

Signal and Image Processing Issues in Molecular and Cellular Imaging

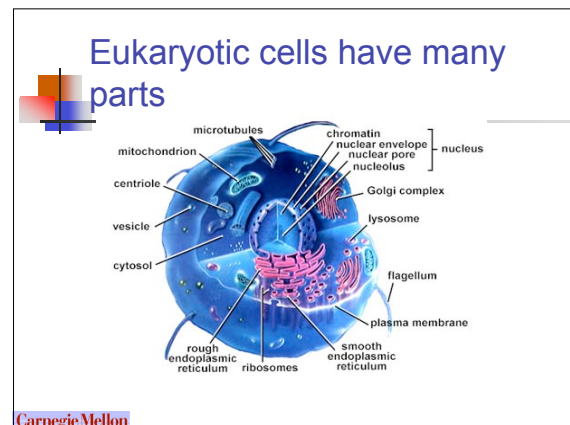
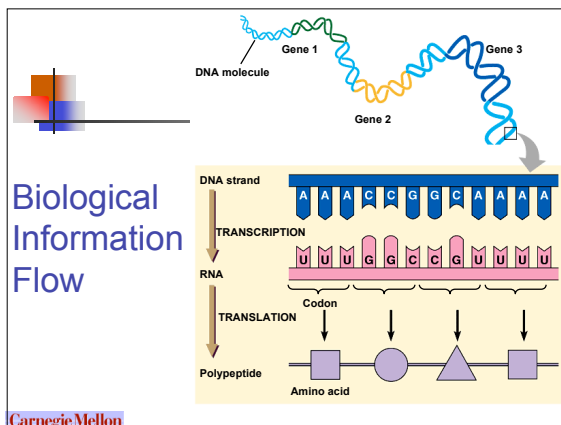
Robert F. Murphy, Depts of Biological Sciences
& Biomedical Engineering, Center for Automated
Learning and Discovery, and Center for Bioimage
Informatics

Christos Faloutsos, Dept of Computer Science
and Center for Automated Learning and Discovery

Carnegie Mellon

A. Introduction to Cell and Molecular Biology of Protein Location

Carnegie Mellon

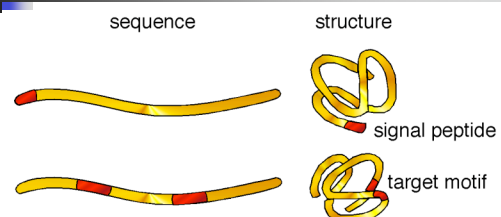


Protein localization

- The sequence of each protein determines where it is localized in cells
- Subsequences (also called "motifs") within the sequence are responsible for targeting the protein to one of the cell parts ("organelles")

Carnegie Mellon

Localization motifs: contiguous in sequence or structure?



Carnegie Mellon

Open questions

- How many distinct locations can proteins be found in? What are they?
- How many distinct motifs direct proteins to those locations? What are they?

Carnegie Mellon

The Omics Revolution: A new paradigm for biology

- The paradigm for biological research for over fifty years was the intensive study by an individual investigator (and his/her students) of a single enzyme, gene, or process, often in a single model system
 - Ion pumping mechanism of Na^+/K^+ -ATPase
 - Transcriptional regulation of *bicoid*
 - Endocytosis in intestinal cells

Carnegie Mellon

The Omics Revolution: A new paradigm for biology

- The success of genome sequencing projects suggested a new “omics” paradigm: analysis of a single property or phenomenon across an entire genome, transcriptome, proteome, etc.
 - Identification of all genes in a given genome
 - Identification of all expressed proteins in a given cell type

Carnegie Mellon

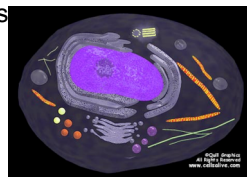
The Omics Revolution: A new paradigm for biology

- Given the large number of proteins, genes, etc. “omics” project usually require high-throughput data collection and automated data analysis
- Key to progress is
 - identification of a new aspect that needs to be analyzed “ome-wide” and
 - development of assays combined with analysis approaches

Carnegie Mellon

Proteomics

- The set of proteins expressed in a given cell type or tissue is called its *proteome*
- Proteomics projects
 - sequence
 - structure
 - activity
 - partners
 - **location**



Carnegie Mellon

Location information in protein databases: Traditional approach

- Conduct experiments of various types
 - Cell fractionation
 - Electron microscopy
 - Fluorescence microscopy
- Describe the results in unstructured text (first in journal articles and then in summaries in databases)
 - “Protein X is located primarily in protrusions from the early endosomal membrane but is also found in the plasma membrane”

