Image Analysis of Subcellular Patterns for High Throughput Screening and 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

Goals of this section of short course

- Introduce image analysis and machine learning methods
- Illustrate in context of development of system for automated learning of subcellular patterns
- Describe utility in basic research and expectation they will incorporated into next generation of screening assays

Center for Bioimage Informatics

from image to knowledge

Carnegie Mellon

Image analysis topics

- Introduction to subcellular pattern analysis and recommendations regarding image acquisition for subsequent automated analysis
- methods for automated segmentation of multi-cell images into single cell regions
- types of features used to describe subcellular patterns and methods for extraction of these features (especially morphological, texture and wavelet features)
- statistical and machine learning methods for comparison, classification and clustering of patterns
- publicly available image database systems

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

Segmentation of Images into Single Cell Regions

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

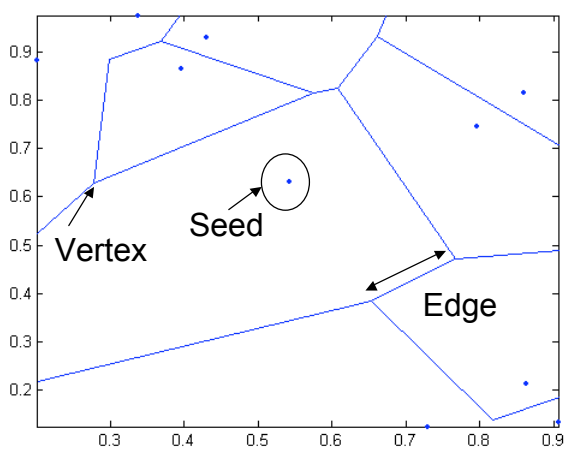
Approaches

- Voronoi
- Watershed
- Seeded Watershed
- Level Set Methods
- Graphical Models

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

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.



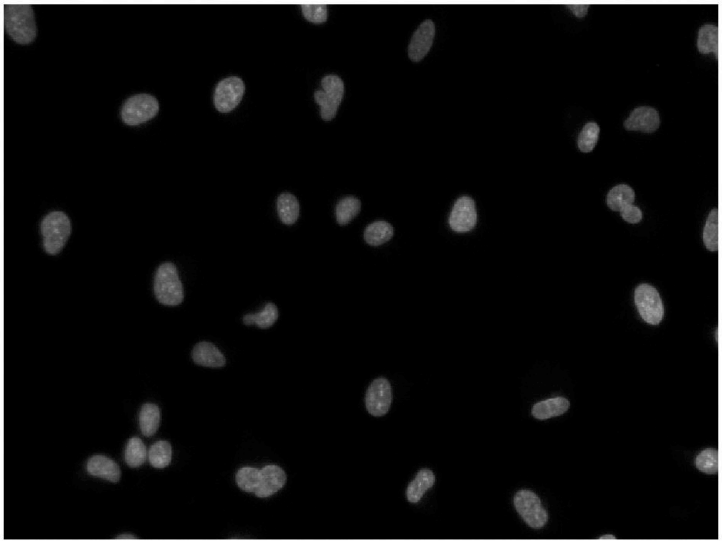
Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

Voronoi Segmentation Process

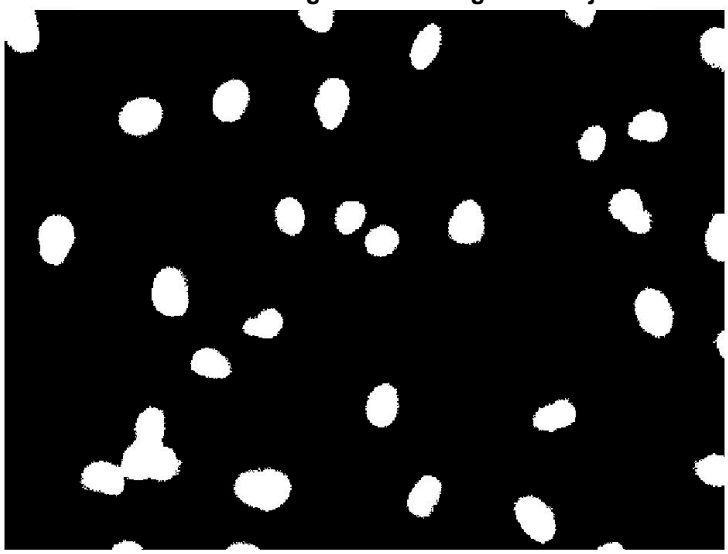
- Threshold DNA image (downsample?)
- Find the objects in the image
- Find the centers of the objects
- Use as seeds to generate Voronoi diagram
- Create a mask for each region in the Voronoi diagram
- Remove regions whose object that does not have intensity/size/shape of nucleus

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

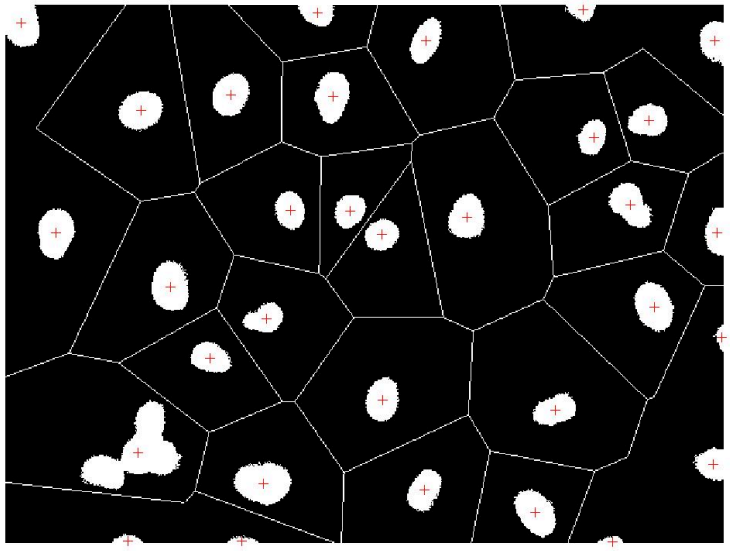
Original DNA image



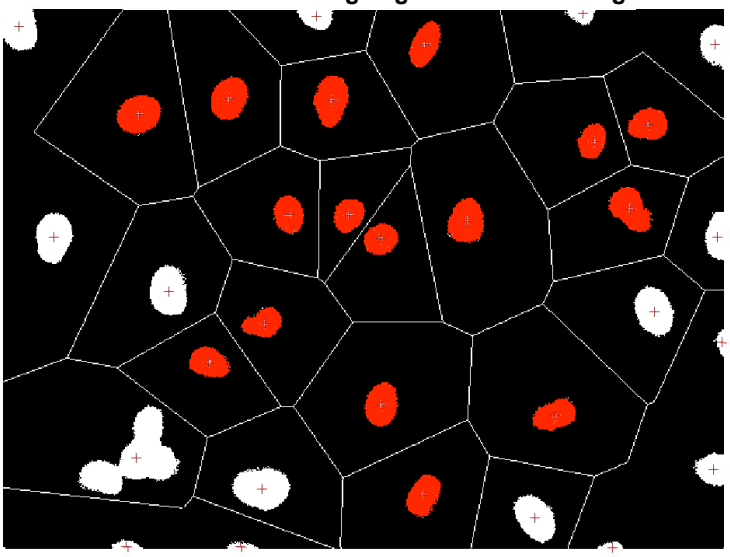
After thresholding and removing small objects



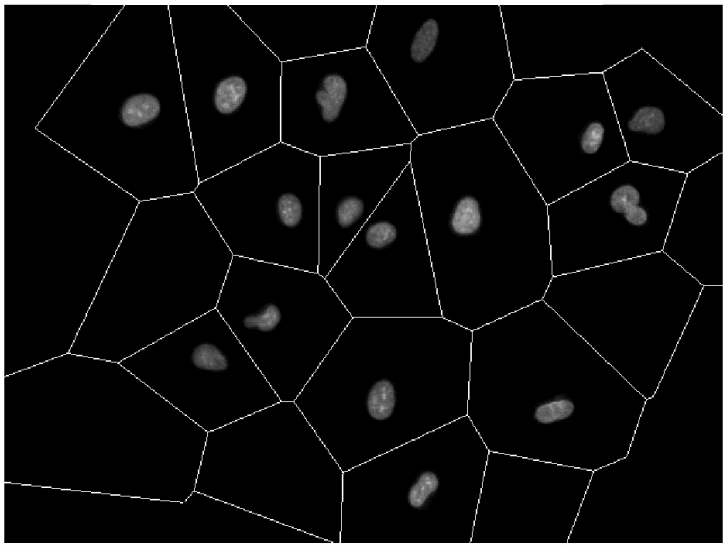
After triangulation



After removing edge cells and filtering

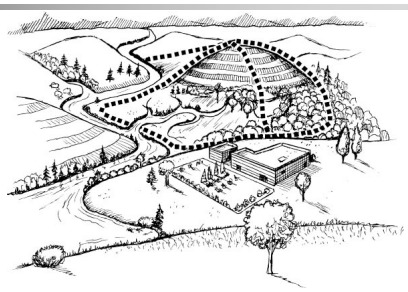


Final regions masked onto original image



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

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

Watershed Segmentation

- If starting image has intensity centered on the cells (e.g., DNA) that you want to segment, invert image so that bright objects are the sources
- If starting image has intensity centered on the boundary between the cells (e.g., plasma membrane protein), don't invert so that boundary runs along high intensity

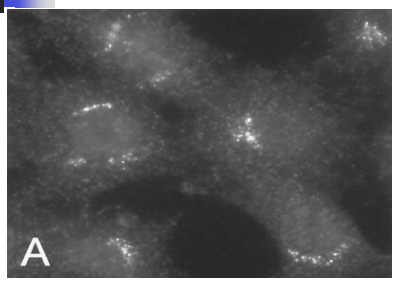
Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

Seeded Watershed Segmentation

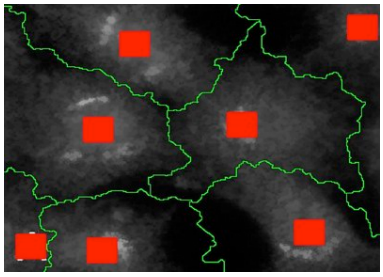
- Drawback is that the number of regions may not correspond to the number of cells
- Seeded watershed allows water to rise only from predefined sources (seeds)
- If DNA image available, can use same approach to generate these seeds as for Voronoi segmentation
- Can use seeds from DNA image but use total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

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

Seeded Watershed Segmentation



Original image



Seeds and boundary

Applied directly to protein image (no DNA image)

