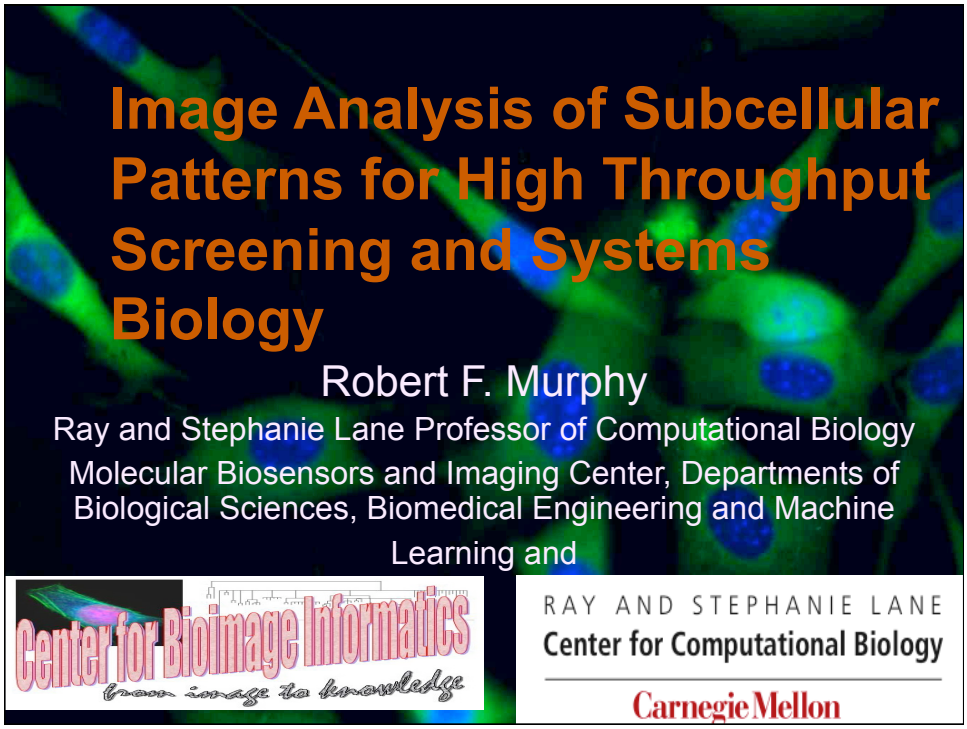


Goals of this section

- 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

Carnegie Mellon

Segmentation of Images into Single Cell Regions

Carnegie Mellon

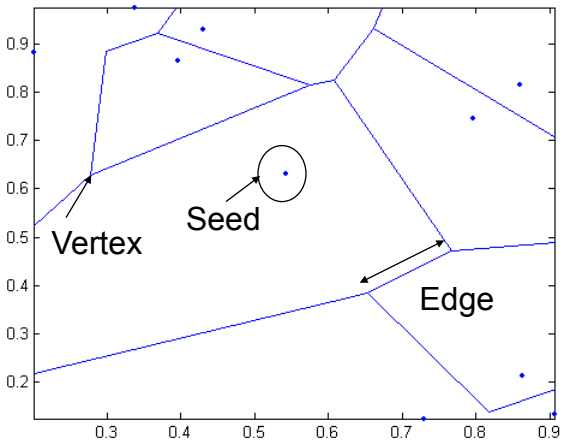
Approaches

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

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.



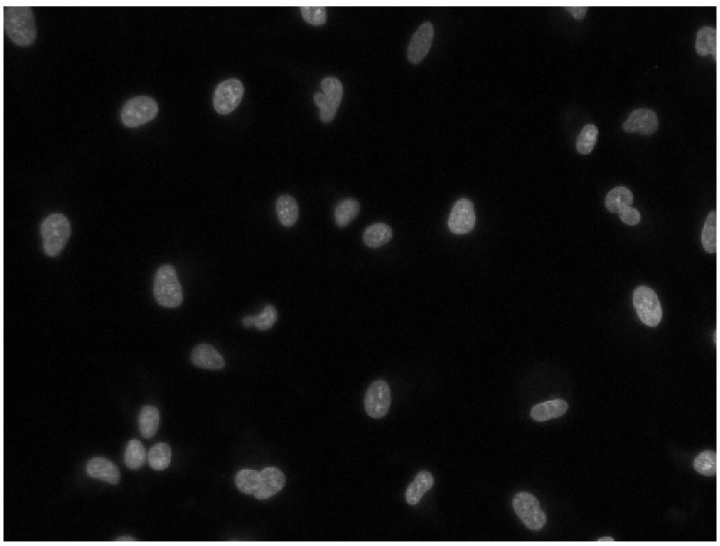
Carnegie Mellon

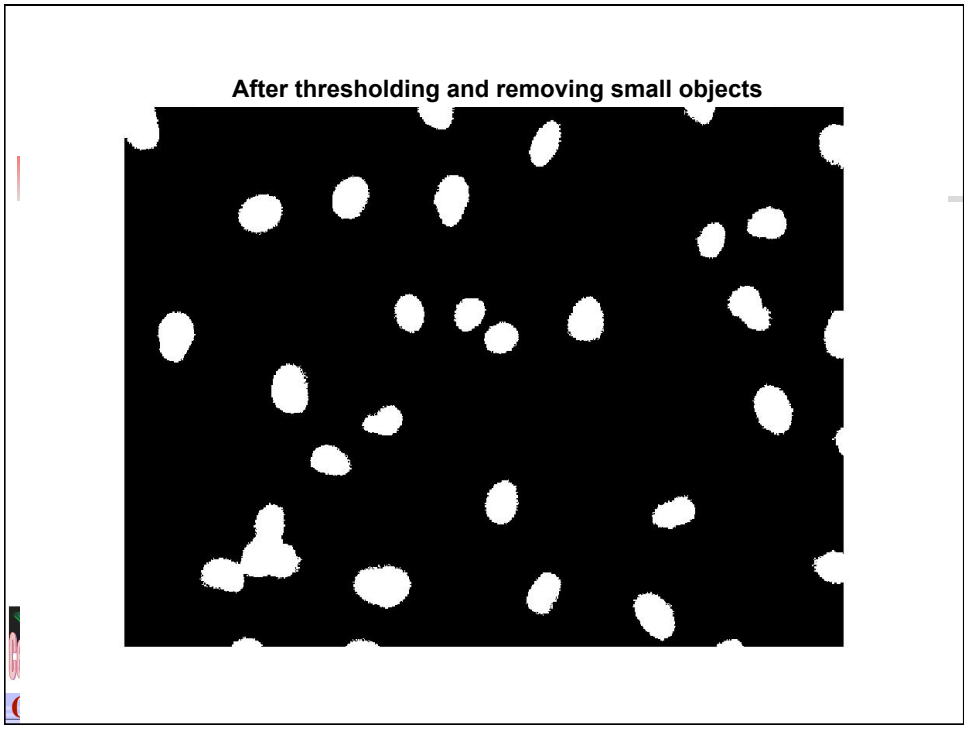
Voronoi Segmentation Process

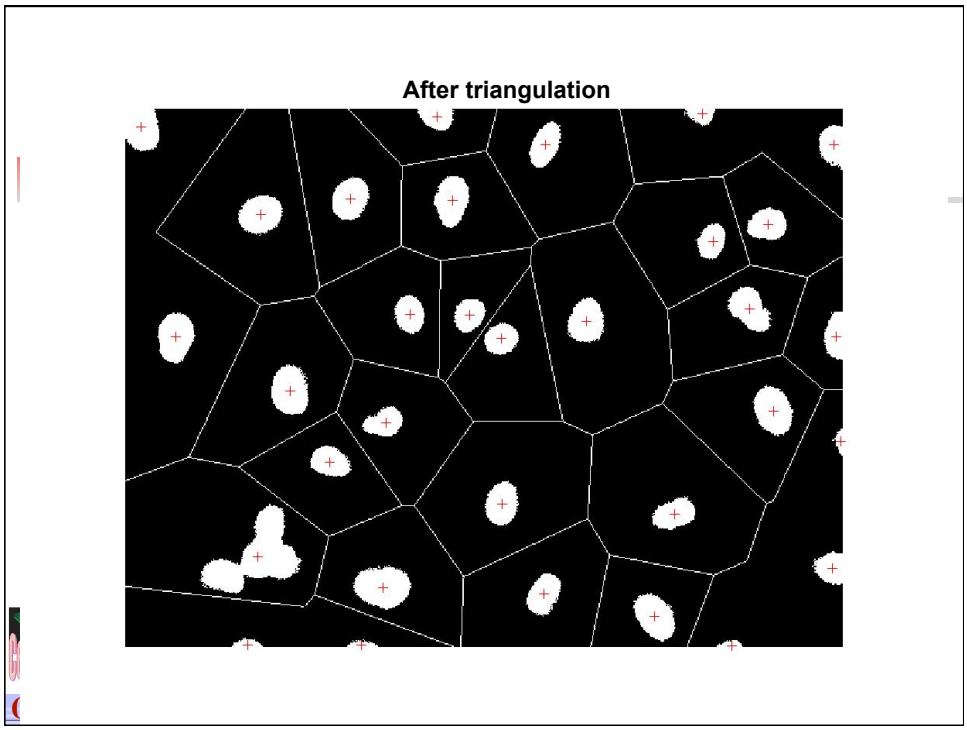
- Threshold DNA image (downsample?)
- Find the objects in the image
- Find the centers of the objects
- Use as seeds to generate Voronoi diagram
- Create a mask for each region in the Voronoi diagram
- Remove regions whose object that does not have intensity/size/shape of nucleus