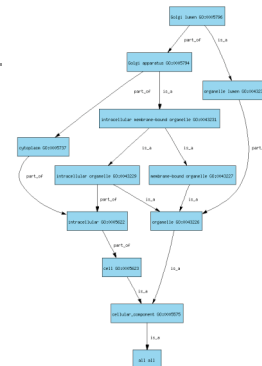
Carnegie Mellon

Location information in protein databases: Ontology approach

- Systematic analysis and comparison of these descriptions were made difficult by both the unstructured nature of the text and the variation in terminology used from one laboratory to another
- To address this problem, a **restricted** vocabulary for cellular components was created by the **Gene Ontology** consortium

Carnegie Mellon

Portion of cellular component ontology



Carnegie Mellon

Use of GO terms

- Databases such as SwissProt use manual curation to assign GO terms to proteins based on reading of relevant literature
- A major problem is consistency of application of terms

Carnegie Mellon

Example comparison of GO terms for two proteins

ID: GIAN_HUMAN STANDARD; PRT; 3259 AA.
AC: Q14789; Q14398;
GN: GOLGB1.
DR: GO; GO:0000139; C:Golgi membrane; TAS.
DR: GO; GO:0005795; C:Golgi stack; TAS.
DR: GO; GO:0016021; C:integral to membrane; TAS.

ID: 000461 PRELIMINARY; PRT; 696 AA.
AC: 000461;
GN: GPP130.
DR: GO; GO:0005810; C:endocytotic transport vesicle; TAS.
DR: GO; GO:0005801; C:Golgi cis-face; TAS.
DR: GO; GO:0005796; C:Golgi lumen; TAS.
DR: GO; GO:0016021; C:integral to membrane; TAS.

Carnegie Mellon

Words are not enough

- We learned that Giantin and GPP130 are both Golgi proteins, but do we know:
 - What part (i.e., cis, medial, trans) of the Golgi complex they each are found in?
 - If they have the same subcellular distribution?
 - If they also are found in other compartments?

Carnegie Mellon

Conclusion

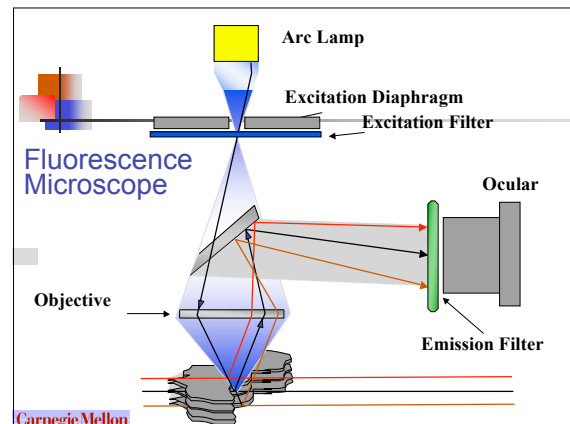
- Current knowledge of subcellular locations of proteins is not sufficiently detailed or systematic
- Systematic description of subcellular locations should be created using a **data-driven** approach rather than a **knowledge-capture** approach

Carnegie Mellon

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

Carnegie Mellon



Widefield Fluorescence Microscopy

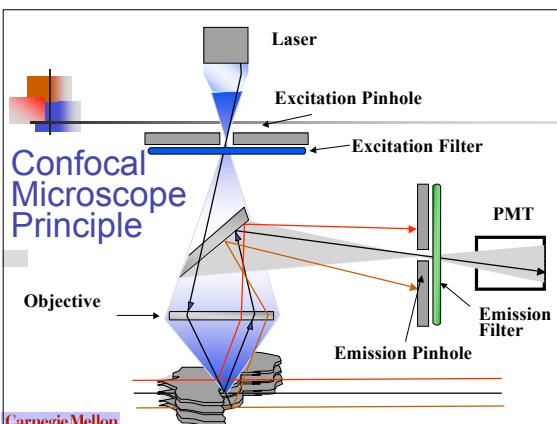
- This type of fluorescence microscope collects light emitted from all points in the specimen (with varying efficiencies depending on position relative to focal plane)
- The result for specimens that are thick relative to the depth of focus of the objective is a blurred image

A square image showing a blurred, red, textured surface, representing a widefield fluorescence image. The image is labeled "Widefield Image" and "Carnegie Mellon".

