Representation in Supervised Machine Learning

Application to Biological Problems

Robert Langlois

Frank Lab

Howard Hughes Medical Institute

&

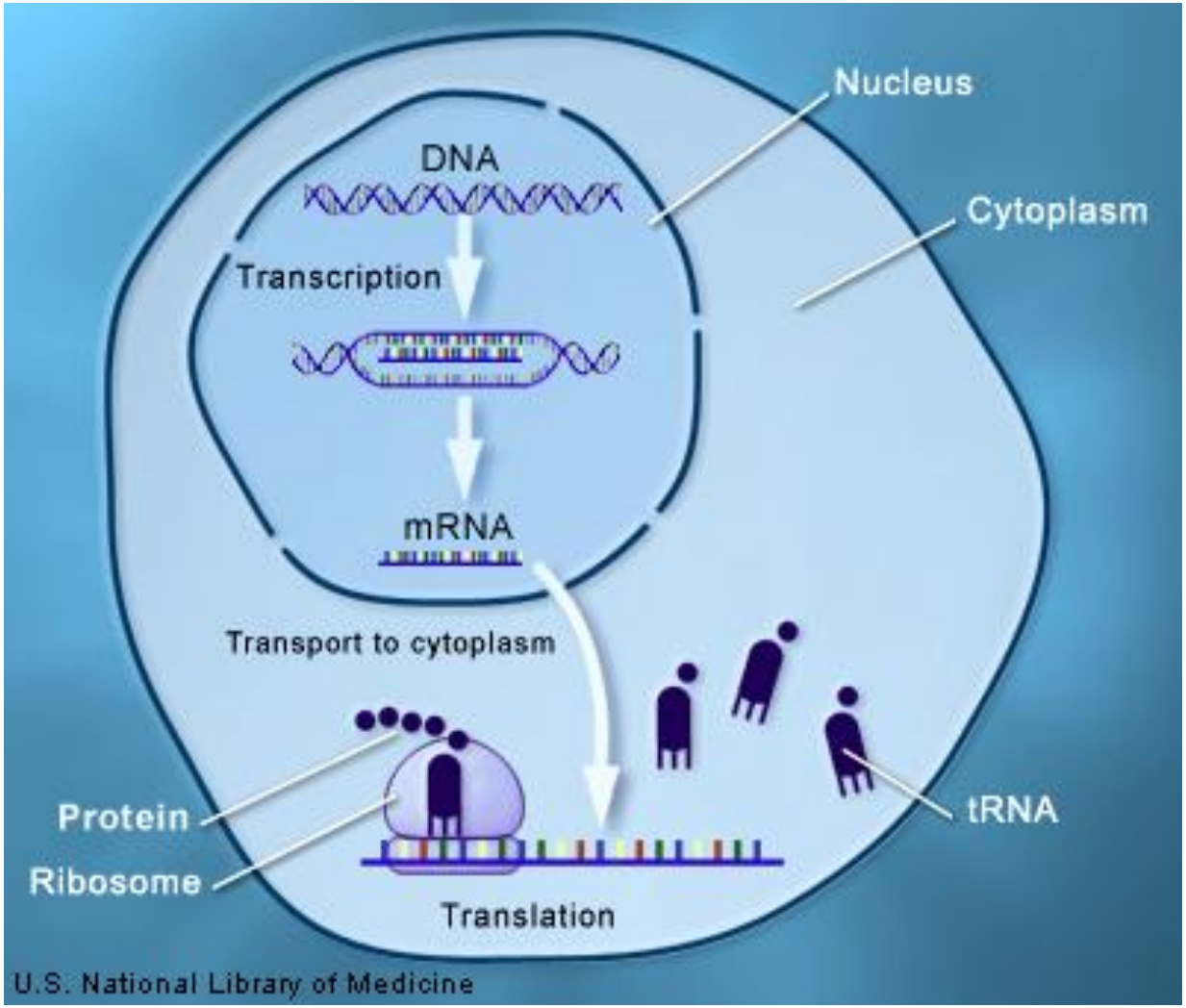
Columbia University

August 9, 2011

What is Machine Learning?

- Computer Algorithms that Learn
  - Think of it as *Programming by Example*
  - Devise a strategy to complete a task
- Example Handwriting
  - Given examples of written letters
  - Optical character recognition
  - U.S. Post Office