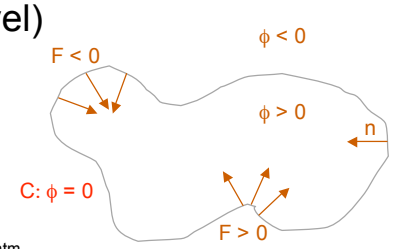
Note non-linear boundaries

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

Level Set Methods

- Level set function $\phi(x,y,t)$
  - Positive inside the contour (mountain)
  - Negative outside the contour (valley)
  - Zero on the contour, C embedded at its zero level (sea level)

http://ranger.uta.edu/~alp/personal/travelImageGallery.htm

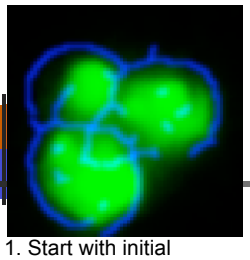


Carnegie Mellon

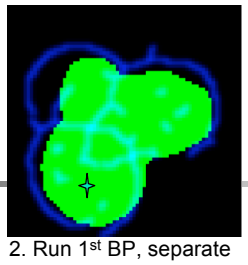
Graphical Model Methods

- Assumptions
  - Two classes of pixels: those part of a cell or part of the background
  - Each pixel is likely to be the same class as its neighbors
  - Have information about where cells are likely to be and where boundaries (edges) are likely to be
  - Probability that two pixels are same class related to probability that there is an edge between them

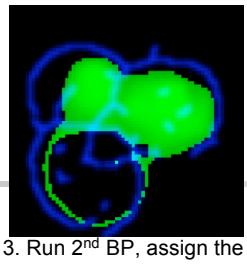
Carnegie Mellon



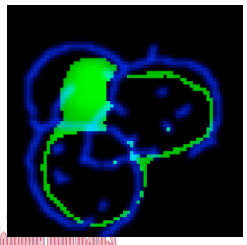
1. Start with initial DNA and edge potential



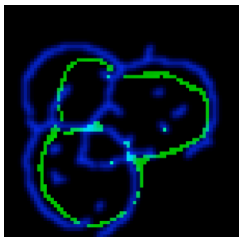
2. Run 1st BP, separate foreground and background. Pick the most confidence foreground pixel p, set its DNA potential high



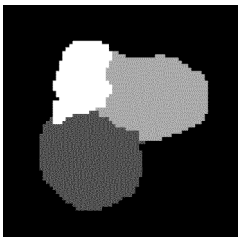
3. Run 2nd BP, assign the pixels with the same class of p to be ***segmented_cell1***, then set these pixels to be background



4. Pick the most confident foreground pixel , Run BP, find another cell, and iterate....



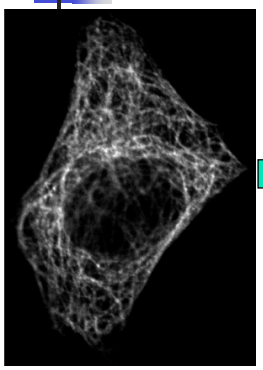
5. Iteration stops when the segmented cell is too small



6. The resulting masks

Feature Extraction for Subcellular Pattern Analysis

Goal

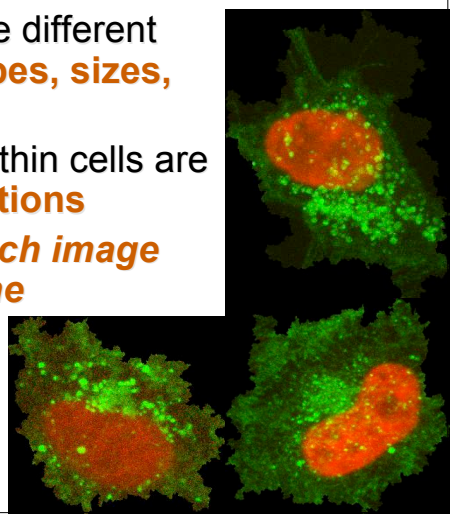


Assign proteins to *major* subcellular structures using fluorescent microscopy

Center for Bioimage Informatics
from image to knowledge
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***



Center for Bioimage Informatics
from image to knowledge
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

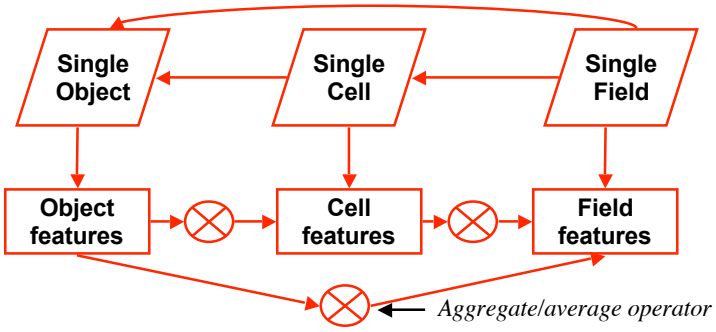
Carnegie Mellon

Subcellular Location Features (SLF)

- Combinations of features of different types that describe different aspects of patterns in fluorescence microscope images have been created
- Motivated in part by descriptions used by biologists (e.g., punctate, perinuclear)
- To ensure that the specific features used for a given experiment can be identified, they are referred to as **S**ubcellular **L**ocation **F**eatures (**SLF**) and defined in sets (e.g., SLF1)

Carnegie Mellon

Feature levels and granularity



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

Thresholding

- First type of feature is morphological
- 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)

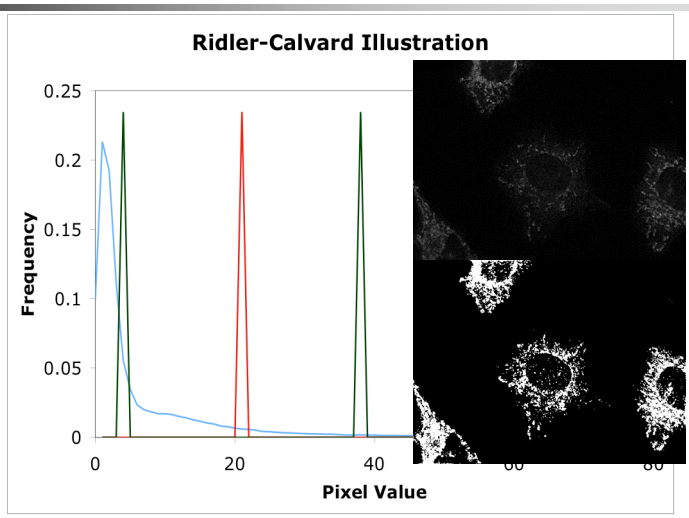
Ridler-Calvard Method

- Find threshold that is equidistant from the average intensity of pixels below and above it
- Ridler, T.W. and Calvard, S. (1978) Picture thresholding using an iterative selection method. *IEEE Transactions on Systems, Man, and Cybernetics* 8:630-632.

Carnegie Mellon

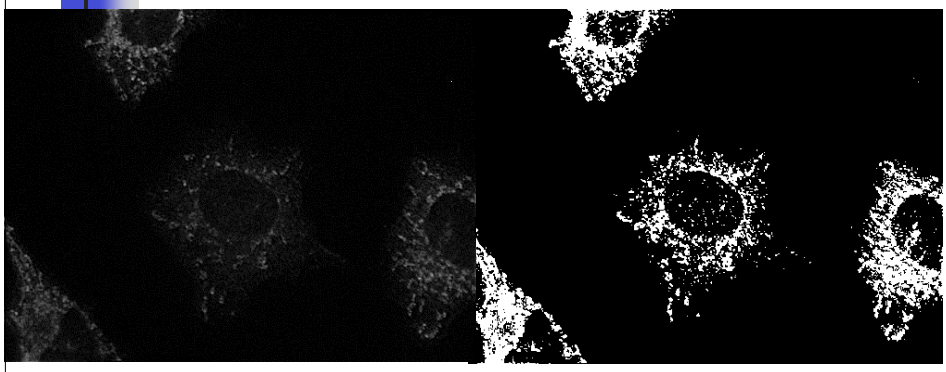
Ridler-Calvard Method

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



Carnegie Mellon

Ridler-Calvard Method



original

thresholded

Carnegie Mellon

Otsu Method

- Find threshold to minimize the variances of the pixels below and above it
- Otsu, N., (1979) A Threshold Selection Method from Gray-Level Histograms, *IEEE Transactions on Systems, Man, and Cybernetics*, 9:62-66.

Carnegie Mellon

Adaptive Thresholding

- Various approaches available
- Basic principle is use automated methods over small regions and then interpolate to form a smooth surface

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

Suitability of Automated Thresholding for Classification

- For the task of subcellular pattern analysis, automated thresholding methods perform quite well in most cases, especially for patterns with well-separated objects
- They do not work well for images with very low signal-noise ratio
- Can tolerate poor behavior on a fraction of images for a given pattern while still achieving good classification accuracies

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

Object finding

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

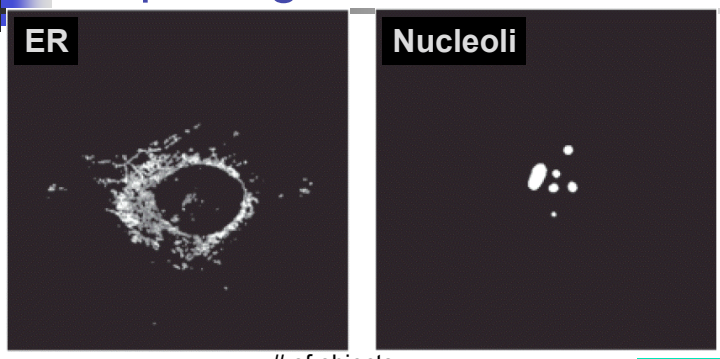
Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

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

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

2D Features
Morphological Features



ER		Nucleoli
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

Carnegie Mellon

Suitability of Morphological Features for Classification

- Images for some subcellular patterns, such as those for cytoskeletal proteins, are not well-segmented by automated thresholding
- When combined with non-morphological features, classifiers can learn to "ignore" morphological features for those classes

Carnegie Mellon

2D Features
DNA Features

DNA features (objects relative to DNA reference)