Confocal Microscopy

- One way to obtain images that better represent the fluorescence distribution just in the focal plane is to use a **confocal** microscope

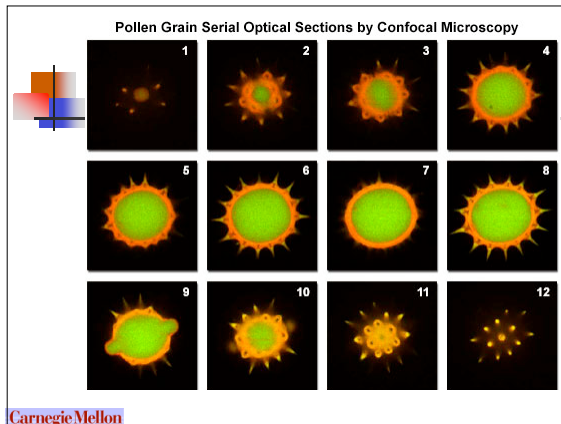
Carnegie Mellon



Two side-by-side images: a "Widefield Image" on the left and a "Confocal Image" on the right. The widefield image is blurred, while the confocal image is sharp and shows distinct cellular structures. Below the images is a control panel with various sliders and buttons. The panel includes a "Focus Loc" checkbox, a "Choose A Specimen" dropdown menu set to "Hamster Cells", a "Magnification" slider set to "60X", and several other sliders for "Focus", "Brightness", "Z-Axis Position", "PMT Red Gain", "Pinhole Aperture Size", "Scan Line Speed", and "PMT Green Gain". The panel is labeled "Carnegie Mellon".

<http://micro.magnet.fsu.edu/primer/confocal/index.html>

(c) Murphy and Faloutsos, 2005



Benefits of Confocal Microscopy

- Reduced blurring of the image from light scattering
- Increased effective resolution
- Improved signal to noise ratio
- Clear examination of thick specimens
- Z-axis scanning
- Depth perception in Z-sectioned images
- Magnification can be adjusted electronically

Carnegie Mellon

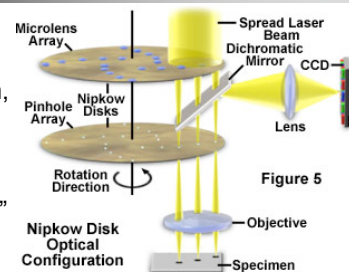
Drawbacks of Confocal Microscopy

- Slower acquisition - need to collect one pixel at a time
- Increased photodamage (photobleaching) due to longer exposure to exciting light

Carnegie Mellon

Spinning Disk Confocal Microscopy

- To allow faster acquisition, do confocal imaging "in parallel"



Carnegie Mellon

Tagging proteins for fluorescence microscopy

- To be visualized in the fluorescence microscope, proteins must be "tagged" with a fluorescent probe (fluorescent probe absorbs one wavelength of light and emits a different, higher wavelength)
- Tagging can be done with antibodies or via gene fusion to a fluorescent protein

Carnegie Mellon

Tagging proteins for fluorescence microscopy

- Immunofluorescence - use
 - "primary" antibody against the target,
 - "secondary" antibody against the "primary" and conjugated with a fluorescent probe
 - Fixed-cells only
- GFP-tagging
 - merge DNA coding for a naturally fluorescent protein with coding sequence of a protein of interest

Carnegie Mellon live-cell possible

Tagging proteins for fluorescence microscopy

- GFP-tagging
 - Can create fusion between GFP and a cDNA, in which case all regulatory sequences that control expression of the corresponding protein is lost
 - Can create fusion between GFP and the genomic sequence of a gene, in which case regulatory sequences preserved
 - Example: CD-tagging

Carnegie Mellon

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

Genomic DNA + CD-cassette
↓
Tagged DNA
↓
Tagged mRNA
↓
Tagged Protein

Carnegie Mellon

Location Proteomics

- Can use **CD-tagging** to randomly tag many proteins: Infect population of cells with a retrovirus carrying a DNA sequence that will produce a "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

Carnegie Mellon

Microscope Datasets for Subcellular Location

- We have collected four datasets of fluorescence microscope images depicting the subcellular location patterns of a number of proteins in three different cell lines
- Available at <http://murphylab.web.cmu.edu>

Carnegie Mellon

Microscope Datasets for Subcellular Location

- 2D Chinese hamster ovary cells
 - Widefield microscopy with numerical deconvolution (100x)
 - 5 different probes
- 2D HeLa
 - Widefield microscopy with numerical deconvolution (100x)
 - 9 different antibodies plus a DNA stain

Carnegie Mellon

Microscope Datasets for Subcellular Location

- 3D HeLa
 - Confocal Microscope (100x)
 - 9 different antibodies plus DNA stain and total protein stain
- 3D 3T3
 - Spinning Disk Confocal Microscope (60x)
 - GFP for a specific protein
 - Also time series

Carnegie Mellon

(c) Murphy and Faloutsos, 2005

Example Images: 2D CHO

- Single color staining for specific protein
- Three 2D slices acquired and numerically deconvolved to yield one in focus 2D slice



Carnegie Mellon

Example Images: 2D HeLa

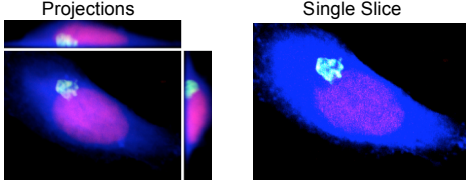


Red=DNA, Green=specific
Three 2D slices acquired & numerically deconvolved to yield one in focus 2D slice
Red and Green semi-automatically registered

Carnegie Mellon

Example Image: 3D HeLa

- Red=DNA, Blue=Total Protein, Green=specific protein
- Acquired as *stack* of 2D slices by changing focal position



Carnegie Mellon

3D HeLa

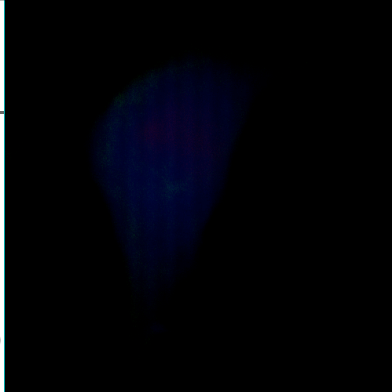
- 2D slices (from bottom to top) for cell labeled for **giantin** (primarily in Golgi)



Carnegie Mellon

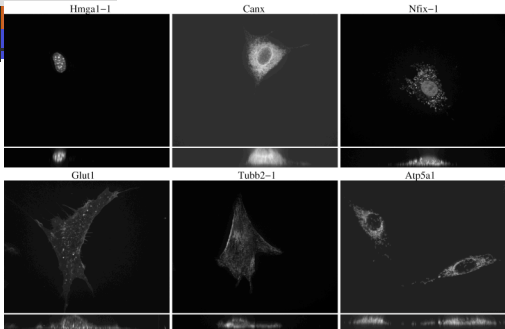
3D HeLa

- 2D slices (from bottom to top) for cell labeled for **tubulin** (major constituent of microtubules)



Carnegie Mellon

3D 3T3



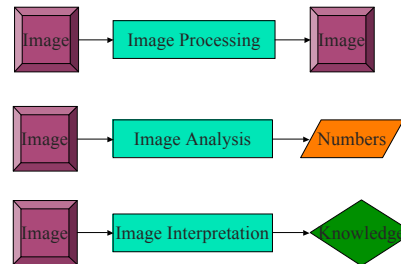
Carnegie Mellon

Automated Interpretation

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

Carnegie Mellon

From Images to Knowledge



Carnegie Mellon

Knowledge from Images

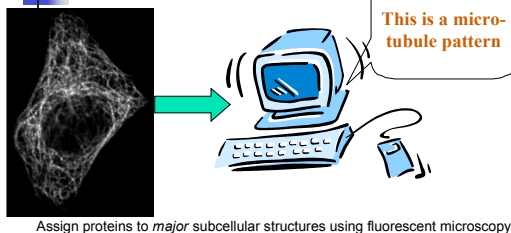
- Statements
 - Protein X shows a tubulin pattern in cell type Y
 - Pattern of Protein X is not affected by Drug Z
 - Drug Z only affects proteins with tubulin pattern
- Models
 - Model Components
 - The distribution of Protein X is best modeled using three states/compartments
 - Model Parameters
 - The diffusion constant for Protein X is C

Carnegie Mellon

B. Results from Supervised and Unsupervised Learning on Fluorescence Microscope Images

Carnegie Mellon

Initial Goal



Carnegie Mellon

The Challenge

- > Classification by direct (pixel-by-pixel) comparison of individual images to known patterns is not useful, since
 - > different cells have different **shapes, sizes, orientations**
 - > organelles within cells are **not found in fixed locations**
- > *Therefore, use feature-based approach*