Short Biology Review



Central Dogma of Proteomics

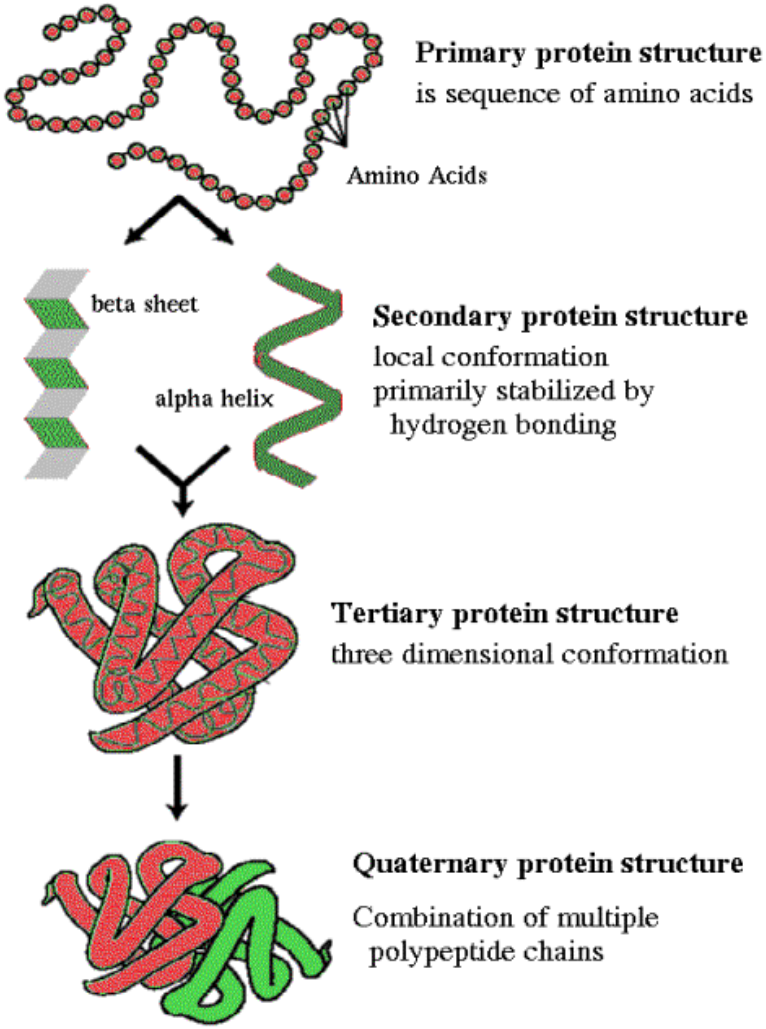
Sequence

↓

Structure

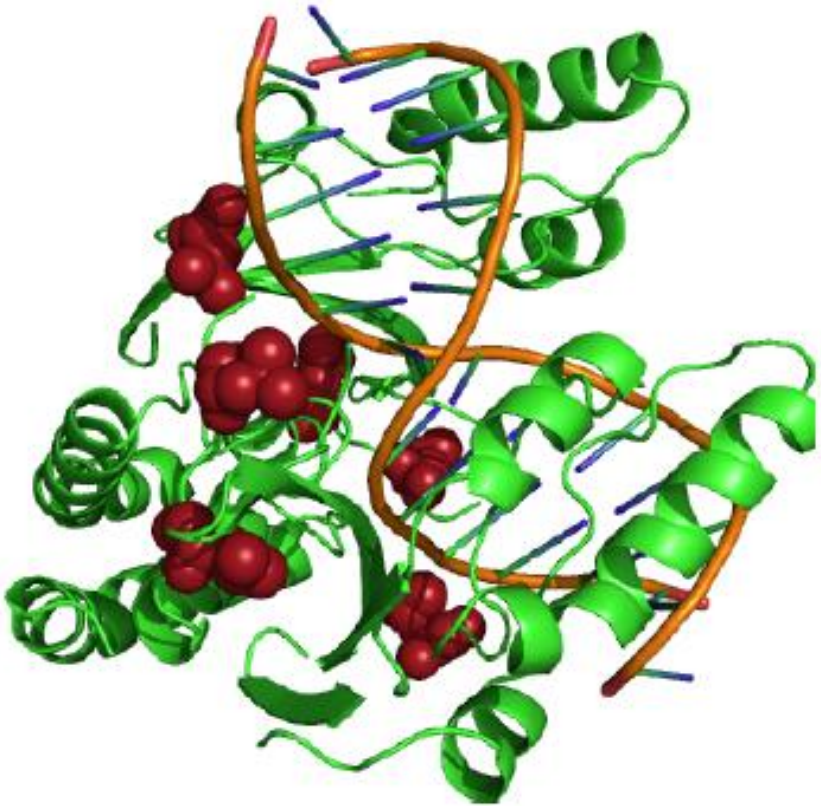
↓

Function



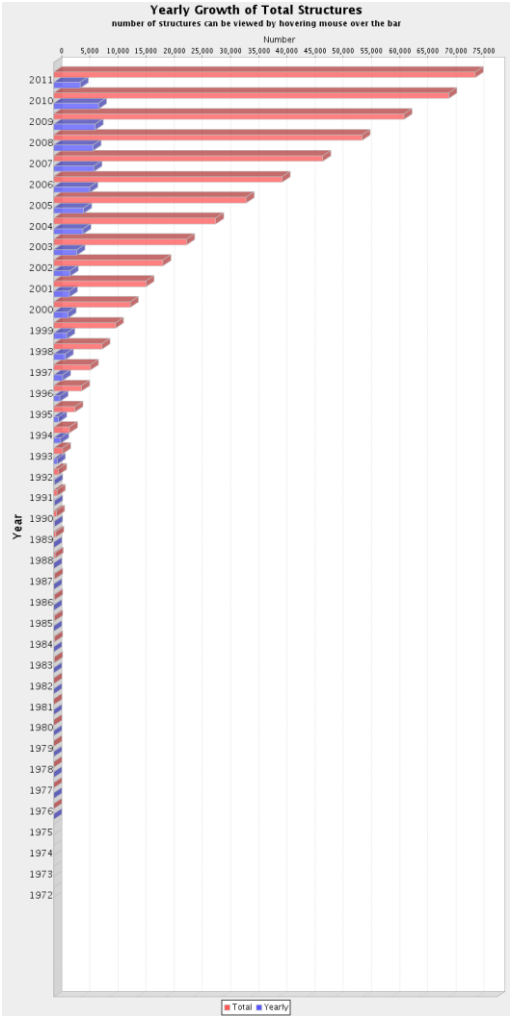
Application: What function does a protein perform?