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

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

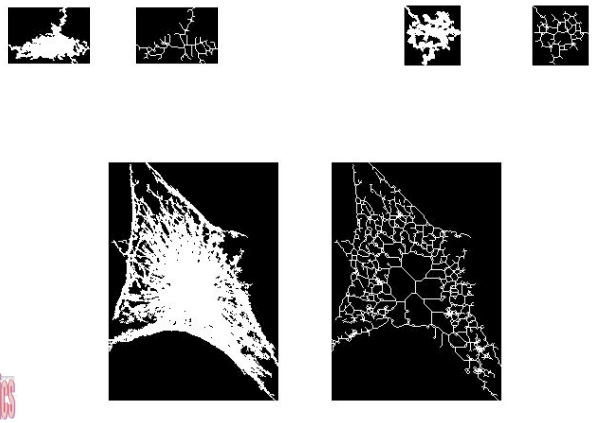
2D Features
Skeleton Features

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

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

Illustration – Skeleton



Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

2D Features
Edge 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

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

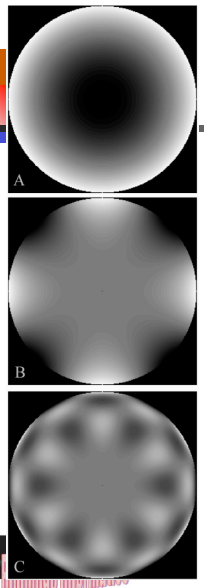
2D Features
Hull Features

Convex hull (geometrical) features

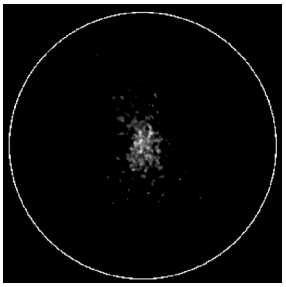
SLF1.14	The fraction of the convex hull area occupied by protein fluorescence
SLF1.15	The roundness of the convex hull
SLF1.16	The eccentricity of the convex hull

Carnegie Mellon

2D Features
Zernike Moment Features
(SLF 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

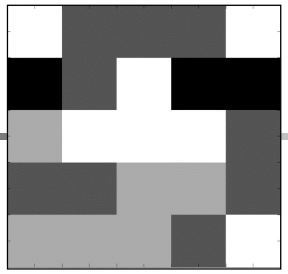
Carnegie Mellon

2D Features
Haralick Texture Features
(SLF7.66-7.78)

- Correlations of adjacent pixels in gray level images
- Start by calculating 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
- Each direction considered separately and then features averaged across all directions

Carnegie Mellon

Example image with 4 gray levels



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

Co-occurrence Matrices

	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

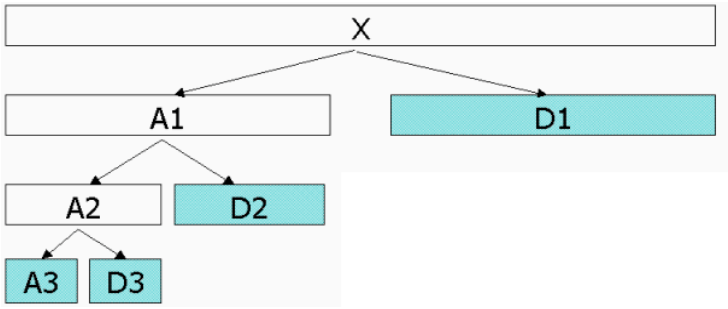
	1	2	3	4
1	0	3	0	1
2	3	0	4	4
3	0	4	0	3
4	1	4	3	2

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

Carnegie Mellon

Wavelet Transformation - 1D



A: approximation (low frequency)

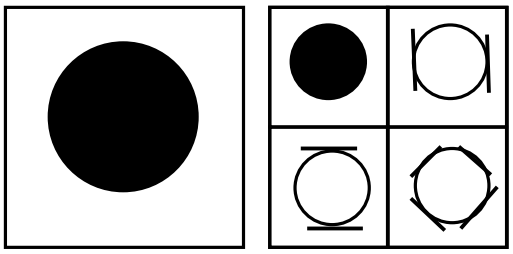
D: detail (high frequency)

X=A3+D3+D2+D1

Carnegie Mellon

2D Wavelets - intuition

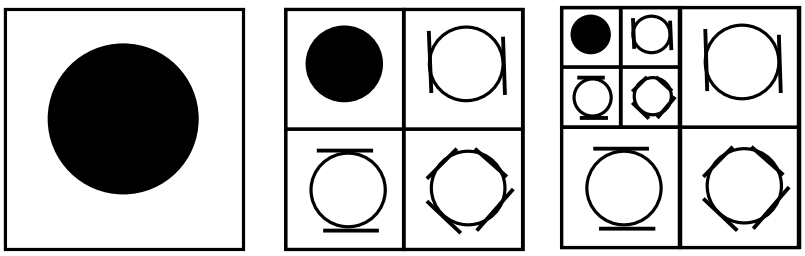
- Apply some filter to detect edges (horizontal; vertical; diagonal)



After Christos Faloutsos

2D Wavelets - intuition

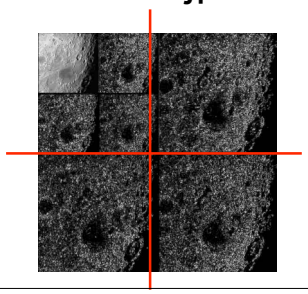
- Recurse



Slide courtesy of Christos Faloutsos

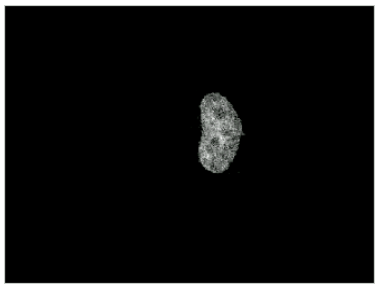
2D Wavelets - intuition

- Many wavelet basis functions (filters):
  - Haar
  - Daubechies (-4, -6, -20)
- http://www331.jpl.nasa.gov/public/wave.html

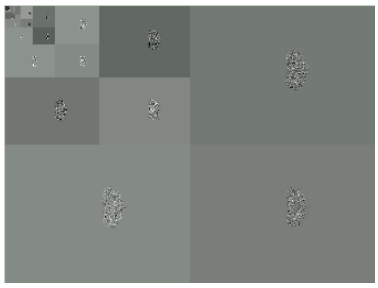


Slide courtesy of Christos Faloutsos

Daubechies D4 decomposition



Original image



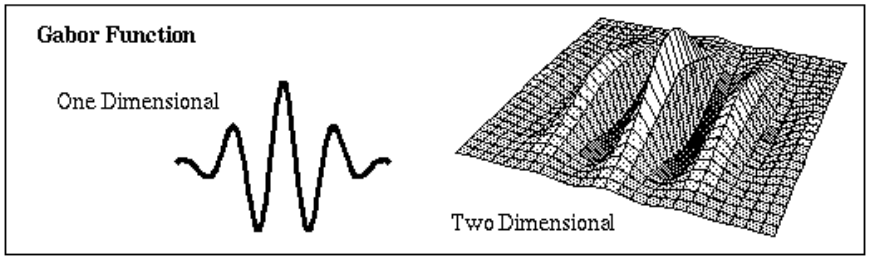
Wavelet Transformation

2D Features
Wavelet Feature Calculation

- Preprocessing
  - Background subtraction and thresholding
  - Translation and rotation
- Wavelet transformation
  - The Daubechies 4 wavelet
  - 10 level decomposition
  - Use the average energy of the three high-frequency components at each level as features

Carnegie Mellon

Gabor Function



Can extend the function to generate Gabor filters by rotating and dilating

Carnegie Mellon

2D Features
Gabor Feature Calculation

- Preprocessing same as Wavelet
- 30 Gabor filters were generated using five different scales and six different orientations
- Convolve an input image with a Gabor filter
- Take the mean and standard deviation of the convolved image
- 60 Gabor texture features

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

3D Features
Morphological (SLF-9)

- 28 features, 14 from protein objects and 14 from their relationship to corresponding DNA images
  - Based on number of objects, object size, object distance to COF
- Corresponding DNA image required

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

SLF-14

- 14 SLF-9 features that do not require DNA images
- 2 Edge features
  - Ratio of above threshold pixel along an edge
  - Ratio of fluorescence along an edge
- 26 3D Haralick texture features
  - Gray level co-occurence matrix for 13 directions
  - Calculate 13 Haralick statistics for each direction
  - Average each statistic over 13 directions and use mean and range as separate features: result is 26 features

Carnegie Mellon

SLF-17

- A feature subset with 7 features selected from SLF-14 at 256 gray levels and 0.4 micron pixel resolution
  - 1 morphological feature
  - 1 edge feature
  - 5 texture features

Carnegie Mellon

Object level features (SOF)

- Subset of SLFs calculated on single objects

Index	Feature Description
SOF1.1	Number of pixels in object
SOF1.2	Distance between object Center of Fluorescence (COF) and DNA COF
SOF1.3	Fraction of object pixels overlapping with DNA
SOF1.4	A measure of eccentricity of the object
SOF1.5	Euler number of the object
SOF1.6	A measure of roundness of the object
SOF1.7	The length of the object's skeleton
SOF1.8	The ratio of skeleton length to the area of the convex hull of the skeleton
SOF1.9	The fraction of object pixels contained within the skeleton
SOF1.10	The fraction of object fluorescence contained within the skeleton
SOF1.11	The ratio of the number of branch points in skeleton to length of skeleton

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

Field level features (SLF21)

- Subset of SLFs that do not require segmentation into single cells
  - Average object features
  - Texture features (on whole field)
  - Edge features (on whole field)

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

Basics of Machine Learning

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

Contents

- The multivariate data matrix and its descriptive statistics
- Comparison: Are two samples the same?
- Classification: Which of a set of known classes should a new sample be assigned to?
- Clustering: What classes are present in a sample?

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

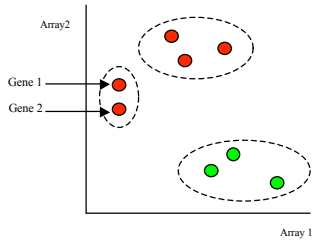
Multivariate Distance

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

Distance at the heart of Machine Learning

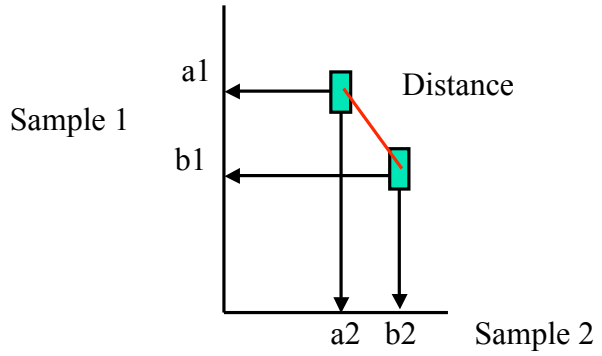
- High dimensionality
- Based on **vector geometry** – how close are two data points?