Carnegie Mellon

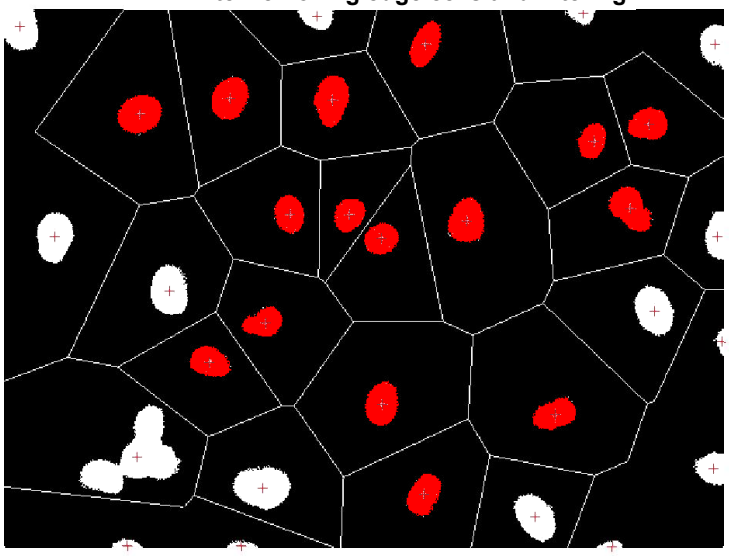
Original DNA image



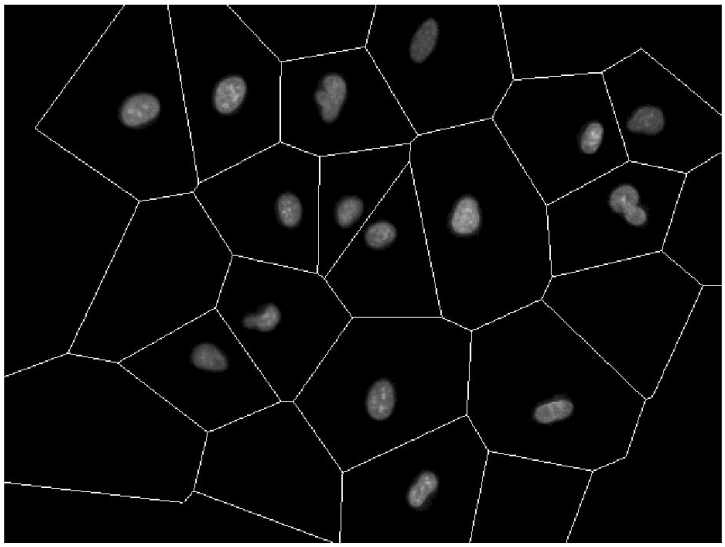

After thresholding and removing small objects


After triangulation

After removing edge cells and filtering

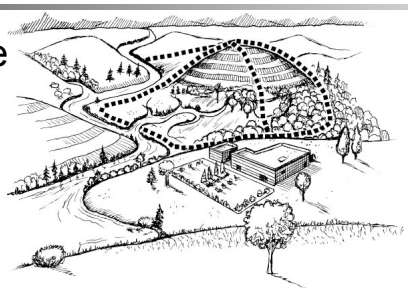


Final regions masked onto original image



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

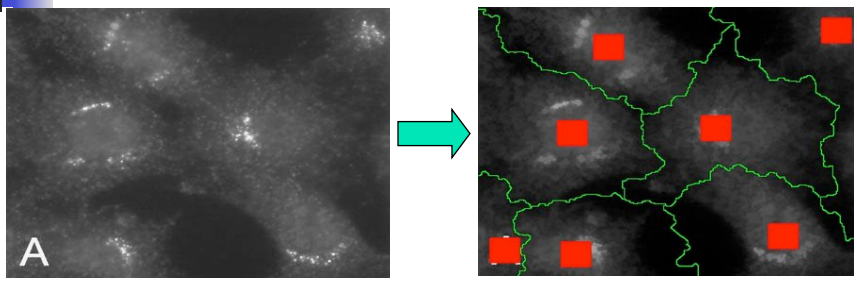
Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

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

Carnegie Mellon

Seeded Watershed Segmentation



Original image Seeds and boundary

Applied directly to protein image (no DNA image)

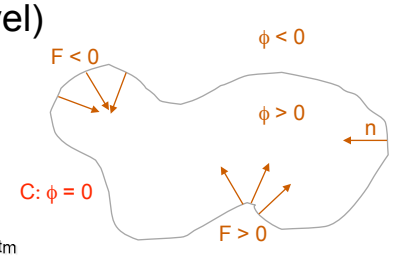
Note non-linear boundaries

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

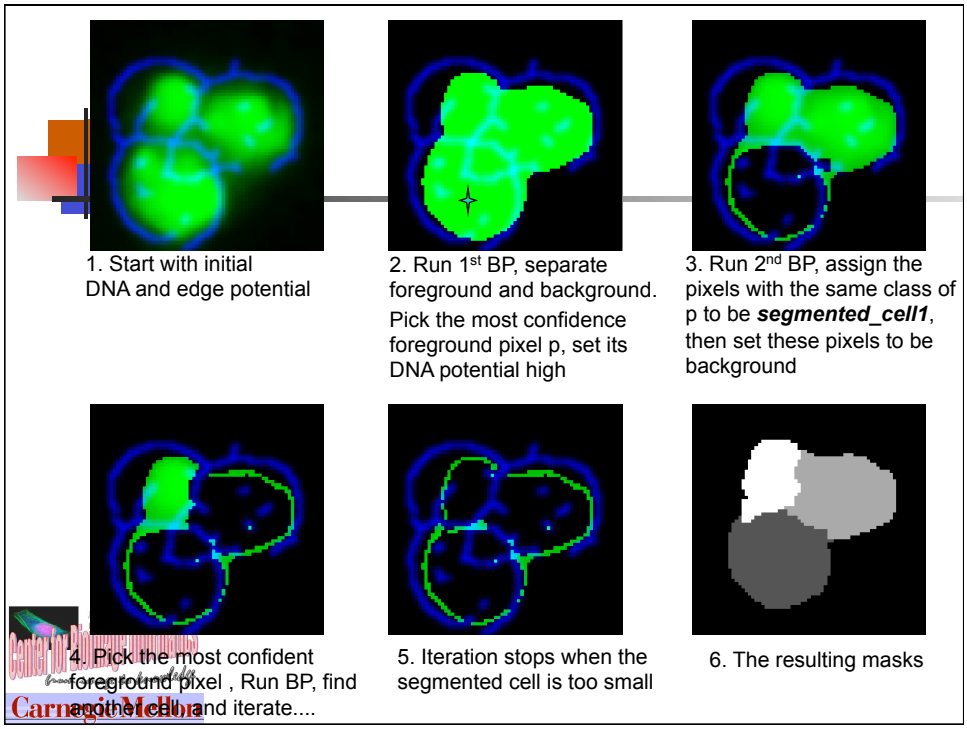


Carnegie Mellon

Graphical Model Methods

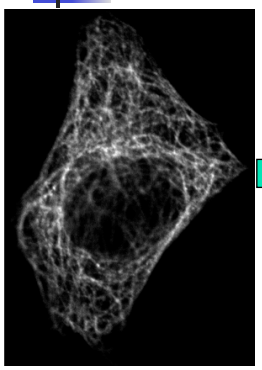
- Assumptions
  - Two classes of pixels: those part of a cell or part of the background
  - Each pixel is likely to be the same class as its neighbors
  - Have information about where cells are likely to be and where boundaries (edges) are likely to be
  - Probability that two pixels are same class related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

Carnegie Mellon



Feature Extraction for Subcellular Pattern Analysis

Goal

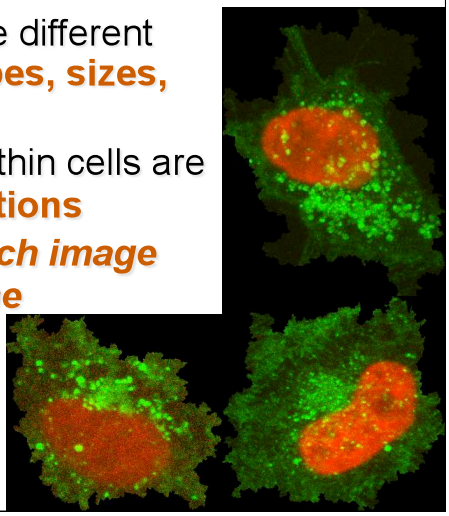


Assign proteins to *major* subcellular structures using fluorescent microscopy

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

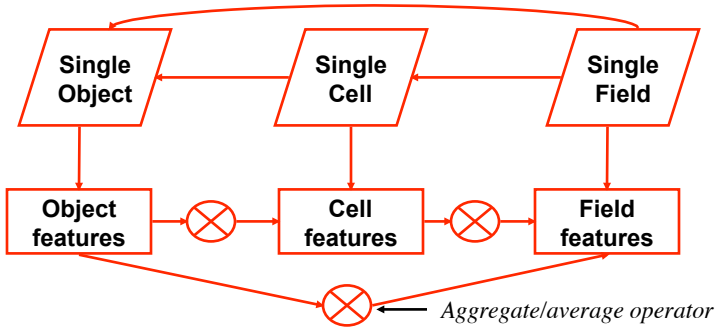
Feature-Based, Supervised Learning Approach

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

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)

Feature levels and granularity



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

Carnegie Mellon

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)

Carnegie Mellon

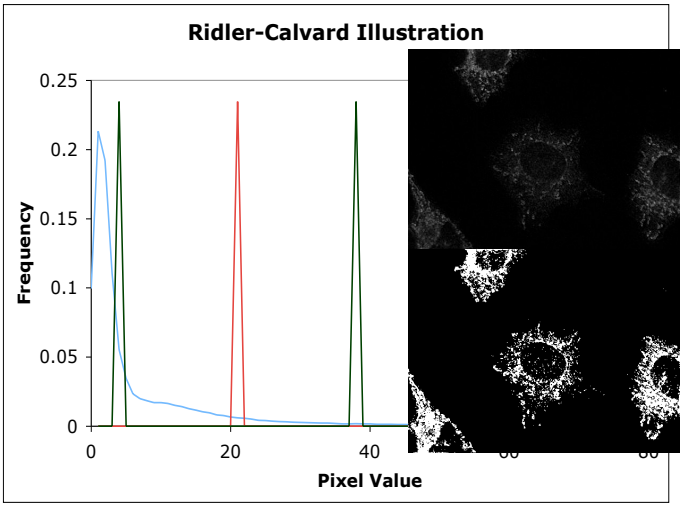
Ridler-Calvard Method

- 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

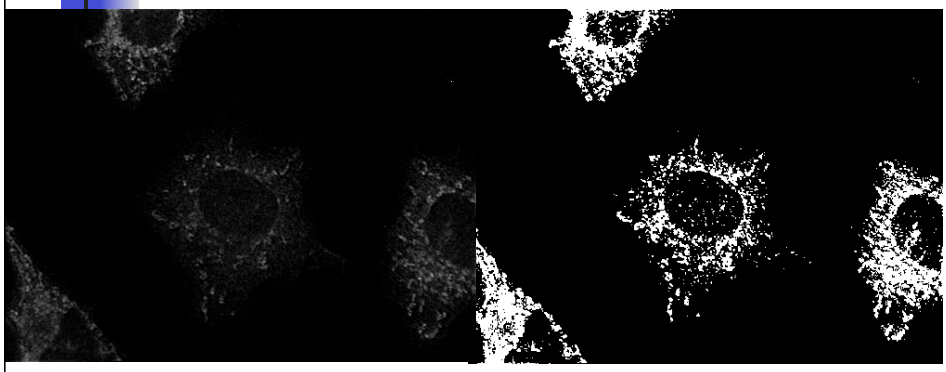
Ridler-Calvard Method

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

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

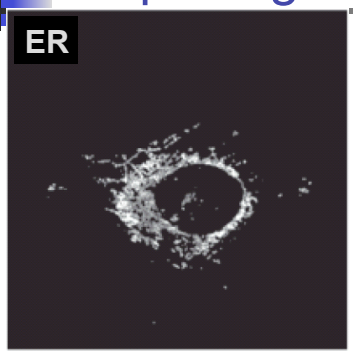
Object finding

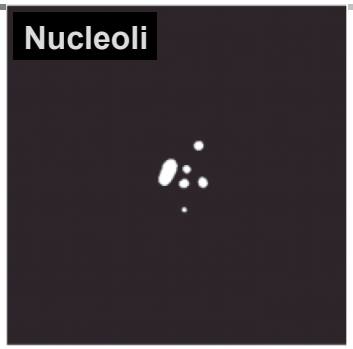
- After choice of threshold, define objects as sets of touching pixels that are above threshold

