Automated Modeling of Subcellular Patterns for Systems Biology

Robert F. Murphy

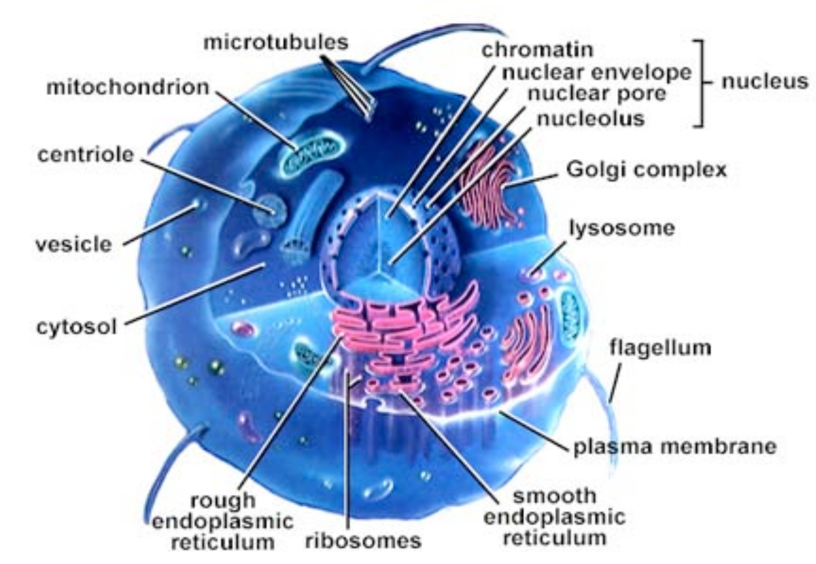
Departments of Biological Sciences, Biomedical Engineering and Machine Learning and

Center for Bioimage Informatics
from image to knowledge

Carnegie Mellon

Importance of Subcellular Location

- Eukaryotic cells are highly compartmentalized, and proper localization of proteins is critical to normal cell behavior
- Systems biology promises understanding of origins and consequences of cell behaviors
- Need 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

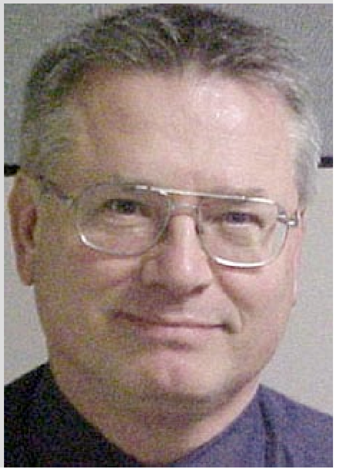


Carnegie Mellon

Subcellular Location in Drug Development

- A number of markers reflect (or cause!) changes in cell state (e.g., disease) by changing subcellular location
- These can be used to identify drugs that might treat or prevent disease
- Automated microscopes can be used to perform screening of a library of drugs
  - High-content screening

Lans Taylor



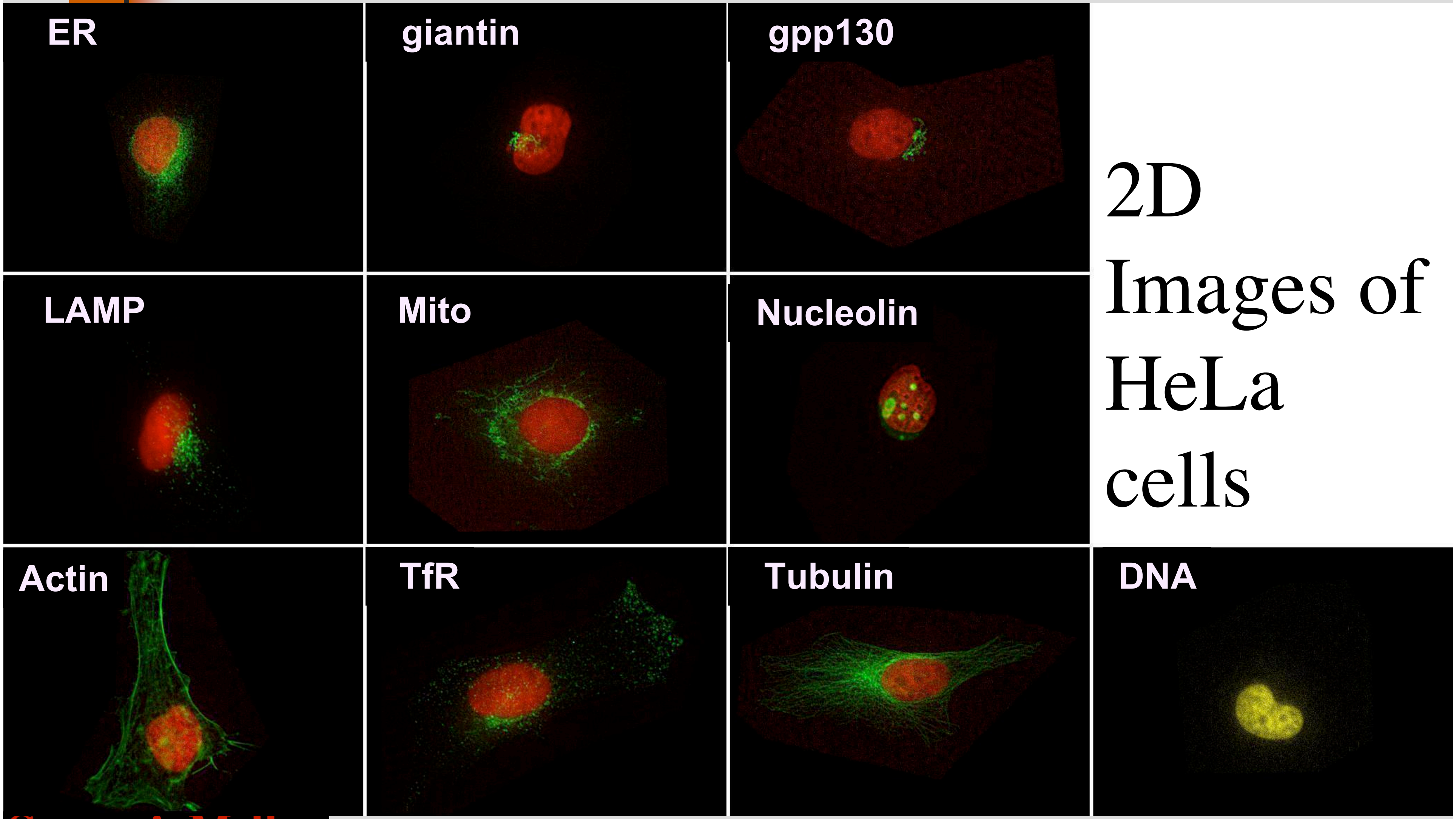
High-Content Screening and Location Proteomics

- Identification of targets for drug development assays typically very slow process driven by traditional biological experiments
- Alternative is to use proteome-wide approach to identify the locations of all proteins, including those that are candidates for disease-specific changes

Automated Interpretation

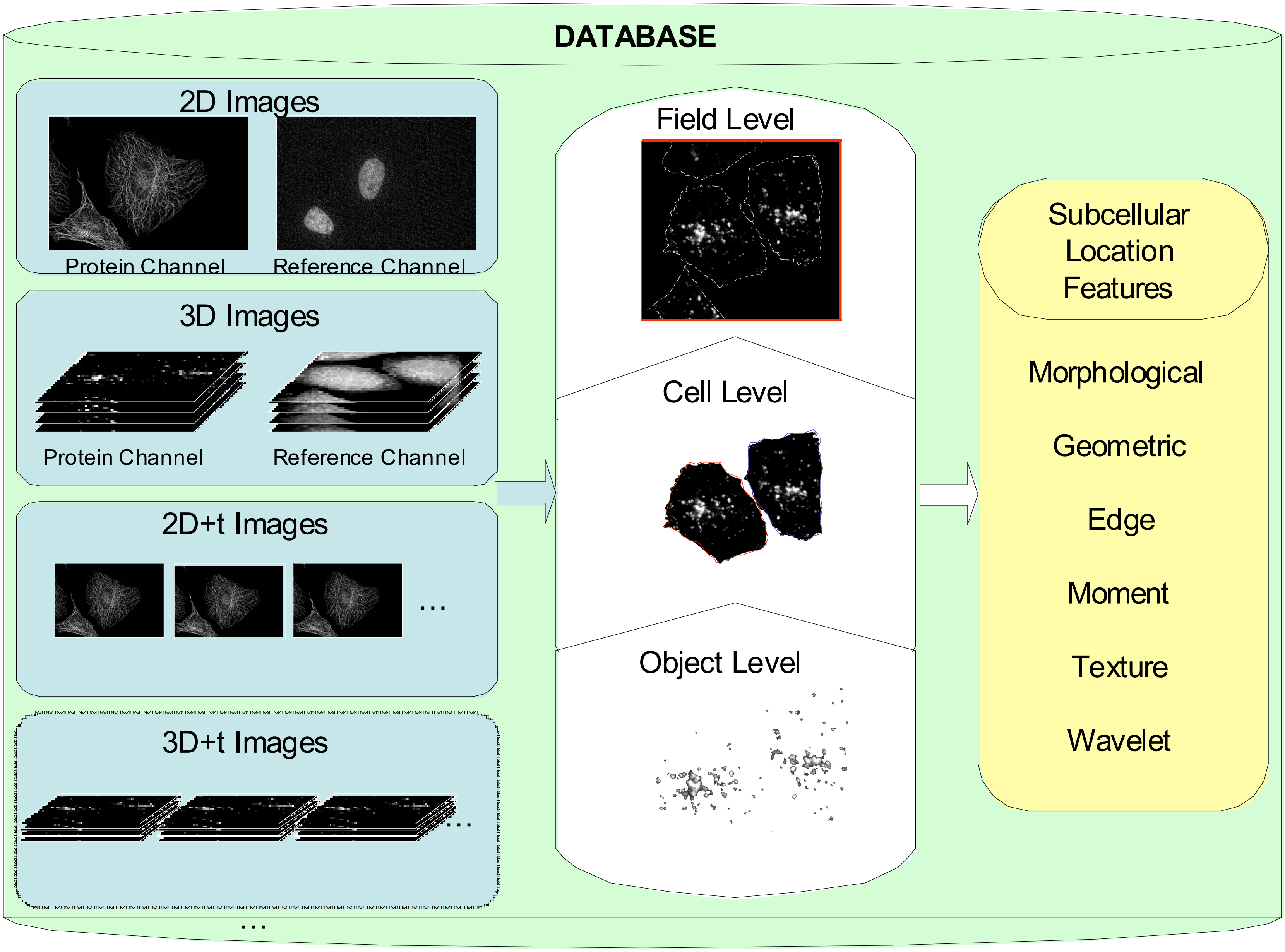
- Traditional analysis of fluorescence microscope images has occurred by visual inspection
- Our goal over the past eleven years has to been to automate interpretation with the ultimate goal of fully automated learning of protein location from images