	Array 1	Array 2
Gene 1	**1**	**4**
Gene 2	**1**	**3**
…		



Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

	Sample 1	sample 2
Event a	a1	a2
Event b	b1	b2



Carnegie Mellon

General Multivariate Dataset

- We are given values of *p* variables for *n* independent observations
- Construct an *n* x *p* matrix **M** consisting of vectors $\mathbf{X}_1$ through $\mathbf{X}_n$ each of length *p*

Carnegie Mellon

Multivariate Sample Mean

- Define mean vector **I** of length *p*

$$\mathbf{I}(j) = \frac{\sum_{i=1}^{n} \mathbf{M}(i,j)}{n}$$

matrix notation

or

$$\mathbf{I} = \frac{\sum_{i=1}^{n} \mathbf{X}_i}{n}$$

vector notation

Carnegie Mellon

Multivariate Variance

- Define variance vector σ^2 of length *p*

$$\sigma^2(j) = \frac{\sum_{i=1}^{n} \left(\mathbf{M}(i,j) - \mathbf{I}(j)\right)^2}{n-1}$$

matrix notation

Carnegie Mellon

Multivariate Variance

- or

$$\sigma^2 = \frac{\sum_{i=1}^{n} \left(\mathbf{X}_i - \mathbf{I}\right)^2}{n-1}$$

vector notation

Carnegie Mellon

Covariance Matrix

- Define a *p* x *p* matrix **cov** (called the covariance matrix) analogous to σ^2

$$\mathbf{cov}(j,k) = \frac{\sum_{i=1}^{n} \left(\mathbf{M}(i,j) - \mathbf{I}(j)\right)\left(\mathbf{M}(i,k) - \mathbf{I}(k)\right)}{n-1}$$

Carnegie Mellon

Covariance Matrix

- Note that the covariance of a variable with itself is simply the variance of that variable

$$\mathbf{cov}(j,j) = \sigma^2(j)$$

Carnegie Mellon

Univariate Distance

- The simple distance between the values of a single variable *j* for two observations *i* and *l* is

$$\mathbf{M}(i,j) - \mathbf{M}(l,j)$$

Carnegie Mellon

Univariate z-score Distance

- To measure distance ***in units of standard deviation*** between the values of a single variable *j* for two observations *i* and *l* we define the z-score distance

$$\frac{\mathbf{M}(i,j)-\mathbf{M}(l,j)}{\sigma(j)}$$

Carnegie Mellon

Bivariate Euclidean Distance

- The most commonly used measure of distance between two observations *i* and *l* on two variables *j* and *k* is the Euclidean distance

$$\sqrt{\left(\mathbf{M}(i,j)-\mathbf{M}(l,j)\right)^2+\left(\mathbf{M}(i,k)-\mathbf{M}(l,k)\right)^2}$$

Carnegie Mellon

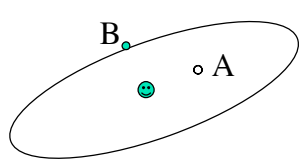
Multivariate Euclidean Distance

- This can be extended to more than two variables

$$\sqrt{\sum_{j=1}^{p}\left(\mathbf{M}(i,j)-\mathbf{M}(l,j)\right)^2}$$

Carnegie Mellon

Effects of covariance on Euclidean distance



The ellipse shows the 50% contour of a hypothetical population.

Points A and B have similar Euclidean distances from the mean, but point B is clearly "more different" from the population than point A.

Carnegie Mellon

Mahalanobis Distance

- To account for differences in variance between the variables, and to account for correlations between variables, we use the Mahalanobis distance

$$D^2 = (\mathbf{X}_i - \mathbf{X}_l)\mathbf{cov}^{-1}(\mathbf{X}_i - \mathbf{X}_l)^{\mathrm{T}}$$

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

Feature Selection and Classification

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

Human Trained Classifiers

- Traditional approach to development of screening assays is to pick one or more features to discriminate between “positive” and “negative”
- Often use hand-developed rules as part of the feature definition and/or the classification process

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

Machine Classifiers

- An alternative is to calculate a large set of features and then use machine learning methods to
  - choose important features and
  - rules to use them to discriminate positives and negatives

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

Feature selection

- Having too many features can confuse a classifier
- Can use comparison of feature distributions between classes to choose a subset of features that gets rid of uninformative or redundant features

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

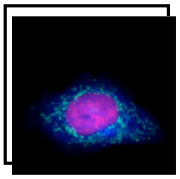
Feature Selection Methods

- Principal Components Analysis
- Non-Linear Principal Components Analysis
- Independent Components Analysis
- Information Gain
- Stepwise Discriminant Analysis
- Genetic Algorithms

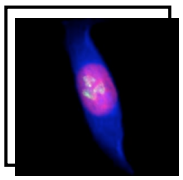
Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

Basic classification problem

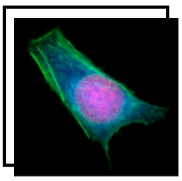
Mitoch.



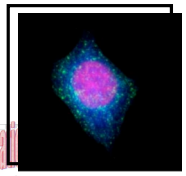
Nucleolar



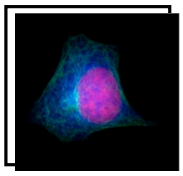
Actin



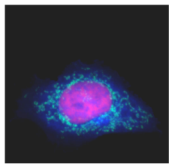
Endosomal



Tubulin



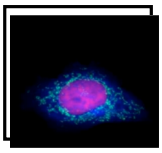
???



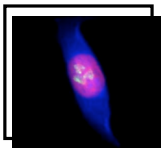
Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

Simple two class problem

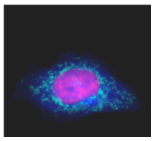
+



-



???

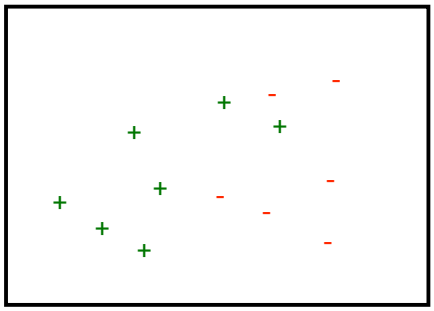


Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

Decision trees

- Pictorially, we have

num. attr#2
(e.g.., brightness)



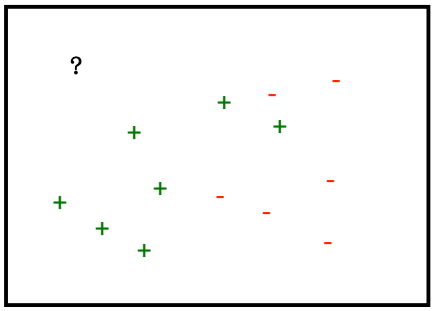
num. attr#1 (e.g.., 'area')

Slide courtesy of Christos Faloutsos

Decision trees

- and we want to label '**?**'

num. attr#2
(e.g.., brightness)

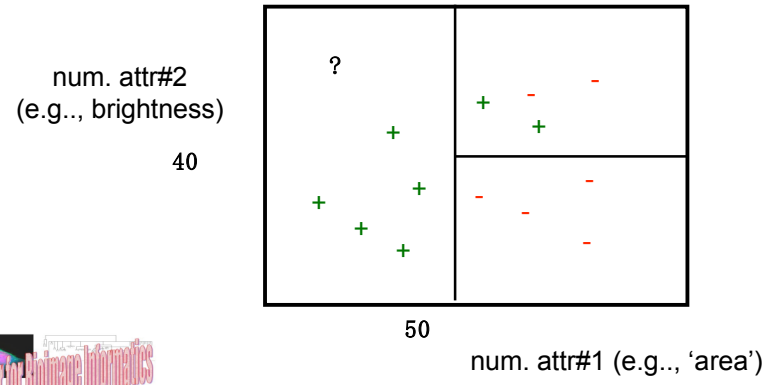


num. attr#1 (e.g.., 'area')

Slide courtesy of Christos Faloutsos

Decision trees

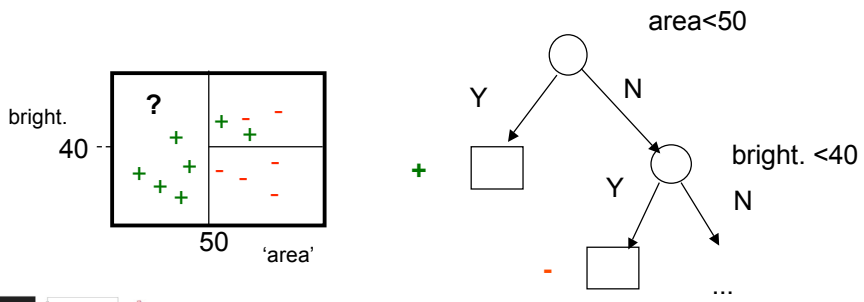
- so we build a decision tree:



Slide courtesy of Christos Faloutsos

Decision trees

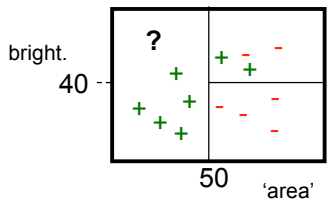
- so we build a decision tree:

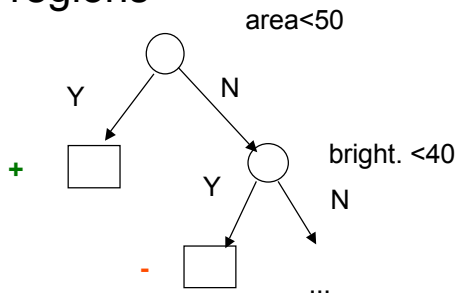


Slide courtesy of Christos Faloutsos

Decision trees

- Goal: split address space in (almost) homogeneous regions





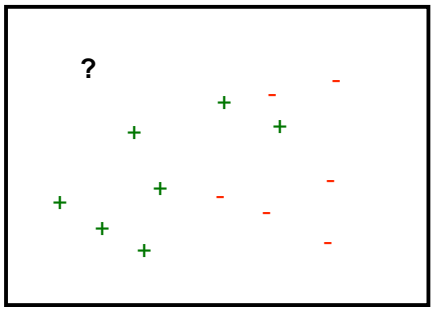
Carnegie Mellon

Slide courtesy of Christos Faloutsos

Problem: Classification

- we want to label '**?**'

num. attr#2
(e.g.., bright.)



num. attr#1 (e.g.., area)

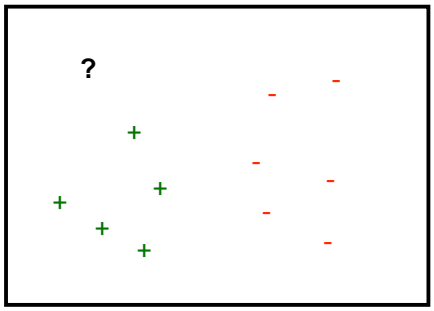
Carnegie Mellon

Slide courtesy of Christos Faloutsos

Support Vector Machines (SVMs)

- we want to label '**?**' - linear separator??

bright.



