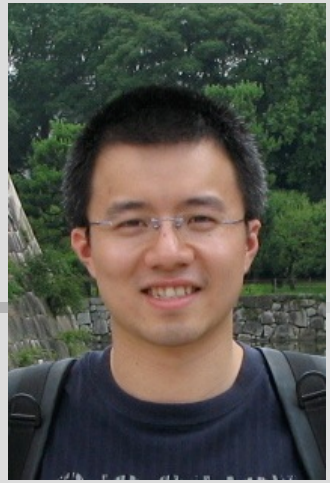
See http://lane.compbio.cmu.edu

2008 Lane Fellows



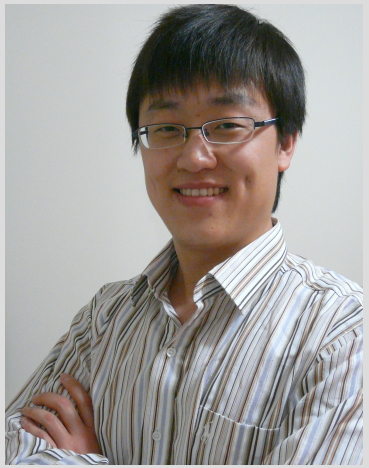
Peter Huggins
Univ California, Berkeley
Lior Pachter / Bernd Sturmfels

Arvind Rao
Univ Michigan (NCIBI)
Alfred Hero / David States / James Engel



Le Song
University of Sydney
Alex Smola

Hiroyuki Kuwahara
University of Utah
Chris J. Meyers



Xin Gao
University of Waterloo
Ming Li

Carnegie Mellon

2009 Lane Fellows