Initial goal: Learn to recognize all major subcellular patterns



2D Images of HeLa cells

**Boland et al 1997; 1998;
Boland & Murphy 2001;
Huang & Murphy 2004**



DATABASE

2D Images

Protein Channel

Reference Channel

3D Images

Protein Channel

Reference Channel

2D+t Images

…

3D+t Images

…

…

Field Level

Cell Level

Object Level

Subcellular Location Features

Morphological

Geometric

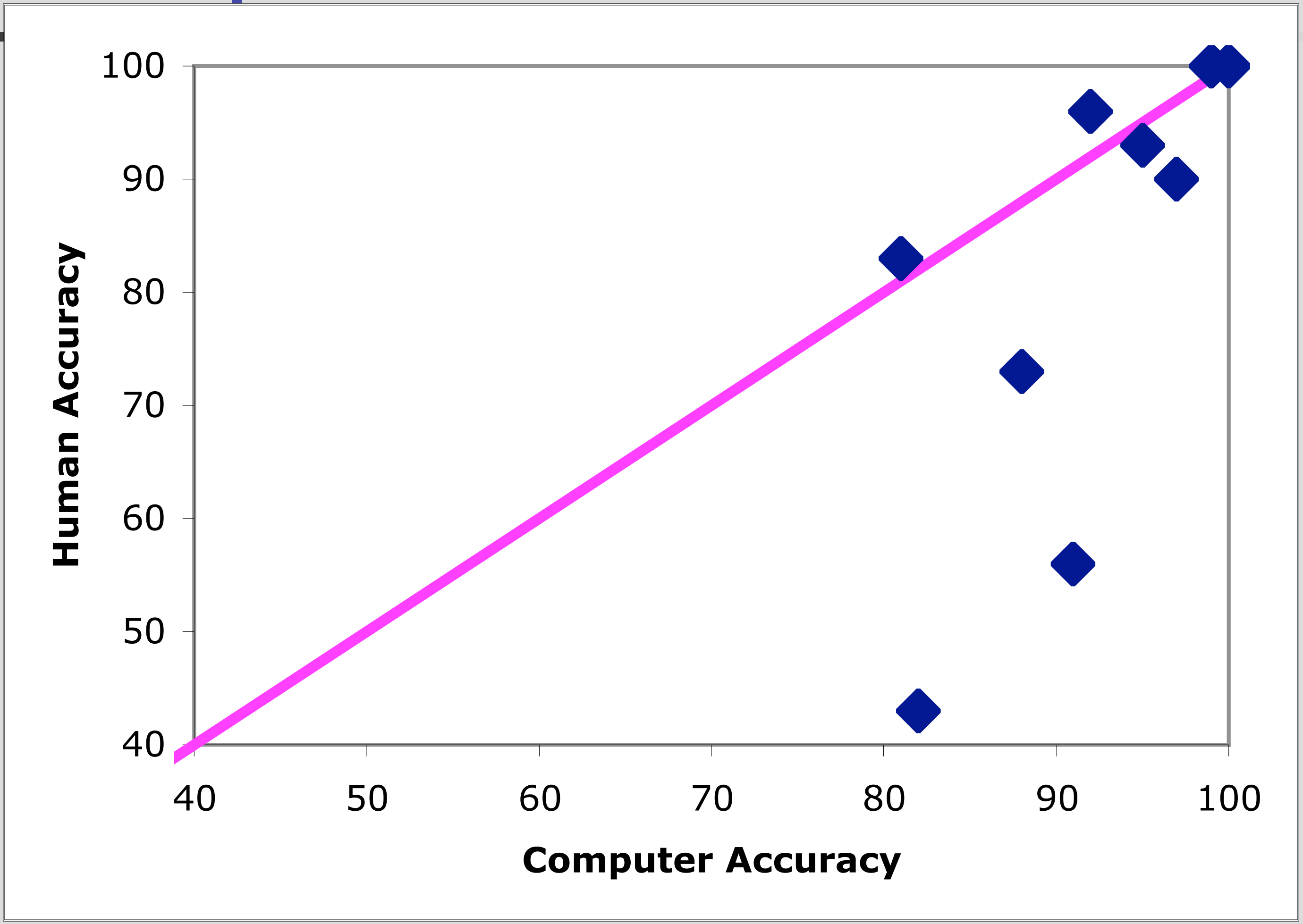
Edge

Moment

Texture

Wavelet

Classification Results: Computer vs. Human



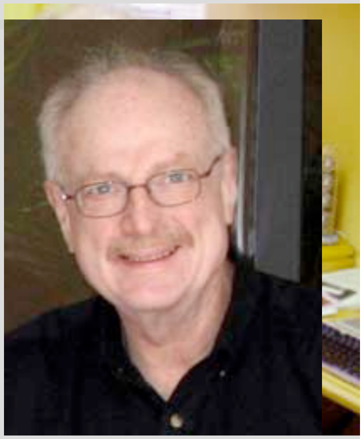
Carnegie Mellon

Supervised vs. Unsupervised Learning

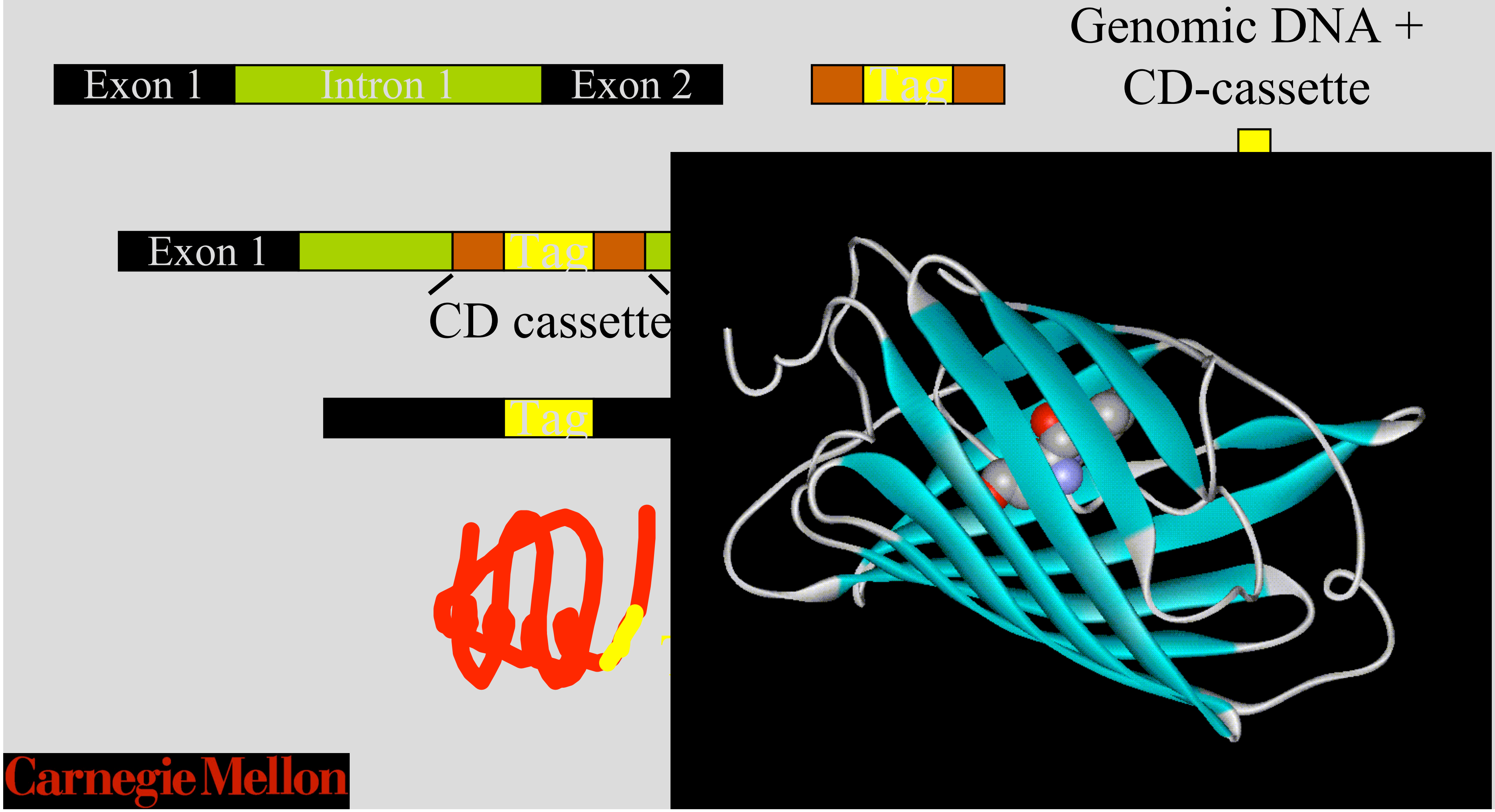
- This work demonstrates the feasibility of using classification methods to assign all proteins to known major classes
- Similar approach being taken in location prediction from sequence
- Do we know all locations? Are assignments to major classes enough?
- Need approach to discover classes

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



Principles of CD-Tagging (Jarvik & Berget) (CD = Central Dogma)