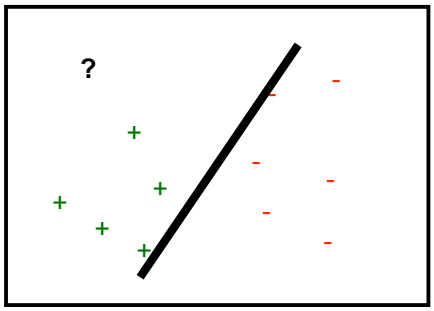
area

Slide courtesy of Christos Faloutsos

Support Vector Machines (SVMs)

- we want to label '**?**' - linear separator??

bright.



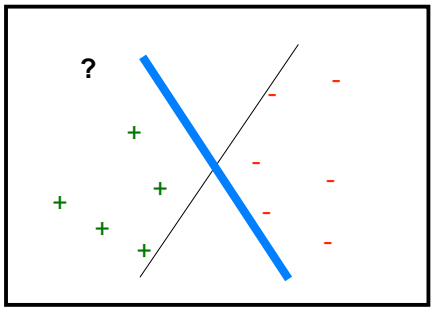
area

Slide courtesy of Christos Faloutsos

Support Vector Machines (SVMs)

- we want to label '**?**' - linear separator??

bright.



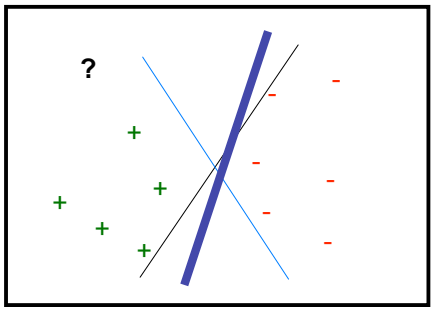
area

Slide courtesy of Christos Faloutsos

Support Vector Machines (SVMs)

- we want to label '**?**' - linear separator??

bright.



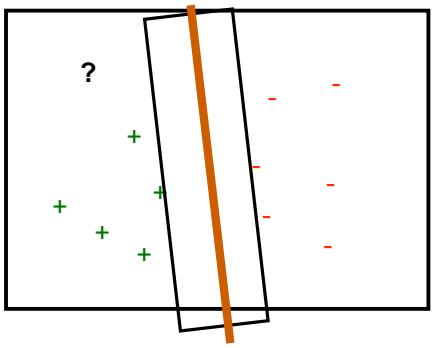
area

Slide courtesy of Christos Faloutsos

Support Vector Machines (SVMs)

- we want to label '**?**' - linear separator??

bright.



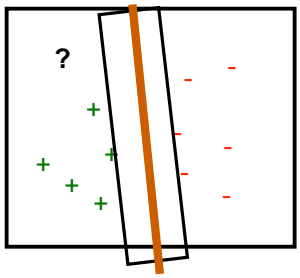
area

Slide courtesy of Christos Faloutsos

Support Vector Machines (SVMs)

- we want to label '**?**' - linear separator??
- A: the one with the widest corridor!

bright.

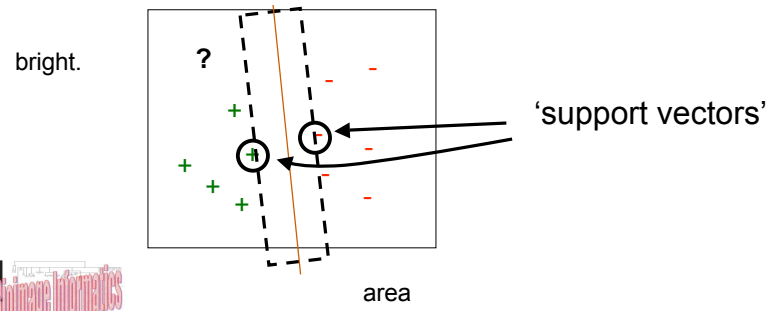


area

Slide courtesy of Christos Faloutsos

Support Vector Machines (SVMs)

- we want to label '**?**' - linear separator??
- A: the one with the widest corridor!



Slide courtesy of Christos Faloutsos

Evaluating Classifiers

- Divide ~100 images for each class into **training** set and **test** set
- Use the **training** set to determine rules for the classes
- Use the **test** set to evaluate performance
- Repeat with different division into training and test
- Evaluate different sets of features chosen as most discriminative by feature selection methods
- Evaluate different classifiers (NN, SVM, MOE)

Flexible assay design

- Same master feature set, same feature selection method, same classification engine can be used for many different assays using supervised learning instead of hand-tuning

Carnegie Mellon

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%

Carnegie Mellon

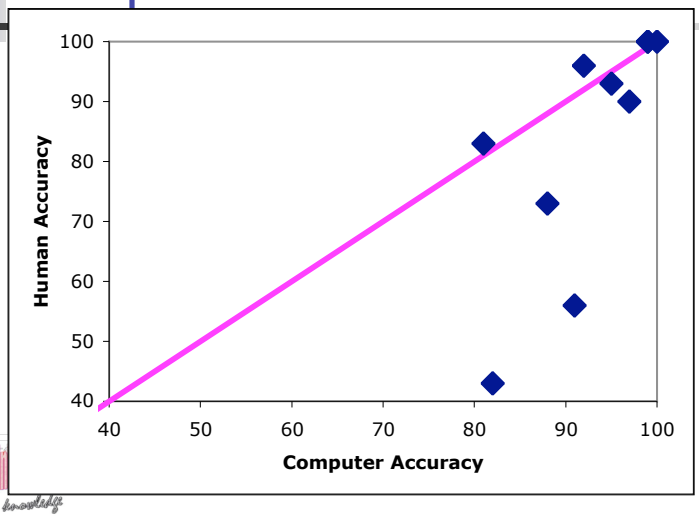
Human Classification Results

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% (92% for major patterns)

Carnegie Mellon

Computer vs. Human



Carnegie Mellon

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%

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

Clustering of Proteins by Subcellular Location

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

Unsupervised clustering algorithms

Many different types:

- Hierarchical clustering
- k – means clustering
- Self-organising maps
- Hill Climbing
- Simulated Annealing

All have the same three basic tasks of:

1. *Pattern representation* – patterns or features in the data.
2. *Pattern proximity* – a measure of the distance or similarity defined on pairs of patterns
3. *Pattern grouping* – methods and rules used in grouping the patterns

Carnegie Mellon

Hierarchical vs. *k*-means clustering

- Hierarchical builds tree sequentially from the closest pair of points (either genes or conditions)
- *k*-means starts with *k* randomly chosen seed points, assigns each remaining point to the nearest seed, and repeats this until no point moves

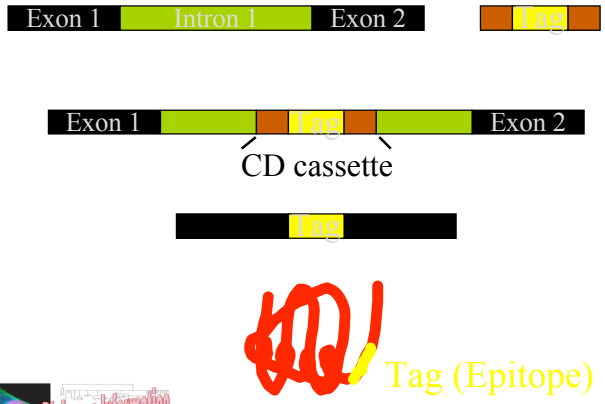
Carnegie Mellon

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

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

Principles of CD-Tagging (CD = Central Dogma)