Carnegie Mellon

Feature-Based, Supervised Learning Approach

1. Create sets of images showing the location of many different proteins (each set defines one **class** of pattern)
2. Reduce each image to a set of numerical values ("**features**") that are insensitive to position and rotation of the cell
3. Use statistical **classification methods** to "learn" how to distinguish each class using the features

Carnegie Mellon

Features: SLF

- Developed sets of **S**ubcellular **L**ocation **F**eatures (**SLF**) containing features of different types
- Motivated in part by descriptions used by biologists (e.g., punctate, perinuclear)
- First type of features derived from **morphological image processing** - finding objects by automated thresholding

Carnegie Mellon

2D Features Morphological Features

SLF No.	Description
SLF1.1	The number of fluorescent objects in the image
SLF1.2	The Euler number of the image
SLF1.3	The average number of above-threshold pixels per object
SLF1.4	The variance of the number of above-threshold pixels per object
SLF1.5	The ratio of the size of the largest object to the smallest
SLF1.6	The average object distance to the cellular center of fluorescence(COF)
SLF1.7	The variance of object distances from the COF
SLF1.8	The ratio of the largest to the smallest object to COF distance

Carnegie Mellon

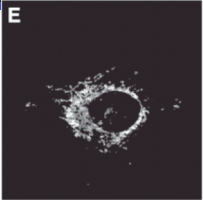
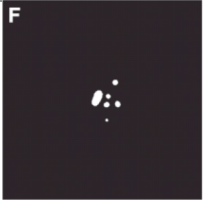
2D Features Morphological Features

Object features (DNA)

SLF No.	Description
SLF2.17	The average object distance from the COF of the DNA image
SLF2.18	The variance of object distances from the DNA COF
SLF2.19	The ratio of the largest to the smallest object to DNA COF distance
SLF2.20	The distance between the protein COF and the DNA COF
SLF2.21	The ratio of the area occupied by protein to that occupied by DNA
SLF2.22	The fraction of the protein fluorescence that co-localizes with DNA

Carnegie Mellon

2D Features Morphological Features

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

(Boland and Murphy, 2001)

Carnegie Mellon

2D Features Edge & Hull Features

Edge features

SLF No.	Description
SLF1.9	The fraction of the non-zero pixels that are along an edge
SLF1.10	Measure of edge gradient intensity homogeneity
SLF1.11	Measure of edge direction homogeneity 1
SLF1.12	Measure of edge direction homogeneity 2
SLF1.13	Measure of edge direction difference

Convex hull (geometrical) features

SLF1.14	The fraction of the convex hull area occupied by protein fluorescence
SLF1.15	The roundness of the convex hull
SLF1.16	The eccentricity of the convex hull

Carnegie Mellon

2D Features

Morphological Features

Skeleton features

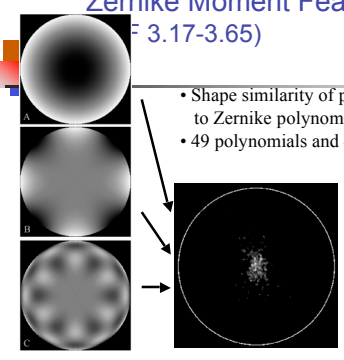
SLF No.	Description
SLF7.80	The average length of the morphological skeleton of objects
SLF7.81	The ratio of object skeleton length to the area of the convex hull of the skeleton, averaged over all objects
SLF7.82	The fraction of object pixels contained within the skeleton
SLF7.83	The fraction of object fluorescence contained within the skeleton
SLF7.84	The ratio of the number of branch points in the skeleton to the length of skeleton

Carnegie Mellon

Zernike Moment Features

(3.17-3.65)

- Shape similarity of protein image to Zernike polynomials $Z(n,l)$
- 49 polynomials and 49 features



left: Zernike polynomials
A: $Z(2,0)$
B: $Z(4,4)$
C: $Z(10,6)$

right: lamp2 image

Carnegie Mellon

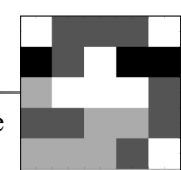
Haralick Texture Features

(SLF7.66-7.78)


- Correlations of adjacent pixels in gray level images
- Co-occurrence matrix P:
N by N matrix, N=number of gray level.
Element $P(i,j)$ is the probability of pixels with value i being adjacent with pixels with value j
- 13 statistical features


Carnegie Mellon


Co-occurrence Matrix




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









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

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

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

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

Carnegie Mellon

BREAK

Carnegie Mellon