


Automated Learning of Protein Subcellular Locations for Modeling of Cell Behavior

Robert F. Murphy
 Ray and Stephanie Lane Professor of Computational Biology
 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



Open questions

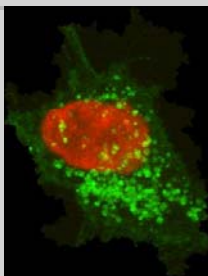
- How many distinct locations can proteins be found in? What are they?

Carnegie Mellon



Determining protein location

- The primary method used to **determine** the subcellular location of a protein is to “tag” it with fluorescent probe and then image its distribution within cells using fluorescence microscopy



Can tag with antibodies or by fusing gene with fluorescent protein such as GFP

Carnegie Mellon

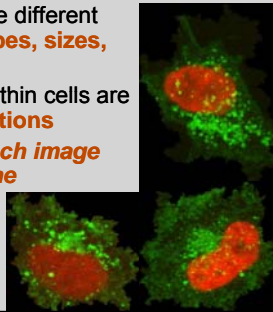
Automated Interpretation

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

Carnegie Mellon

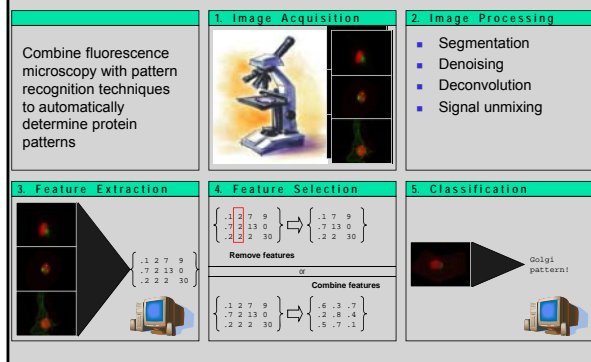
The Challenge

- Problem is hard because different cells have different **shapes, sizes, orientations**
- Organelles/structures within cells are **not found in fixed locations**
- Therefore, describe each image numerically and use the descriptors**



Carnegie Mellon

Approach



Initial goal: Learn to recognize all major subcellular patterns

2D Images of HeLa cells

Carnegie Mellon

Classification Results: Computer vs. Human

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

Organelle	Computer Accuracy (%)	Human Accuracy (%)
Lysosomes	~85	~75
Giantin (Golgi)	~90	~60
Gpp130 (Golgi)	~85	~45

Note: Even better results for 3D images!

Carnegie Mellon

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

Carnegie Mellon

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

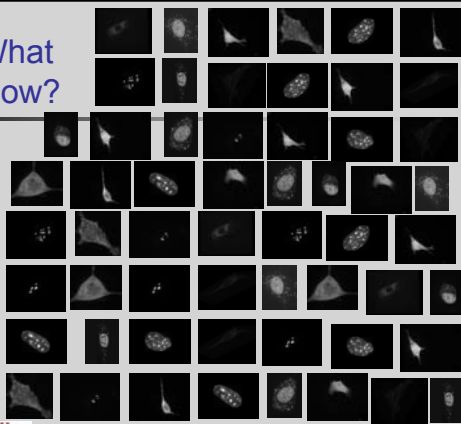
Jarvik et al 2002

Use RT-PCR to **identify tagged gene** in each clone
Collect **many live cell images** for each clone using spinning disk confocal fluorescence microscopy

Carnegie Mellon

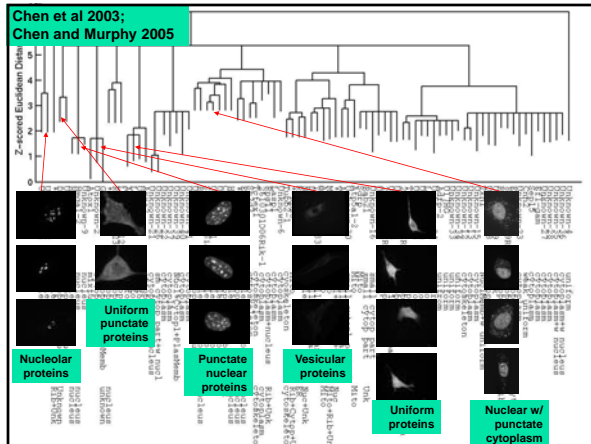
What Now?

Group ~90 tagged clones by pattern




Carnegie Mellon

Chen et al 2003;
Chen and Murphy 2005

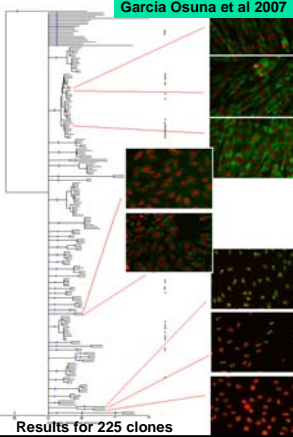


CD-tagging project

- Running ~100 clones/wk
- Automated imaging



Elvira Garcia Osuna



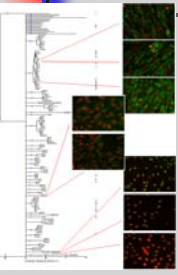
Garcia Osuna et al 2007

Results for 225 clones

Carnegie Mellon

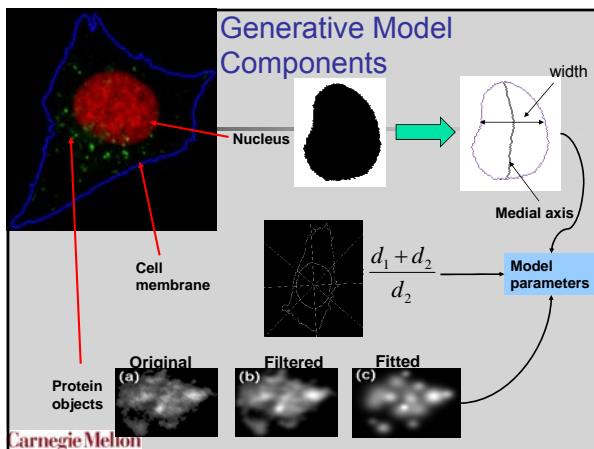
Subcellular Location Families and Generative Models

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



Carnegie Mellon

Generative Model Components



Nucleus

Cell membrane

Protein objects

Original (a)

Filtered (b)

Fitted (c)

width

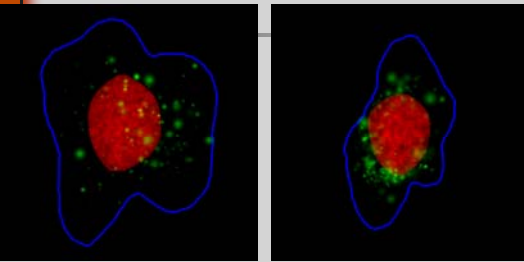
Medial axis

Model parameters

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

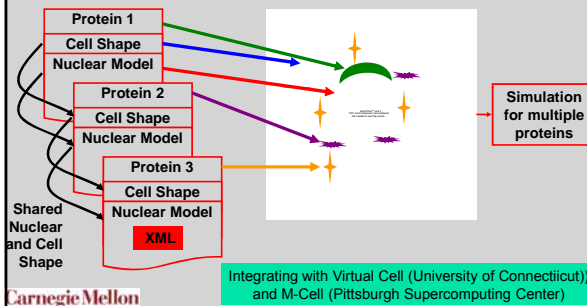
Carnegie Mellon

Synthesized Images



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

Combining Models for Cell Simulations



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**

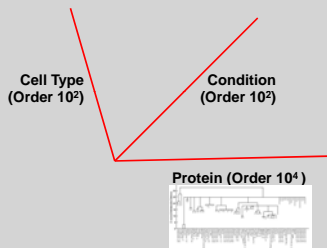
Carnegie Mellon

Examples of other subcellular location projects

- Pepperkok group (Heidelberg) - human (MCF7 cells)
 - GFP-tagged cDNAs
- Teasdale group (Brisbane) - mouse
 - GFP-tagged cDNAs
- 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

Carnegie Mellon

The future of subcellular location analysis



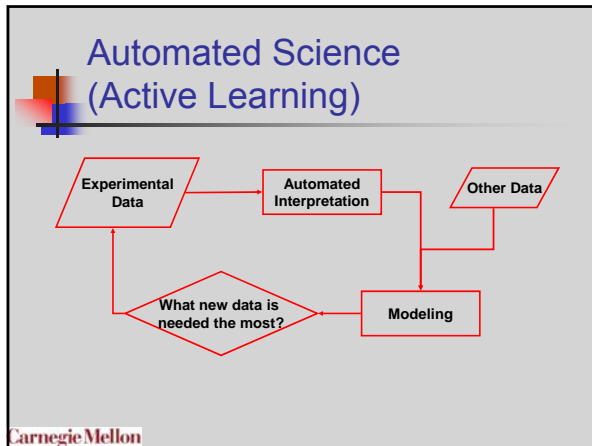
Plus: Time scale from subsecond to years