Genomic DNA + CD-cassette

↓

Tagged DNA

↓

Tagged mRNA

↓

Tagged Protein

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

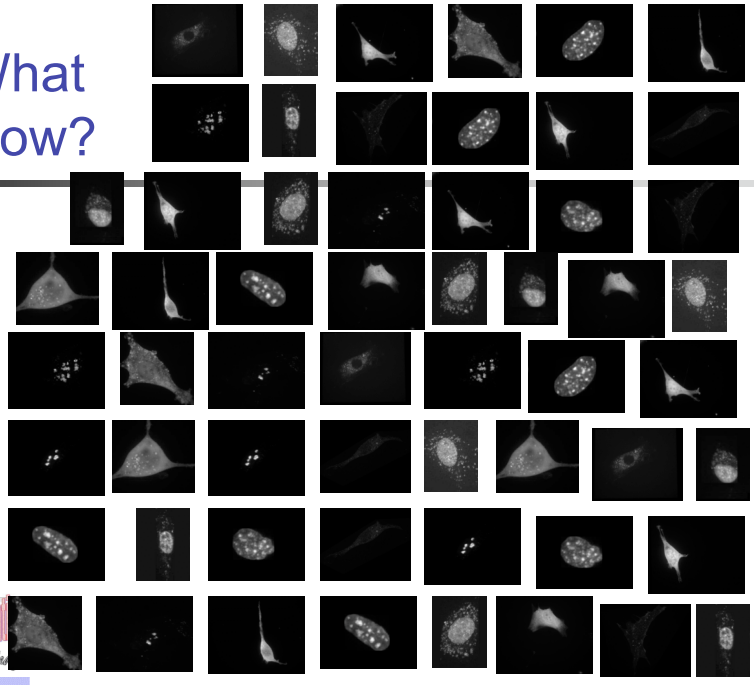
Location Proteomics

- **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

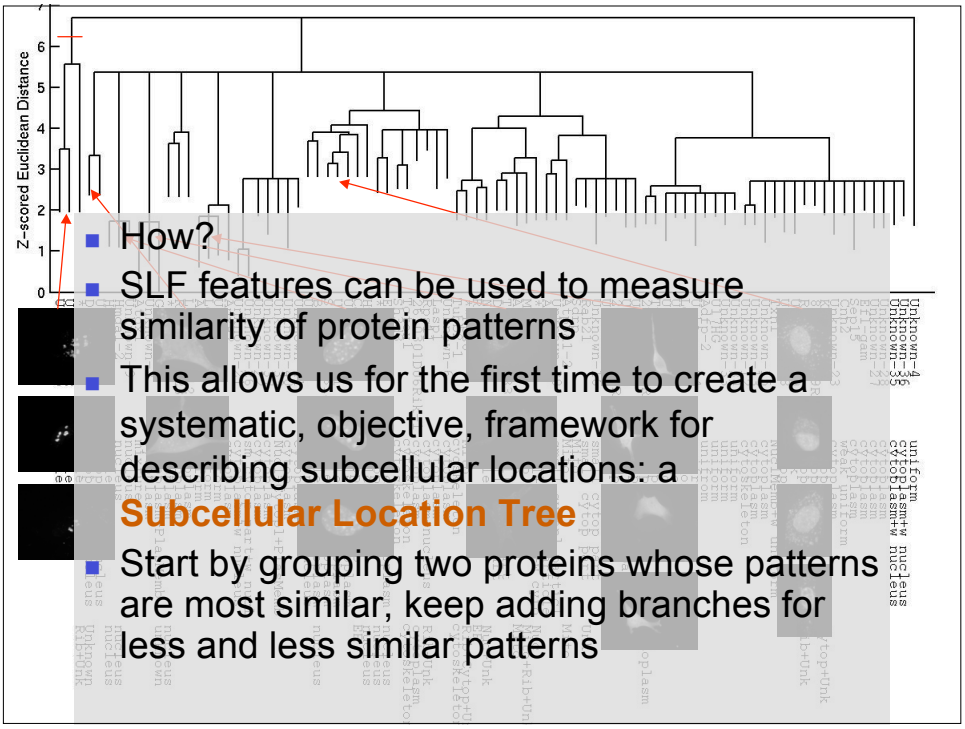
Carnegie Mellon

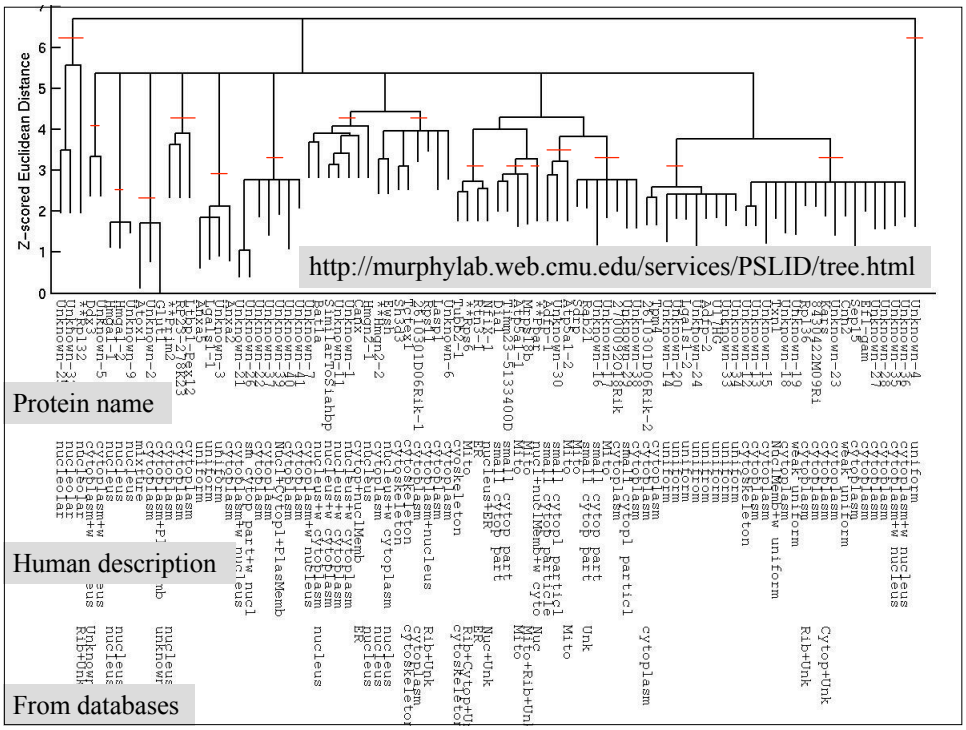
What Now?

Group ~90 tagged clones by pattern



Carnegie Mellon

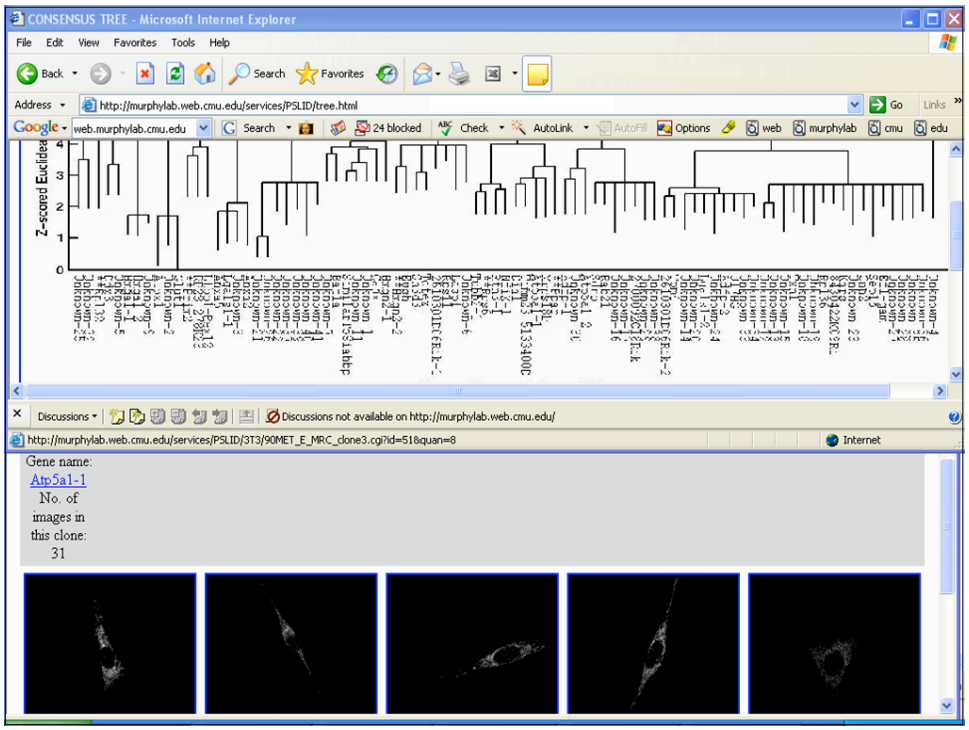

Z-scored Euclidean Distance
How?
SLF features can be used to measure similarity of protein patterns
This allows us for the first time to create a systematic, objective, framework for describing subcellular locations: a
Subcellular Location Tree
Start by grouping two proteins whose patterns are most similar, keep adding branches for less and less similar patterns

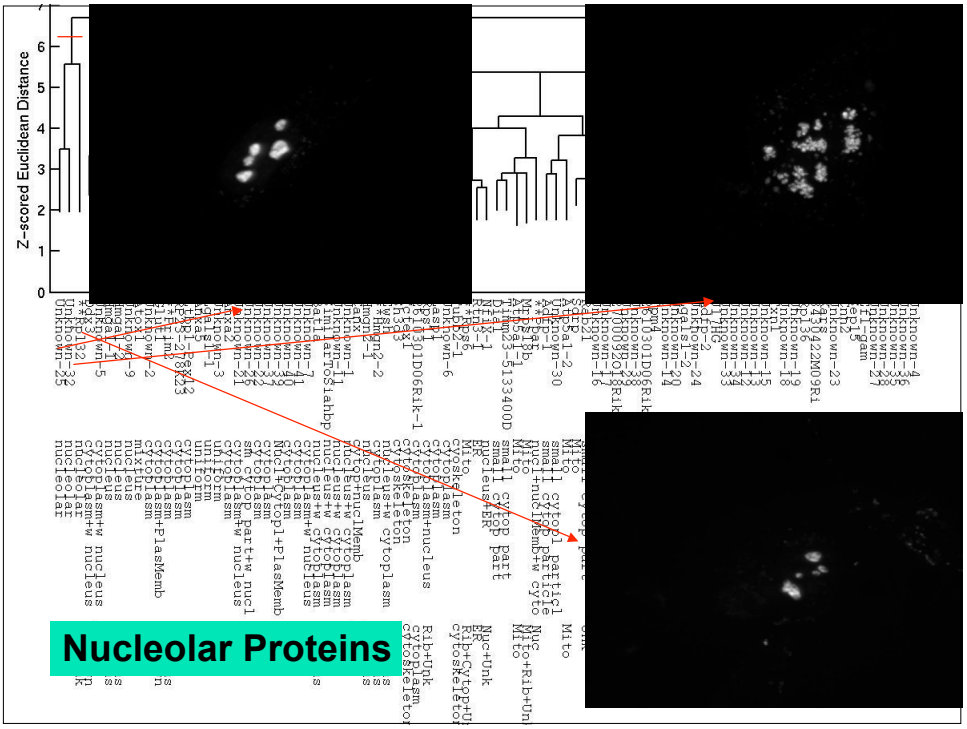

Z-scored Euclidean Distance
http://murphylab.web.cmu.edu/services/PSLID/tree.html
Protein name
Human description
From databases

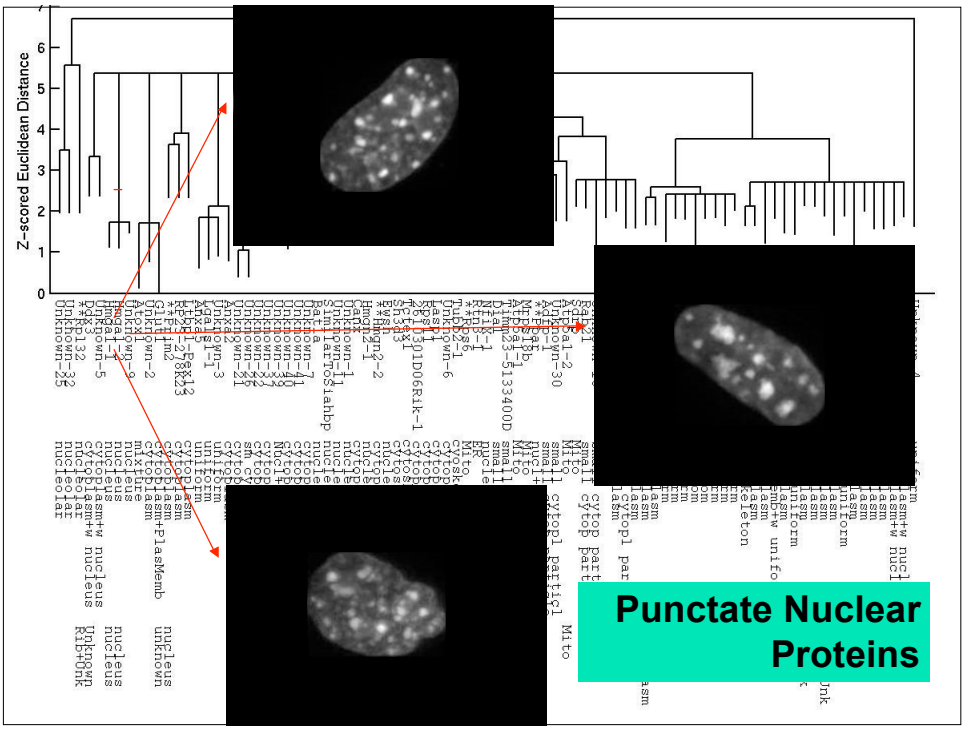
Clustering Protein Subcellular Location Patterns