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

2D Features
Morphological Features





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

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

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)

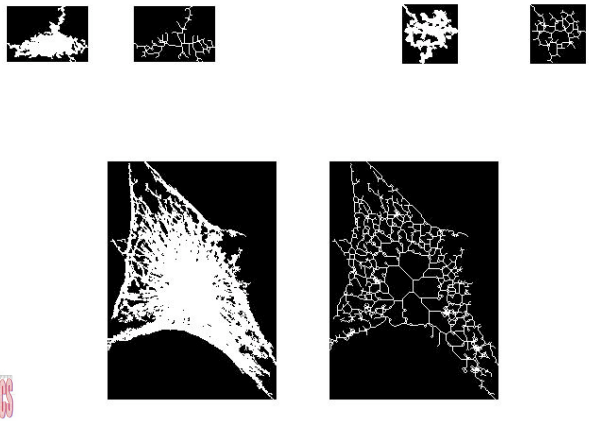
SLF No.	Description
SLF2.17	The average object distance from the COF of the DNA image
SLF2.18	The variance of object distances from the DNA COF
SLF2.19	The ratio of the largest to the smallest object to DNA COF distance
SLF2.20	The distance between the protein COF and the DNA COF
SLF2.21	The ratio of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

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

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

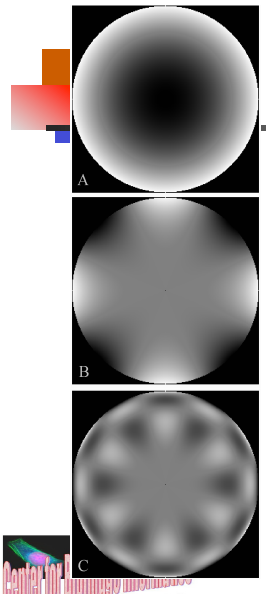
2D Features
Hull Features

Convex hull (geometrical) features

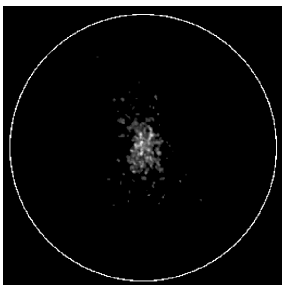
SLF1.14	The fraction of the convex hull area occupied by protein fluorescence
SLF1.15	The roundness of the convex hull
SLF1.16	The eccentricity of the convex hull

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

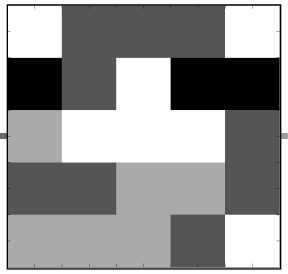
- Correlations of adjacent pixels in gray level images
- Start by calculating co-occurrence matrix P:

  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

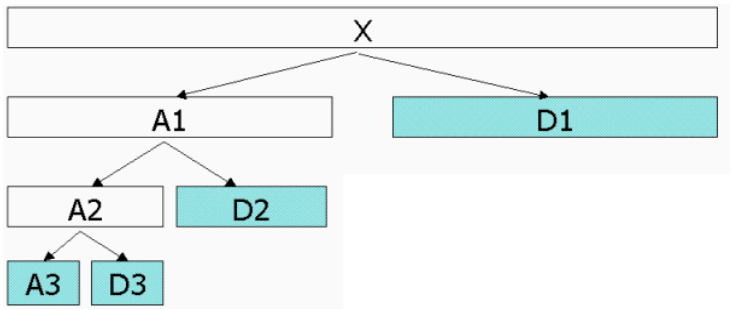
	1	2	3	4
1	0	3	0	1
2	3	0	4	4
3	0	4	0	3
4	1	4	3	2

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)

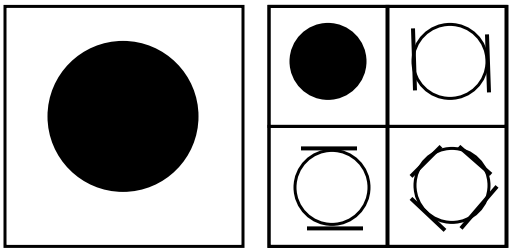
D: detail (high frequency)

X=A3+D3+D2+D1

Carnegie Mellon

2D Wavelets - intuition

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

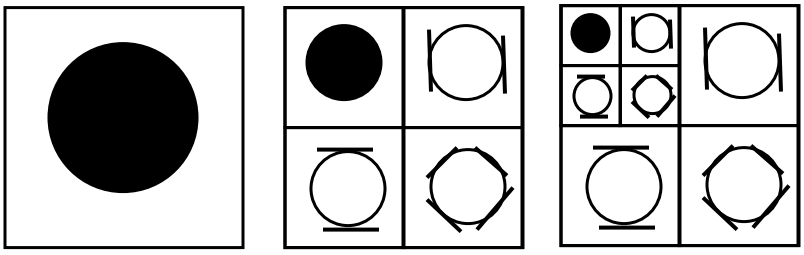


Carnegie Mellon

After Christos Faloutsos

2D Wavelets - intuition

- Recurse

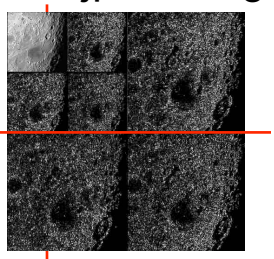


Carnegie Mellon

Slide courtesy of Christos Faloutsos

2D Wavelets - intuition

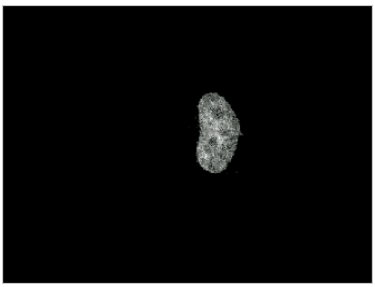
- Many wavelet basis functions (filters):
  - Haar
  - Daubechies (-4, -6, -20)
- http://www331.jpl.nasa.gov/public/wave.html



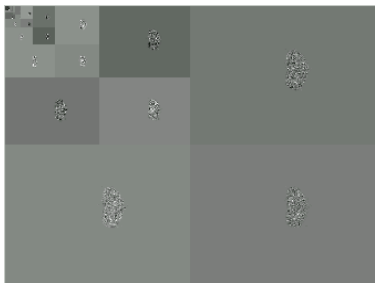
Carnegie Mellon

Slide courtesy of Christos Faloutsos

Daubechies D4 decomposition



Original image



Wavelet Transformation

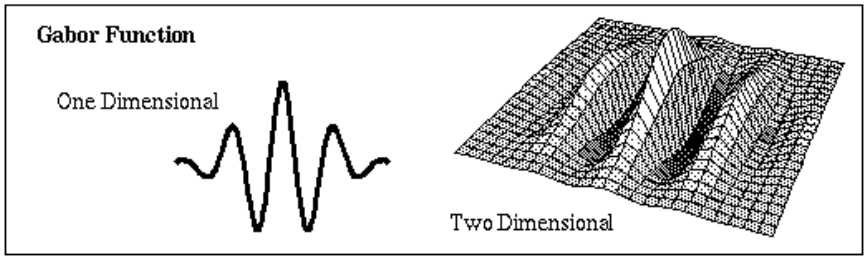
Carnegie Mellon

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

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

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)

2Dt or 3Dt Features
Temporal Texture Features

- **Haralick texture features** describe the correlation in intensity of pixels that are next to each other in **space**.
  - These have been valuable for classifying static patterns.
- **Temporal texture features** describe the correlation in intensity of pixels in the same position in images next to each other over **time**.

Temporal Textures based on Co-occurrence Matrix

- Temporal co-occurrence matrix P:
  N_{level} by N_{level} matrix, Element P[i, j] is the probability that a pixel with value i has value j in the next image (time point).
- Thirteen statistics calculated on P are used as features

Carnegie Mellon

Temporal Textures based on Co-occurrence Matrix

Image at t0

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

Image at t1

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

Temporal co-occurrence matrix (for image that does not change)

	1	2	3	4
1	3	0	0	0
2	0	9	0	0
3	0	0	6	0
4	0	0	0	7

Carnegie Mellon

Image at t0

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

Image at t1

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

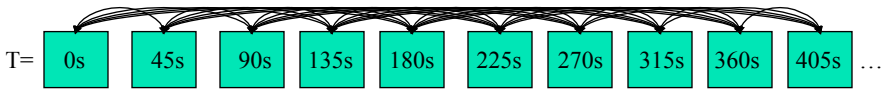
Temporal co-occurrence matrix (for image that changes)

	1	2	3	4
1	1	0	2	0
2	2	1	1	5
3	0	5	0	1
4	0	3	3	1

Carnegie Mellon

Implementation of Temporal Texture Features

- Compare image pairs with different time interval ,compute 13 temporal texture features for each pair.



- Use the average and variance of features in each kind of time interval, yields 13*5*2=130 features

Carnegie Mellon

Feature Selection and Classification

Robert F. Murphy

Departments of Biological Sciences, Biomedical Engineering, and Machine Learning

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

Feature selection

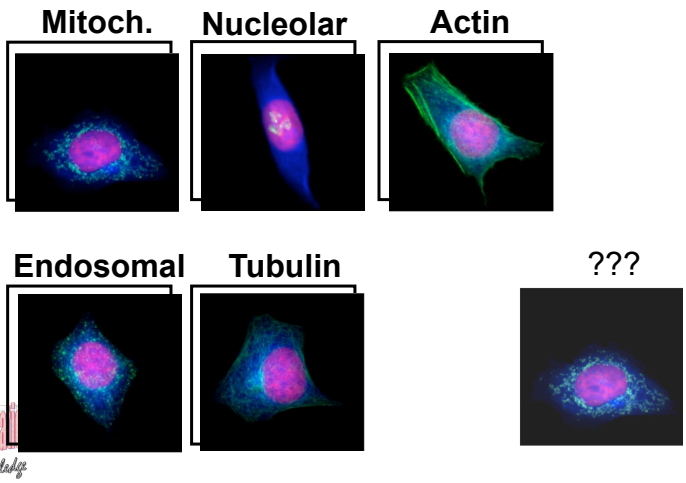
- Having too many features can confuse a classifier
- Can use comparison of feature distributions between classes to choose a subset of features that gets rid of uninformative or redundant features

Feature Selection Methods

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

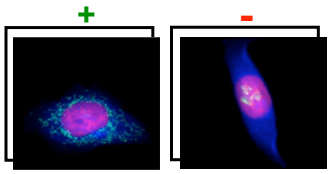
Carnegie Mellon

Basic classification problem

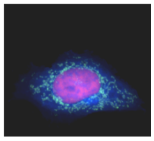


Carnegie Mellon

Simple two class problem



???

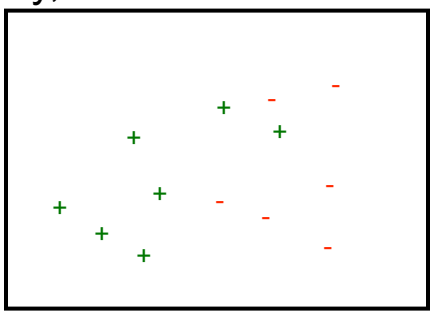


Carnegie Mellon

Decision trees

- Pictorially, we have

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



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

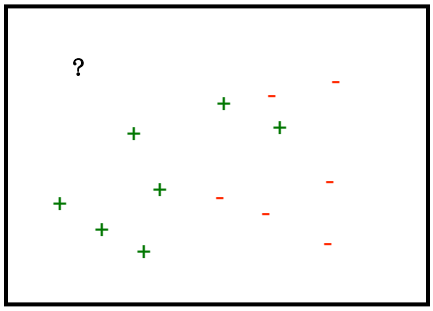
Carnegie Mellon

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

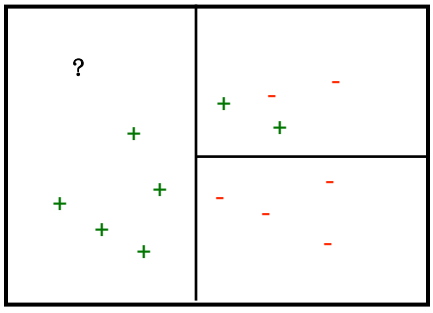
Carnegie Mellon

Decision trees

- so we build a decision tree:

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

40



50

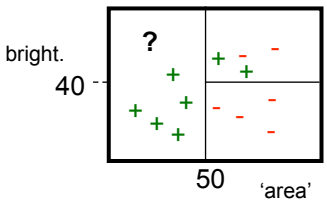
num. attr#1 (e.g.., 'area')

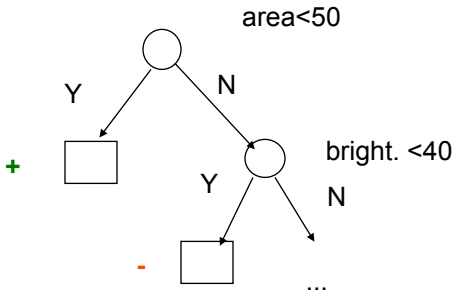
Slide courtesy of Christos Faloutsos

Carnegie Mellon

Decision trees

- so we build a decision tree:



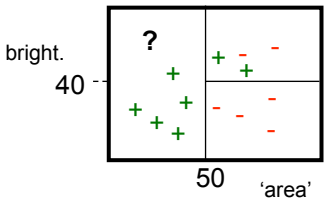


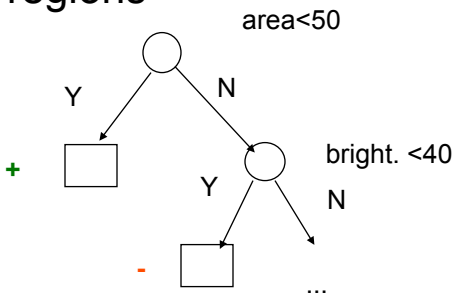
Carnegie Mellon

Slide courtesy of Christos Faloutsos

Decision trees

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





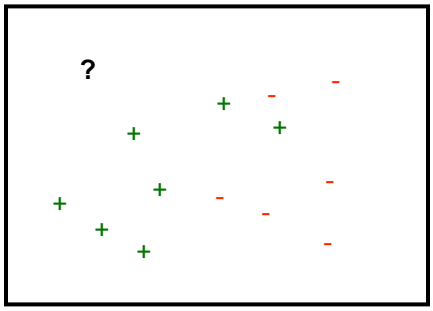
Carnegie Mellon

Slide courtesy of Christos Faloutsos

Problem: Classification

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

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



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

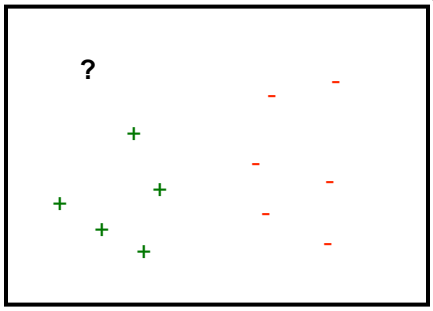
Carnegie Mellon

Slide courtesy of Christos Faloutsos

Support Vector Machines (SVMs)

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

bright.



area

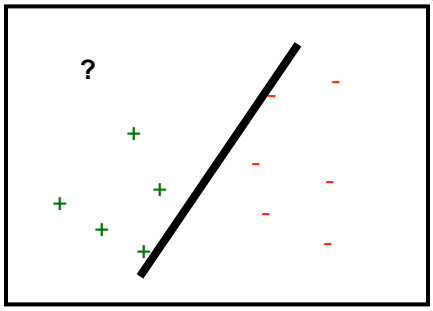
Carnegie Mellon

Slide courtesy of Christos Faloutsos

Support Vector Machines (SVMs)

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

bright.



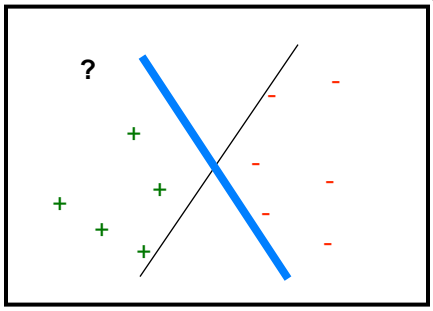
area

Slide courtesy of Christos Faloutsos

Support Vector Machines (SVMs)

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

bright.



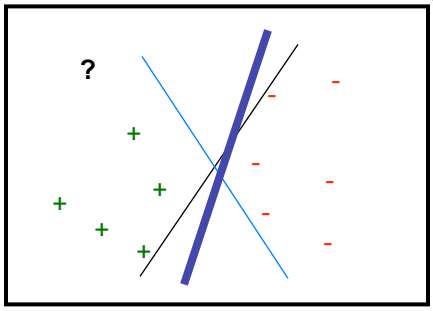
area

Slide courtesy of Christos Faloutsos

Support Vector Machines (SVMs)

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

bright.



area

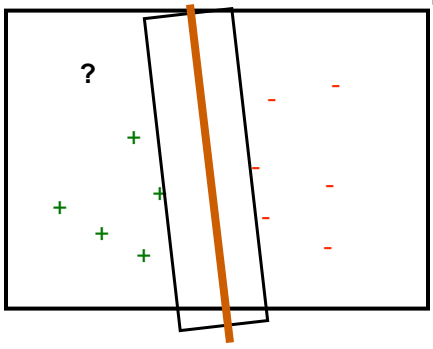
Carnegie Mellon

Slide courtesy of Christos Faloutsos

Support Vector Machines (SVMs)

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

bright.



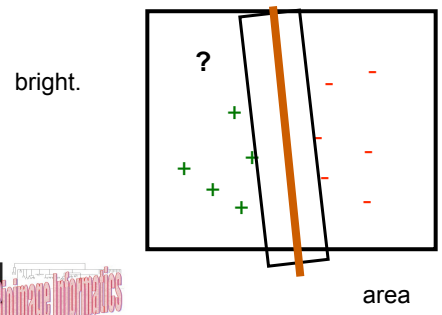
area

Carnegie Mellon

Slide courtesy of Christos Faloutsos

Support Vector Machines (SVMs)

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

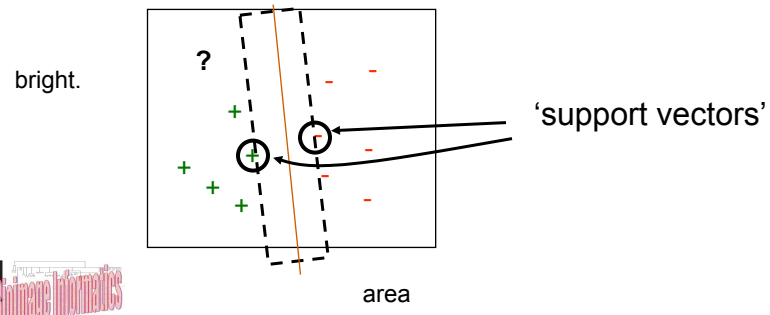


Carnegie Mellon

Slide courtesy of Christos Faloutsos

Support Vector Machines (SVMs)

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



Carnegie Mellon

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

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

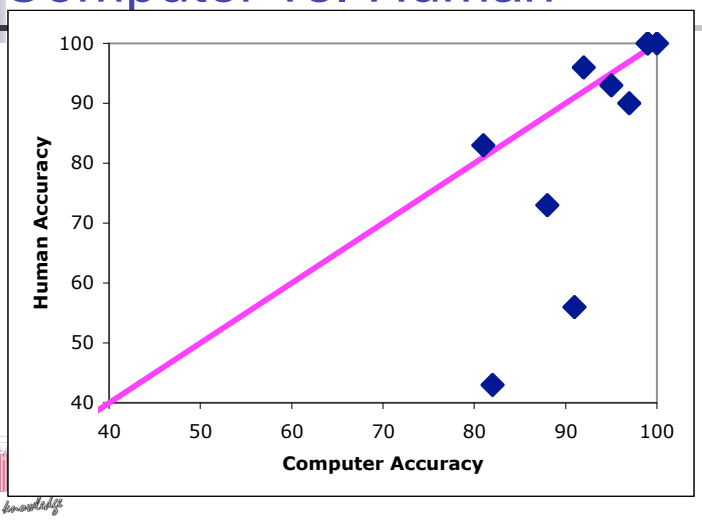
Human Classification Results

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%

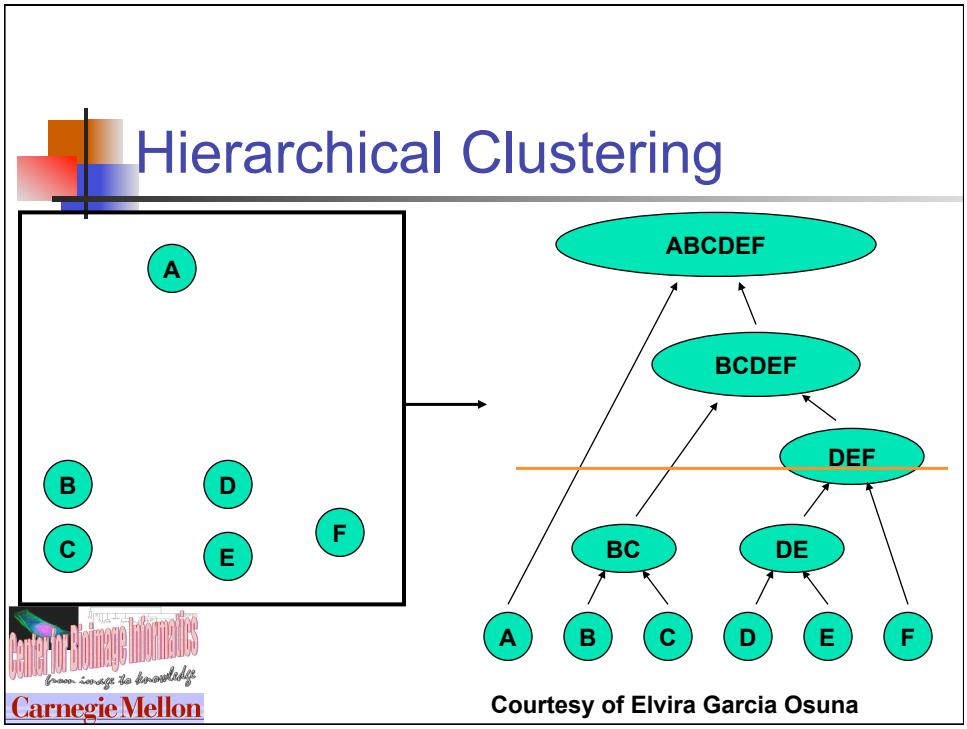
Carnegie Mellon

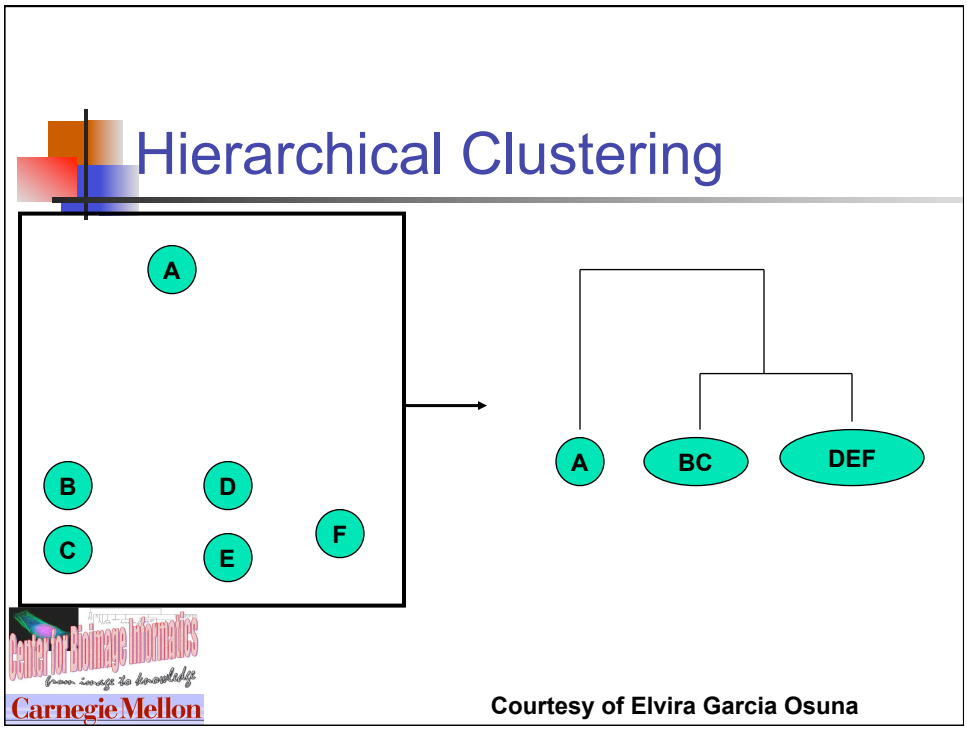
Cluster analysis

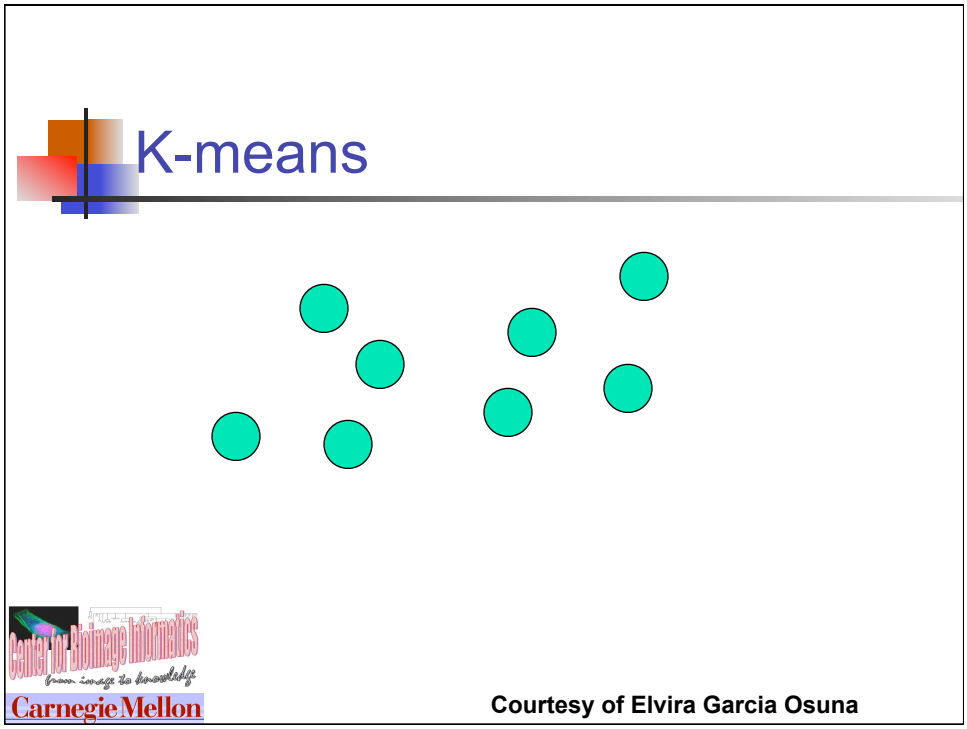
- Supervised learning (Classification) assumes classes are known
- Unsupervised learning (Cluster analysis) seeks to discover the classes

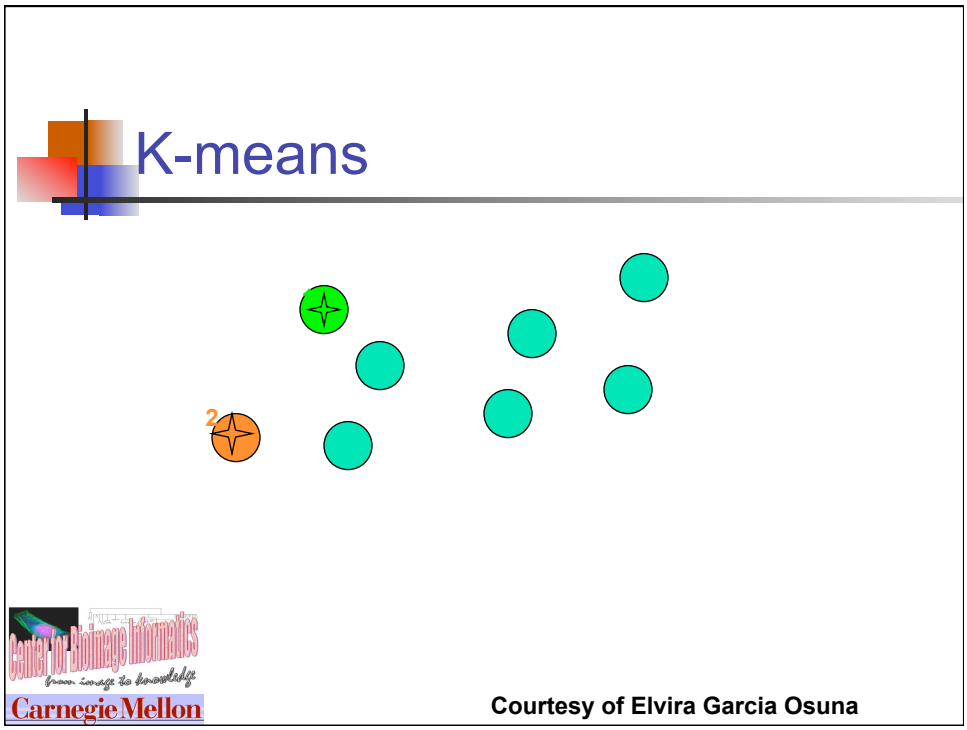
Hierarchical vs. *k*-means clustering

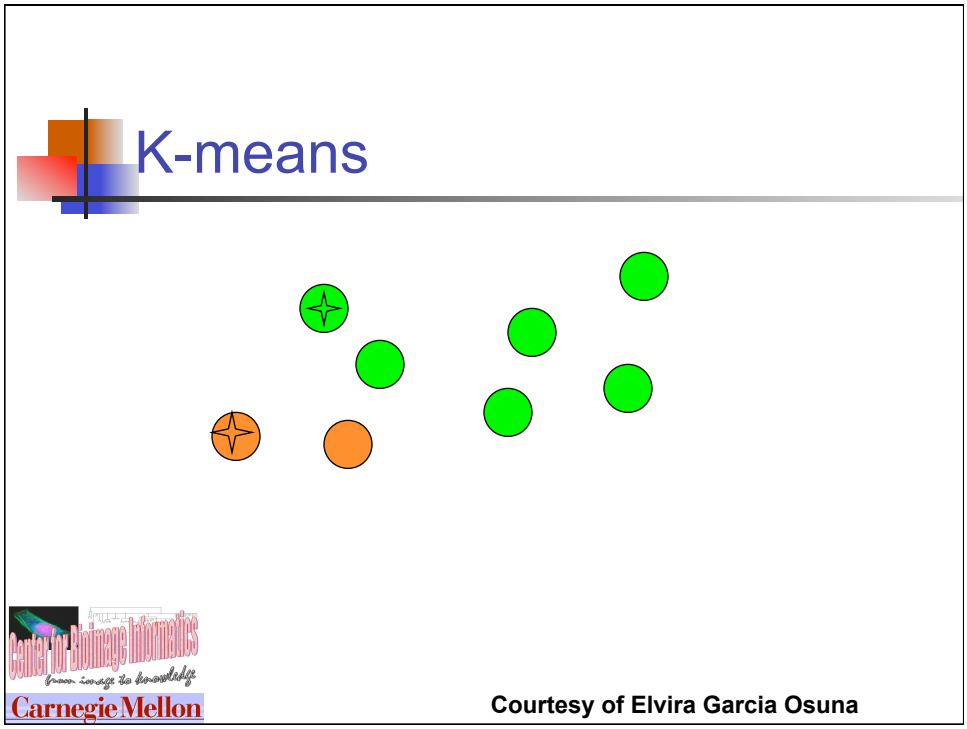
- Two most popular clustering algorithms
- Hierarchical builds tree sequentially from the closest pair of points (wells/ cells/probes/conditions)
- *k*-means starts with *k* randomly chosen seed points, assigns each remaining point to the nearest seed, and repeats this until no point moves

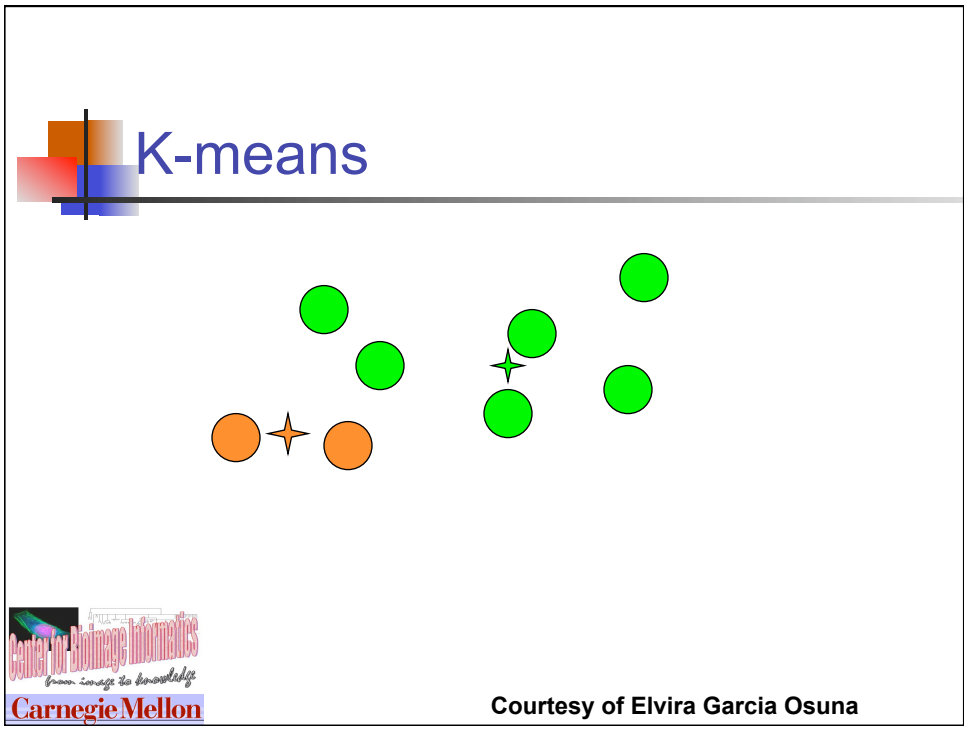

Hierarchical Clustering
A
B
C
D
E
F
ABCDEF
BCDEF
DEF
BC
DE
A
B
C
D
E
F
Carnegie Mellon
Courtesy of Elvira Garcia Osuna

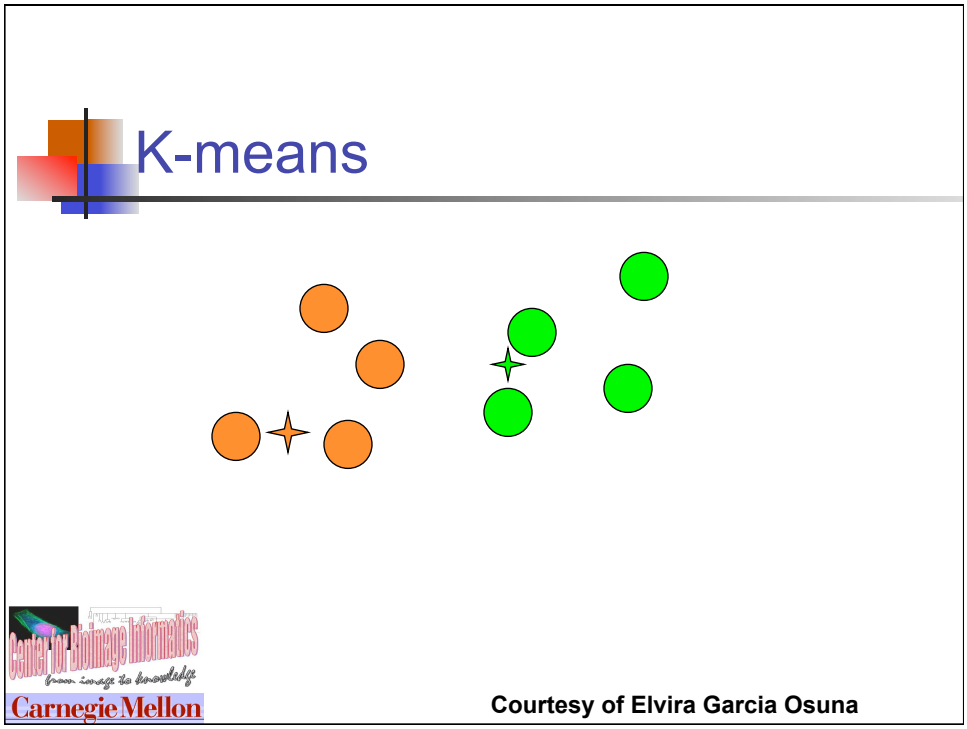

Hierarchical Clustering
A
B
C
D
E
F
A
BC
DEF
Carnegie Mellon
Courtesy of Elvira Garcia Osuna

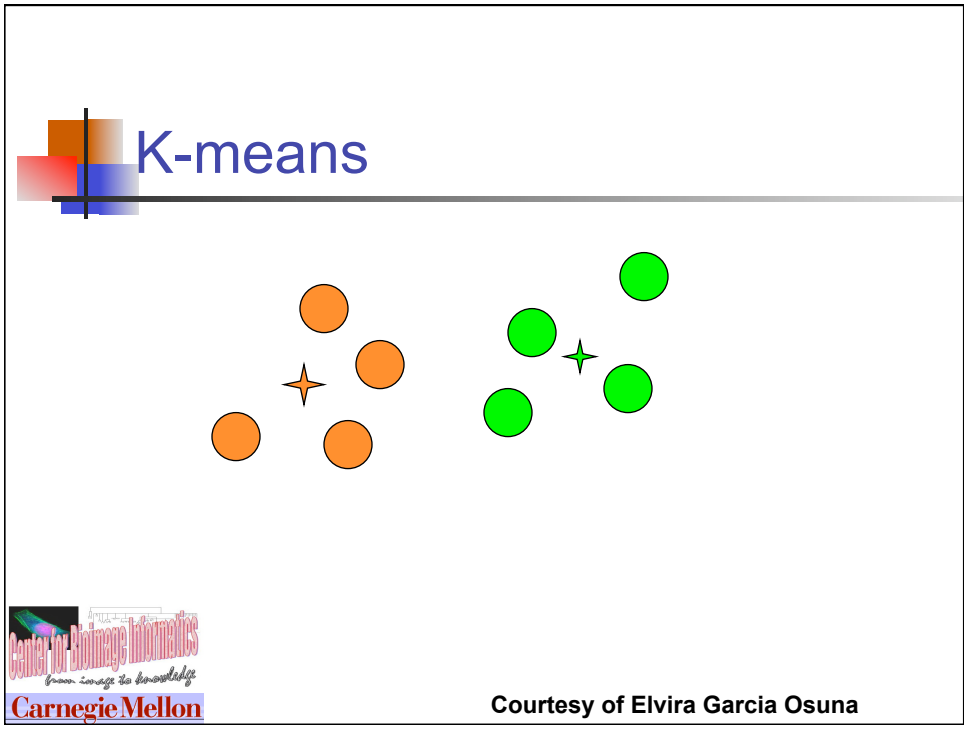

K-means
Carnegie Mellon
Courtesy of Elvira Garcia Osuna


K-means
2
Carnegie Mellon
Courtesy of Elvira Garcia Osuna

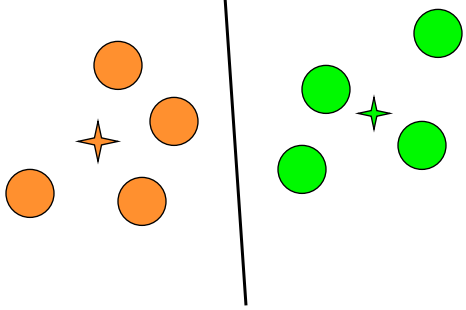

K-means
Carnegie Mellon
Courtesy of Elvira Garcia Osuna


K-means
Carnegie Mellon
Courtesy of Elvira Garcia Osuna


K-means
Carnegie Mellon
Courtesy of Elvira Garcia Osuna


K-means
Carnegie Mellon
Courtesy of Elvira Garcia Osuna

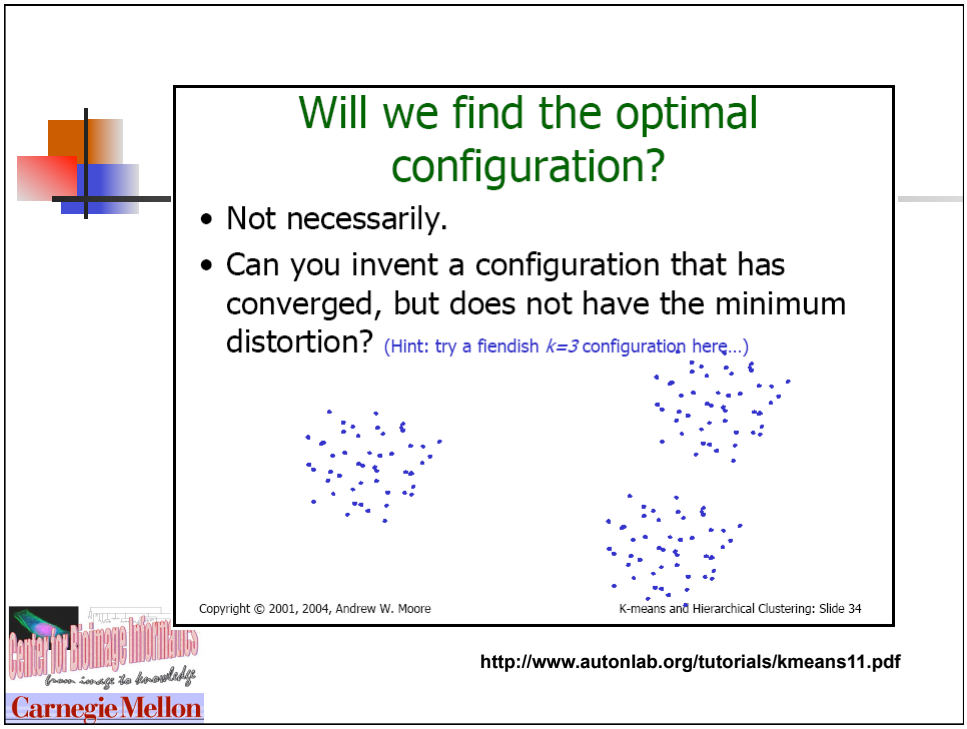
K-means



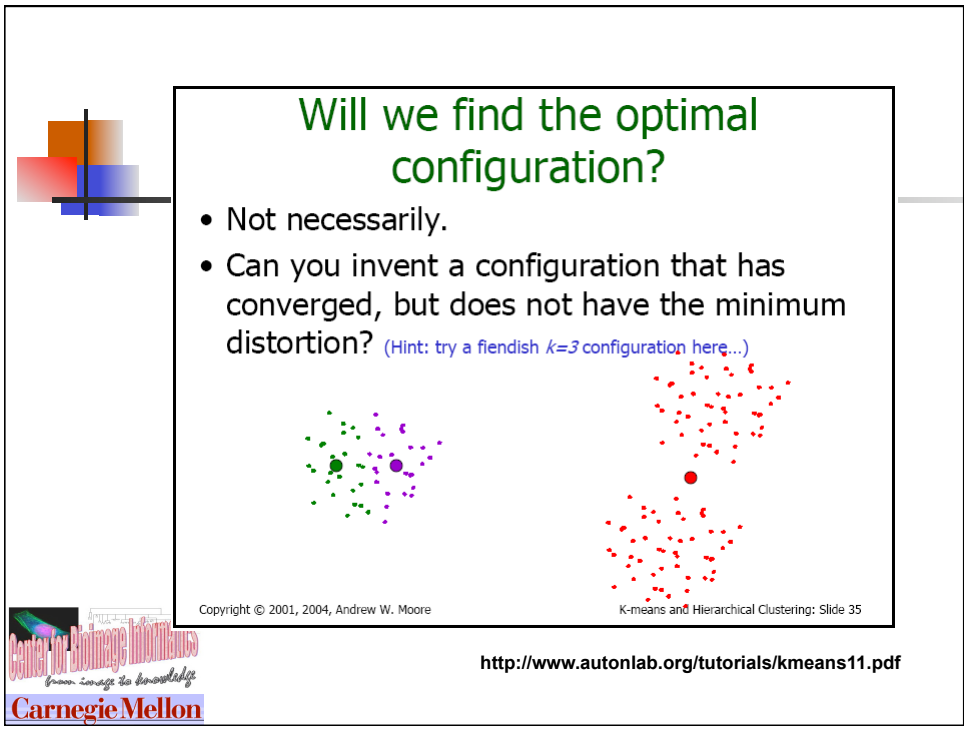
Courtesy of Elvira Garcia Osuna

K-means Questions

- What is it trying to optimize?
- Are we sure it will terminate?
- Are we sure it will find an optimal clustering?
- How should we start it?
- How could we automatically choose the number of centers?



http://www.autonlab.org/tutorials/kmeans11.pdf



http://www.autonlab.org/tutorials/kmeans11.pdf

Choosing the number of Centers

- A difficult problem
- Most common approach is to try to find the solution that minimizes the Bayesian Information Criterion

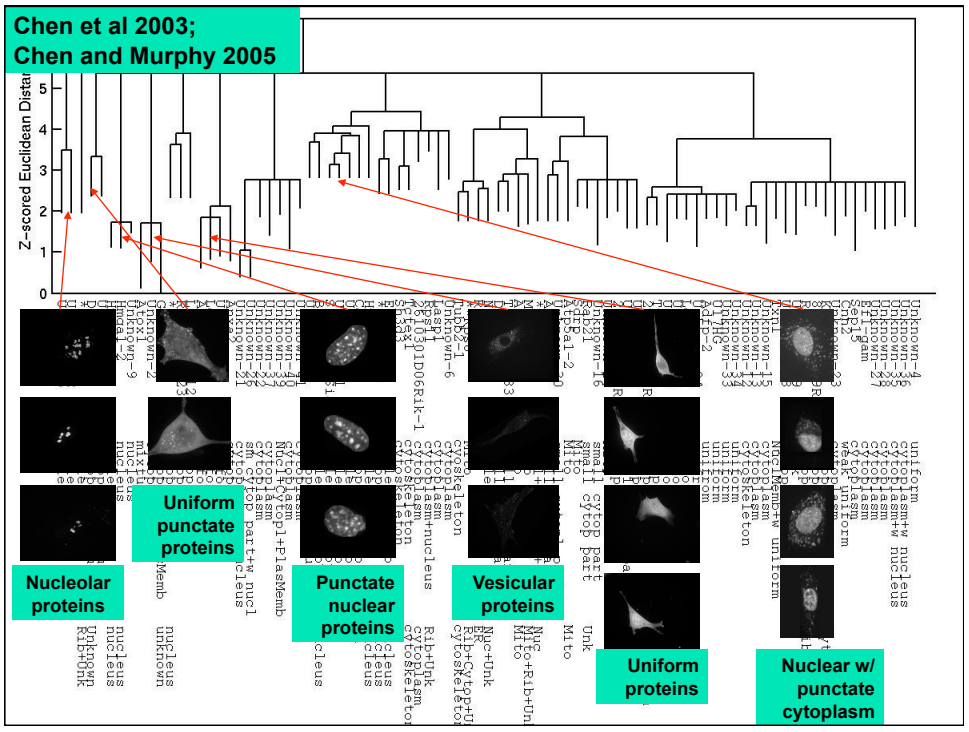
$$BIC = -2\ln L + k\ln(n)$$

L = the likelihood function for the estimated model

K = # of parameters

n = # of samples

Carnegie Mellon



Graphical Models for Subcellular Pattern Analysis

Robert F. Murphy

Departments of Biological Sciences, Biomedical Engineering, and Machine Learning

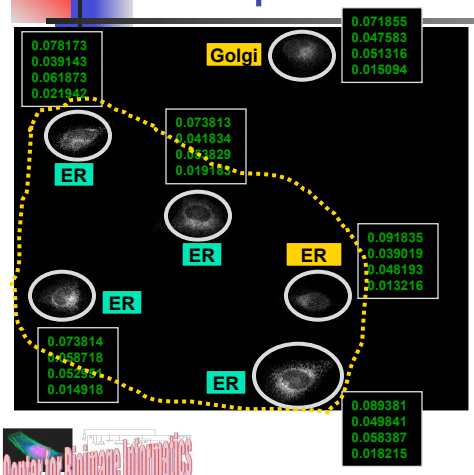
Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

Graphical Models for Improving Pattern Recognition

- Since cells with same location pattern are often clustered together, considering *multiple cells* may improve the discrimination of similar location patterns.
- We developed a *novel graphical model* to describe the relationship between multiple cells in a field.
- The classification of a cell is influenced by the classification results of neighboring cells.

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

Multiple Cells in an Image



1. Segmentation
2. Feature Extraction
3. Cell Classification
   - ❑ Individually
   - ❑ Dependently
     - o Majority Voting
       Homogeneous Field
       accuracy: 98%
       (Boland and Murphy, 2001)
     - o Local Dependence
       Heterogenous Field

Carnegie Mellon

Bayes Decision Theory

Bayes Rule

$$p(w_j \mid x) = \frac{p(x \mid w_j)p(w_j)}{p(x)}$$

x: features
w_j: j_{th} class

$$posterior = \frac{likelihood \times prior}{evidence}$$

Carnegie Mellon

Bayes Decision Theory

Training

$$p(w_j \mid x) = \frac{p(x \mid w_j)p(w_j)}{p(x)}$$

x: features
w_j: j_{th} class

Train a classifier given training images of each class

Carnegie Mellon

Bayes Decision Theory

Testing

$$p(w_j \mid x) = \frac{p(x \mid w_j)p(w_j)}{p(x)}$$

x: features
w_j: j_{th} class

Assign x to the class with max posterior probability

Carnegie Mellon

Bayes Decision Theory

Testing

$$p(w_j \mid x) = \frac{p(x \mid w_j)p(w_j)}{p(x)}$$

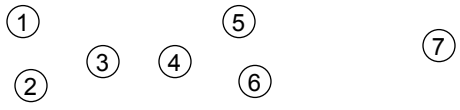
x: features
w_j: j_{th} class

Normally, prior distribution assumed or determined ahead of time (prior!). Our idea: adjust priors to reflect the neighbors of a cell (iteratively).
> The posterior probability changes to reflect neighbors

Carnegie Mellon

Graphical Cell Model

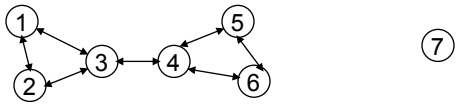
Consider multiple cells in a field



Carnegie Mellon

Graphical Cell Model

Connect cells if they are close enough
(either in physical space or feature space)

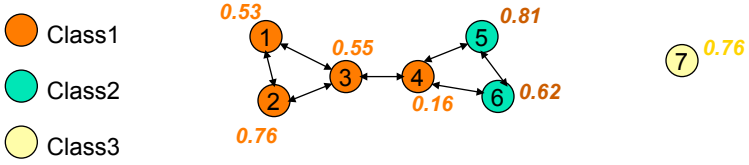


Links are decided by d_{cutoff}

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

Graphical Cell Model

Assign each cell a label and a confidence measure

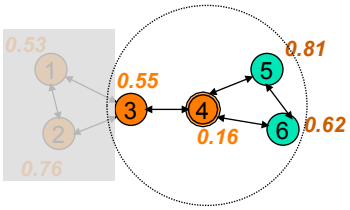


Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

Graphical Cell Model

Assign each cell a label and a confidence measure

Class1
Class2
Class3



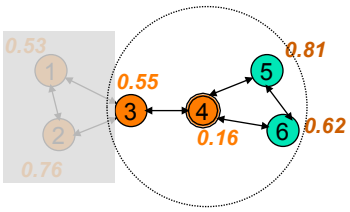
Consider 4 Influenced by 3 5 6 not by 7

Carnegie Mellon

Graphical Cell Model

Assign each cell a label and a confidence measure

Class1
Class2
Class3



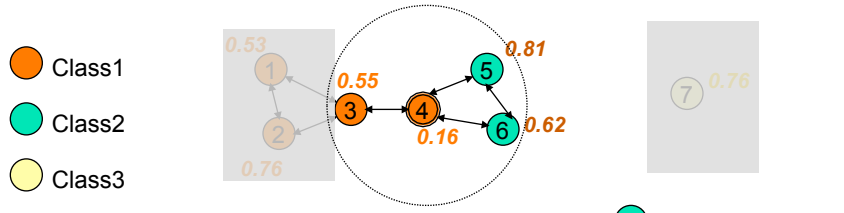
Consider 4 Influenced by 3 5 6 not by 7

	Class1	Class2	Class3
Initial Priors(Uniform)	0.33	0.33	0.33
New Priors	↑	↑↑	↓

Carnegie Mellon

Graphical Cell Model

Assign each cell a label and a confidence measure



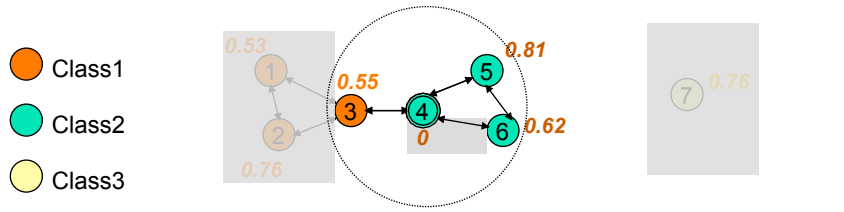
Consider 4 **Influenced by** 3 5 6 **not by** 7

	Class1	Class2	Class3
Initial Priors(Uniform)	*0.33*	*0.33*	*0.33*
New Priors	*0.40*	*0.58*	*0.02*

Carnegie Mellon

Graphical Cell Model

Classify the cell with the new priors

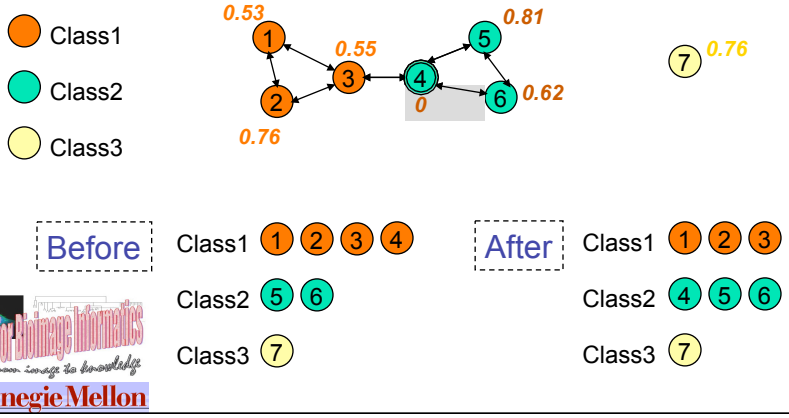


	Class1	Class2	Class3
New Priors	*0.40*	*0.58*	*0.02*

Carnegie Mellon

Graphical Cell Model

Iterate until no label changes



Before

Class1 1 2 3 4

Class2 5 6

Class3 7

After

Class1 1 2 3

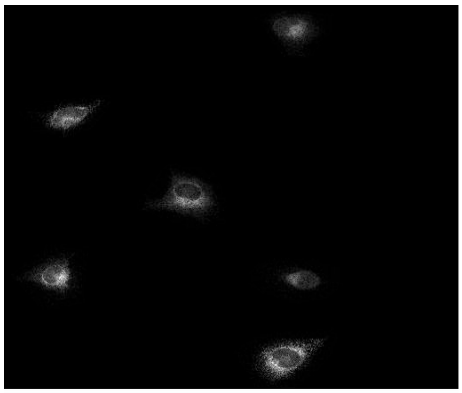
Class2 4 5 6

Class3 7

Carnegie Mellon

Evaluating Prior Updating Scheme

- Use the 10 class 2D HeLa data set to create synthetic multi-cell images where the class of each individual cell is known
- Compare performance to base (single cell) classifier (SVM)



Carnegie Mellon

Chen & Murphy, 2006

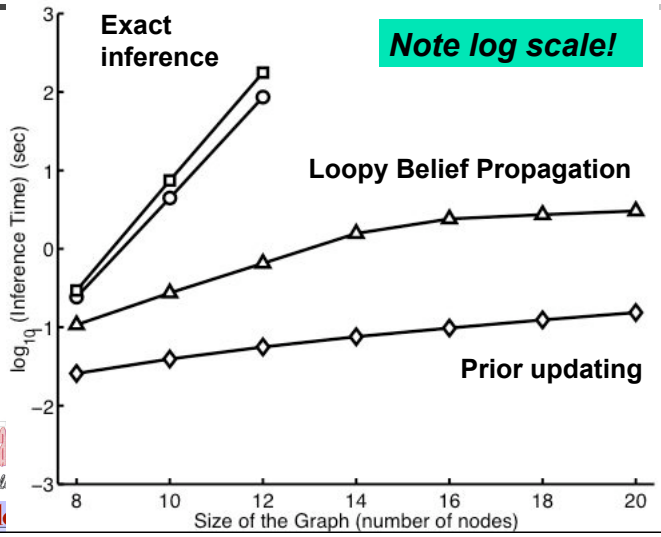
Classification accuracies for multicell images from two classes

No. training images	Without updating (%)	With updating (%)	Improvement (%)
50	90.1	95.6	5.5
40	89.2	95.1	5.9
30	88.1	94.6	6.5
20	86.4	93.8	7.4
10	82.9	90.3	7.4

Improvement is greater for weaker base classifier (more room to improve)

Carnegie Mellon

Prior updating much faster than previous methods



Carnegie Mellon

Prior Updating Conclusions

- Graphical models can be used to improve accuracy of classification of heterogeneous images
- Each individual cell is still classified, and minor or unusual cells are not “lost”
- Appropriate for cell array experiments (e.g., RNAi) where heterogeneity expected
- Appropriate for tissue images

References on Automated Interpretation of Subcellular Patterns

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

Review Articles

- K. Huang and R. F. Murphy (2004). From Quantitative Microscopy to Automated Image Understanding. *J. Biomed. Optics 9:*893-912.
- 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.
- E. Glory and R.F. Murphy (2007). Automated Subcellular Location Determination and High Throughput Microscopy. *Developmental Cell 12*:7-16.

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.

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.

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.

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.

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*.

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.
- E. Garcia Osuna, J. Hua, N.W. Bateman, T. Zhao, P.B. Berget and R.F. Murphy (2007). Large-Scale Automated Analysis of Protein Subcellular Location Patterns in Randomly-Tagged 3T3 Cells. *Annals Biomed. Eng., in press*

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

Graphical models for multi-cell classification

- 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.

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

Cell Segmentation

- L. Coulot, H. Kirschner, A. Chebira, J.M.F. Moura, J. Kovacevic, E. Garcia Osuna, and R.F. Murphy (2006). Topology Preserving STACS Segmentation of Protein Subcellular Location Images. *Proceedings of the 2006 IEEE International Symposium on Biomedical Imaging (ISBI 2006)*, pp. 566-569.
- S.-C. Chen, T. Zhao, G.J. Gordon, and R.F. Murphy (2006). A Novel Graphical Model Approach to Segmenting Cell Images. *Proceedings of the 2006 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB '06)*, 1079.

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.

Center for Bioimage Informatics
from image to knowledge
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.

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.

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.

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.
- W.W. Cohen, R. Wang and R.F. Murphy (2003). Understanding Captions in Biomedical Publications. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003)*, pp. 499-504.

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

SLIF - Subcellular Location Image Finder

- 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.
- Z. Kou, W.W. Cohen and R.F. Murphy (2005). High-recall protein entity recognition using a dictionary. *Bioinformatics 21(suppl_1):*i266-i273.
- Z. Kou, W.W. Cohen, and R.F. Murphy (2007). A Stacked Graphical Model for Associating Information from Text and Images in Figures. *Pacific Symposium on Biocomputing 12*:257-268.

Center for Bioimage Informatics
from images to knowledge
Carnegie Mellon