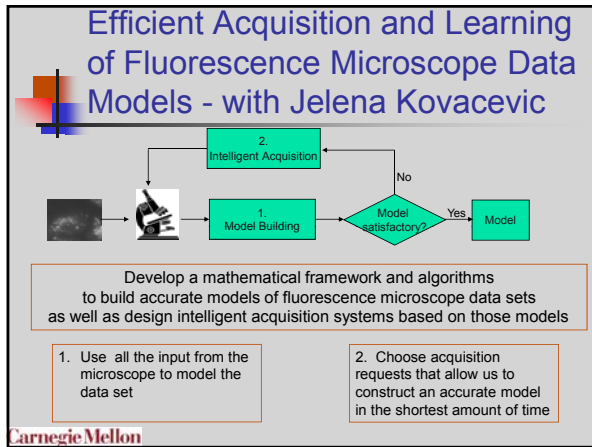
Carnegie Mellon

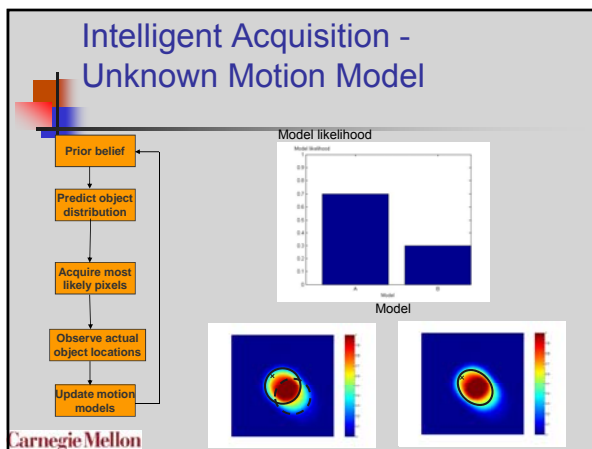
How do we really analyze subcellular location?

- 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

Carnegie Mellon

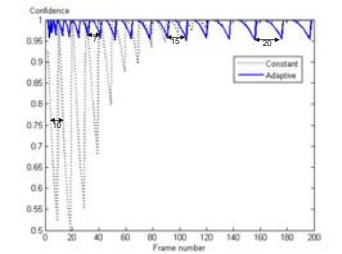






Intelligent Acquisition - Frame Rate

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



Carnegie Mellon

Intelligent Acquisition - Efficient Learning of Motion Models

- Learn the motion model, not the trajectory
- Fisher information: Amount of information that an observed random variable X contains about an unknown parameter θ

$$I(\theta) = E\left[\left(\frac{\partial}{\partial \theta} \ln f(X; \theta)\right)^2 \middle| \theta\right]$$