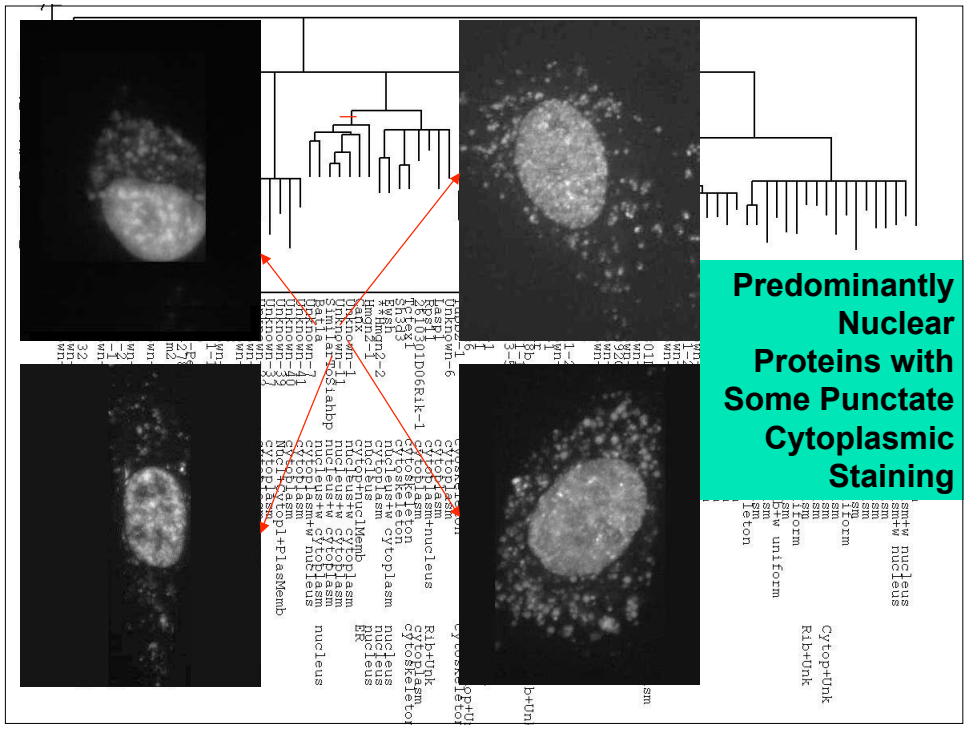
- Image acquisition
- Feature calculation
- Feature selection
- Distance selection
- Clustering/partitioning
- Evaluation

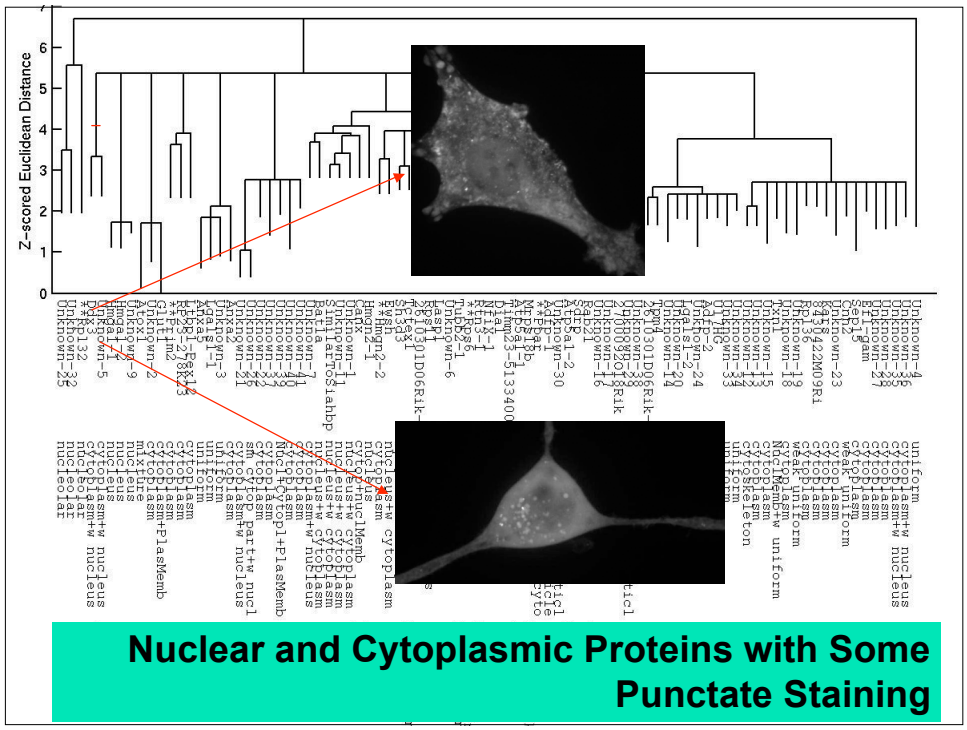
Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