Genomic DNA + CD-cassette

Exon 1 | Intron 1 | Exon 2

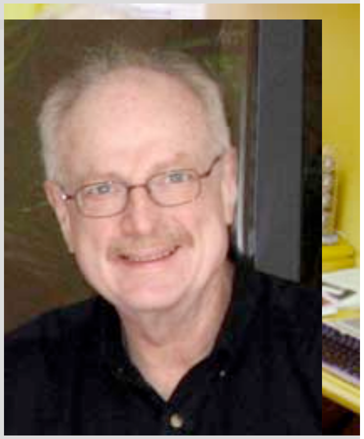
Tag

Exon 1 | Tag

CD cassette

Tag

Location Proteomics

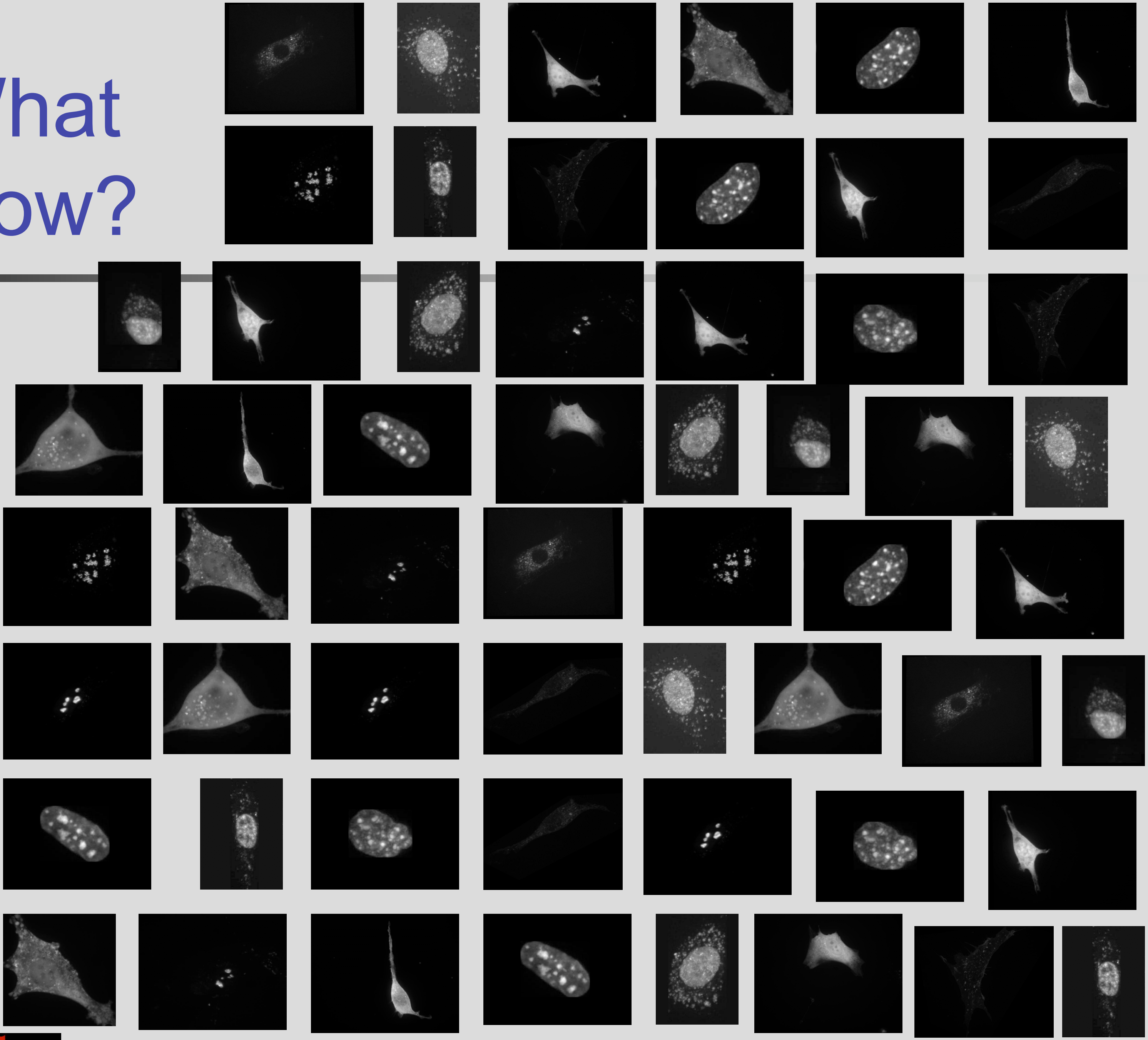


- **Tag** many proteins
  - We have used **CD-tagging** (developed by Jonathan Jarvik and Peter Berget): Infect population of cells with a retrovirus carrying DNA sequence that will "tag" in a random gene
- Isolate separate **clones**, each of which produces express one tagged protein
- Use RT-PCR to **identify tagged gene** in each clone
- Collect **many live cell images** for each clone using spinning disk confocal fluorescence microscopy

Jarvik et al 2002

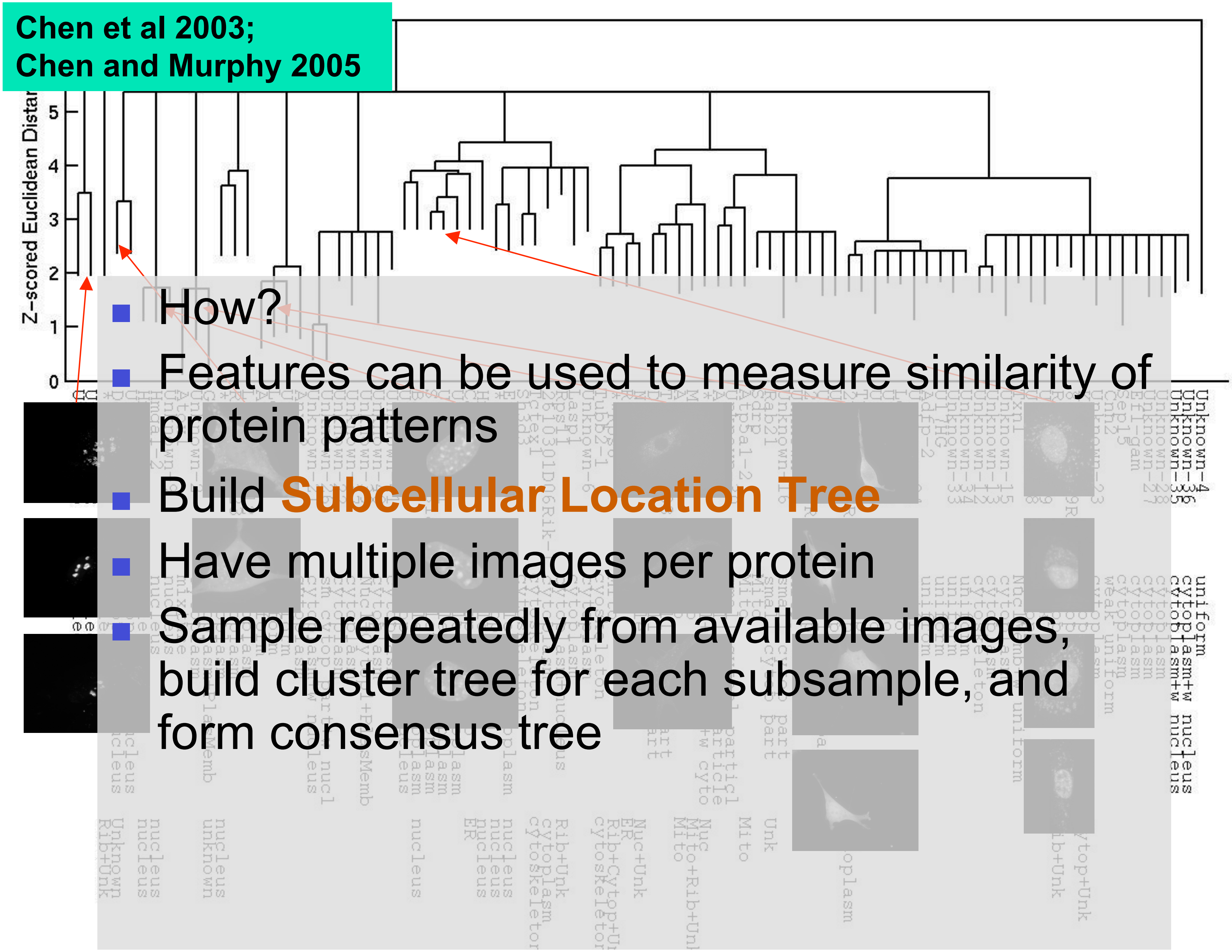
What Now?

Group ~90 tagged clones by pattern



Carnegie Mellon

Chen et al 2003;
Chen and Murphy 2005

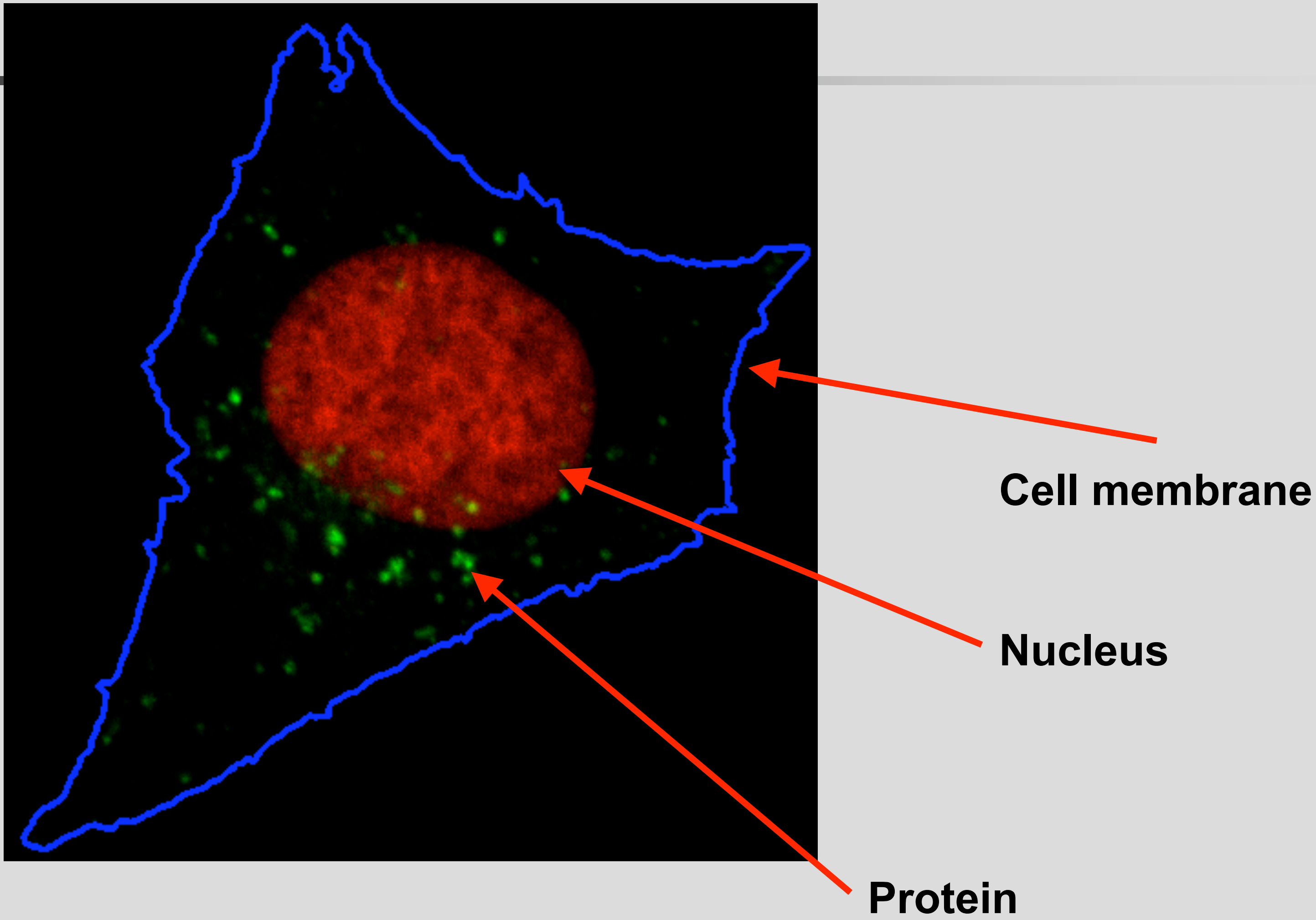


- How?
- Features can be used to measure similarity of protein patterns
- Build **Subcellular Location Tree**
- Have multiple images per protein
- Sample repeatedly from available images, build cluster tree for each subsample, and form consensus tree

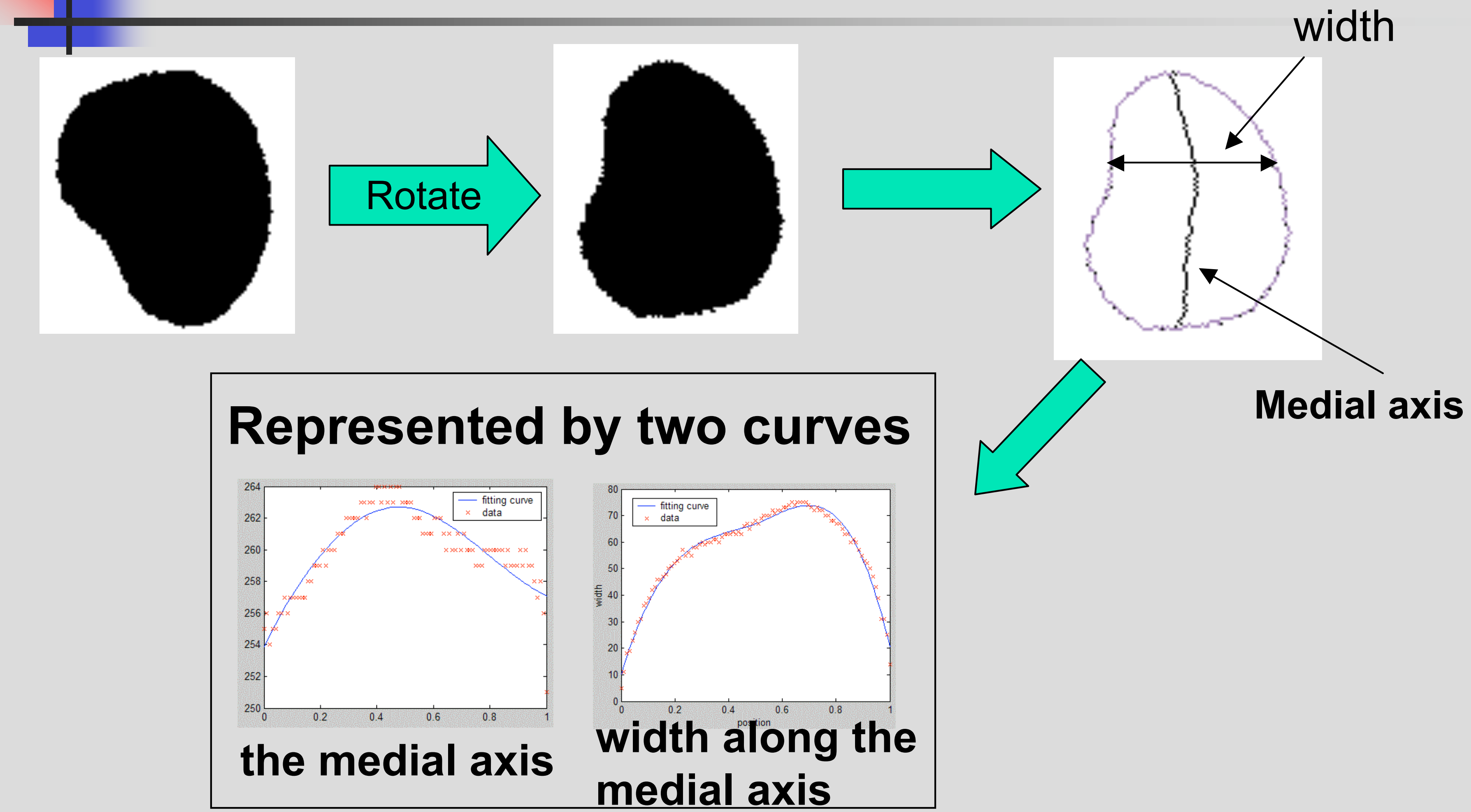
Need

- How do we communicate results of clustering patterns?
- Show all images from a given cluster?
  - Long download
  - No ability to generalize
- Proposal: Use generative models

LAMP2 pattern



Nuclear Shape - Medial Axis Model

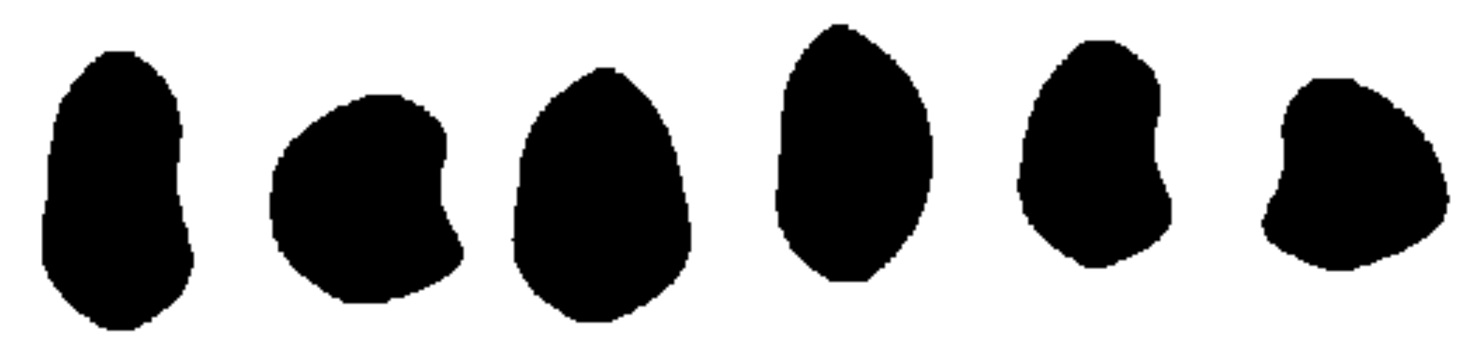


Represented by two curves

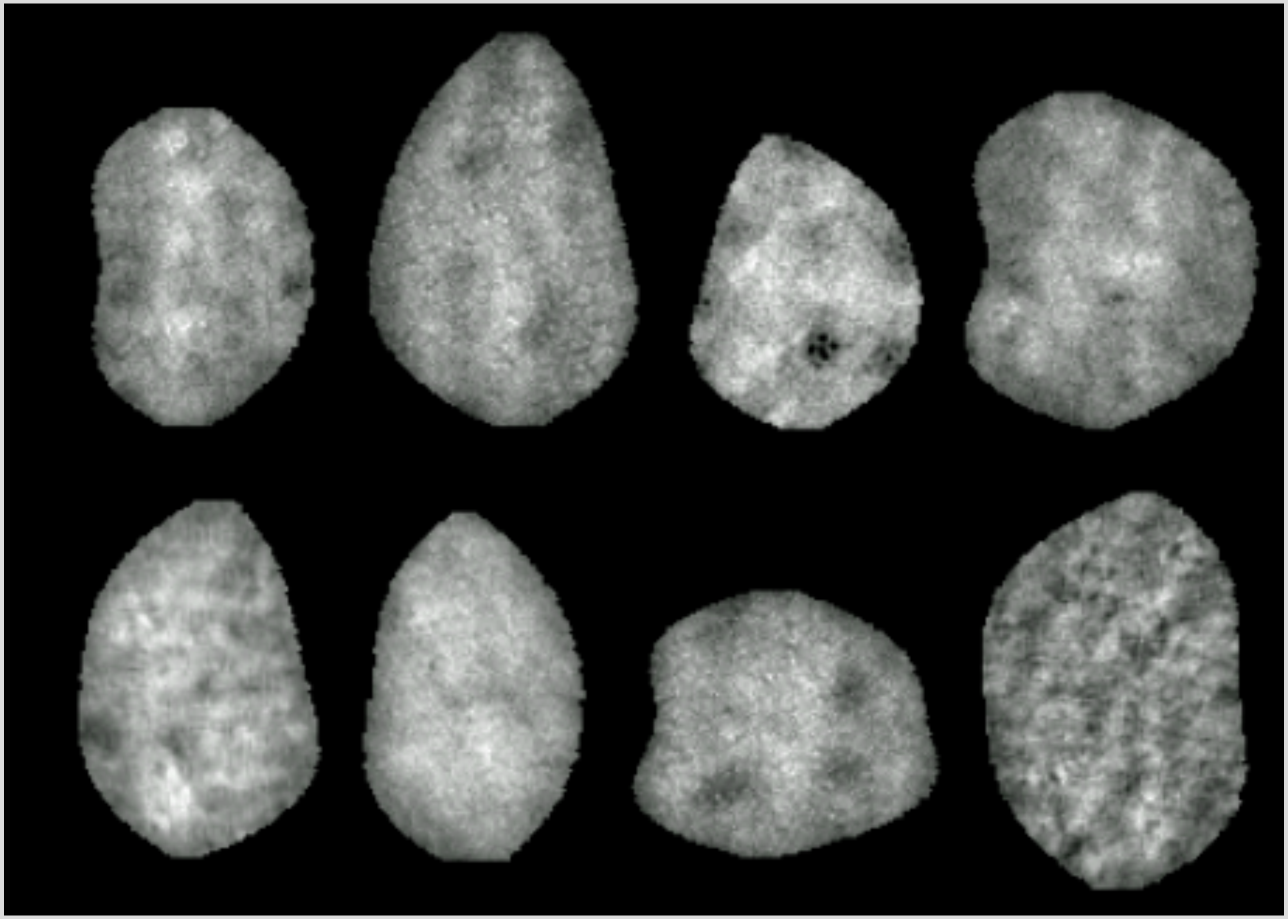
the medial axis

width along the medial axis

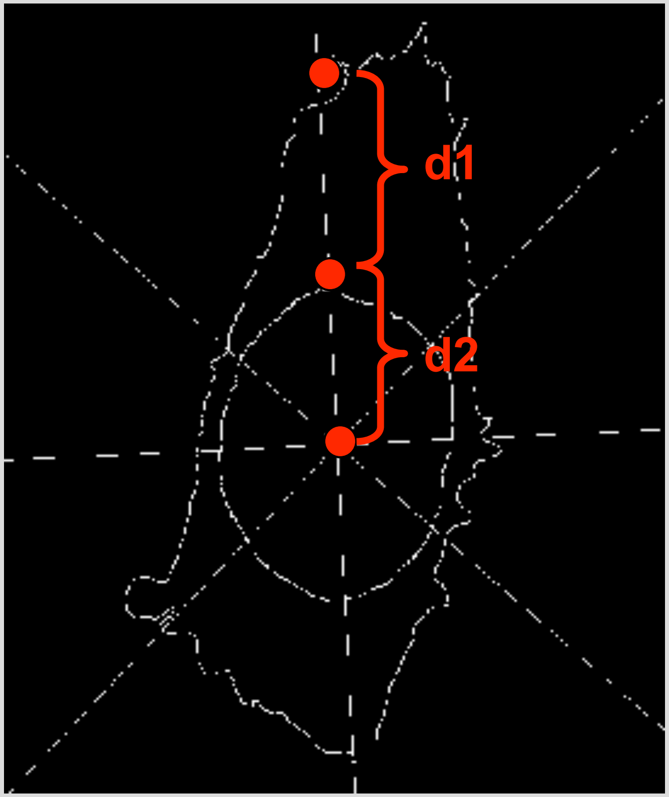
Synthetic Nuclear Shapes



Synthetic nuclei generated by learned model



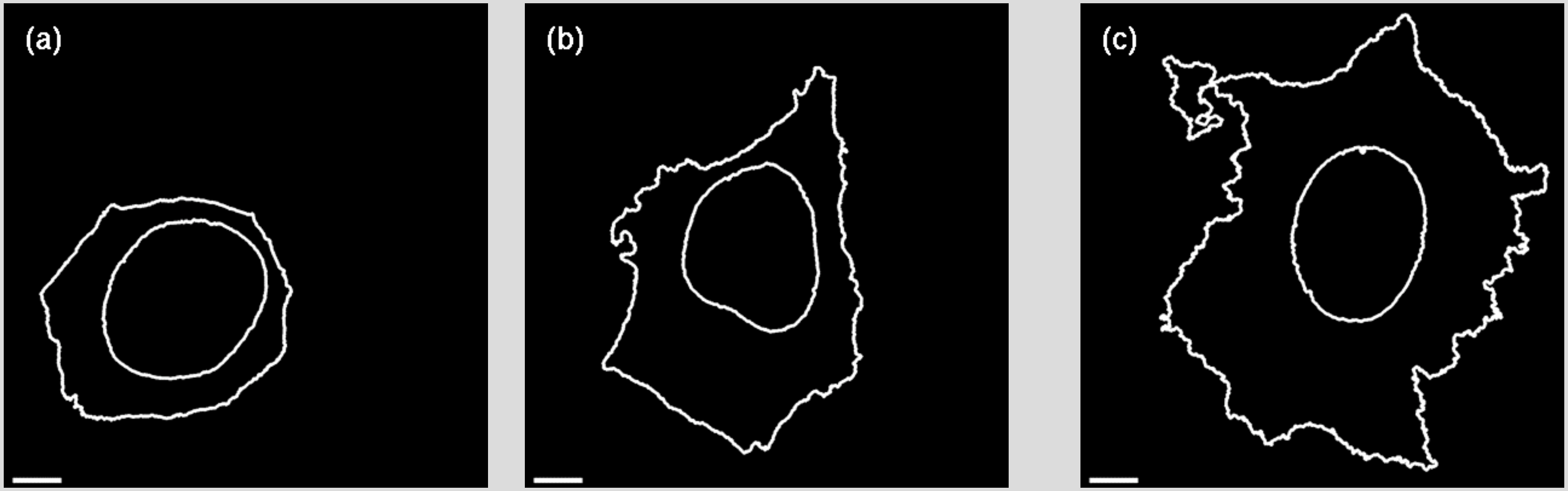
Cell Shape Description: Distance Ratio



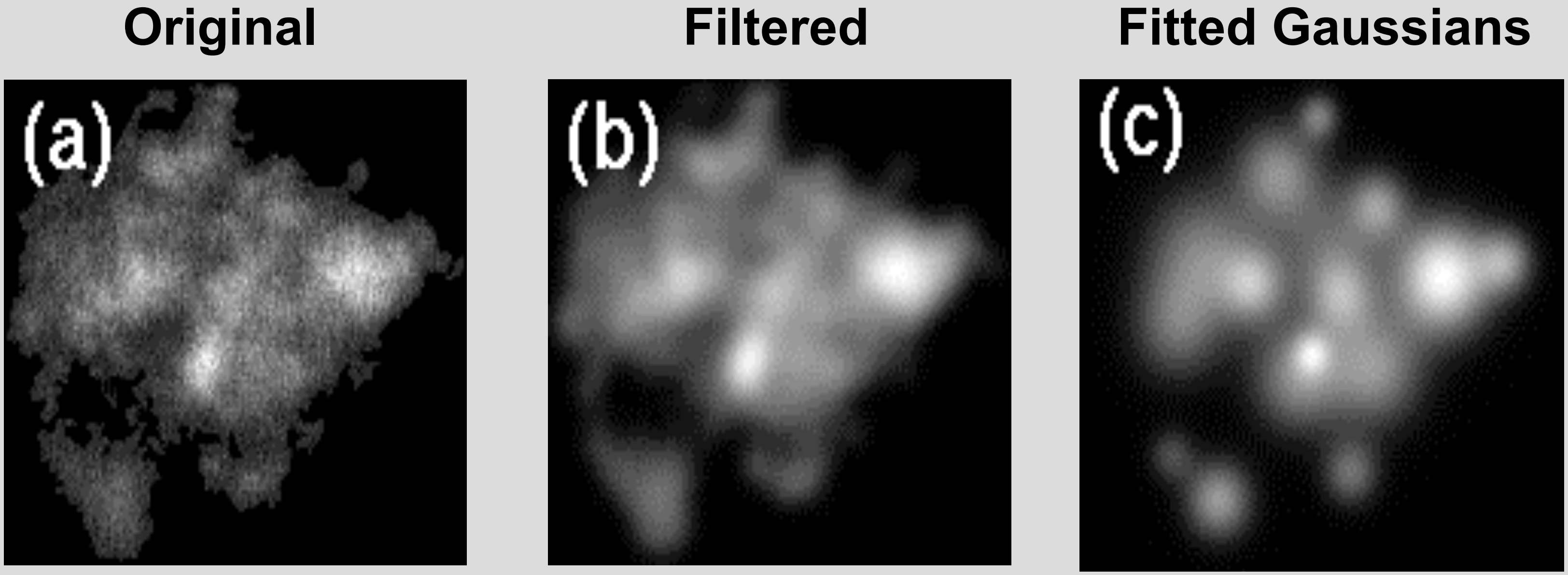
$$r = \frac{d_1 + d_2}{d_2}$$

Capture variation in the model

Examples of natural variation in cell shape



Modeling Vesicular Organelles

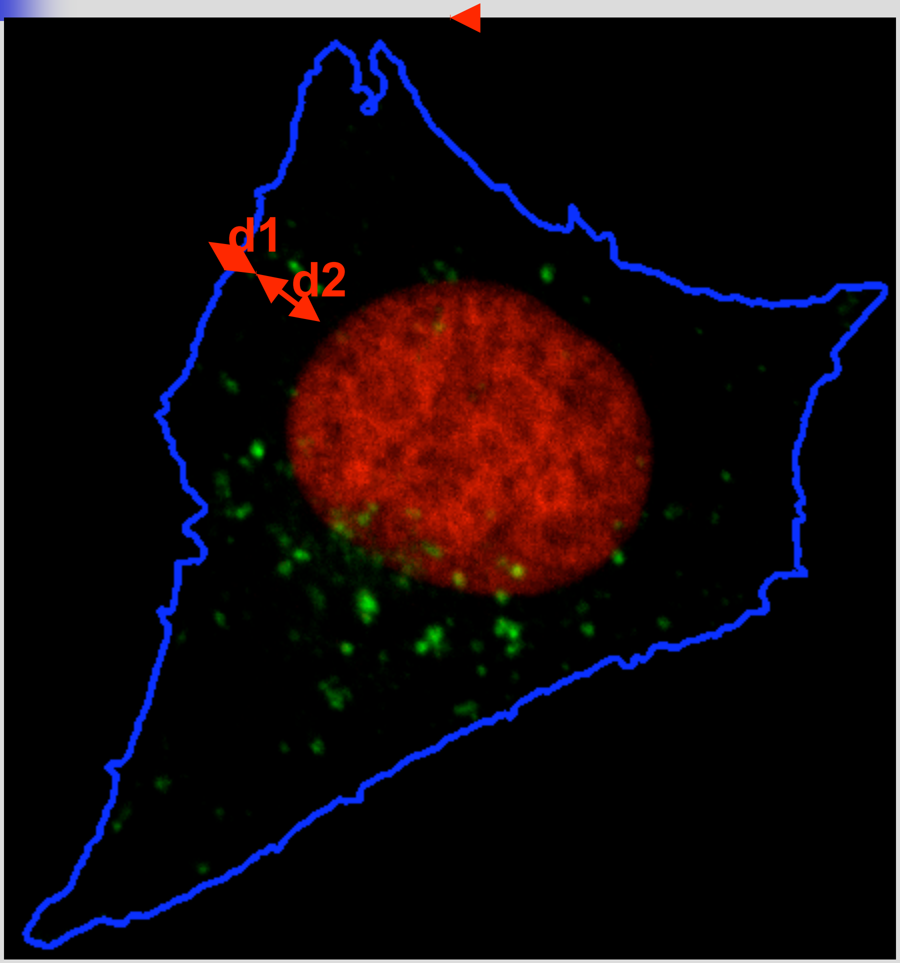


Original

Filtered

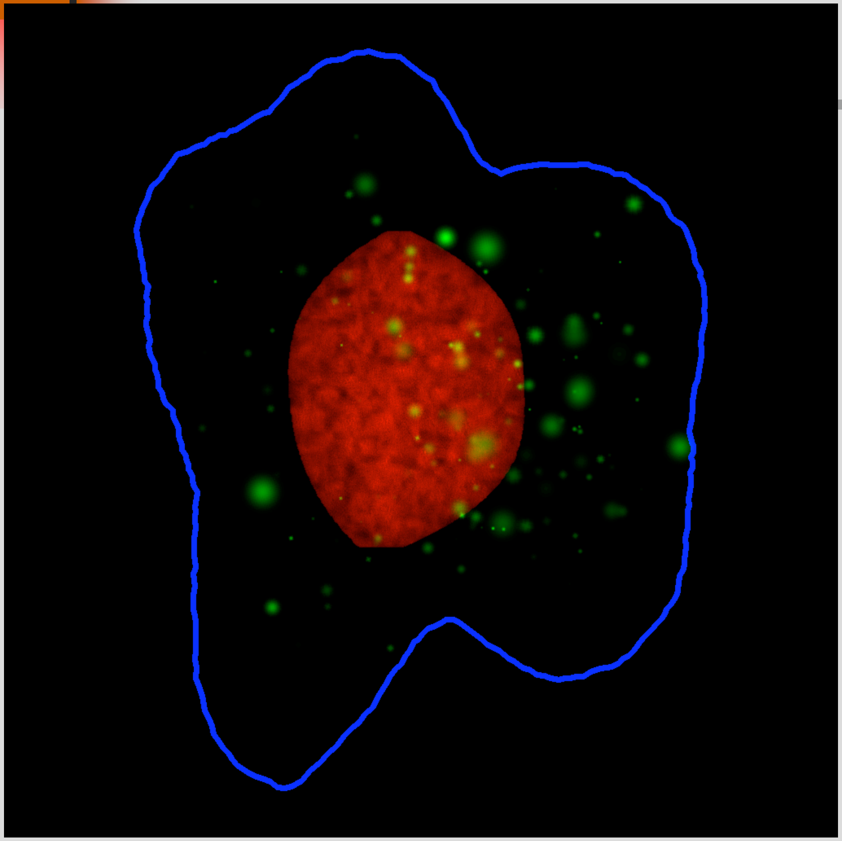
Fitted Gaussians

Object Positions

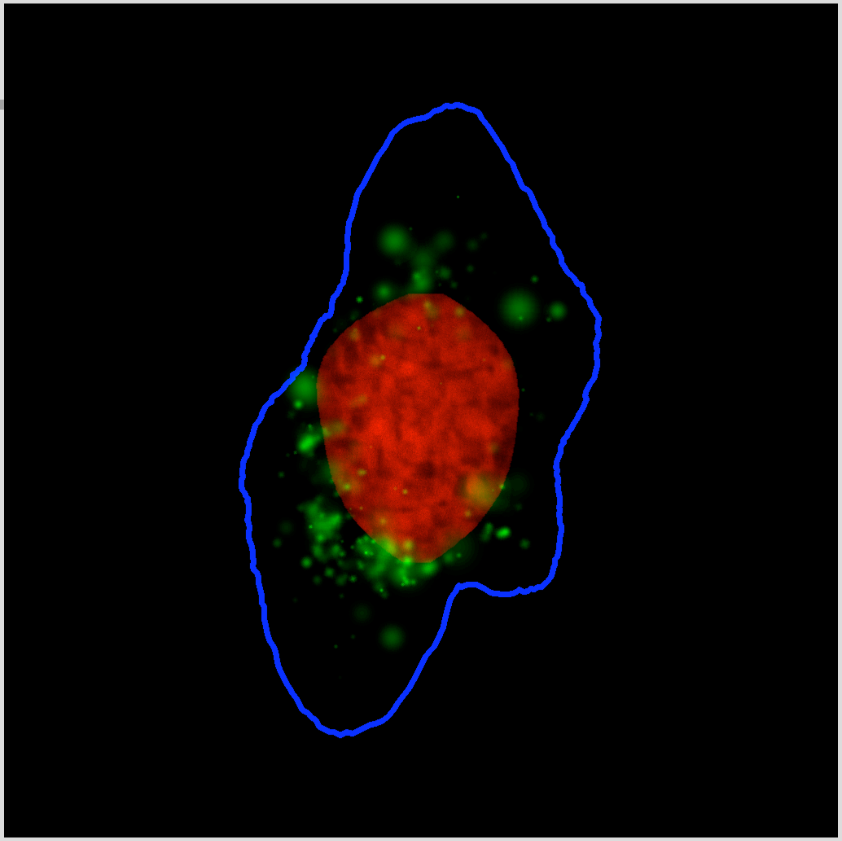


$$r = \frac{d_2}{d_1 + d_2}$$

Synthesized Images



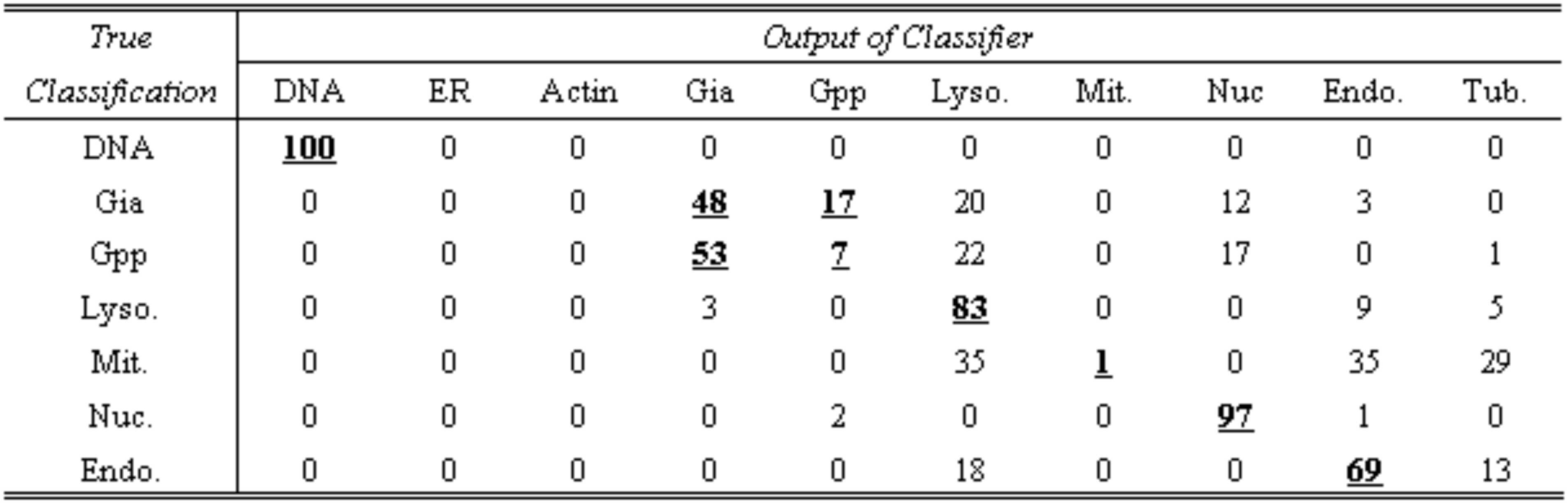
Lysosomes



Endosomes

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

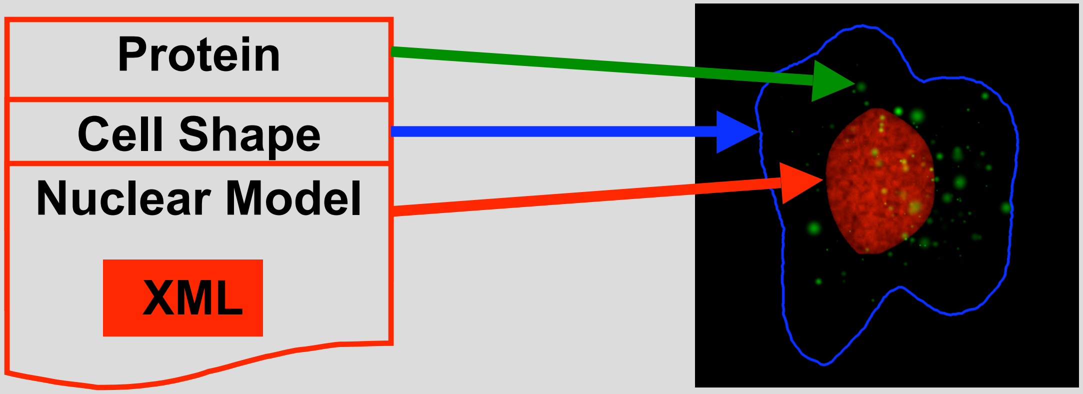


True	*Output of Classifier*									
Classification	DNA	ER	Actin	Gia	Gpp	Lyso.	Mit.	Nuc	Endo.	Tub.
DNA	**100**	0	0	0	0	0	0	0	0	0
Gia	0	0	0	**48**	**17**	20	0	12	3	0
Gpp	0	0	0	**53**	**7**	22	0	17	0	1
Lyso.	0	0	0	3	0	**83**	0	0	9	5
Mit.	0	0	0	0	0	35	**1**	0	35	29
Nuc.	0	0	0	0	2	0	0	**97**	1	0
Endo.	0	0	0	0	0	18	0	0	**69**	13

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



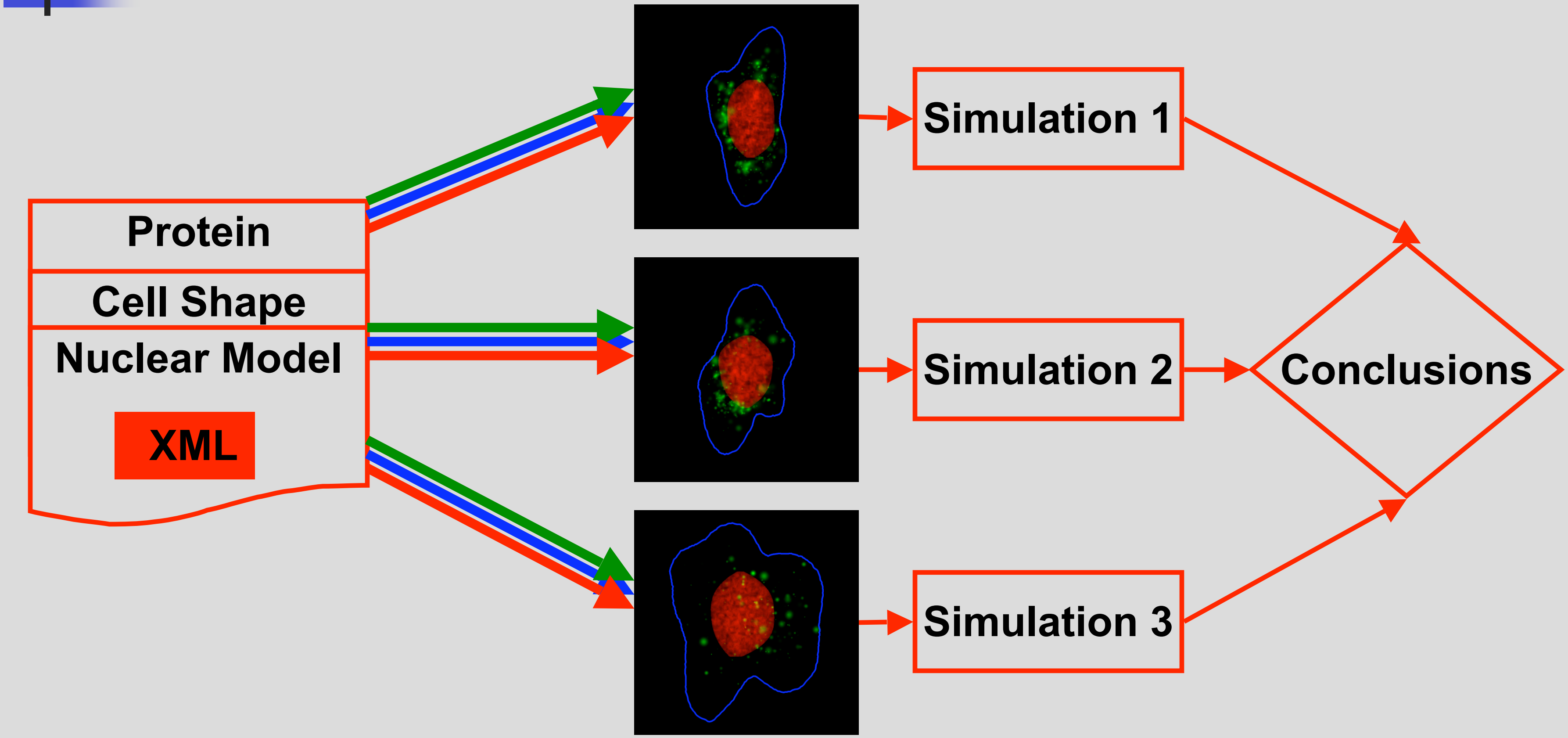
Protein

Cell Shape

Nuclear Model

XML

Generating Multiple Distributions for Simulations



Protein

Cell Shape

Nuclear Model

XML

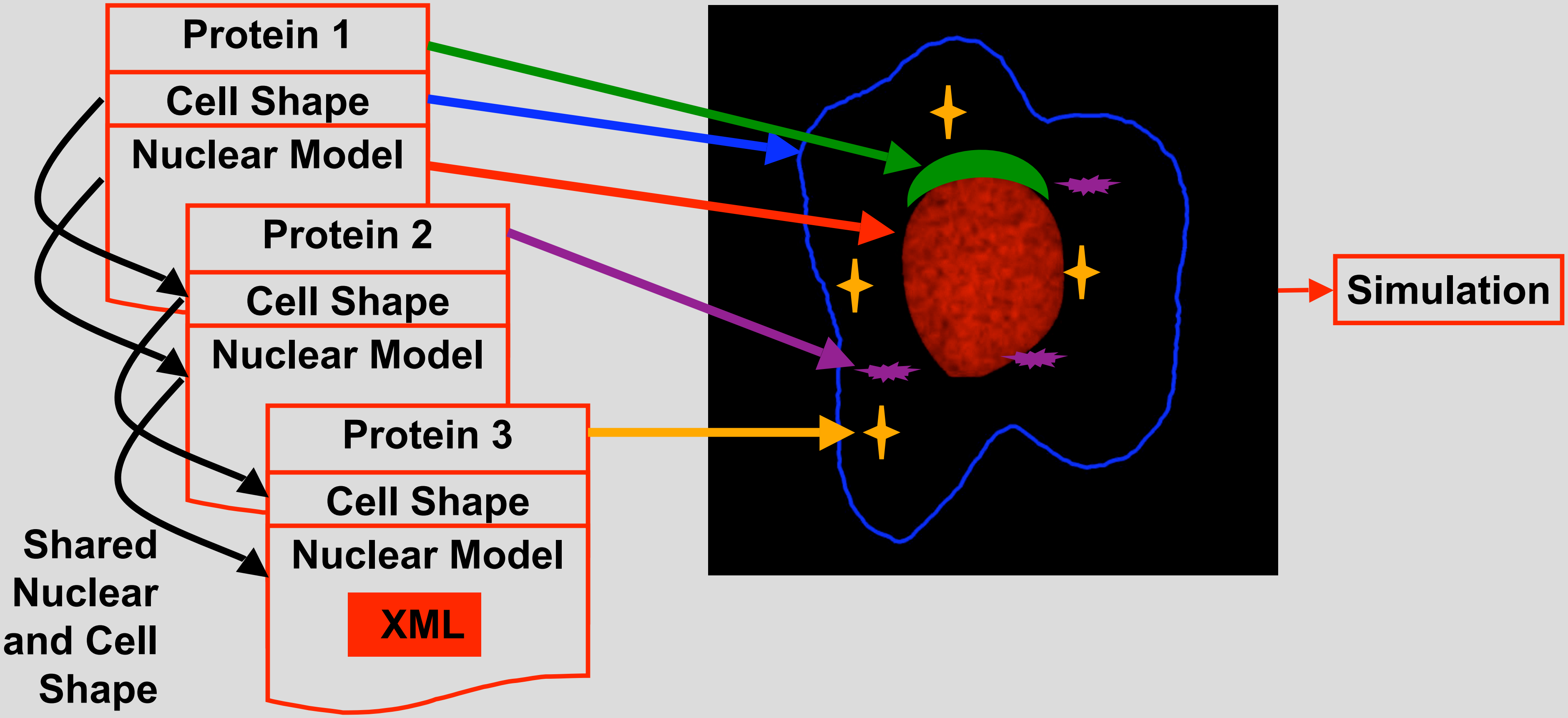
Simulation 1

Simulation 2

Simulation 3

Conclusions

Combining Models for Cell Simulations



Protein 1

Cell Shape

Nuclear Model

Protein 2

Cell Shape

Nuclear Model

Protein 3

Cell Shape

Nuclear Model

XML

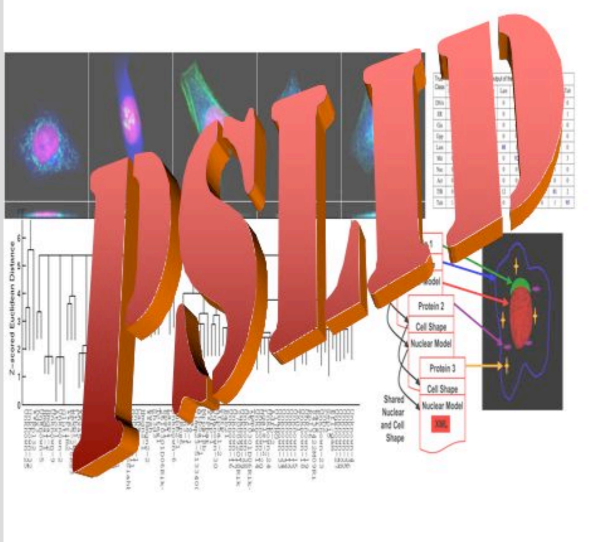
Shared Nuclear and Cell Shape

Simulation

Huang et al 2002; Huang et al 2007

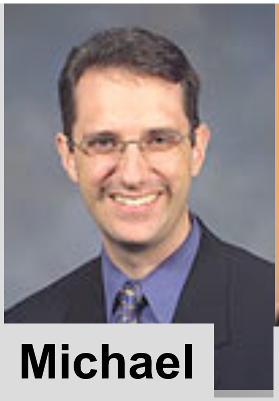
PSLID: Protein Subcellular Location Image Database

- Publicly accessible image database at http://pslid.cbi.cmu.edu
  - Version 3 released February 2, 2007
  - 2D and 3D images (single cell regions defined)
  - Two cell types, HeLa and 3T3
  - Over 120,000 images/
    3000 unique fields/14,000 cells
  - 111 classes; 55 known proteins;
    11 targeting mutants of one protein
  - Programmatic search via URL



Protein Subcellular Location Image Database

Acknowledgments


Michael

Mia

Greg

Meel

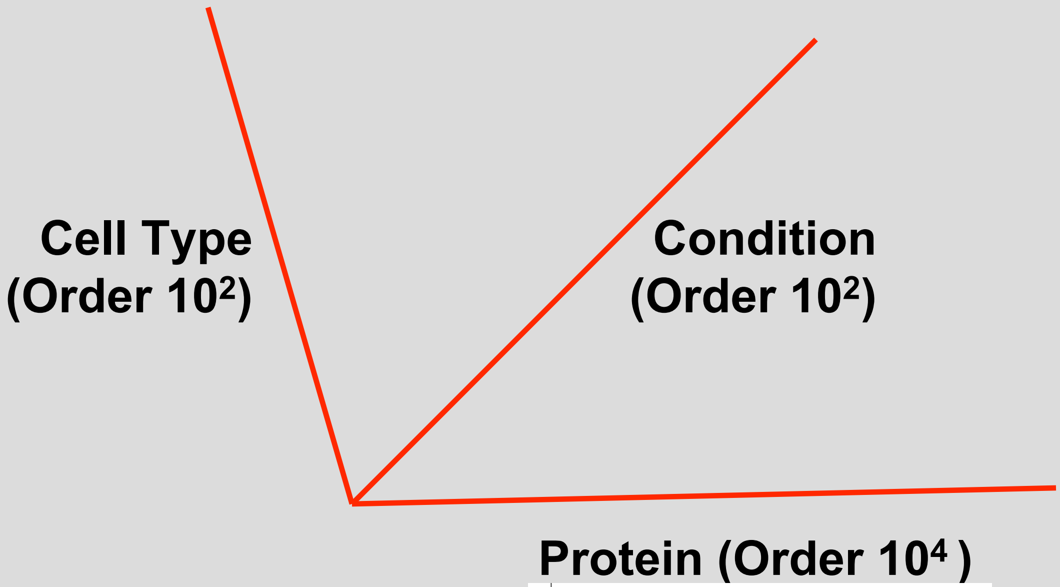
- Students
  - Dr. Michael Boland
  - Dr. Mia Markey (ugrad)
  - Gregory Porreca (ugrad)
  - Dr. Meel Velliste
  - Dr. Kai Huang
  - **Dr. Xiang Chen**
  - **Ting Zhao**
  - **Shann-Ching Chen**
  - **Juchang Hua**
- Funding
  - NSF, NIH, Commonwealth of Pennsylvania
- Collaborators/Consultants
  - David Casasent, Simon Watkins, **Jon Jarvik, Peter Berget, Jack Rohrer**, Tom Mitchell, Christos Faloutsos, Jelena Kovacevic, William Cohen, **Geoff Gordon**
  - NSF ITR: B. S. Manjunath Ambuj Singh

Kai

Xiang

Carnegie Mellon

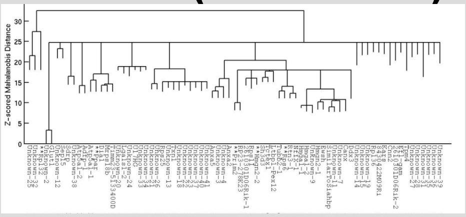
The future of subcellular location analysis



Cell Type (Order 10^2)

Condition (Order 10^2)

Protein (Order 10^4)



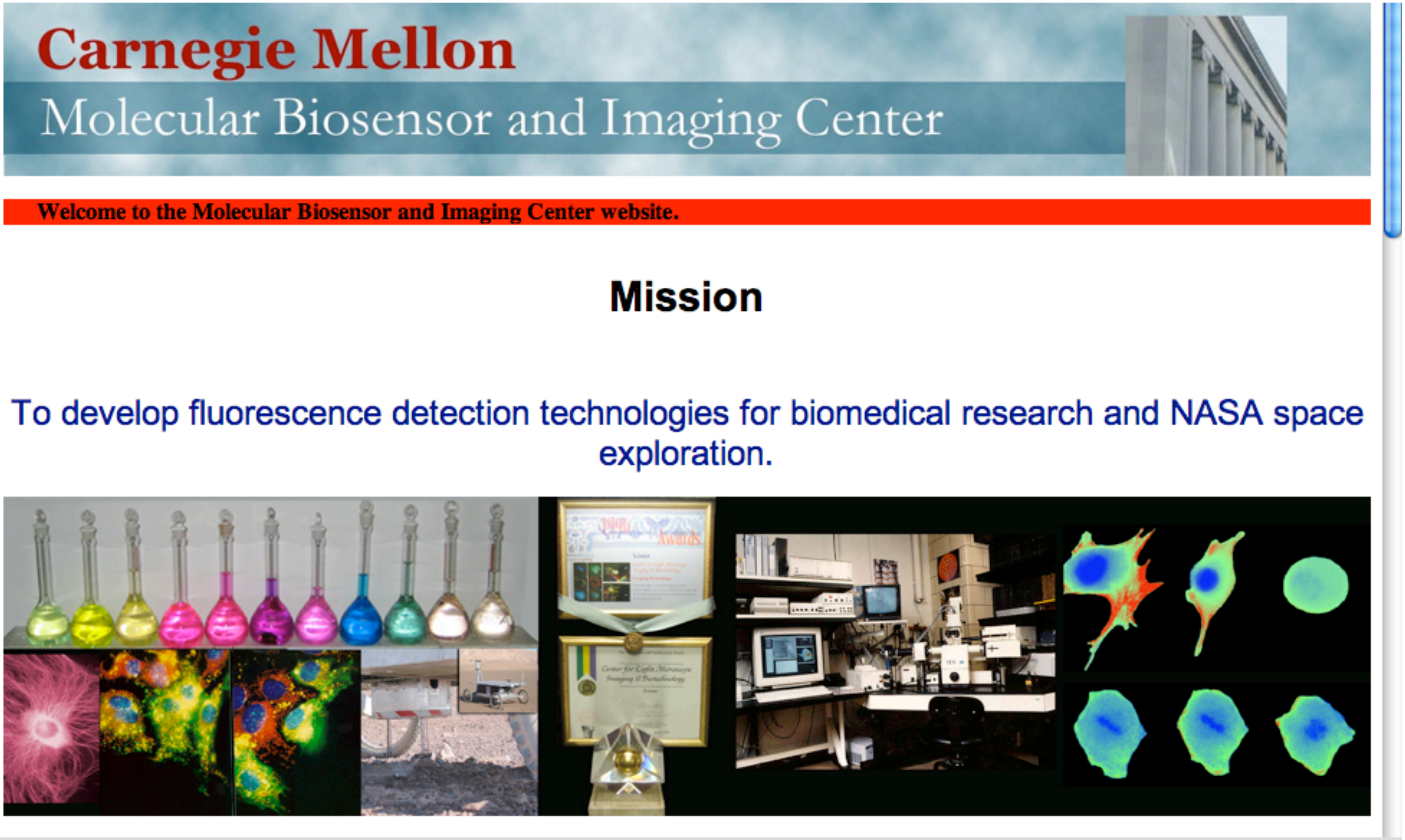
Plus: Time scale from subsecond to years

Other subcellular location projects

- Pepperkok group - human (MCF7 cells)
  - GFP-tagged cDNAs
  - GFP and DNA images
- Uhlen group (Protein Atlas) - human
  - Immunohistochemistry with monospecific antibodies
  - DAB and hematoxylin images
  - Fixed tissues
- Schubert group (MELK technology)
  - Cycles of immunofluorescence, imaging and bleaching
  - Fixed tissues

How do we really analyze subcellular location?

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



Carnegie Mellon

Molecular Biosensor and Imaging Center

Welcome to the Molecular Biosensor and Imaging Center website.

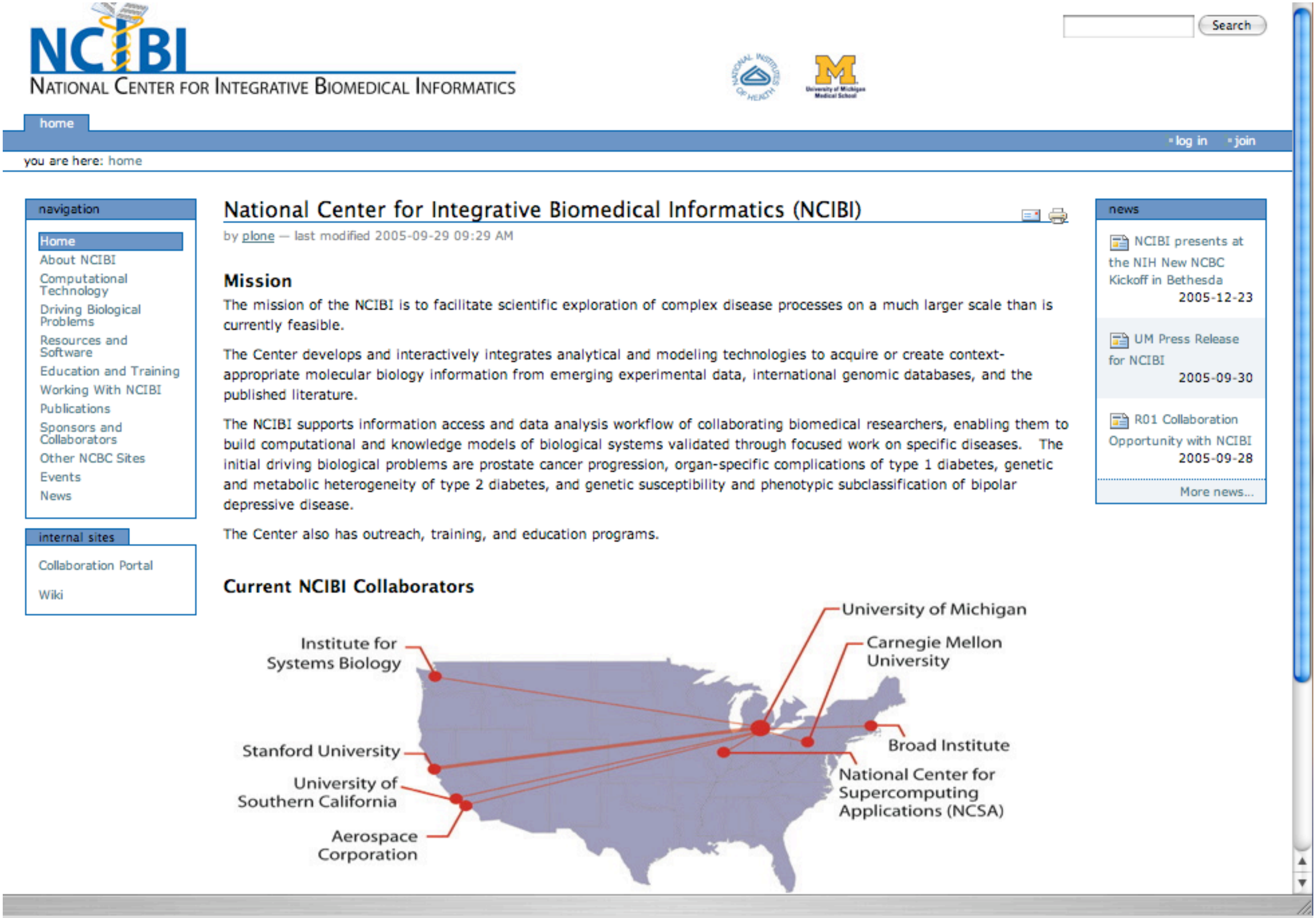
Mission

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

NIH Technology Center for Networks and Pathways

Carnegie Mellon

Alan Waggoner



you are here: home

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

University of Michigan
Carnegie Mellon University
Institute for Systems Biology
Stanford University
University of Southern California
Aerospace Corporation
Broad Institute
National Center for Supercomputing Applications (NCSA)

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...

Brian Athey (UMich), CMU: Bob Murphy

XXIV International Congress

Cytometry in the Age of Systems Biology

ISAC
International Society for Analytical Cytology

May 17-21, 2008
Budapest Sportarena
Budapest, Hungary

Mark your calendar and watch for details as they develop.
www.isac-net.org

Pre-Congress Introductory Courses for Flow Cytometry and for Cellular Imaging, May 16-17

Polychromatic flow Cytometry • High-content analysis • High-throughput analysis • Automated cellular imaging • Translational medicine applications • Proteomics and cytomics