Observation from the acquired region The unknown parameters of the motion model

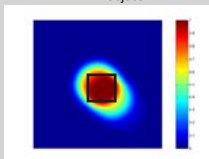
Carnegie Mellon

Intelligent Acquisition - Efficient Learning of Motion Models

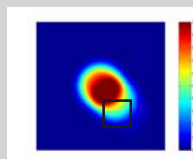
- Maximize benefit to cost ratio

$$C(D) = \int_D f(x) dx + \tau t$$

Probability of detecting object Cost per unit time Time taken to acquire a frame

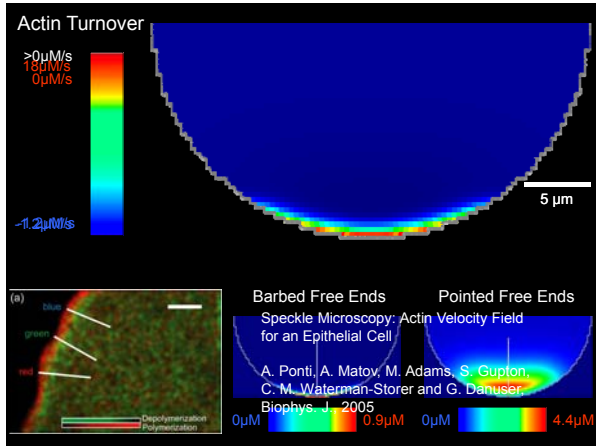


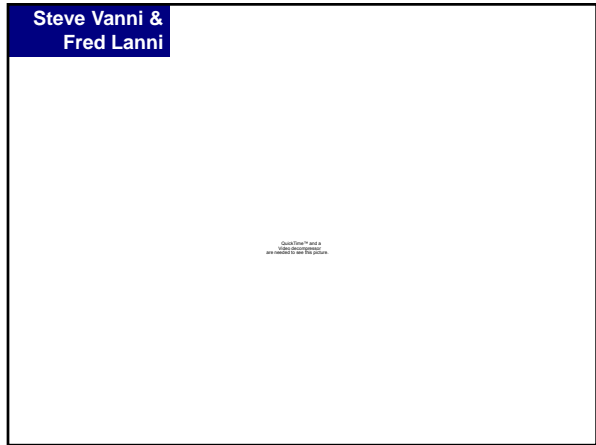
High information
High cost

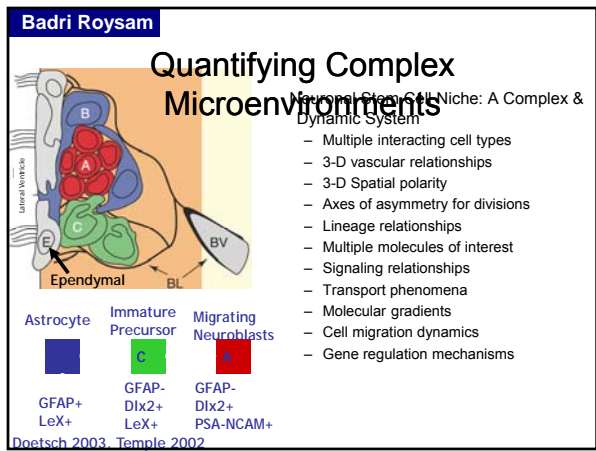


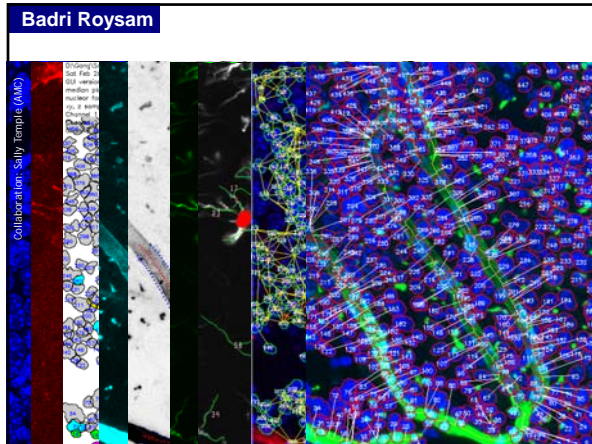
Low information
Low cost

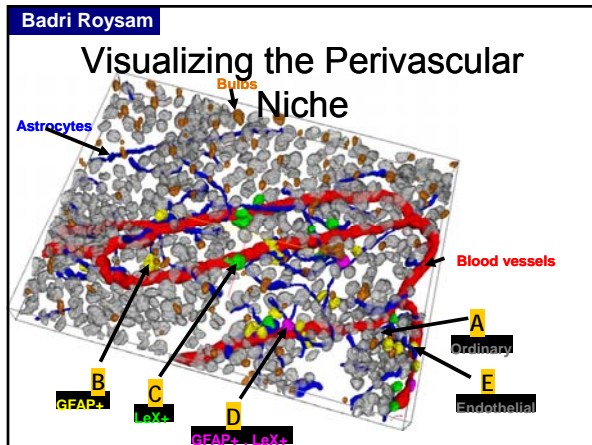
Carnegie Mellon











Acknowledgments

- Past and Present Students and Postdocs
 - Michael Boland (Hopkins), Mia Markey (UT Austin), Gregory Porreca (Harvard), Meel Velliste (U Pitt), Kai Huang, Xiang Chen (Yale), Yanhua Hu, Juchang Hua, Ting Zhao (HHMI Janelia Farm), Shann-Ching Chen (Scripps), Elvira Garcia Osuna (CMU), Justin Newberg, Estelle Glory, Tao Peng, Luis Coelho
- Funding
 - NSF, NIH, Commonwealth of Pennsylvania
- Collaborators/Consultants
 - David Casasent, Simon Watkins, Jon Jarvik, Peter Berget, Jack Rohrer, Tom Mitchell, Christos Faloutsos, Jelena Kovacevic, William Cohen, Geoff Gordon, B. S. Manjunath, Ambuj Singh, Les Loew, Ion Moraru, Jim Schaff, Paul Campagnola
- Slides/Data
 - Jelena Kovacevic, Fred Lanni, Les Loew, Badri Roysam
- 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

Recruiting postdocs and faculty!

Our mission:
To realize the potential of
machine learning for
understanding complex
biological systems
To advance cancer
diagnosis and treatment