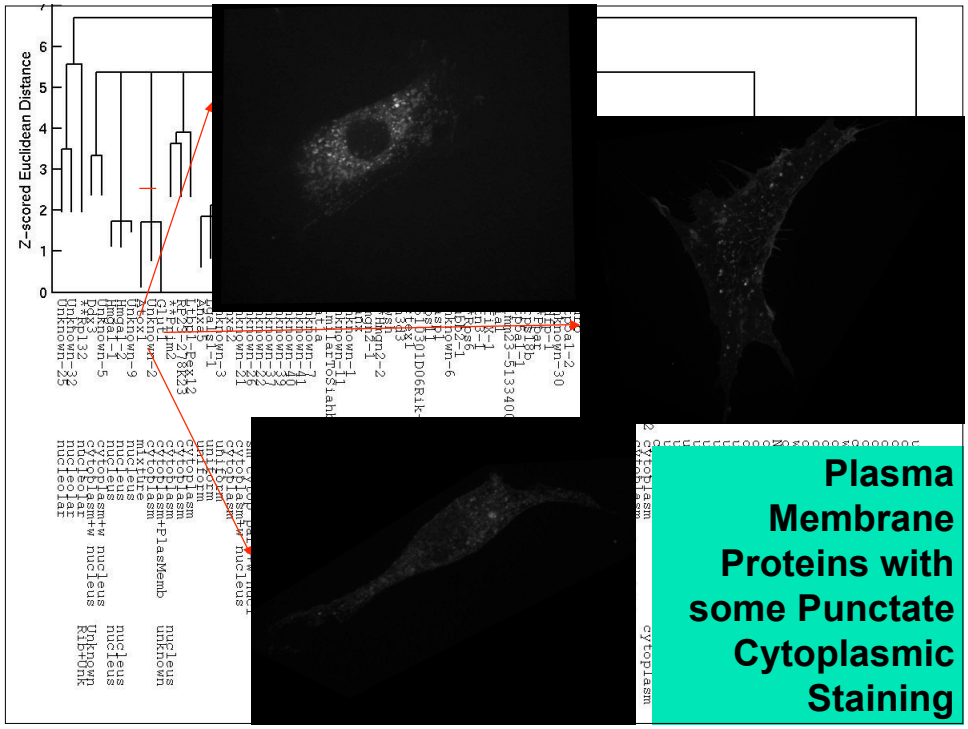


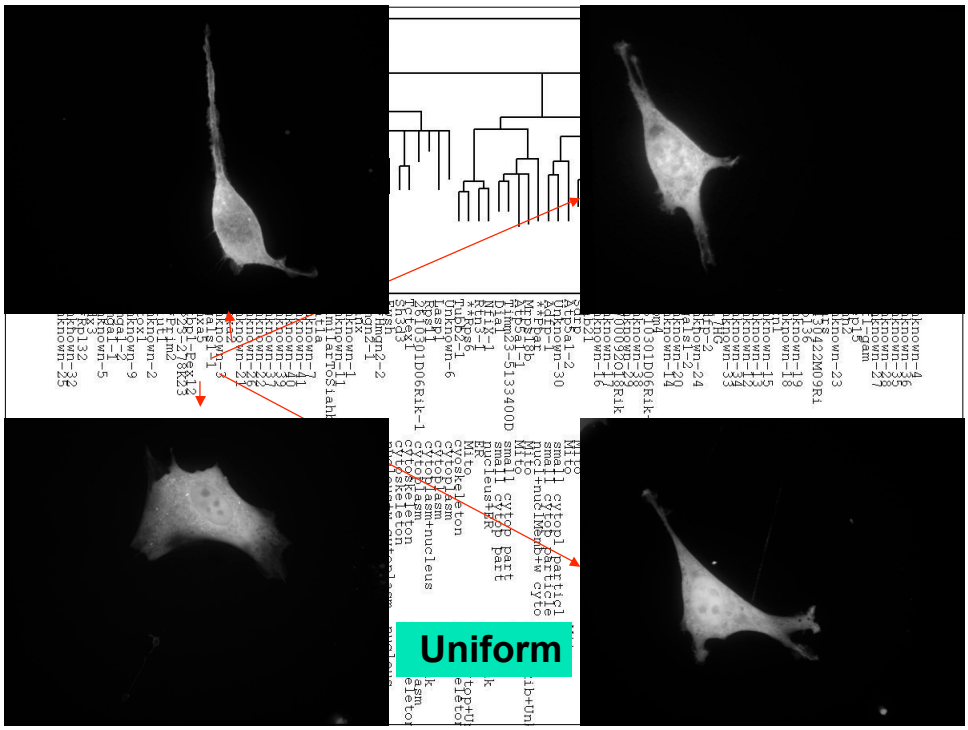

Z-scored Euclidean Distance
Nucleolar Proteins

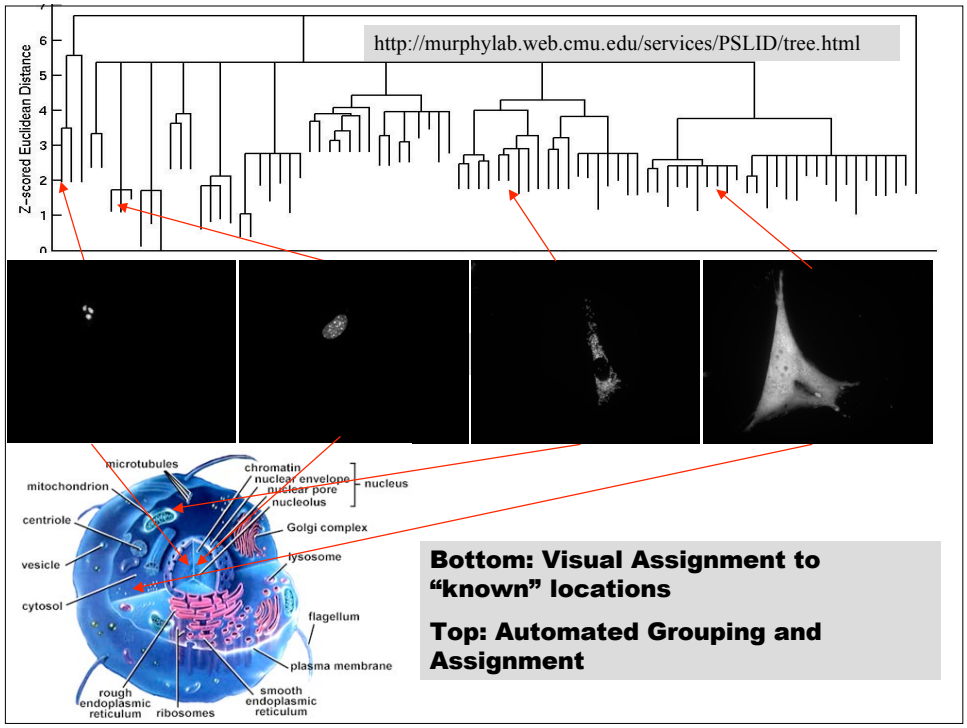

Z-scored Euclidean Distance
Punctate Nuclear Proteins


Predominantly Nuclear Proteins with Some Punctate Cytoplasmic Staining


Z-scored Euclidean Distance
Nuclear and Cytoplasmic Proteins with Some Punctate Staining


Plasma Membrane Proteins with some Punctate Cytoplasmic Staining
Z-scored Euclidean Distance


Uniform



Significance

- Can ***subdivide*** clusters by observing response to drugs, oncogenes, etc.
- These represent protein location ***states***
- Base knowledge required for ***modeling*** (systems biology)
- Can be used to identify potential protein ***interactions***

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

References on Automated Interpretation of Subcellular Patterns

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

http://murphylab.web.cmu.edu/publications

Review Articles

- Y. Hu and R. F. Murphy (2004). Automated Interpretation of Subcellular Patterns from Immunofluorescence Microscopy. *J. Immunol. Methods 290*:93-105.
- K. Huang and R. F. Murphy (2004). From Quantitative Microscopy to Automated Image Understanding. *J. Biomed. Optics 9:*893-912.
- R.F. Murphy (2005). Location Proteomics: A Systems Approach to Subcellular Location. *Biochem. Soc. Trans. 33:*535-538.
- R.F. Murphy (2005). Cytomics and Location Proteomics: Automated Interpretation of Subcellular Patterns in Fluorescence Microscope Images. *Cytometry 67A:1-3*.
- X. Chen, and R.F. Murphy (2006). Automated Interpretation of Protein Subcellular Location Patterns. *International Review of Cytology 249*:194-227.
- X. Chen, M. Velliste, and R.F. Murphy (2006). Automated Interpretation of Subcellular Patterns in Fluorescence Microscope Images for Location Proteomics. *Cytometry 69A:*631-640,

First published system for recognizing subcellular location patterns - 2D CHO (5 patterns)

- M. V. Boland, M. K. Markey and R. F. Murphy (1997). Automated Classification of Cellular Protein Localization Patterns Obtained via Fluorescence Microscopy. *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 594-597.
- M. V. Boland, M. K. Markey and R. F. Murphy (1998). Automated Recognition of Patterns Characteristic of Subcellular Structures in Fluorescence Microscopy Images. *Cytometry 33*:366-375.

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

http://murphylab.web.cmu.edu/publications

http://murphylab.web.cmu.edu/publications

2D HeLa pattern classification (10 major patterns)

- R. F. Murphy, M. V. Boland and M. Velliste (2000). Towards a Systematics for Protein Subcellular Location: Quantitative Description of Protein Localization Patterns and Automated Analysis of Fluorescence Microscope Images. *Proc Int Conf Intell Syst Mol Biol 8*:251-259.
- M. V. Boland and R. F. Murphy (2001). A Neural Network Classifier Capable of Recognizing the Patterns of all Major Subcellular Structures in Fluorescence Microscope Images of HeLa Cells. *Bioinformatics 17:*1213-1223.

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

http://murphylab.web.cmu.edu/publications

3D HeLa pattern classification (11 major patterns)

- M. Velliste and R.F. Murphy (2002). Automated Determination of Protein Subcellular Locations from 3D Fluorescence Microscope Images. *Proceedings of the 2002 IEEE International Symposium on Biomedical Imaging (ISBI 2002)*, pp. 867-870.

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

http://murphylab.web.cmu.edu/publications

Improving features, feature selection, classification method

- R.F. Murphy, M. Velliste, and G. Porreca (2003). Robust Numerical Features for Description and Classification of Subcellular Location Patterns in Fluorescence Microscope Images. *J. VLSI Sig. Proc. 35:* 311-321.
- K. Huang, M. Velliste, and R. F. Murphy (2003). Feature reduction for improved recognition of subcellular location patterns in fluorescence microscope images. *Proc. SPIE 4962:* 307-318.

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

http://murphylab.web.cmu.edu/publications

Improving features, feature selection, classification method

- K. Huang and R.F. Murphy (2004). Boosting accuracy of automated classification of fluorescence microscope images for location proteomics. *BMC Bioinformatics 5:*78.
- X. Chen and R.F. Murphy (2004). Robust Classification of Subcellular Location Patterns in High Resolution 3D Fluorescence Microscope Images. *Proceedings of the 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. *1632-1635*.

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

http://murphylab.web.cmu.edu/publications

Classification of multi-cell images

- K. Huang and R. F. Murphy (2004). Automated Classification of Subcellular Patterns in Multicell images without Segmentation into Single Cells. *Proceedings of the 2004 IEEE International Symposium on Biomedical Imaging (ISBI 2004),* pp. 1139-1142.
- S.-C. Chen, and R.F. Murphy (2006). A Graphical Model Approach to Automated Classification of Protein Subcellular Location Patterns in Multi-Cell Images. *BMC Bioinformatics* 7:90.
- S.-C. Chen, G. Gordon, and R.F. Murphy (2006). A Novel Approximate Inference Approach to Automated Classification of Protein Subcellular Location Patterns in Multi-Cell Images. *Proceedings of the 2006 IEEE International Symposium on Biomedical Imaging (ISBI 2006),* pp. 558-561.

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

http://murphylab.web.cmu.edu/publications

Temporal Texture Features

- Y. Hu, J. Carmona, and R.F. Murphy (2006). Application of Temporal Texture Features to Automated Analysis of Protein Subcellular Locations in Time Series Fluorescence Microscope Images. *Proceedings of the 2006 IEEE International Symposium on Biomedical Imaging (ISBI 2006)*, pp. 1028-1031.

Carnegie Mellon

http://murphylab.web.cmu.edu/publications

Temporal Texture Features

- Y. Hu, J. Carmona, and R.F. Murphy (2006). Application of Temporal Texture Features to Automated Analysis of Protein Subcellular Locations in Time Series Fluorescence Microscope Images. *Proceedings of the 2006 IEEE International Symposium on Biomedical Imaging (ISBI 2006)*, pp. 1028-1031.

Carnegie Mellon

http://murphylab.web.cmu.edu/publications

Subcellular Location Trees - 3D 3T3 CD-tagged images

- X. Chen, M. Velliste, S. Weinstein, J.W. Jarvik and R.F. Murphy (2003). Location proteomics - Building subcellular location trees from high resolution 3D fluorescence microscope images of randomly-tagged proteins. *Proc. SPIE 4962:* 298-306.
- X. Chen and R. F. Murphy (2005). Objective Clustering of Proteins Based on Subcellular Location Patterns. *Journal of Biomedicine and Biotechnology 2005*: 87-95.

Carnegie Mellon

http://murphylab.web.cmu.edu/publications

Subcellular Location Trees - Analysis of Location Mutants

- P. Nair, B.E. Schaub, K. Huang, X. Chen, R.F. Murphy, J.M. Griffith, H.J. Geuze, and J. Rohrer (2005). Characterization of the TGN Exit Signal of the human Mannose 6-Phosphate Uncovering Enzyme. *J. Cell Sci. 118:*2949-2956.

Carnegie Mellon

http://murphylab.web.cmu.edu/publications

PSLID - Protein Subcellular Location Image Database

- K. Huang, J. Lin, J.A. Gajnak, and R.F. Murphy (2002). Image Content-based Retrieval and Automated Interpretation of Fluorescence Microscope Images via the Protein Subcellular Location Image Database. *Proceedings of the 2002 IEEE International Symposium on Biomedical Imaging (ISBI 2002)*, pp. 325-328.

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

http://murphylab.web.cmu.edu/publications

SLIF - Subcellular Location Image Finder

- R. F. Murphy, M. Velliste, J. Yao, and G. Porreca (2001). Searching Online Journals for Fluorescence Microscope Images Depicting Protein Subcellular Location Patterns. *Proceedings of the 2nd IEEE International Symposium on Bio-Informatics and Biomedical Engineering (BIBE 2001)*, pp. 119-128.
- R. F. Murphy, Z. Kou, J. Hua, M. Joffe, and W. W. Cohen (2004). Extracting and Structuring Subcellular Location Information from On-line Journal Articles: The Subcellular Location Image Finder. *Proceedings of the IASTED International Conference on Knowledge Sharing and Collaborative Engineering (KSCE 2004)*, pp. 109-114.

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon