Machine Learning Approaches to Information Extraction from Text and Images in Biomedical Journal Articles

Robert F. Murphy
Departments of Biological Sciences, Biomedical Engineering and Machine Learning and
Center for Bioimage Informatics
Carnegie Mellon

Goal of tutorial

- Introduce problem of automated interpretation of articles containing text and images
- Describe relevant methods, mostly in context of SLIF (Subcellular Location Image Finder) system
- Describe future directions for field

Ultimate Goal of the field

- Machine understanding of biological journal articles (text and image)
- Criteria for success: Turing test - have machine be able to answer questions about an article as well as a human scientist

Intermediate Goal

- Extract information from combination of text and any kind of image in biological journal article
- Criteria for success: Achieve high precision and recall for extracted assertions (compared to expert scientist)

Immediate Goal (SLIF)

- Extract information about subcellular location from captions and figures containing fluorescence microscope images in biological journal articles
- Criteria for success: Achieve high precision and recall for extracted assertions (compared to expert scientist)

State of art: Bio Journal Information Extraction

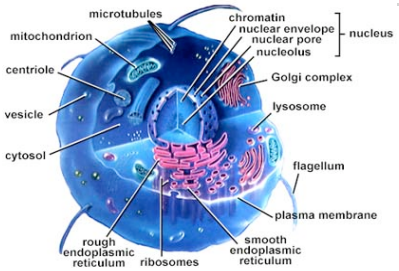
- A number of systems to index literature via extracted terms
- A few systems to index image content in literature
- A few systems for document classification

Practices in Biological Journal Articles

- Articles not monolithic: they can support more than one biological conclusion
- Different types of data often combined in one article and in one figure
- Assume knowledge of basic biology
- Captions should be understandable without reference to paper
- Materials often defined in separate section

Introduction to Protein Subcellular Location

Eukaryotic cells have many parts



Protein localization

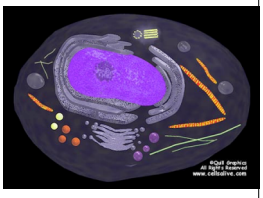
- The sequence of each protein determines where it is localized in cells
- Subsequences ("motifs") within a protein's sequence are responsible for targeting it to one (or more) locations (structures/organelles)

Open questions

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

Proteomics

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



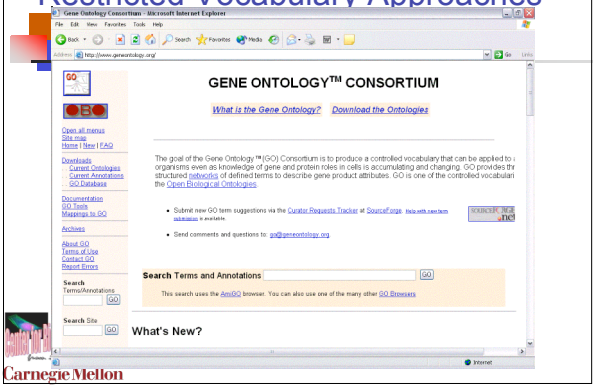
Location information in protein databases: Traditional approach

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

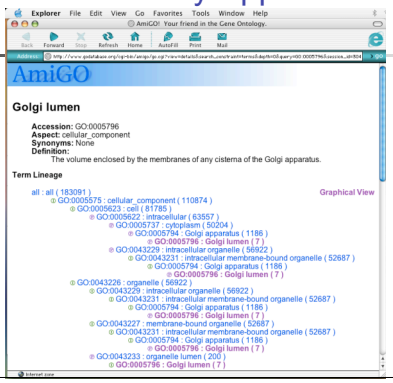
Location information in protein databases: Ontology approach

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

Restricted Vocabulary Approaches



Restricted Vocabulary Approaches



Use of GO terms

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

Comparison of GO terms for two proteins

GolgB1	GPP130
Integral to membrane; Golgi membrane; Golgi stack;	Integral to membrane; Golgi cis-face; Golgi lumen; endocytotic transport vesicle

Source: SwissProt

Determining protein location

- The primary method used to **determine** the subcellular location of a protein is to "tag" it with a fluorescent probe and then image its distribution within cells using fluorescence microscopy (abbreviate resulting **F**luorescence **M**icroscope **I**mage as FMI)

Tagging proteins for fluorescence microscopy

- Immunofluorescence
  - "primary" antibody against the target,
  - "secondary" antibody against the "primary" and conjugated with a fluorescent probe
  - Fixed-cells only
- Gene/cDNA-tagging
  - merge DNA coding for a naturally fluorescent protein (or vital probe binding sequence) with coding sequence of a protein of interest
  - Live-cell possible

Tagging proteins for fluorescence microscopy

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

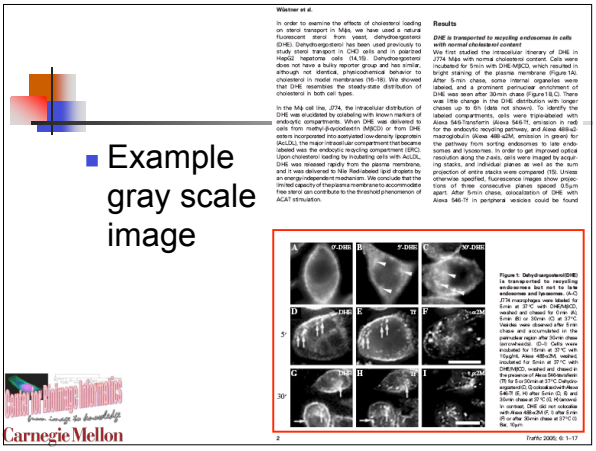
Major information to extract for FMI in article figures

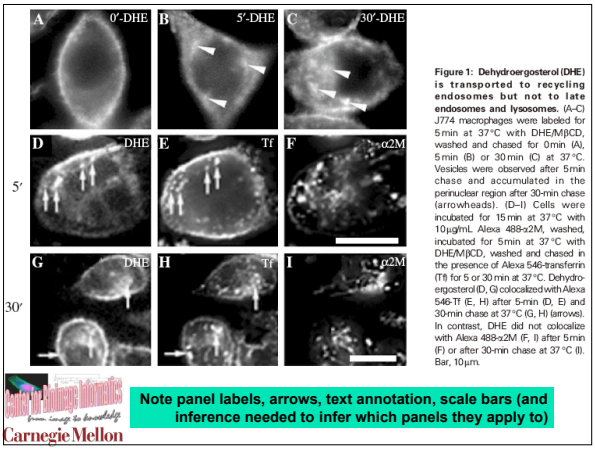
- Sample
  - Cell or tissue type
  - Treatments (drug addition, fixation)
  - Probes (fluorophores, targets)
- Acquisition
  - Microscope type
  - Magnification
- Display
  - Color mapping
  - Internal Annotations
  - Panel labels

Analysis of example paper

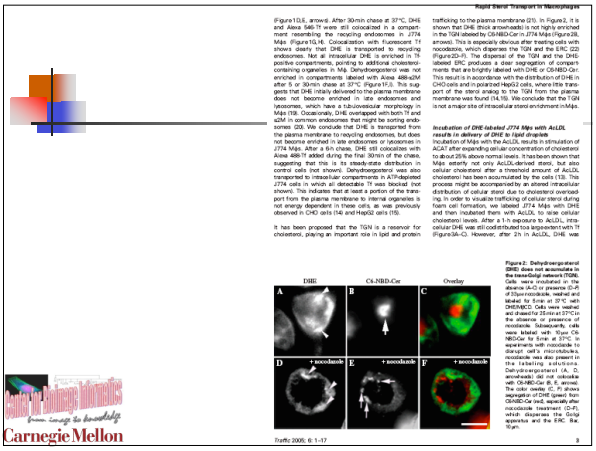
Direct Observation of Rapid Internalization and Intracellular Transport of Sterol by Macrophage Foam Cells

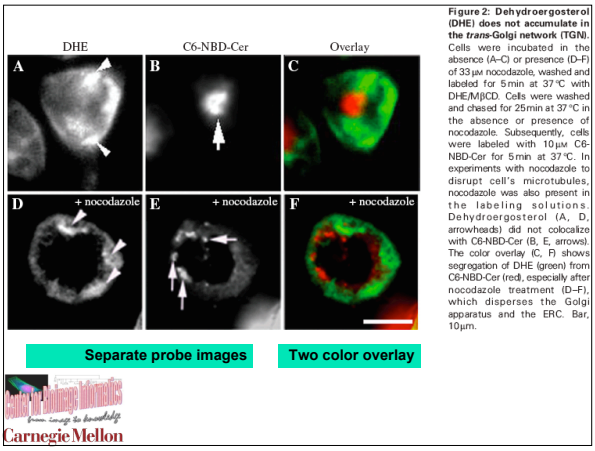
- Example gray scale image





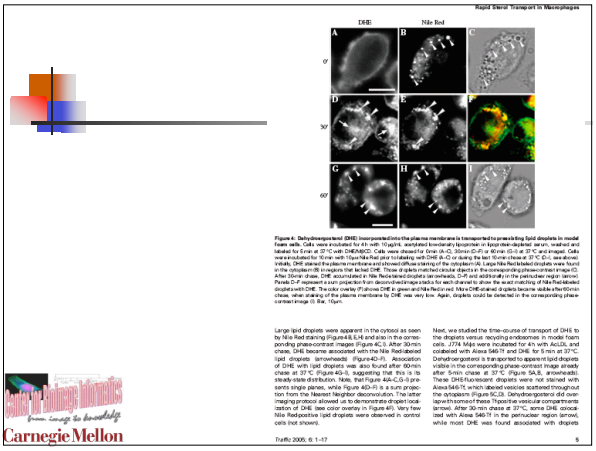
Note panel labels, arrows, text annotation, scale bars (and inference needed to infer which panels they apply to)

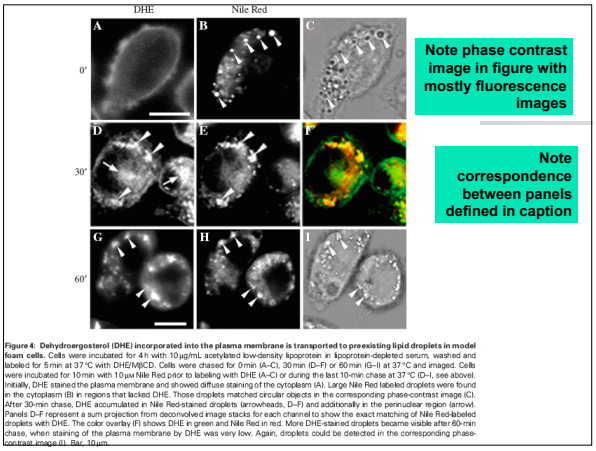




Separate probe images

Two color overlay

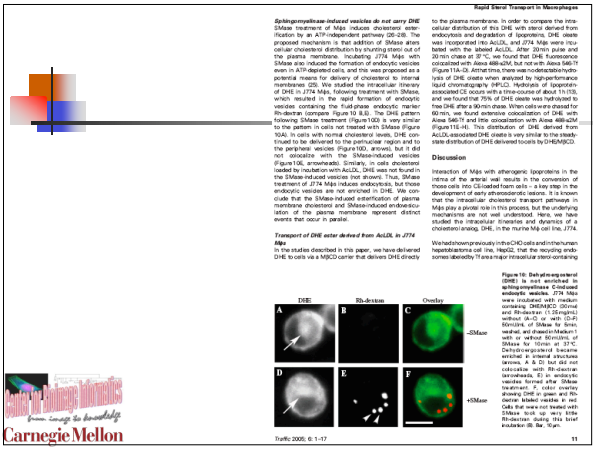


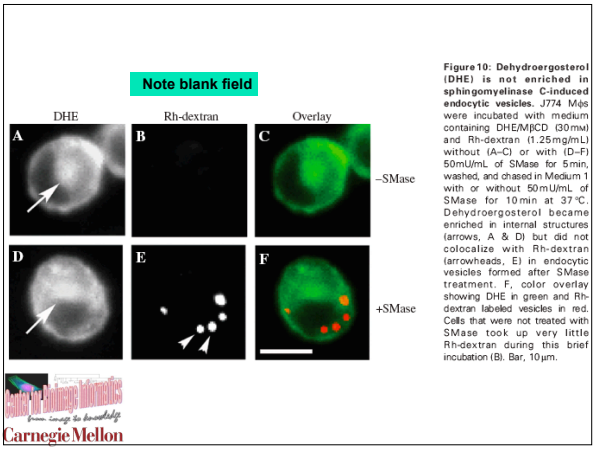


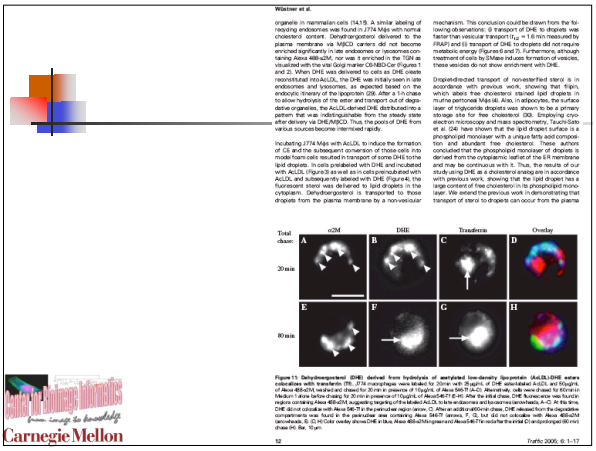
Note phase contrast image in figure with mostly fluorescence images

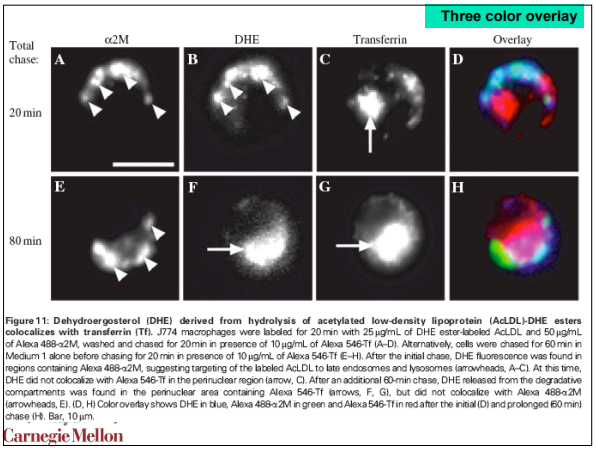
Note correspondence between panels defined in caption

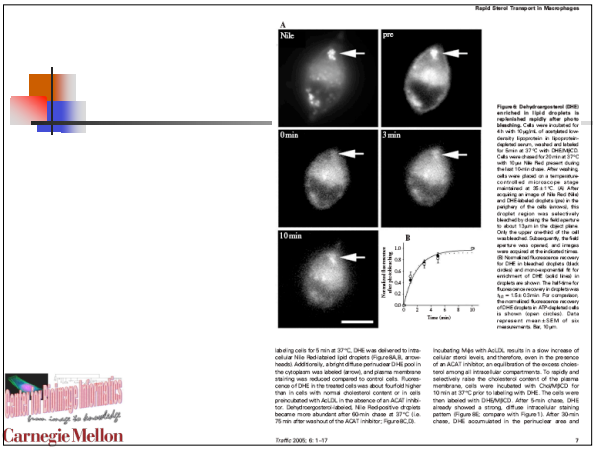
Information Extraction from Image and Text in Journal Articles

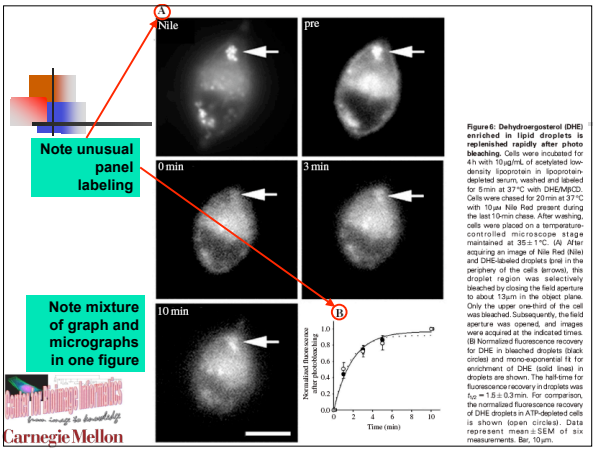


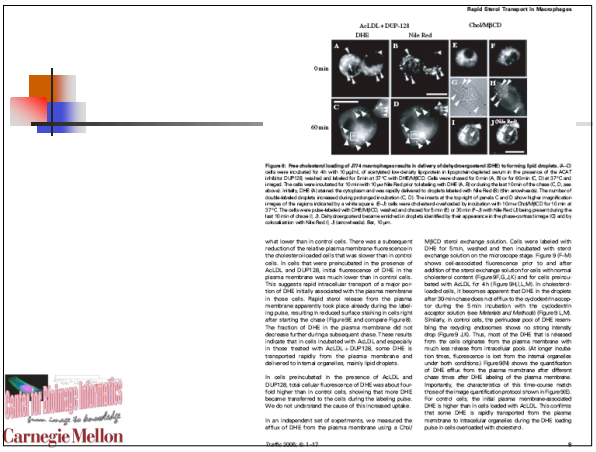












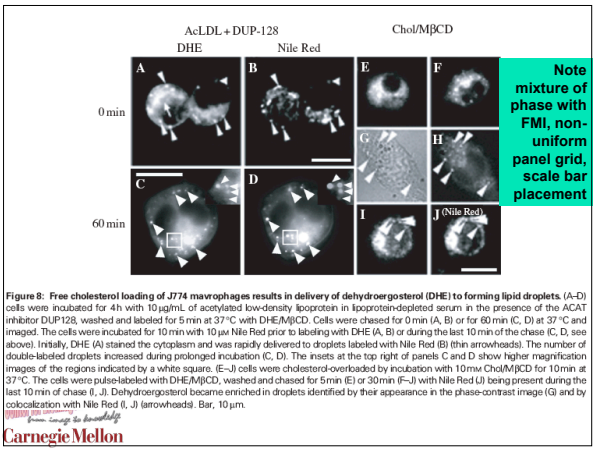


Figure 8: Free cholesterol loading of J774 macrophages results in delivery of dehydroergosterol (DHE) to forming lipid droplets. (A–D) cells were incubated for 4h with 10 μg/mL of acetylated low-density lipoprotein in lipoprotein-depleted serum in the presence of the ACAT inhibitor DUP128, washed and labeled for 5 min at 37 °C with DHE/MβCD. Cells were chased for 0 min (A, B) or for 60 min (C, D) at 37 °C and imaged. The cells were incubated for 10 min with 10 μM Nile Red prior to labeling with DHE (A, B) or during the last 10 min of the chase (C, D, see above). Initially, DHE (A) stained the cytoplasm and was rapidly delivered to droplets labeled with Nile Red (B) (thin arrowheads). The number of double-labeled droplets increased during prolonged incubation (C, D). The insets at the top right of panels C and D show higher magnification images of the regions indicated by a white square. (E–J) cells were cholesterol-overloaded by incubation with 10 mM Chol/MβCD for 10 min at 37 °C. The cells were pulse-labeled with DHE/MβCD, washed and chased for 5 min (E) or 30 min (F–J) with Nile Red (J) being present during the last 10 min of chase (I, J). Dehydroergosterol became enriched in droplets identified by their appearance in the phase-contrast image (G) and by colocalization with Nile Red (I, J) (arrowheads). Bar, 10 μm.

Inputs for automated paper interpretation

Data Sources

- All journals published electronically
- Many biological journals are open access
  - Pubmed Central collects them in one place
  - Biomed Central collection contains a number of journals in same style
- Many others have *delayed* open access
- Some have *initial* open access
- Those without open access have subscription access

Paper Formats

- All(?) journals use Publishing XML
- All provide PDF version

Biological Databases

- Many biological database containing structure information, especially about gene and protein names, sequences, structures, interactions

Basics of Supervised Machine Learning: Feature Selection and Classification

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

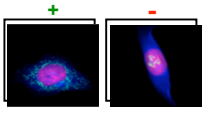
Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

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

Simple two class problem



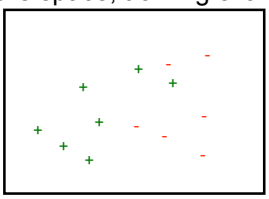
???

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

k-Nearest Heighbor (kNN)

- In feature space, training examples are

Feature #2
(e.g.., roundness)



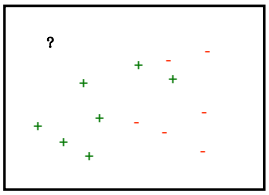
Feature #1 (e.g.., 'area')

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

k-Nearest Heighbor (kNN)

- We want to label '**?**'

Feature #2
(e.g.., roundness)

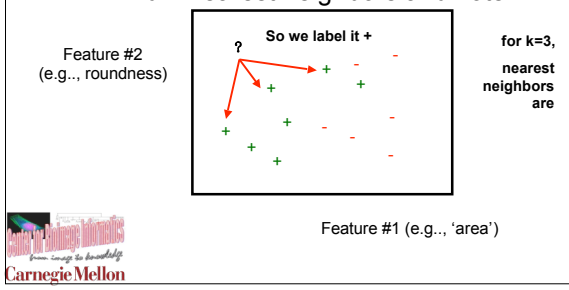


Feature #1 (e.g.., 'area')

Center for Bioimage Informatics
from image to knowledge
Carnegie Mellon

k-Nearest Heighbor (kNN)

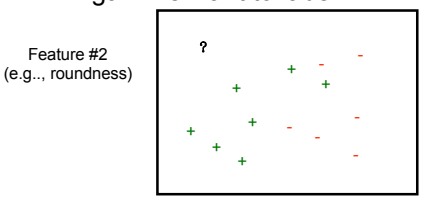
- Find k nearest neighbors and vote



Carnegie Mellon

Decision trees

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

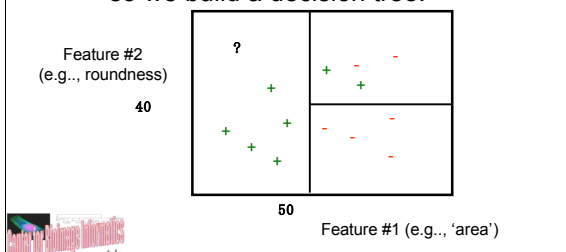


Carnegie Mellon

Slide courtesy of Christos Faloutsos

Decision trees

- so we build a decision tree:

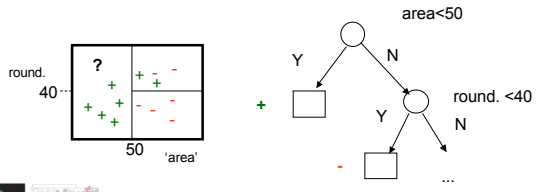


Carnegie Mellon

Slide courtesy of Christos Faloutsos

Decision trees

- so we build a decision tree:

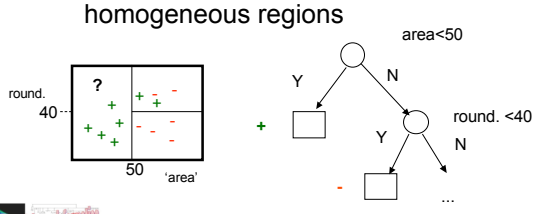


Carnegie Mellon

Slide courtesy of Christos Faloutsos

Decision trees

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

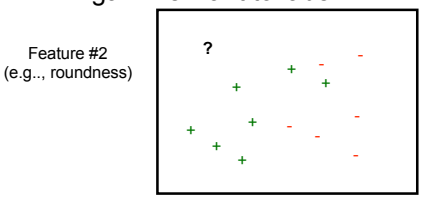


Carnegie Mellon

Slide courtesy of Christos Faloutsos

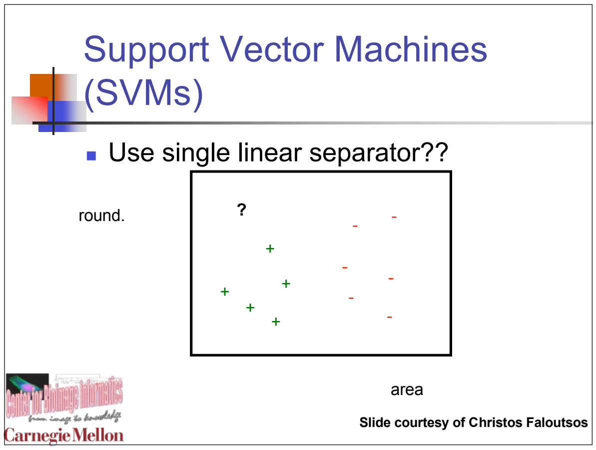
Support vector machines

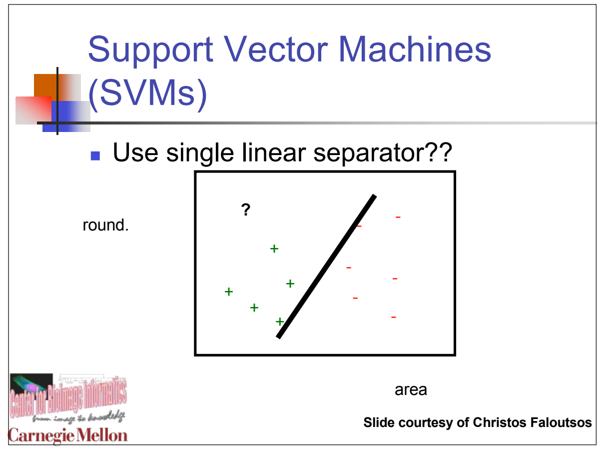
- Again we want to label '**?**'

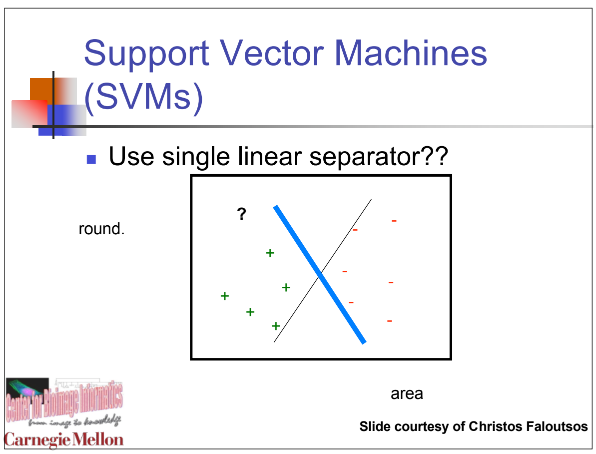


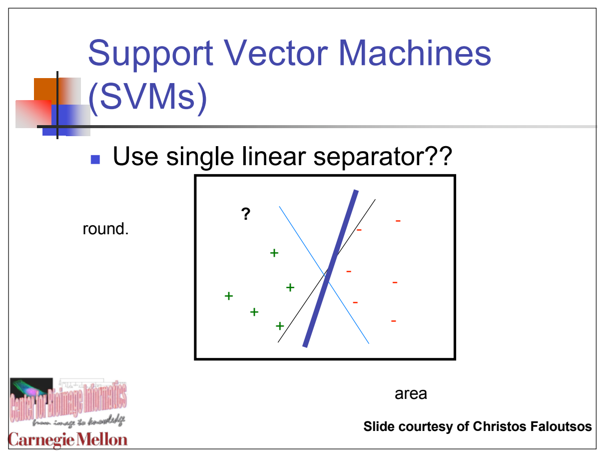
Carnegie Mellon

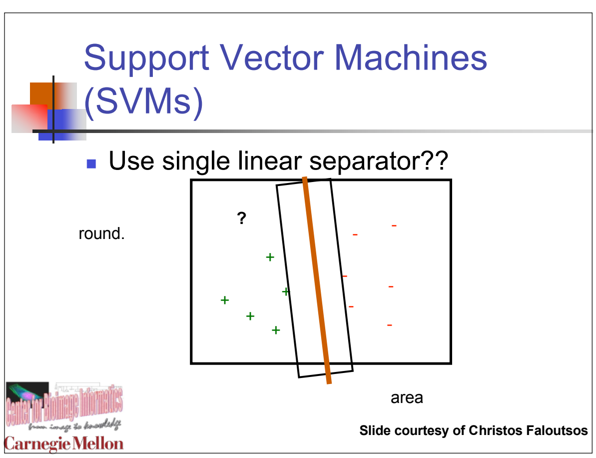
Slide courtesy of Christos Faloutsos

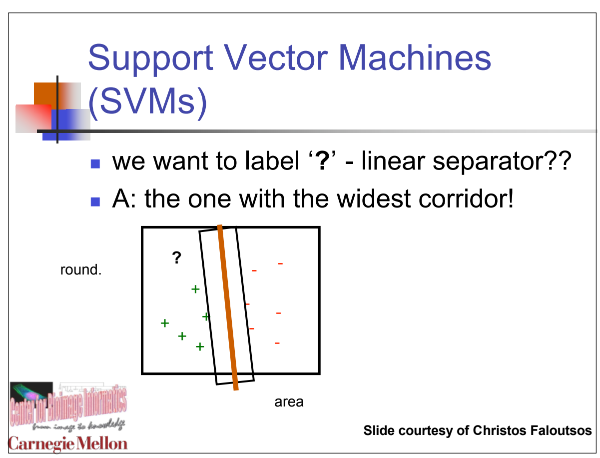

Support Vector Machines (SVMs)
Use single linear separator??
round.
?
area
Carnegie Mellon
Slide courtesy of Christos Faloutsos


Support Vector Machines (SVMs)
Use single linear separator??
round.
?
area
Carnegie Mellon
Slide courtesy of Christos Faloutsos


Support Vector Machines (SVMs)
Use single linear separator??
round.
?
area
Carnegie Mellon
Slide courtesy of Christos Faloutsos

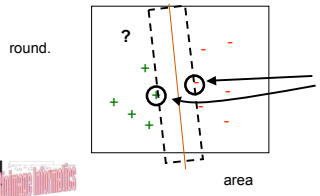

Support Vector Machines (SVMs)
Use single linear separator??
round.
?
area
Carnegie Mellon
Slide courtesy of Christos Faloutsos


Support Vector Machines (SVMs)
Use single linear separator??
round.
?
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!
round.
?
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!



Slide courtesy of Christos Faloutsos

Cross-Validation

- If we train a classifier to minimize error on a set of data, have no ability to generalize error that will be seen on new dataset
- To calculate *generalizable* accuracy, we use *n*-fold cross-validation
- Divide images into *n* sets, train using n-1 of them and test on the remaining set
- Repeat until each set is used as test set and average results across all trials

Describing classifier errors

- For multi-class classifiers, typically report
  - Accuracy = $\frac{\text{\# test images correctly classified}}{\text{\# test images}}$
- For binary classifiers (positive or negative), define
  - TP = true positives, FP = false positives
  - TN = true negatives, FN = false negatives
  - Recall = TP / (TP + FN)
  - Precision = TP / (TP + FP)
  - F-measure= 2*Recall*Precision/(Recall + Precision)

Design Issues

System structure considerations

- Even immediate goal requires complex mixture of functions to process papers
- Some functions require outputs of other functions as inputs
- Inputs and outputs may change as system evolves
- Functions may be written in different languages
- System uses and creates large number of images

System structure considerations

- Incremental nature of project argues for flexible pipeline system
  - Good choices available (not when we started SLIF project!)
- Large numbers of papers and processing times for images argue for ability to compute (or recompute) only some results
- Large numbers and sizes argue for storage of images on disk rather than inside database
- Desire for modules using heterogenous languages argues for use of scripting language to manage system

Labeling and evaluation

- Hand label as many cases as possible for each step to enable machine learning for that step and evaluation of effectiveness of each step in pipeline

SLIF design

- Preprocessing job to take PXML or PDF files and convert to "standard" organization
- Pipeline to process each paper and store results on disk and in relational database
  - Use machine learning as much as possible
- Web application to interface between user and database

SLIF Preprocessor

- Can handle small differences between input formats
- Spiders source directories
  - creating a directory for each paper it finds
  - remembering Pubmed ID for each paper
  - creating subdirectories for each figure it finds
    - extracting figure as JPEG image
    - extracting caption as plain text

SLIF Pipeline

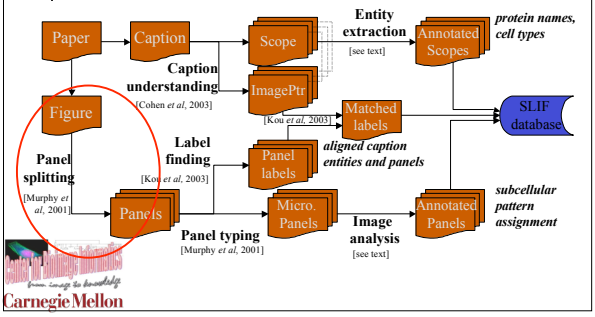
- Master Controller script in Perl
- Inputs and outputs for each module defined in terms of files that they need or create
- Controller can be asked to make any target
- Order that modules are run defined by dependencies
- Processing of each paper independent so compute cluster can be used for collection
- Results stored in Postgresql database

SLIF Web Application

- Java Server Pages to define queries and display results
- Programmatic access support through modifiers on URL
  - SOAP interface written and being tested

SLIF Pipeline

SLIF Pipeline components



Carnegie Mellon

Panel Splitting [image]

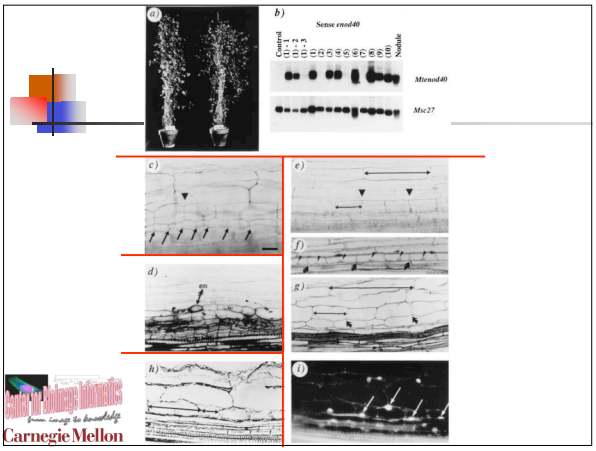
- Difficult task in general case
- SLIF focuses on images, so chose approach with high precision and recall for images
- Recursive detection of light areas between panels with trimming

Carnegie Mellon

Panel Splitting [image]

- Find horizontal or vertical line through figure with lowest average intensity
- If lowest is above threshold, stop
- Cut figure into two pieces
- Trim horizontal or vertical lines from edges of pieces if those lines have average intensity close to white or black
- If piece too small, discard
- Recurse on resulting pieces

Carnegie Mellon



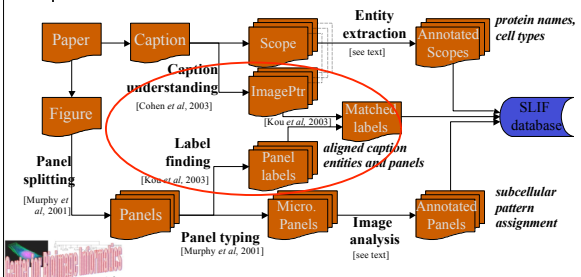
Carnegie Mellon

Semi-automated labeling tool

1. Initialize list of previously labeled results to empty; Initialize panel splitter parameters
2. Run initial panel splitter on some figures; output is coordinates of each putative panel
3. Compare to list of previously labeled putative panels
4. If match, assign previous label (correct or incorrect)
5. If not, display figure/panel and get label
6. If desired, change algorithm/parameters and go to step 2
7. Run again on new set of figures and just save initial results as unbiased estimate of accuracy

Carnegie Mellon

SLIF Pipeline components

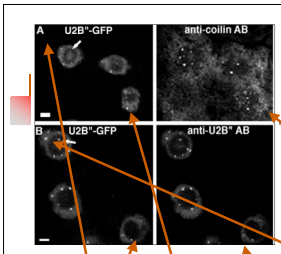


Carnegie Mellon

Image pointer detection [text]

- Parse caption using set of rules to identify potential image pointers
- Single letters followed by period or comma
- Single letters or short phrases followed or surrounded by parentheses or brackets

Carnegie Mellon



Identify all *image pointers:* Substrings that refer to parts of the image

Figure 1. (A) Single confocal optical section of BY-2 cells expressing U2B 0-GFP, double labeled with GFP (left panel) and autoantibody against p80 coilin (right panel). Three nuclei are shown, and the bright GFP spots colocalize with bright foci of anti-coilin labeling. There is some labeling of the cytoplasm by anti-p80 coilin. (B) Single confocal optical section of BY-2 cells expressing U2B 0 -GFP, double labeled with GFP (left panel) and 4G3 antibody (right panel). Three nuclei are shown. Most coiled bodies occur in the nucleoplasm, but occasionally are seen in the nucleolus (arrows). All coiled bodies that contain U2B 0 also express the U2B 0-GFP fusion. Bars, 5 m m. Movement of Coiled Bodies Vol. 10, July 1999 2299

Carnegie Mellon

Identifying Image Pointers: Learning *vs* Hand-coded Heuristics

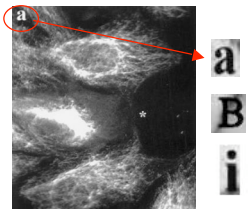
	HC-1	HC-2	ABWI (W=2)	ABWI + NA	SABWI + NA
Precis.	98.5	74.5	89.7	85.9	88.6
Recall	45.6	98.0	91.0	92.2	93.8
F1	62.3	84.6	90.3	89.0	**91.1**

Hand-coded methods **Learned filters on hand-coded candidate generator**

Carnegie Mellon

Panel Label Finding [image and text]

- *Finding* annotations is not difficult
  - look for sharp edges,etc
- *Interpreting* annotations (what letter is it?) is hard
  - complex backgrounds
  - partially occluded letters
- Method:
  - find *candidate* regions (using position & size)
  - enhance, rescale, binarize
  - apply OCR to regions
  - match possible label **patterns** to labels from text



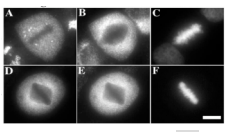
Carnegie Mellon

Label matching

[Kou *et al*, BioKDD 2003]

- Labels from caption (sorted): ABCDEF
- OCR candidate patterns, based on layout
  - ADB_GF (column-major)
  - ABGD_F (row-major)
- Closest match by dynamic programming:
  - ABGD_F ~ ABCDEF

C E



A B **C**

D **E** F

correct using best alignment with respect to Needleman-Wunsch edit distance, using model of common OCR errors to set weights

Carnegie Mellon

Evaluation

[Kou *et al*, BioKDD 2003]

# panels	# text regions
427	380

OCR directly on panels		
#	Prec.	Recall
15	3.9%	3.5%

OCR on intensity-normalized text regions		
#	Prec.	Recall
271	71.3%	63.5%

OCR on enhanced text regions		
#	Prec.	Recall
302	79.1%	70.7%

OCR on enhanced text regions, after string-match corrections		
#	Prec.	Recall
316	83.2%	74.0%

Carnegie Mellon

Limitations

- Only looks for labels *within* panels (misses labels next to panel)
- Can't assign same label to *set* of panels
- Only recognizes single letter labels (does not recognize "control")

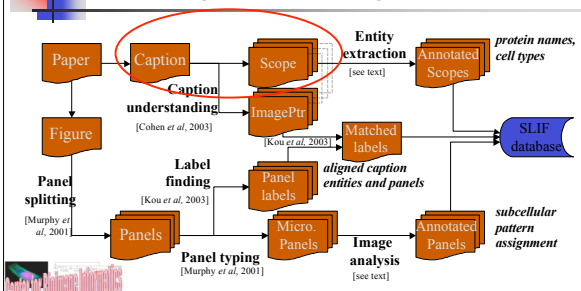
Annotation removal [image]

- All candidate annotations (including panel labels) are removed (set to background)
- Future: could define filters to recognize non-alpha symbols (arrows)

Scale bar finding [image and text]

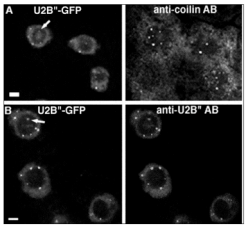
- In image, look for solid, horizontal black or white bars
- In text, look for strings of form "(Bb)ar" followed by number followed by "m"
- Assume number is in μm (microns)
- Scale in microns per pixel is number divided by length of bar in pixels

SLIF Pipeline components



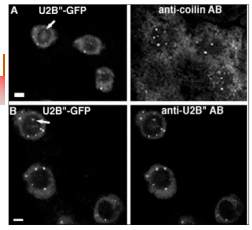
Caption scoping [text]

- Goal is to try to determine which words in the caption refer to which parts of the figure



Classify image pointers as **citation-style** or **bullet-style.**

Figure 1. A) Single confocal optical section of BY-2 cells expressing U2B 0-GFP, double labeled with GFP (left panel) and autoantibody against p80 coilin (right panel). Three nuclei are shown, and the bright GFP spots colocalize with bright foci of anti-coilin labeling. There is some labeling of the cytoplasm by anti-p80 coilin. (B) Single confocal optical section of BY-2 cells expressing U2B 0 -GFP, double labeled with GFP (left panel) and 4G3 antibody (right panel). Three nuclei are shown. Most coiled bodies also in the nucleoplasm, but occasionally are seen in the nucleolus (arrows). All coiled bodies that contain U2B 0 also express the U2B 0-GFP fusion. Bars, 5 m m. Movement of Coiled Bodies Vol. 10, July 1999 2299

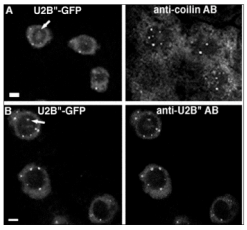


Style determines **scope:**
- The *scope* of a **bullet-style** image pointer is all words between it and the next "bullet"

Figure 1. (A) Single confocal optical section of BY-2 cells expressing U2B 0-GFP, double labeled with GFP against p80 coilin (right panel). Three nuclei are shown, a[illegible]calize with bright foci of anti-coilin labeling. There is some labeling of the cytoplasm by anti-p80 coilin. (B) Single confocal optical section of BY-2 cells expressing U2B 0 -GFP, double labeled with GFP (left panel) and 4G3 antibody (right panel). Three nuclei are shown. Most coiled bodies [illegible] in the nucleoplasm, but occasionally are seen in the nucleolus (arrows). All coiled bodies that contain U2B 0 also express the U2B 0-GFP fusion. Bars, 5 m m. Movement of Coiled Bodies Vol. 10, July 1999 2299

scope of (A)

Carnegie Mellon

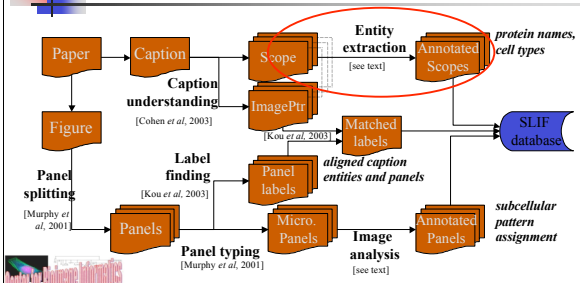


Style determines **scope:**
- The *scope* of a **bullet-style** image pointer is all words between it and the next "bullet"
- The scope of a **citation-style** image pointer is some set of words nearby it (heuristically determined by separating words and punctuation)

Figure 1. (A) Single confocal optical section of BY-2 cells expressing U2B 0-GFP, double labeled with GFP (left panel) and autoantibody against p80 coilin (right panel). Three nuclei are shown, and the bright GFP spots colocalize with bright foci of anti-coilin labeling. There is some labeling of the cytoplasm by anti-p80 coilin. (B) Single confocal optical section of BY-2 cells expressing U2B 0 -GFP, double labeled with GFP (left panel) and 4G3 antibody (right panel). Three nuclei are shown. Most coiled bodies [illegible] in the nucleoplasm, but occasionally are seen in the nucleolus (arrows). All coiled bodies that contain U2B 0 also express the U2B 0-GFP fusion. Bars, 5 m m. Movement of Coiled Bodies Vol. 10, July 1999 2299

Carnegie Mellon

SLIF Pipeline components



Carnegie Mellon

Named entity recognition (NER) [text]

- Need to match results of image analysis of panel contents with words describing the image
- Name of protein visualized, cell type used, etc.
- Very hard task because names of biological entities not used consistently

Carnegie Mellon

Protein Name Recognition

Two potentially oncogenic cyclins, cyclin A and cyclin D1, share common properties of subunit configuration, tyrosine phosphorylation and physical association with the Rb protein.

Carnegie Mellon

Protein Name Recognition

Two potentially oncogenic cyclins, cyclin A and cyclin D1, share common properties of subunit configuration, tyrosine phosphorylation and physical association with the Rb protein.

Carnegie Mellon

Use cases

- Possible query: "find all images of *some protein involved in ribosome assembly* that appears to be located in the cytoplasm"
  - "Proteins involved in ribosome assembly" determined by membership in a database (eg PIR,…)
- A high recall protein name extractor is preferred
- We care most about proteins from databases of all known proteins

Carnegie Mellon

Dictionary based algorithms for protein name recognition

Two potentially oncogenic cyclins, cyclin A and cyclin D1…..

Dictionary	Pattern Dictionary
alpha tubulin	Greek tubulin
…	…
cyclin A	cyclin CapitalLetter
cyclin D1	cyclin CapitalLetter+Digit
……	……

Carnegie Mellon

Problems with dictionary based algorithms

- Words in a dictionary may not always be proteins, particularly after generalization to a pattern (e.g., "AT", "fragment", …)
  - Dictionaries must be first *curated* by removing such words
  - Constructing patterns requires engineering

Two potentially oncogenic cyclins, cyclin A and cyclin D1…..

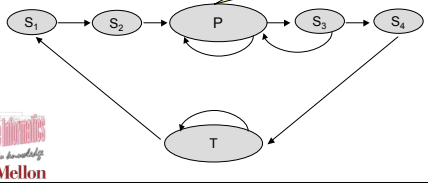
Dictionary	Pattern Dictionary
alpha tubulin	Greek tubulin
…	…
cyclin A	cyclin CapitalLetter
cyclin D1	cyclin CapitalLetter+Digit
……	……

Carnegie Mellon

Context based algorithms

- Context based algorithms a… machine learning algorithms
- Hidden Markov Models(HM… extract names from text

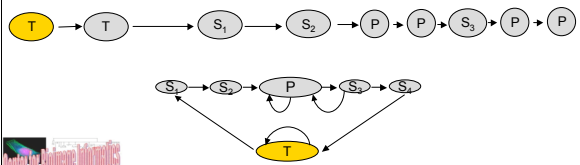
alpha	.000002
tubulin	.000013
…	
cyclin	.000009
……	



Carnegie Mellon

An HMM for protein name extraction

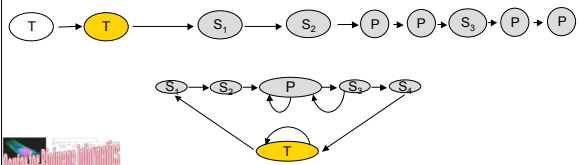
Two potentially oncogenic cyclins, cyclin A and cyclin D1



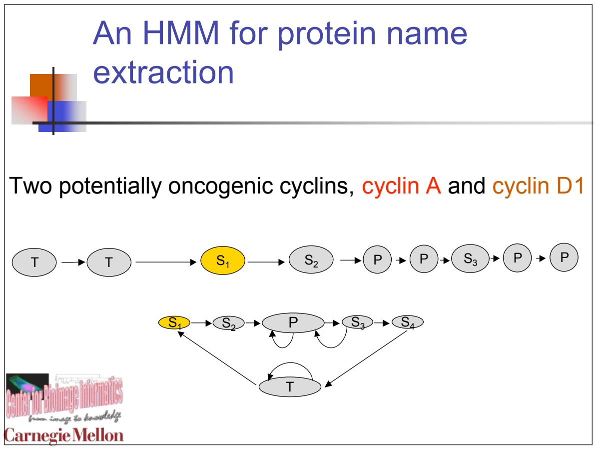
Carnegie Mellon

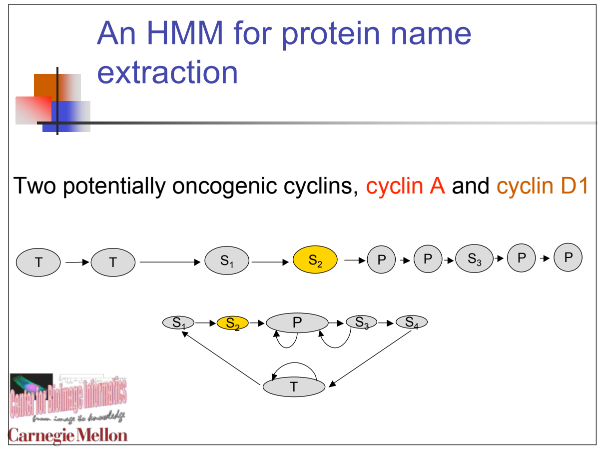
An HMM for protein name extraction

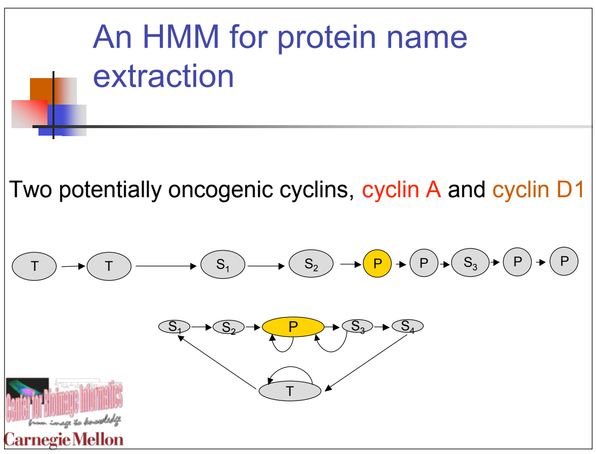
Two potentially oncogenic cyclins, cyclin A and cyclin D1

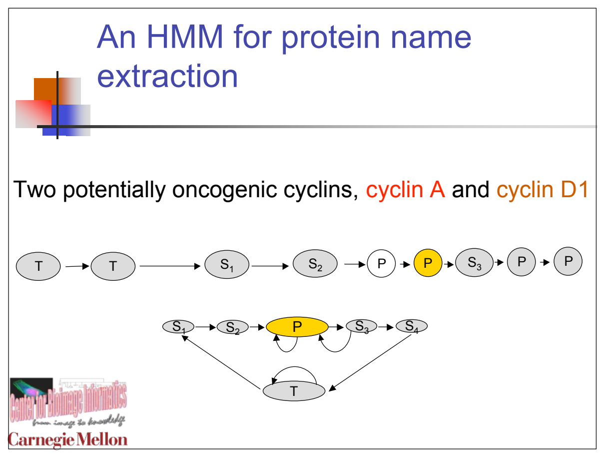


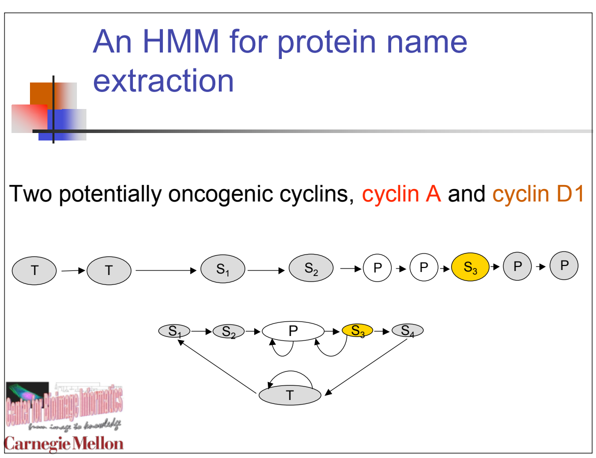
Carnegie Mellon

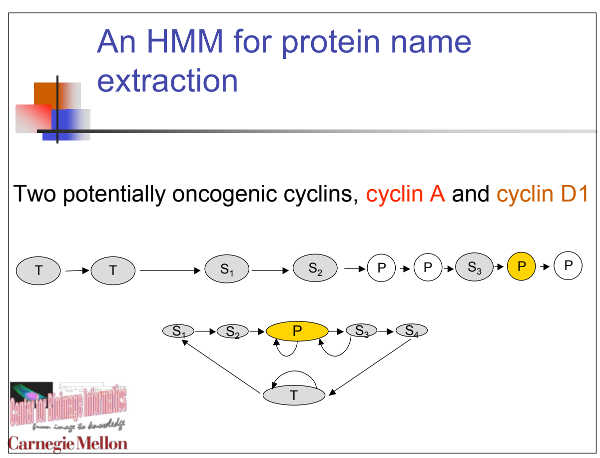

An HMM for protein name extraction
Two potentially oncogenic cyclins, cyclin A and cyclin D1
Carnegie Mellon


An HMM for protein name extraction
Two potentially oncogenic cyclins, cyclin A and cyclin D1
Carnegie Mellon


An HMM for protein name extraction
Two potentially oncogenic cyclins, cyclin A and cyclin D1
Carnegie Mellon

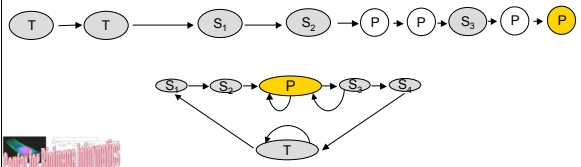

An HMM for protein name extraction
Two potentially oncogenic cyclins, cyclin A and cyclin D1
Carnegie Mellon


An HMM for protein name extraction
Two potentially oncogenic cyclins, cyclin A and cyclin D1
Carnegie Mellon


An HMM for protein name extraction
Two potentially oncogenic cyclins, cyclin A and cyclin D1
Carnegie Mellon

An HMM for protein name extraction

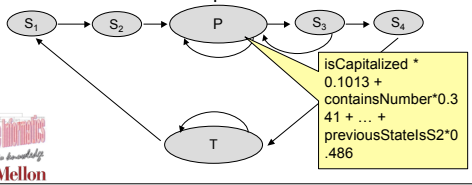
Two potentially oncogenic cyclins, cyclin A and cyclin D1



Carnegie Mellon

Discriminative versions of HMMS (CRFs, MEMMs/MaxEnt Taggers)

- New HMM-like methods:
  - Each token can have many features associated with it (isCapitalized, containsNumber, containsGreekLetter) as well as an "identity" ("alpha-3")
  - State is predicted with a linear weighting scheme that considers features and previous state

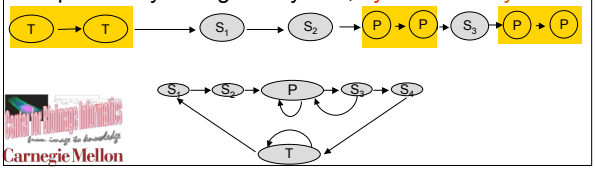


Carnegie Mellon

SemiCRFs

- Semi-markov version of CRFs
- Viterbi search replaced with search for best sequence of ***segments***
- Distance to dictionary is feature of segments

Two potentially oncogenic cyclins, cyclin A and cyclin D1



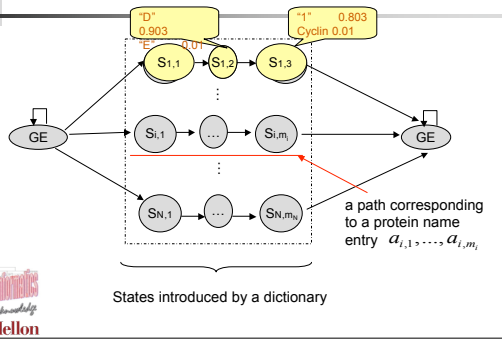
Carnegie Mellon

Combining a dictionary with a hidden Markov model (Dictionary-HMM)

- Dictionary based algorithms can take advantage of existing resources, such as protein names in PIR database
- Context based algorithms do not in principle need updating
- Dictionary-HMM: learn how to do a soft match based on a small number of training data

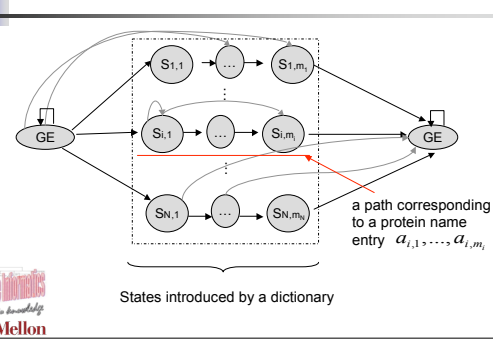
Carnegie Mellon

Combining a dictionary with a hidden Markov model (Dictionary-HMM)



Carnegie Mellon

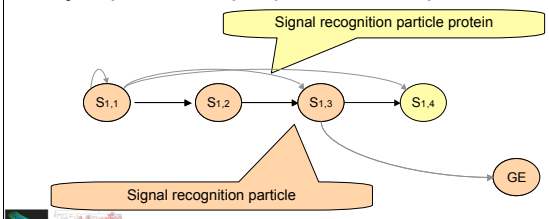
Combining a dictionary with a hidden Markov model (Dictionary-HMM)



Carnegie Mellon

Soft match to a path

With jumps and loops, path is like a profile-HMM

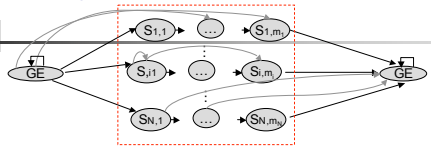


Dictionary-HMM

We need to specify:

- Structure: states and transitions
- Alphabet: set of emissions
- Initial Probability, Transition matrix, Emission matrix

Building the structure of the dictionary- HMM

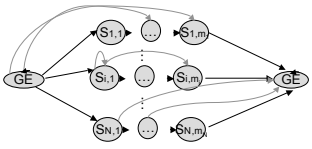


- Strategies of introducing paths
  - Integrate the whole dictionary: huge structure will bring huge transition and emission matrix
  - Use heuristics to choose a small number of likely paths

Building the alphabet

- Emissions we have
  - Tokens from training data
  - Tokens from dictionary
    - Subsampling to avoid too many emitted words
  - Unknown token

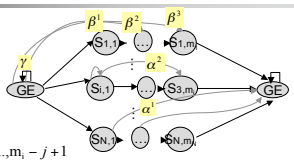
Initial probability



- Initial probability
  - Learn from data
  - $\pi\{GE\} = \pi_0$
  - $\pi\{S_{i,1}\} = (1 - \pi_0) / N$

Transition matrix A:

- *Depends on a small number of parameters a,b,g*



$$P(S_{i,j+k} \mid S_{i,j}) = \frac{\alpha^k}{Z_1}$$

$k=1,\ldots,m_i - j + 1$

$$P(GE \mid S_{i,m_i}) = 1$$

$$P(GE \mid GE) = \gamma$$

$$P(S_{i,k} \mid GE) = \frac{\beta^k}{Z_2}\frac{(1-\gamma)}{N}$$

$$0 < \alpha, \beta, \gamma < 1$$

Z_1, Z_2 are for normalization, N is the number of paths

Emission matrix B

- P(W_l|GE): estimate from the training data
  - $P(W_l \mid S_{i,j}) = \delta\, P(W_l \mid Dict)$ — W_l is a word only in the dictionary of protein names, except $a_{i,1}, \ldots, a_{i,m_i}$
  - $P(W_l \mid S_{i,j}) = \varepsilon\, P(W_l \mid GE)$ — W_l is a word only observed in GE
  - $P(a_{i,l} \mid S_{i,j}) = \frac{1-\varepsilon-\delta}{m_i}$ — $a_{i,l}$ is any token in $a_{i,1}, \ldots, a_{i,m_i}$

Carnegie Mellon

Learning the parameters

- EM approach based on Baum-Welch
  - E-step: run B-W on the test data to learn A, B, then estimate the average parameters $\alpha, \beta, \gamma, \varepsilon, \delta$ from A, B.
  - M-step: Use these estimated $\alpha, \beta, \gamma, \varepsilon, \delta$ to recalculate A,B

Carnegie Mellon

Experiments

- Available datasets
  - Univ. of Texas: 700 Medline abstracts
  - GENIA 3.04: 2000 Medline abstracts
  - Yapex: 200 Medline abstracts
- None of these is completely appropriate for us
  - Contains non-dictionary as well as dictionary proteins
- Baseline methods
  - CRFs, MaxEnt
  - Competitive previously published method on same dataset
- Features (for CRF,MaxEnt) and tokenization (for dictHMM)

Carnegie Mellon

Performance of different algorithms on different datasets

	Precision/Recall/F-measure (%)		
	U. of Texas	GENIA	YAPEX
Previously published methods	73.4 / 47.8 / 57.9 (Bunescu et al., 2004)	49.2 / 66.4 / 56.5 (Kazama, et al., 2002)	67.8/ 66.4/**67.1** (Franzén, et al., 2002)
Bunescu's dictionary-based method	62.3 / 45.9 / 52.8 (Bunescu et al., 2004)	-	-
MaxEnt	87.2 / 57.3 / 69.1	67.3 / 65.4 / 66.2	69.3/ 58.1/ 63.2
CRFs	83.5 / 66.1 / 73.8	75.0 / 67.6 / 71.1	76.0/ 59.5/ 66.7
SemiCRFs	83.1 / 66.8 / **73.9**	74.8 / 68.3 / **72.3**	76.1/ 58.9/ 66.1
Dict-HMM	46.0 / 69.2 / 55.2	44.8 / 70.1 / 54.7	42.4/ 64.1/ 51.0

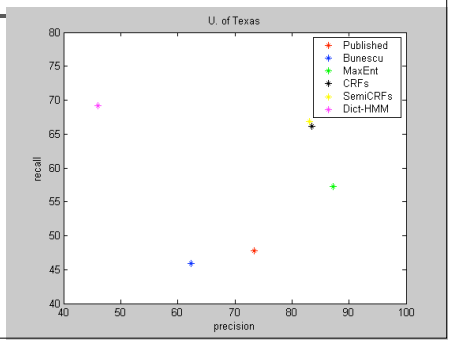
Carnegie Mellon

Performance of different algorithms on different datasets

	Precision/Recall/F-measure (%)		
	U. of Texas	GENIA	YAPEX
Previously published methods	73.4 / 47.8 / 57.9 (Bunescu et al., 2004)	49.2 / 66.4 / 56.5 (Kazama, et al., 2002)	67.8/ 66.4/**67.1** (Franzén, et al., 2002)
Bunescu's dictionary-based method	62.3 / 45.9 / 52.8 (Bunescu et al., 2004)	-	-
MaxEnt	87.2 / 57.3 / 69.1	67.3 / 65.4 / 66.2	69.3/ 58.1/ 63.2
CRFs	83.5 / 66.1 / 73.8	75.0 / 67.6 / 71.1	76.0/ 59.5/ 66.7
SemiCRFs	83.1 / 66.8 / **73.9**	74.8 / 68.3 / **72.3**	76.1/ 58.9/ 66.1
Dict-HMM	46.0 / 69.2 / 55.2	44.8 / 70.1 / 54.7	42.4/ 64.1/ 51.0

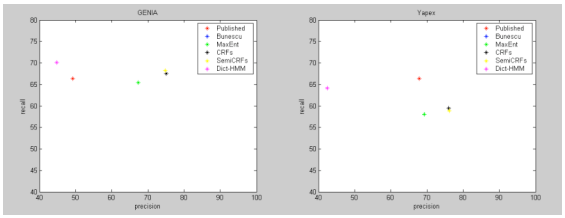
Carnegie Mellon

Performance on U. of Texas dataset



Carnegie Mellon

Performance on GENIA and Yapex datasets



Carnegie Mellon

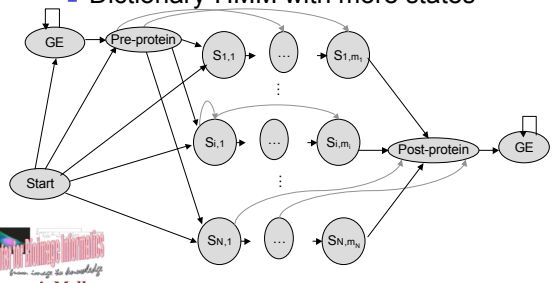
Two strategies to improve the Dictionary-HMM (1)

- Boosting-like strategy:
  - Step 1. build a Dictionary-HMM on a test sentence. If no protein found, end.
  - Step 2. learn the dictionary-HMM and calculate the optimal state sequence. Find the single protein path with highest likelihood and report it.
  - Step 3. remove the protein found in step 2 from test sentence. Go to step 1 with the reduced test sentence.

Carnegie Mellon

Two strategies to improve the Dictionary-HMM (2)

- Dictionary-HMM with more states



Carnegie Mellon

Performance of improved Dict-HMMs

	Precision/Recall/F-measure (%)		
	U. of Texas	GENIA	YAPEX
CRFs	83.5 / 66.1 / 73.8	75.0 / 67.6 / 71.1	76.0/ 59.5/ **66.7**
SemiCRFs	83.1 / 66.8 / **73.9**	74.8 / 68.3 / **72.3**	76.1/ 58.9/ 66.1
Dict-HMM	46.0 / 69.2 / 55.2	44.8 / 70.1 / 54.7	42.4/ 64.1/ 51.0
Dict-HMM + boosting-like method	49.8 / 74.3 / 59.6	48.3 / 73.9 / 58.5	45.1/ 69.7/ 54.8
Dict-HMM + additional states	51.8 / 72.3 / 60.4	51.3 / 72.4 / 60.1	45.1/ 65.7/ 53.5

Carnegie Mellon

Performance on words that match dictionary

- Many putative protein names by CRFs or semiCRFs are poor matches to dictionary entries
- Can measure similarity of a putative name to its closest match in dictionary using TFIDF (term frequency * inverse document frequency)
- Calculate as number of words in common divided by total number of words in both (weighted by frequency of words overall)
- Examine only putative protein names with TFIDF score greater than 0.9

Carnegie Mellon

Evaluation for protein names with TFIDF > 0.9

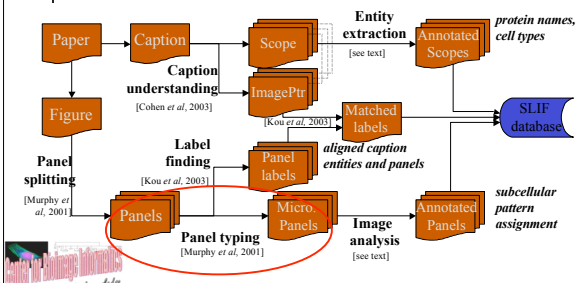
	Precision/Recall/F-measure (%)		
	U. of Texas	GENIA	YAPEX
CRFs	84.7 / 68.5 / 75.7	76.9 / 67.3 / 71.8	78.5/ 60.3/ 68.2
SemiCRFs	85.3 / 69.8 / 76.8	77.9 / 73.6 / 75.7	80.1/ 61.9/ 69.8
Dict-HMM	69.1 / 99.3 / **81.5**	65.8 / 98.7 / **79.0**	64.3/ 100/ **78.3**

Carnegie Mellon

Conclusions

- SemiCRFs have higher precision, lower recall
- Dictionary-HMM has higher recall, lower precision
- Dictionary-HMMs have high recall for dictionary-like protein names

SLIF Pipeline components



Panel typing [image and text]

- Goal is to identify the general type of each panel
- Possibilities are graph, cartoon, electron micrograph, light micrograph, fluorescence micrograph, gel picture

Observations/Assumptions

- Graphs and cartoons have very high contrast (black on white)
- Electron micrographs and light micrographs have gray background and little contrast
- Fluorescence micrographs and gel pictures have near black backgrounds and full range of gray levels

Initial approach (2001)

- Downloaded PDF files from Pubmed Central
- Extracted figures, split into panels
- Labeled 1586 panels as either FMI or non-FMI by viewing panel
- Made 64-bin histogram of gray levels for each panel

Initial approach

- Used 64 values as features to "train" *k*-nearest neighbor classifier for FMI vs. non-FMI
- Used labeled examples with leave-one-out cross validation to choose best *k*
- Calculate number of neighbors that are FMI
- Choose threshold on this number to trade precision vs. recall

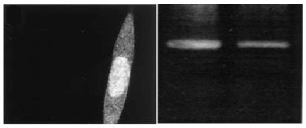
Initial approach

- Best *k* was 9
  - Obtained recall of 70% and precision of 100% for high threshold
  - Obtained recall of 92% and precision of 97% for lower threshold
- Tested for another set of 100 panels
  - For *k*=11 and *T*=5, obtained recall of 90% and precision of 100%

Carnegie Mellon

Second approach

- For new collection of figures from PNAS, precision not as good (~50%)
- Especially observed gel pictures frequently being classified as FMI



Carnegie Mellon

Second approach

- Labeled 1993 panels (one panel each from 898 figures and all panels from 175 figures)
- Displayed both figure and caption during labeling to increase accuracy
- Initial labeling by one person, checked by another
- 41% were FMI, 19% were gels

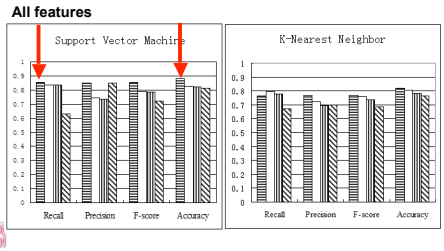
Carnegie Mellon

Second approach

- Calculated 64 histogram features
- Added 7 edge features measuring fraction of edge, homogeneity of edge direction and horizontal and vertical edge content
- Added "bag of words" text features
- One feature for each word found in all of the training examples (20,767 words)
- For each panel, words in the scope of that panel and words in the scope of the entire caption were counted

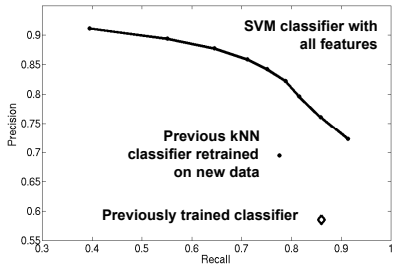
Carnegie Mellon

Performance with different feature sets



Carnegie Mellon

Second approach



Carnegie Mellon

Cotraining

Experiments		Recall	Precision	Error Rate
50% training	SVM	0.829	0.836	0.132
	Co-training	0.826	0.828	0.137
10% training	SVM	0.561	0.791	0.229
	Co-training	0.666	0.849	0.179

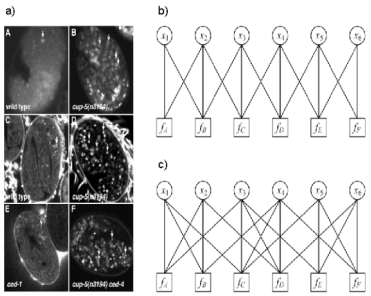
Carnegie Mellon

Cotraining

- Conclusion is that representation of classes among labeled examples is good

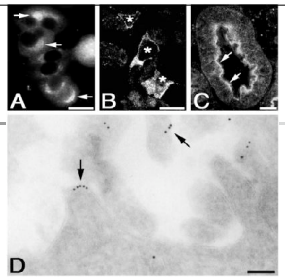
Carnegie Mellon

Graphical model classification



Carnegie Mellon

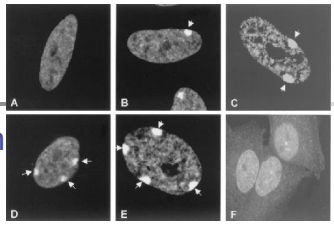
Example where original classes correct



Actual panel class	Initial label probabilities		Final label probabilities ($\alpha = 0, \lambda = 2$)		Final label probabilities ($\alpha = 0.5, \lambda = 2$)		Final label probabilities ($\alpha = 1, \lambda = 2$)	
	FMI	Non-FMI	FMI	Non-FMI	FMI	Non-FMI	FMI	Non-FMI
FMI	0.740	0.260	0.838	0.162	0.882	0.119	0.883	0.117
FMI	0.704	0.296	0.809	0.191	0.835	0.165	0.863	0.138
FMI	0.695	0.305	0.800	0.200	0.762	0.238	0.742	0.258
Non-FMI	0.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000

Carnegie Mellon

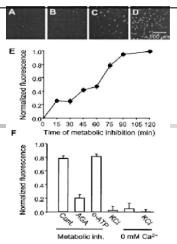
Example with one panel wrong



Actual panel class	Initial label probabilities		Final label probabilities ($\alpha = 0, \lambda = 2$)		Final label probabilities ($\alpha = 0.5, \lambda = 2$)		Final label probabilities ($\alpha = 1, \lambda = 2$)	
	FMI	Non-FMI	FMI	Non-FMI	FMI	Non-FMI	FMI	Non-FMI
FMI	0.792	0.209	0.958	0.042	0.946	0.054	0.938	0.062
FMI	0.784	0.216	0.956	0.044	0.948	0.052	0.946	0.054
FMI	0.718	0.282	0.939	0.061	0.928	0.072	0.921	0.079
FMI	0.796	0.204	0.959	0.042	0.942	0.058	0.932	0.068
FMI	0.731	0.269	0.925	0.075	0.916	0.085	0.926	0.074
FMI	**0.492**	**0.508**	**0.797**	**0.203**	**0.726**	**0.274**	**0.672**	**0.328**

Carnegie Mellon

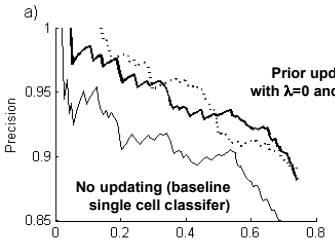
Example with panels of different classes and one wrong



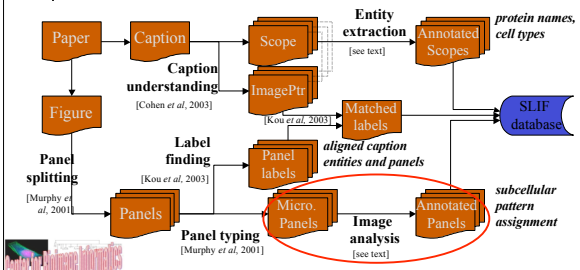
Actual panel class	Initial label probabilities		Final label probabilities ($\alpha = 0, \lambda = 2$)		Final label probabilities ($\alpha = 0.5, \lambda = 2$)		Final label probabilities ($\alpha = 1, \lambda = 2$)	
	FMI	Non-FMI	FMI	Non-FMI	FMI	Non-FMI	FMI	Non-FMI
FMI	0.877	0.123	0.947	0.053	0.955	0.045	0.972	0.028
FMI	0.869	0.131	0.940	0.060	0.953	0.047	0.974	0.026
FMI	0.810	0.190	0.905	0.095	0.917	0.083	0.940	0.060
FMI	**0.491**	**0.509**	**0.664**	**0.336**	**0.667**	**0.333**	**0.675**	**0.325**
Non-FMI	0.038	0.962	0.045	0.955	0.025	0.975	0.006	0.994
Non-FMI	0.018	0.982	0.004	0.996	0.003	0.997	0.003	0.997

Carnegie Mellon

Precision Recall Analysis



SLIF Pipeline components



Pattern classification [image]

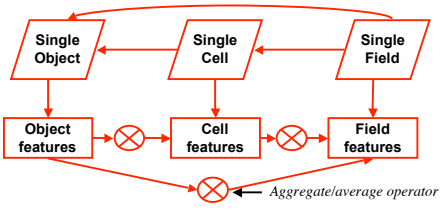
- For each panel that has an identified scale bar, calculate subset of Subcellular Location Features that do not require segmentation into single cells

Approaches to classify protein patterns

- Fluorescence micrographs can contain subcellular region, single cell, or multiple cells/tissues

Approaches to classify protein patterns

- Features can be calculated at each level and aggregated to higher levels

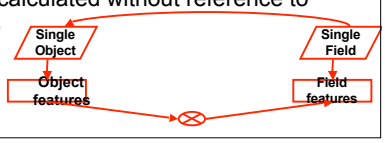


Approaches to classify protein patterns

- Analyzing patterns at single cell level requires segmenting multi-cell images
- Not easy in general case (algorithms usually customized to type of data available)

Field-level classification

- Alternative: assume entire field has same subcellular pattern (mostly true)
- Use features that
  - don't require cell segmentation
  - are not sensitive to number of cells in field
  - can be calculated without reference to nucleus



Carnegie Mellon

Field-level classification

- Object features (object size, shape)
- Edge features
- Texture features

Carnegie Mellon

Scale normalization

- Images in figures have widely varying scales
- Use of features for classification requires scale to be the same
- Can use pixel size to rescale images to common size

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

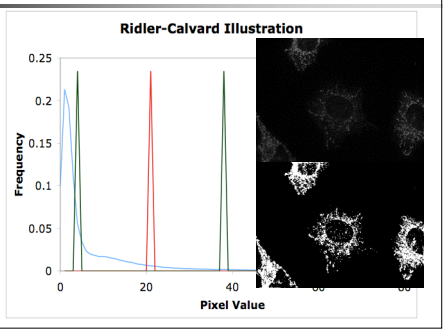
Ridler-Calvard Method

- 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

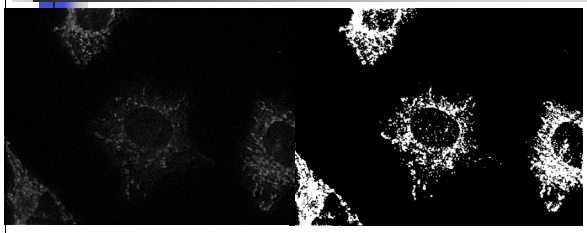
Ridler-Calvard Method

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

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

Carnegie Mellon

Object finding

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

Carnegie Mellon

2D Features Morphological Features

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

Carnegie Mellon

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

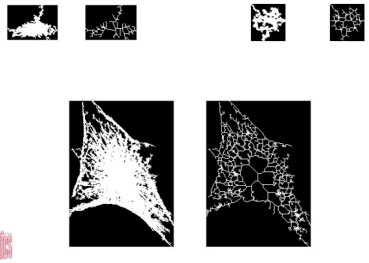
2D Features Object Skeleton Features

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

Carnegie Mellon

Illustration – Skeleton



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

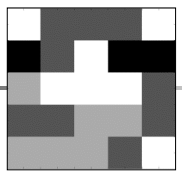
Carnegie Mellon

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

Summary

Carnegie Mellon

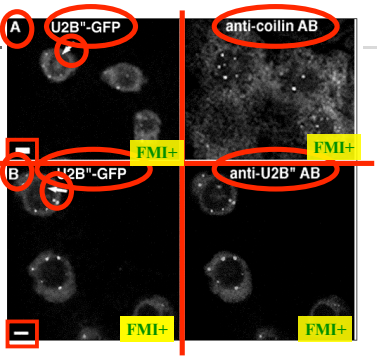
Overview: Image processing tasks

Segment into "panels"

Detect & remove **annotations**

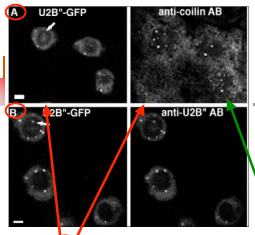
Classify panels

Find scale bars



Carnegie Mellon

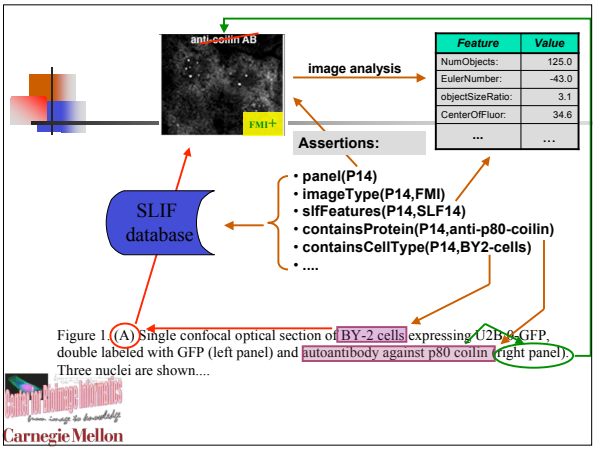
Overview: Text processing tasks



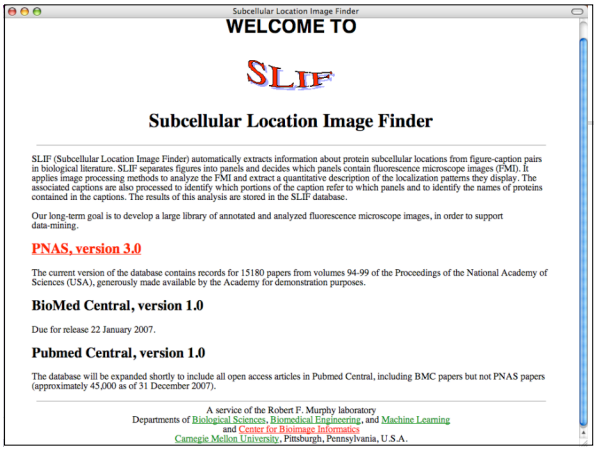
- Find *entity names* in text, and *panel labels* in text and the image.
- *Match* panels labels in text to panel labels on the image.
- *Associate* entity names to textual panel labels using *scoping* rules.

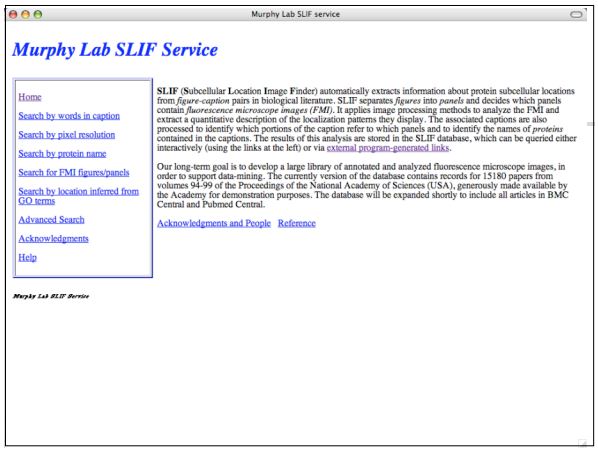
Figure 1. (A) Single confocal optical section of BY-2 cells expressing U2B 0-GFP, double labeled with GFP (left panel) and autoantibody against p80 coilin (right panel). Three nuclei are shown, and the bright GFP spots colocalize with bright foci of anti-coilin labeling. There is some labeling of the cytoplasm by anti-p80 coilin. (B) Single confocal optical section of BY-2 cells expressing U2B 0 -GFP, double labeled with GFP (left panel) and 4G3 antibody (right panel). Three nuclei are shown. Most coiled bodies are in the nucleoplasm, but occasionally are seen in the nucleolus (arrows). All coiled bodies that contain U2B 0 also express the U2B 0-GFP fusion. Bars, 5 m m. Movement of Coiled Bodies Vol. 10, July 1999 2299

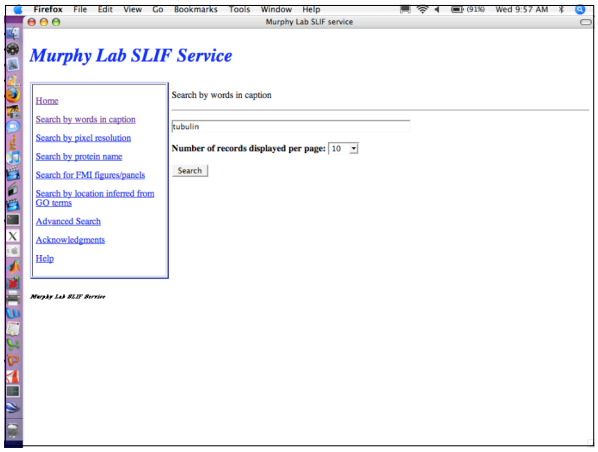
Carnegie Mellon

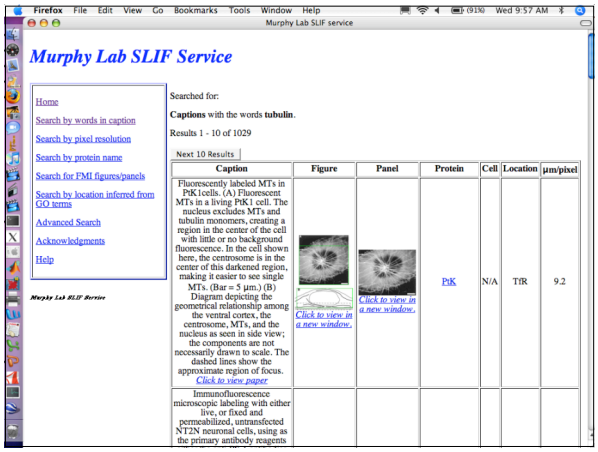


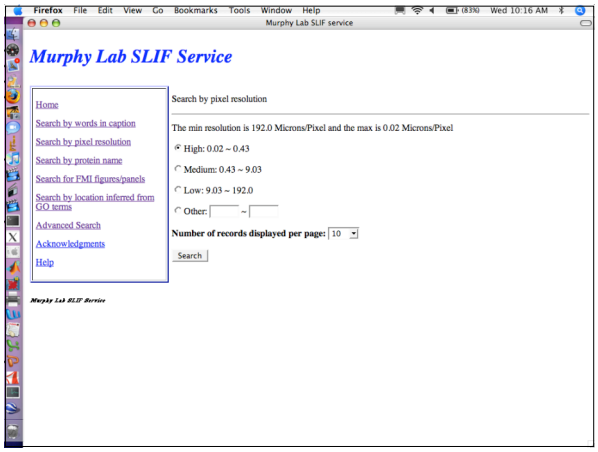
Carnegie Mellon

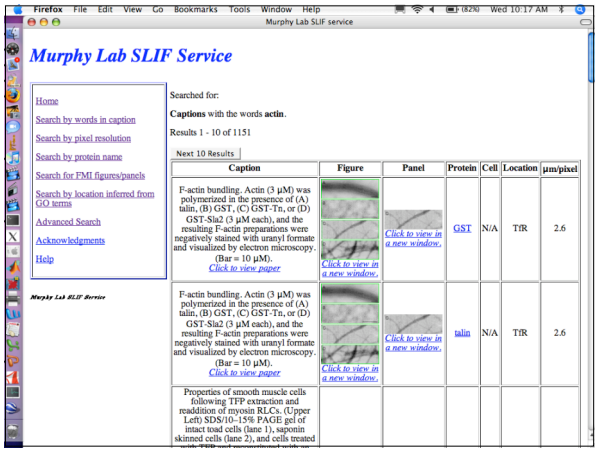


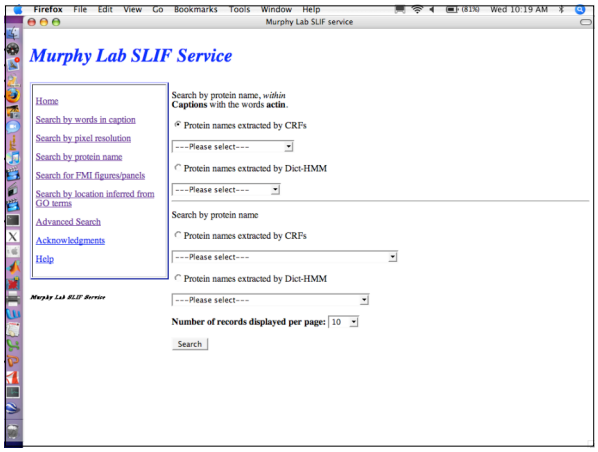

Murphy Lab SLIF service
Murphy Lab SLIF Service
Home
Search by words in caption
Search by pixel resolution
Search by protein name
Search for FMI figures/panels
Search by location inferred from GO terms
Advanced Search
Acknowledgments
Help
SLIF (Subcellular Location Image Finder) automatically extracts information about protein subcellular locations from figure-caption pairs in biological literature. SLIF separates figures into panels and decides which panels contain fluorescence microscope images (FMI). It applies image processing methods to analyze the FMI and extract a quantitative description of the localization patterns they display. The associated captions are also processed to identify which portions of the caption refer to which panels and to identify the names of proteins contained in the captions. The results of this analysis are stored in the SLIF database, which can be queried either interactively (using the links at the left) or via external program-generated links.
Our long-term goal is to develop a large library of annotated and analyzed fluorescence microscope images, in order to support data-mining. The currently version of the database contains records for 15180 papers from volumes 94-99 of the Proceedings of the National Academy of Sciences (USA), generously made available by the Academy for demonstration purposes. The database will be expanded shortly to include all articles in BMC Central and Pubmed Central.
Acknowledgments and People
Reference

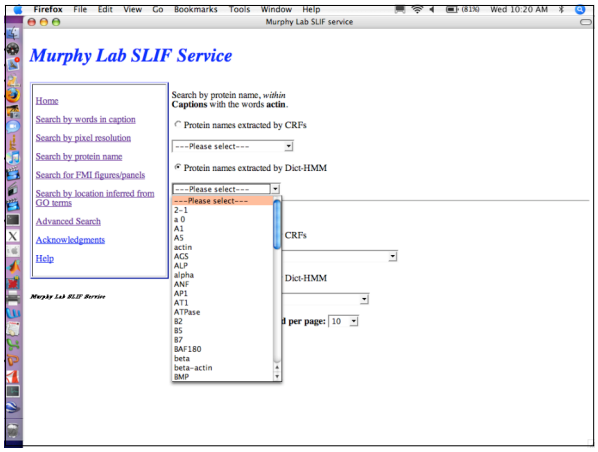

Murphy Lab SLIF Service
Search by words in caption
tubulin
Number of records displayed per page: 10
Search

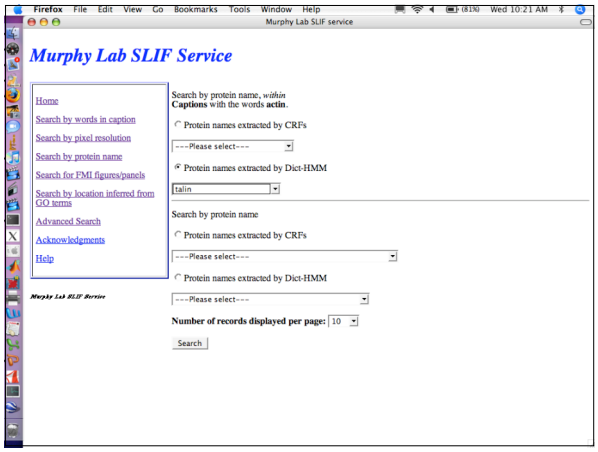

Murphy Lab SLIF Service
Searched for:
Captions with the words tubulin.
Results 1 - 10 of 1029
Next 10 Results
Caption
Figure
Panel
Protein
Cell
Location
µm/pixel
PtK
N/A
TfR
9.2

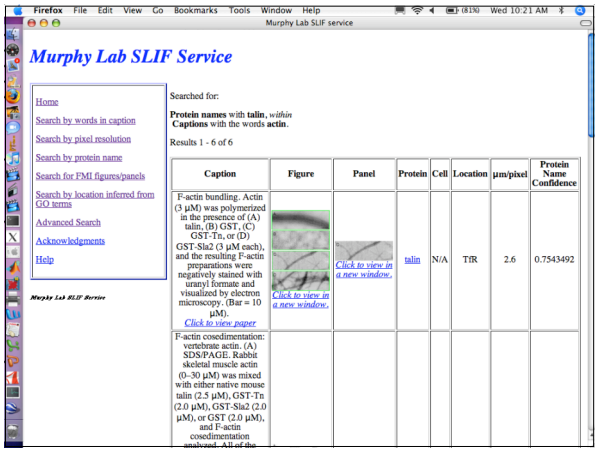

Murphy Lab SLIF Service
Search by pixel resolution
The min resolution is 192.0 Microns/Pixel and the max is 0.02 Microns/Pixel
High: 0.02 ~ 0.43
Medium: 0.43 ~ 9.03
Low: 9.03 ~ 192.0
Other:
Number of records displayed per page: 10
Search

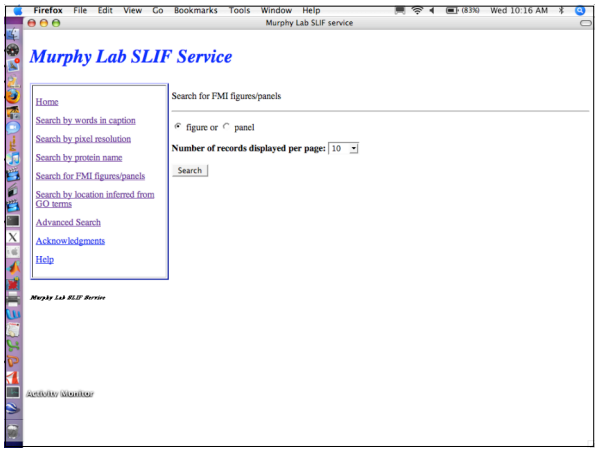

Murphy Lab SLIF Service
Searched for:
Captions with the words actin.
Results 1 - 10 of 1151
Next 10 Results
Caption
Figure
Panel
Protein
Cell
Location
µm/pixel
GST
N/A
TfR
2.6
talin
N/A
TfR
2.6

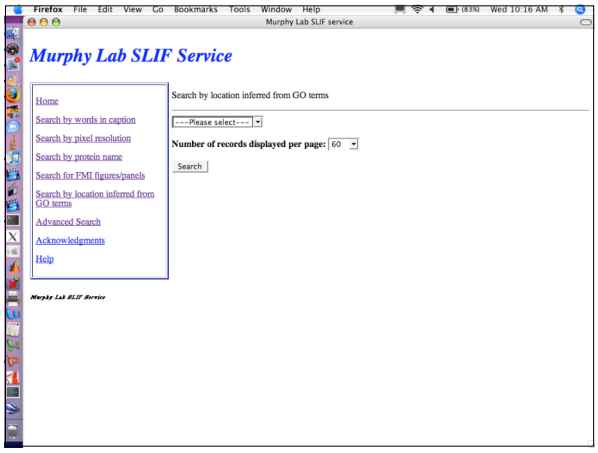

Murphy Lab SLIF Service
Search by protein name, within Captions with the words actin.
Protein names extracted by CRFs
---Please select---
Protein names extracted by Dict-HMM
Search by protein name
Number of records displayed per page: 10
Search

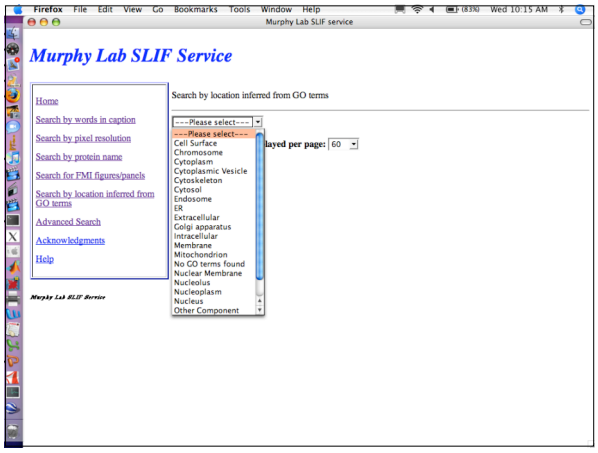

Murphy Lab SLIF Service
Home
Search by words in caption
Search by pixel resolution
Search by protein name
Search for FMI figures/panels
Search by location inferred from GO terms
Advanced Search
Acknowledgments
Help
Search by protein name, within Captions with the words actin.
Protein names extracted by CRFs
Protein names extracted by Dict-HMM
---Please select---
2-1
a 0
A1
A5
actin
ACS
ALP
alpha
ANF
AP1
AT1
ATPase
B2
B5
B7
BAF180
beta
beta-actin
BMP
CRFs
Dict-HMM
per page: 10

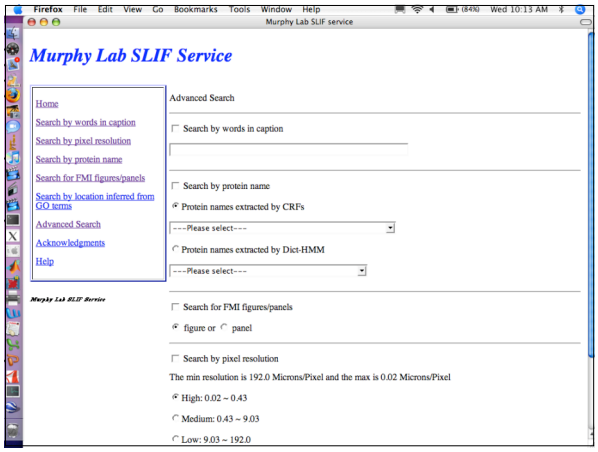

Murphy Lab SLIF Service
Home
Search by words in caption
Search by pixel resolution
Search by protein name
Search for FMI figures/panels
Search by location inferred from GO terms
Advanced Search
Acknowledgments
Help
Search by protein name, within Captions with the words actin.
Protein names extracted by CRFs
Protein names extracted by Dict-HMM
talin
Search by protein name
Protein names extracted by CRFs
Protein names extracted by Dict-HMM
Number of records displayed per page: 10
Search


Murphy Lab SLIF Service
Home
Search by words in caption
Search by pixel resolution
Search by protein name
Search for FMI figures/panels
Search by location inferred from GO terms
Advanced Search
Acknowledgments
Help
Searched for:
Protein names with talin, within Captions with the words actin.
Results 1 - 6 of 6
Caption
Figure
Panel
Protein
Cell
Location
µm/pixel
Protein Name Confidence
F-actin bundling. Actin (3 µM) was polymerized in the presence of (A) talin, (B) GST, (C) GST-Tn, or (D) GST-Sla2 (3 µM each), and the resulting F-actin preparations were negatively stained with uranyl formate and visualized by electron microscopy. (Bar = 10 µM).
Click to view paper
Click to view in a new window.
Click to view in a new window.
talin
N/A
TfR
2.6
0.7543492
F-actin cosedimentation: vertebrate actin. (A) SDS/PAGE: Rabbit skeletal muscle actin (0–30 µM) was mixed with either native mouse talin (2.5 µM), GST-Tn (2.0 µM), GST-Sla2 (2.0 µM), or GST (2.0 µM), and F-actin cosedimentation analyzed. All of the


Murphy Lab SLIF Service
Home
Search by words in caption
Search by pixel resolution
Search by protein name
Search for FMI figures/panels
Search by location inferred from GO terms
Advanced Search
Acknowledgments
Help
Search for FMI figures/panels
figure or panel
Number of records displayed per page: 10
Search
Activity Monitor

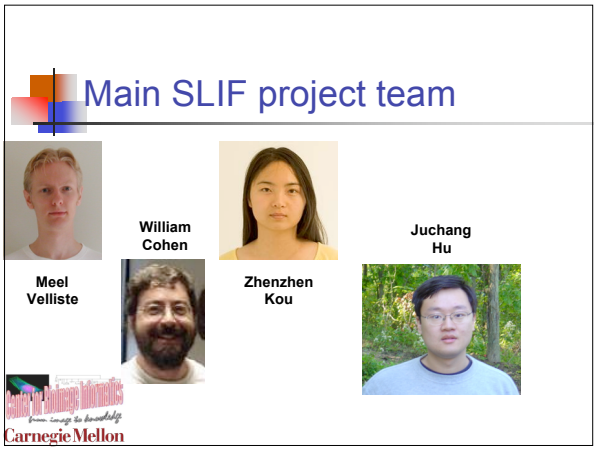

Murphy Lab SLIF Service
Home
Search by words in caption
Search by pixel resolution
Search by protein name
Search for FMI figures/panels
Search by location inferred from GO terms
Advanced Search
Acknowledgments
Help
Search by location inferred from GO terms
---Please select---
Number of records displayed per page: 60
Search

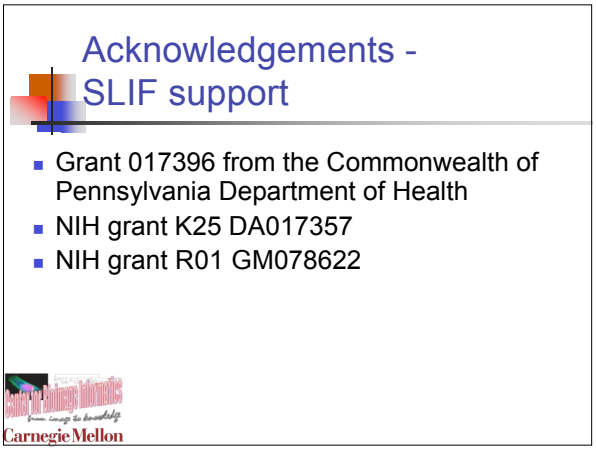

Murphy Lab SLIF Service
Home
Search by words in caption
Search by pixel resolution
Search by protein name
Search for FMI figures/panels
Search by location inferred from GO terms
Advanced Search
Acknowledgments
Help
Search by location inferred from GO terms
---Please select---
Cell Surface
Chromosome
Cytoplasm
Cytoplasmic Vesicle
Cytoskeleton
Cytosol
Endosome
ER
Extracellular
Golgi apparatus
Intracellular
Membrane
Mitochondrion
No GO terms found
Nuclear Membrane
Nucleolus
Nucleoplasm
Nucleus
Other Component
layed per page: 60

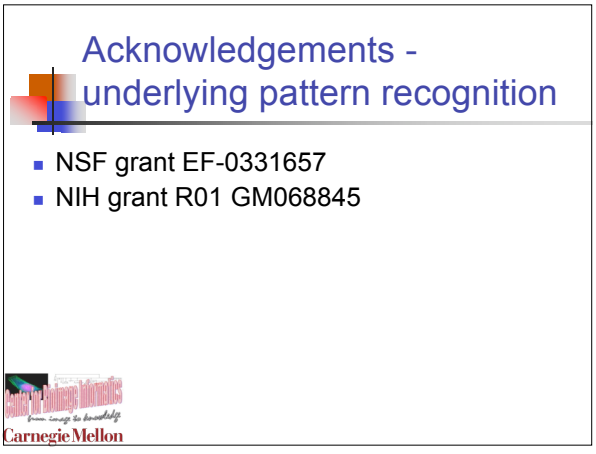

Murphy Lab SLIF Service
Home
Search by words in caption
Search by pixel resolution
Search by protein name
Search for FMI figures/panels
Search by location inferred from GO terms
Advanced Search
Acknowledgments
Help
Advanced Search
Search by words in caption
Search by protein name
Protein names extracted by CRFs
Protein names extracted by Dict-HMM
Search for FMI figures/panels
figure or panel
Search by pixel resolution
The min resolution is 192.0 Microns/Pixel and the max is 0.02 Microns/Pixel
High: 0.02 ~ 0.43
Medium: 0.43 ~ 9.03
Low: 9.03 ~ 192.0


Murphy Lab SLIF Service
Home
Search by words in caption
Search by pixel resolution
Search by protein name
Search for FMI figures/panels
Search by location inferred from GO terms
Advanced Search
SLIF stands for Subcellular Location Image Finder, which automatically extracts information about protein subcellular locations from figure-caption pairs in biological literature. SLIF applies image processing methods to analyze fluorescence microscope images and extract a quantitative description of the localization patterns of the tagged proteins. Detailed textual annotations of the images and associated proteins are also generated by analyzing the accompanying captions.
Our long-term goal is to develop a large library of annotated and analyzed fluorescence microscope images, in order to support data-mining.


Acknowledgments
Michael
Mia
Greg
Meel
Students
Dr. Michael Boland
Dr. Mia Markey (ugrad)
Gregory Porreca (ugrad)
Dr. Meel Velliste
Dr. Kai Huang
Dr. Xiang Chen
Ting Zhao
Shann-Ching Chen
Juchang Hua
Kai
Xiang
Collaborators/Consultants
CMU: David Casasent, Simon Watkins, Jo Jarvik, Peter Berget, Jack Rohrer, Tom Mitchell, Christos Faloutsos, Jelena Kovacevic, William Cohen, Geoff Gordon
UCSB: B. S. Manjunath, Ambuj Singh
Juchang
Sam Ting
Carnegie Mellon


Main SLIF project team
Meel Velliste
William Cohen
Zhenzhen Kou
Juchang Hu
Carnegie Mellon


Acknowledgements - SLIF support
Grant 017396 from the Commonwealth of Pennsylvania Department of Health
NIH grant K25 DA017357
NIH grant R01 GM078622
Carnegie Mellon


Acknowledgements - underlying pattern recognition
NSF grant EF-0331657
NIH grant R01 GM068845
Carnegie Mellon

References

All available from 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.

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.

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.

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.

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

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.

Carnegie Mellon

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.

Carnegie Mellon