- Protein → Molecular machines
  - Catalyze reactions
  - Mechanical work
  - Structural components
- Proteins that bind DNA
  - Transcription Regulation
  - Epigenetics (Encode behavior)
  - DNA Utility: Repair

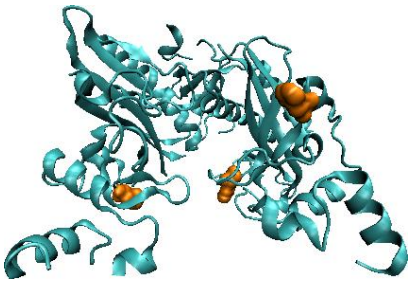


Motivation: Protein Function Prediction

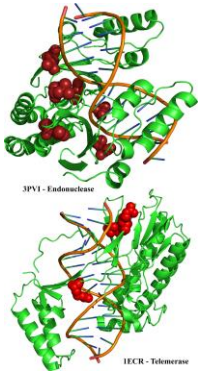
- Large amount of sequence data
  - Number of solved genomes
  - Personal genomics (10^6 bases < $0.50)
- Large amount of structural data
  - Protein databank: Exponential Growth
  - Cryo-EM Databank: Protein complexes
- Training data (Experimental Function)
  - Filter binding assays
  - Chromatin Immuno-precipitation



Binary Classification



- Protein Function Annotation
  - Does a protein bind DNA?
  - Given two sets of structures
    1. Binds DNA
    2. Not bind DNA

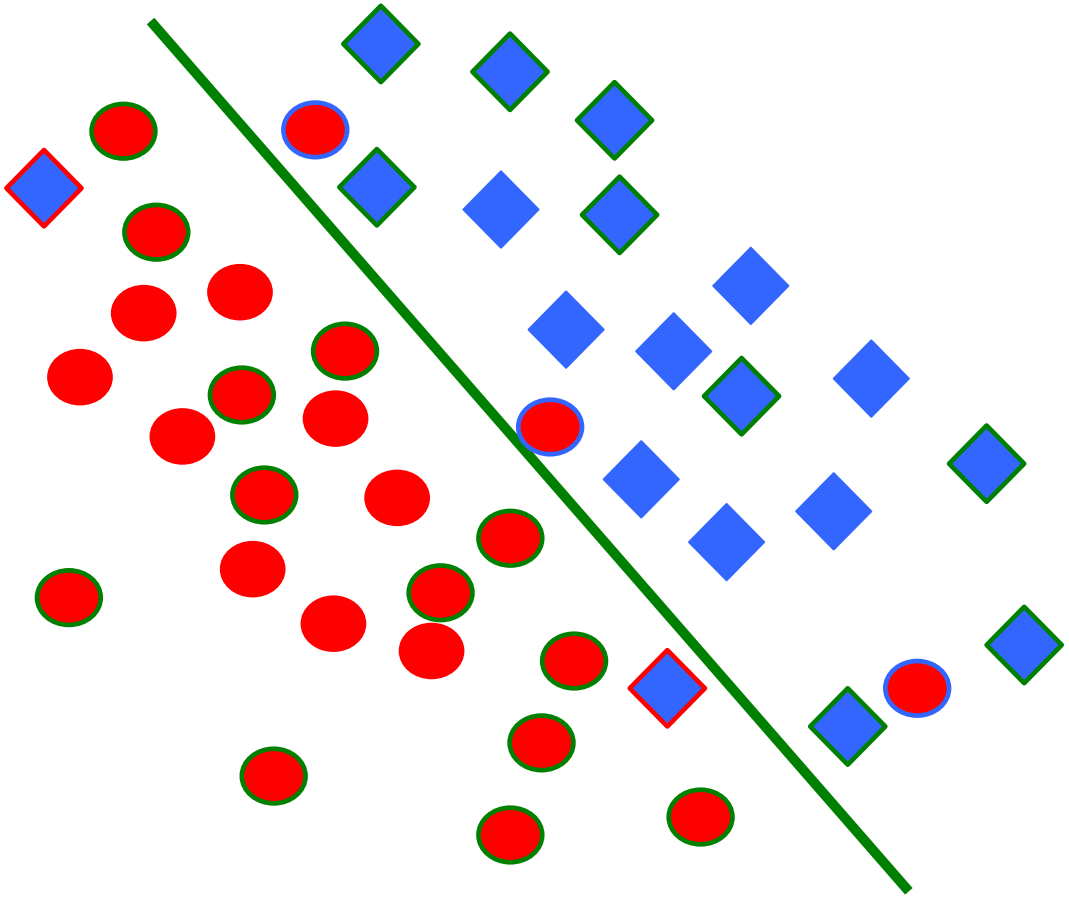


Labeled training examples → Machine Learning Algorithm → Prediction Rule

Unseen Example → Prediction Rule → Predicted Label

- Learning
  - Minimize number of mistakes

Learning Process (Binary Classification)



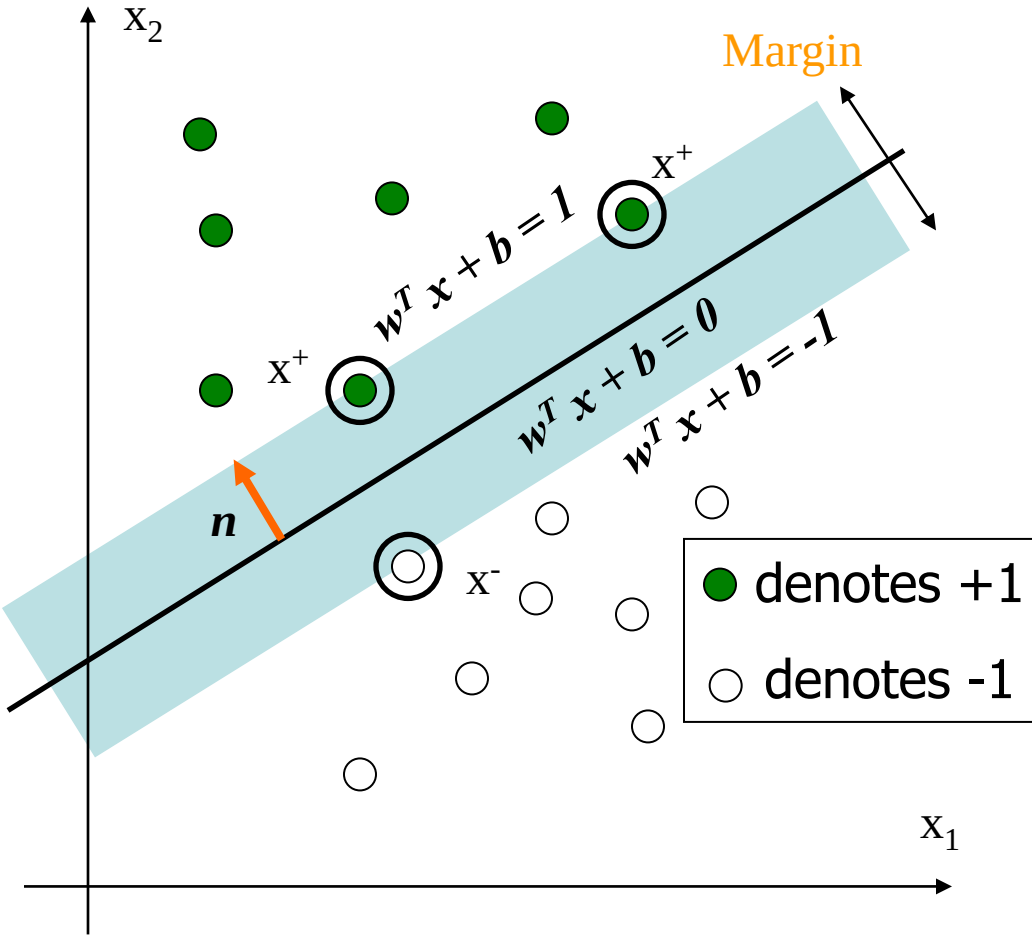
Large Margin Linear Classifier: Support Vector Machine

- Formulation:

$$\text{minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2$$

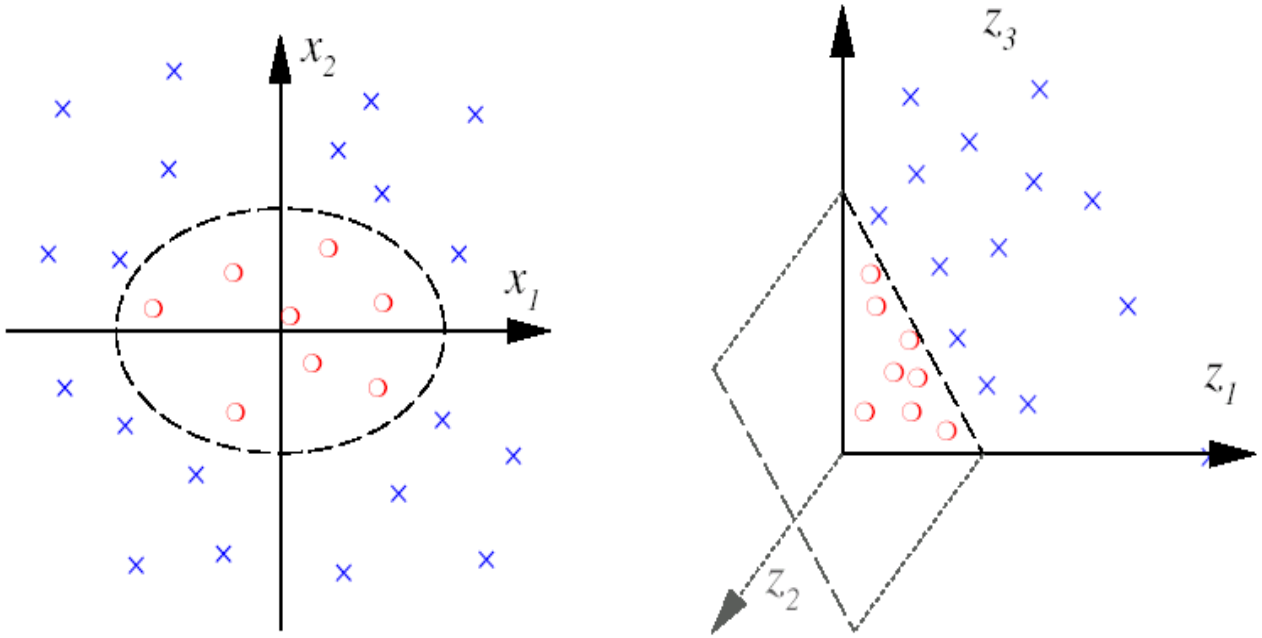
such that

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$$



Non-linear Support Vector Machines

$$\Phi : \mathbb{R}^2 \to \mathbb{R}^3$$
$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{2}\, x_1 x_2, x_2^2)$$



B. Schölkopf, Canberra, February 2002

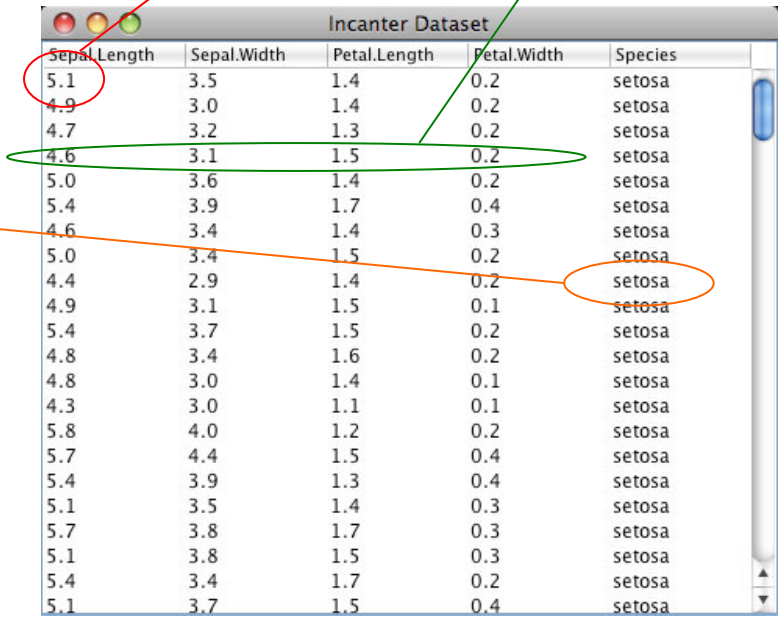
Representation

- Attribute: Numerical description of an object
  - Where j=1…m
- Feature Vector : Fixed length set of attributes
  - Where i=1…n
- Special Attributes
  - Class: Target or unknown attribute

$$(\vec{x}_i, y_i) \in S$$

$x_j \in \vec{x}_i$

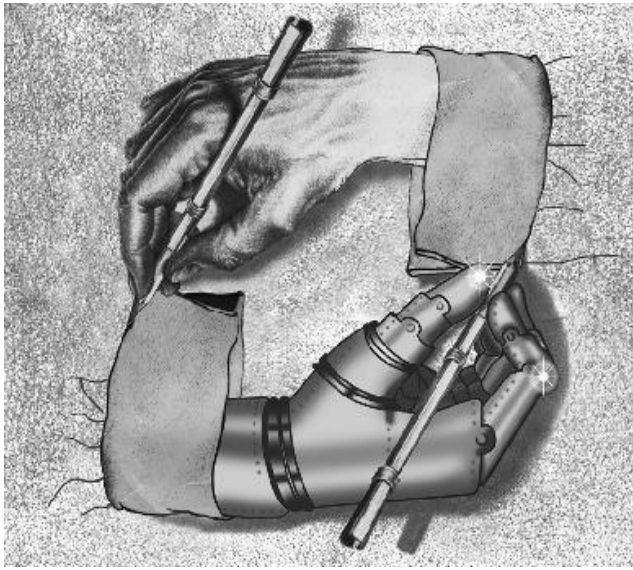
y_i



http://data-sorcery.org/category/pca

Outline

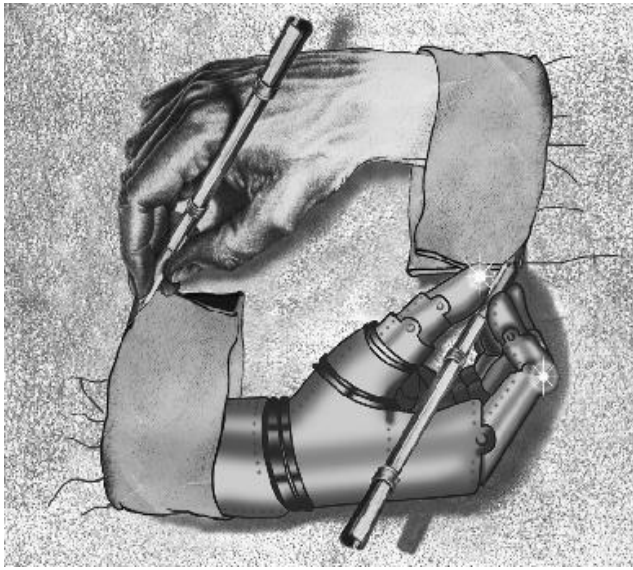
- Can a positive patch solve the problem?
  - Binary Classification
  - Represent entire sequence or structure
- Is there a more generic representation?
  - Binary Classification
  - Represent by residue content
- Most generic representation
  - Multiple-instance Learning
  - Represent each residue



http://ce.sharif.ir/courses/87-88/1/ce717/syllabus/Logo

Outline

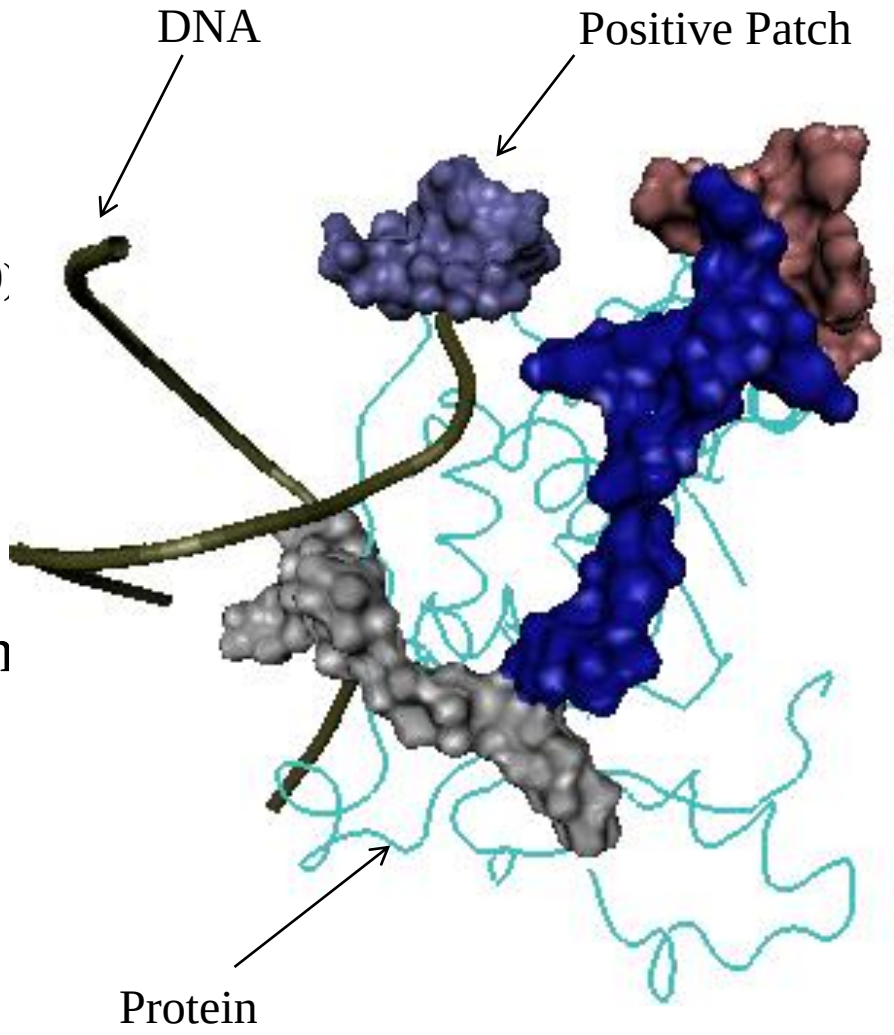
- Can a positive patch solve the problem?
  - Binary Classification
  - Represent entire sequence or structure
- Is there a more generic representation?
  - Binary Classification
  - Represent by residue content
- Most generic representation
  - Multiple-instance Learning
  - Represent each residue



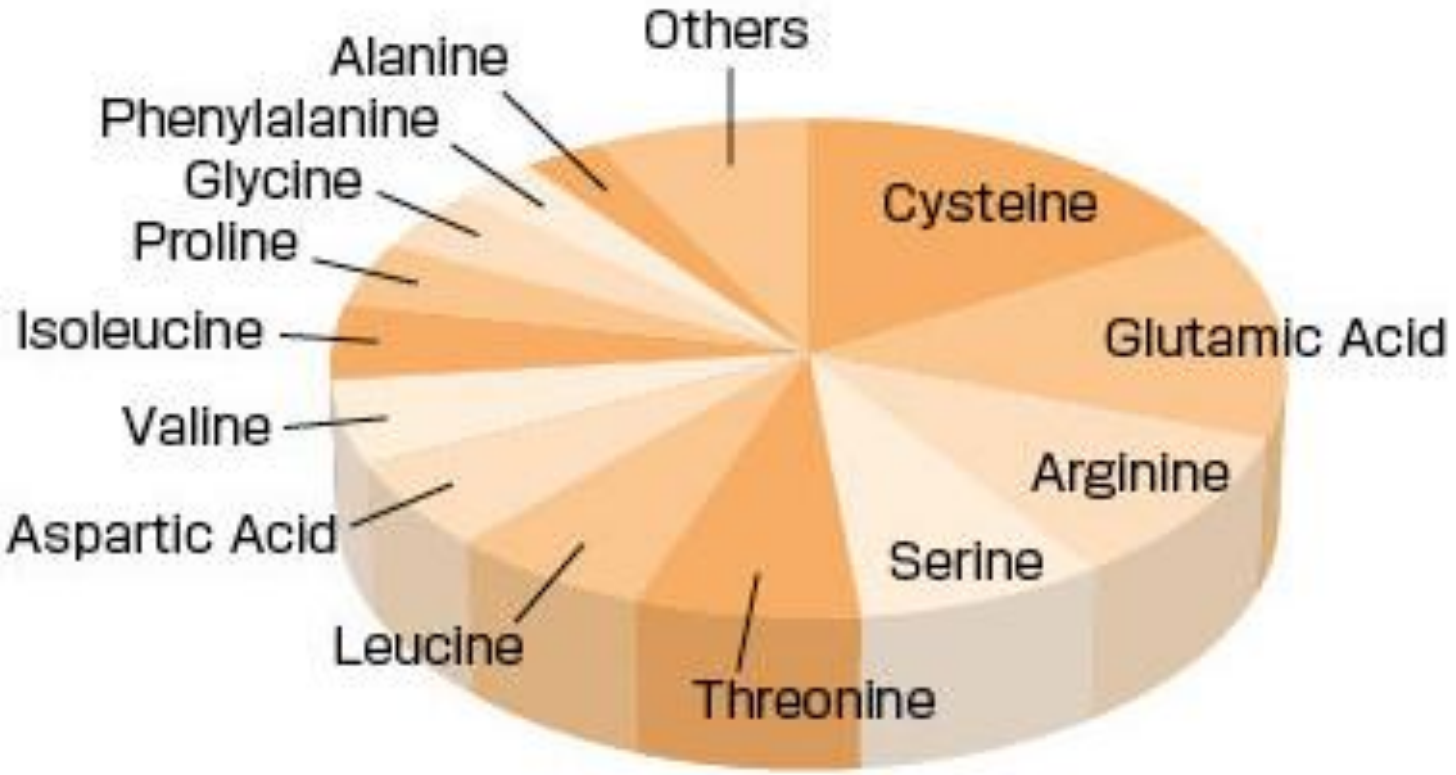
http://ce.sharif.ir/courses/87-88/1/ce717/syllabus/Logo

Protein Representation

- Structure
  - Size of largest positive patch (1)
  - Surface amino acid composition (20)
- Sequence
  - Charge (1)
  - Amino acid composition (20)
- Represent each protein with
  - 42 numeric features
  - -10, 1, 0, 0, 11, 5 … 3



Composition



Performance of Positive Patch + 41 other features

- Support vector machines
  - Outperforms sequence analysis techniques like Blast
- Boosted Trees
  - Compares well compared to Support Vector Machines (SVM)
  - 10-fold cross-validation

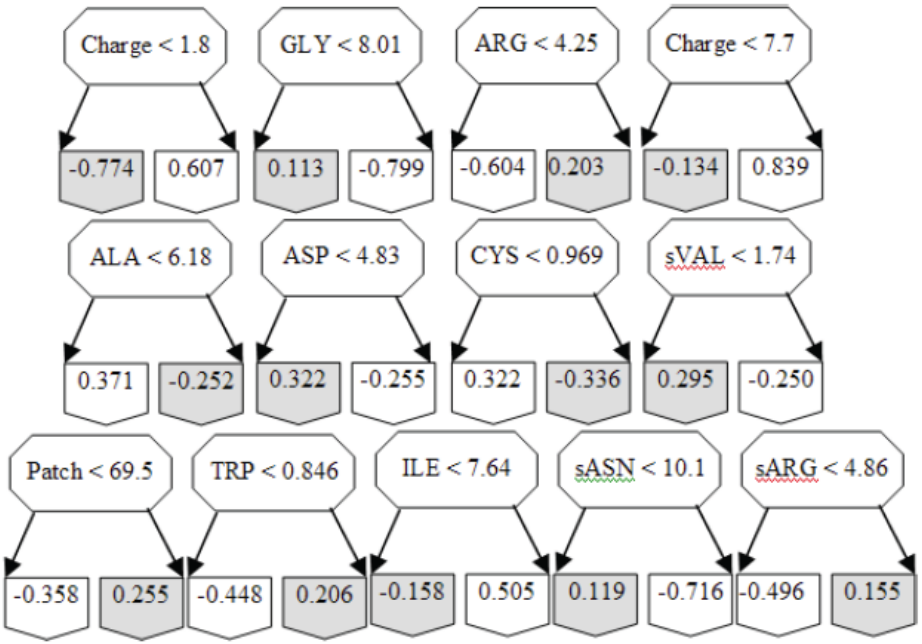
	Accuracy	Sensitivity	Specificity	AUC
SVM	83.0	68.5	88.0	83.5
Boosted Tree	88.2	64.4	96.3	89.8
ADTree	85.1	66.7	91.3	85.3

Related Work: Classification with simple features

- Structural motifs (Shanahan 2005)
- Structural motifs and HMM (Pellegrini-Calace 2005)
- Structure attributes and NN (Ahmad 2002, 2004)
- Low resolution homology and LR (Szilagyi 2006)
- Structure attributes and SVM (Bhardwaj, **Langlois**, 2005)
  - Natures Highlights
- Structure attributes and AB,Tree,SVM (**Langlois**,2006)

What Rules are Used?

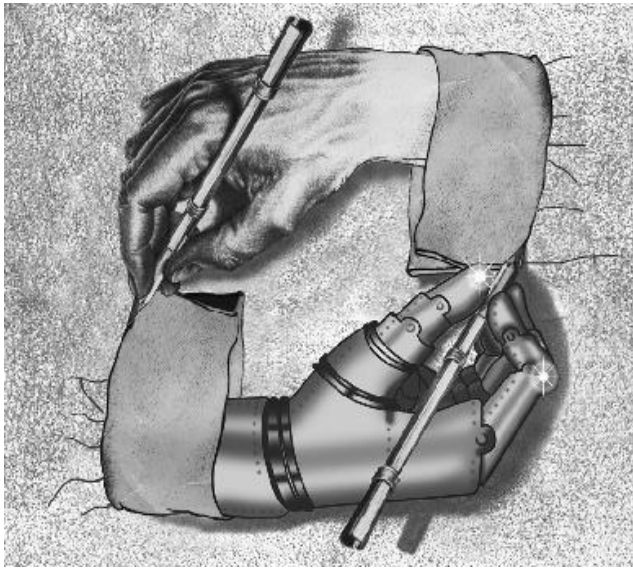
- Unexpected Results
  - Mostly sequence features
  - Exclusionary rules
  - Patch does not dominate
- Caveats
  - Only an approximation
  - Still performs fairly well
- Back to the drawing board



Langlois, R., M. Carson, N. Bhardwaj and H. Lu. Learning to translate sequence and structure to function: Identifying DNA binding and membrane binding proteins. Annals of Biomedical Engineering. 35:1043-1052, 2007.

Outline

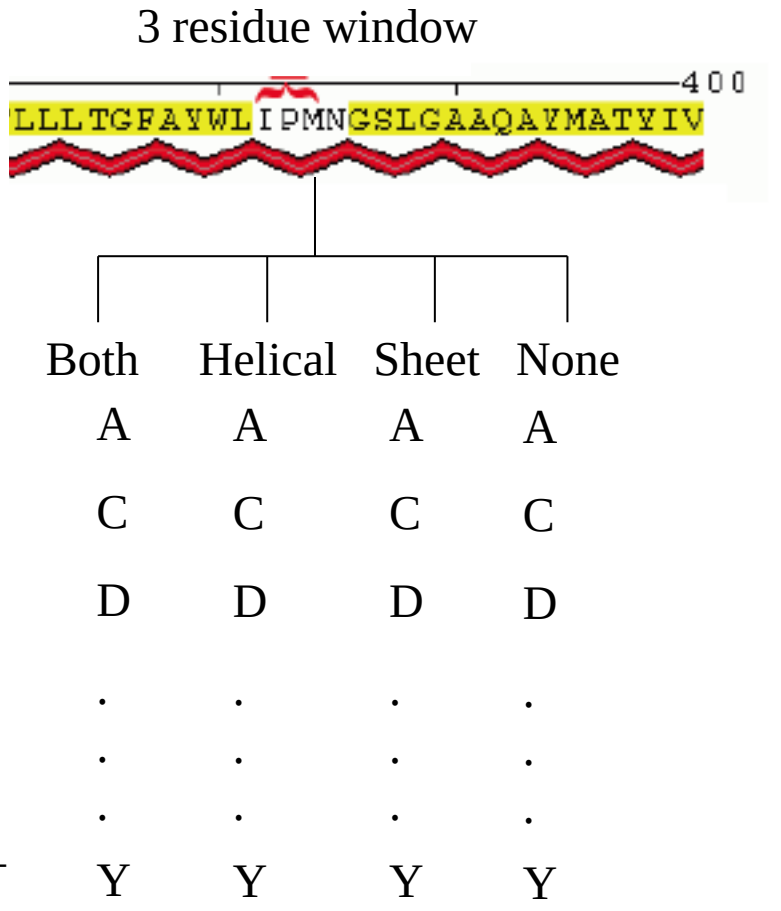
- Can a positive patch solve the problem?
  - Binary Classification
  - Represent entire sequence or structure
- Is there a more generic representation?
  - Binary Classification
  - Represent by residue content
- Most generic representation
  - Multiple-instance Learning
  - Represent each residue



http://ce.sharif.ir/courses/87-88/1/ce717/syllabus/Logo

Conditional Amino Acid Composition

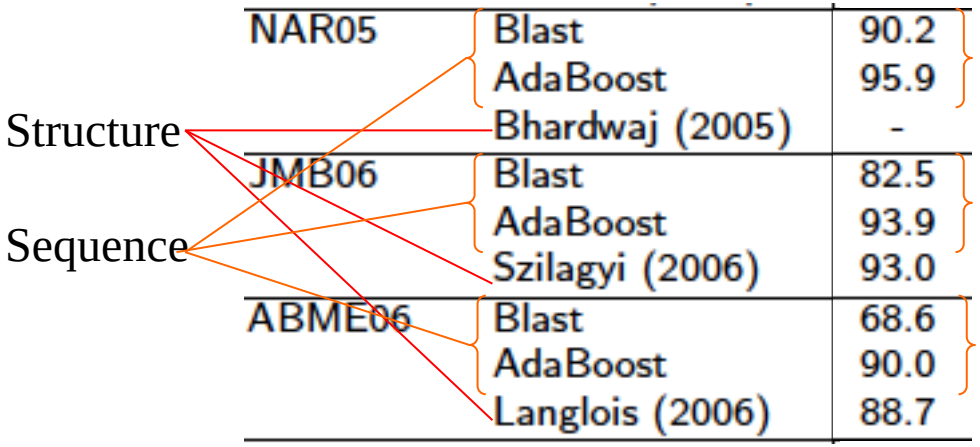
- Composition of both
  - Type
  - Environment
- Composition of Proline given
  - Helix environment
  - Sheet environment
  - Neither
  - Both



Robert E Langlois, Hui Lu (2010) Boosting the prediction and understanding of DNA-binding domains from sequence, 3149-3158. In Nucleic Acids Research 38 (10).

Performance of generic sequence features

- Performance with Boosted Trees

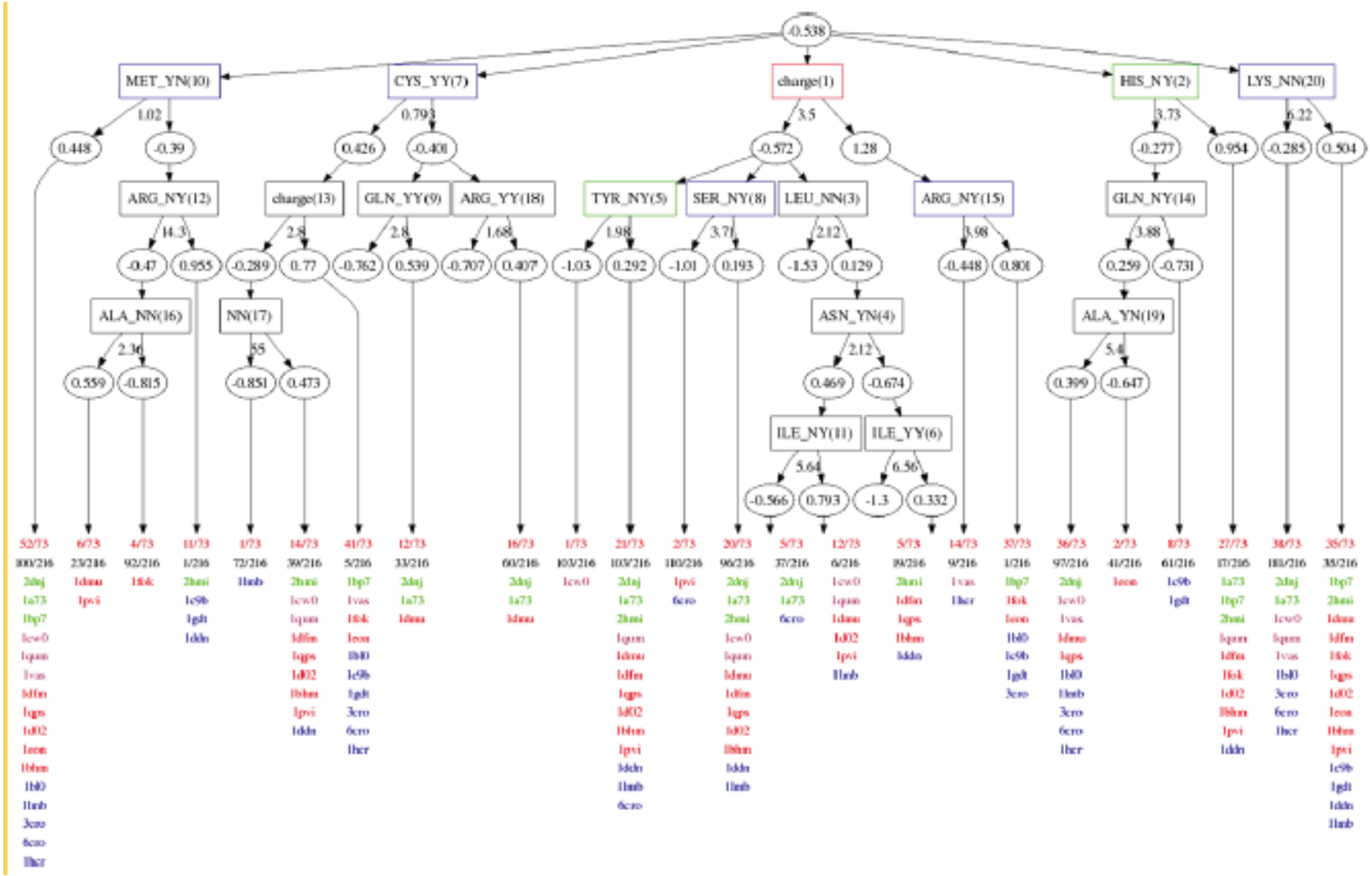


NAR05	Blast	90.2
	AdaBoost	95.9
	Bhardwaj (2005)	-
JMB06	Blast	82.5
	AdaBoost	93.9
	Szilagyi (2006)	93.0
ABME06	Blast	68.6
	AdaBoost	90.0
	Langlois (2006)	88.7

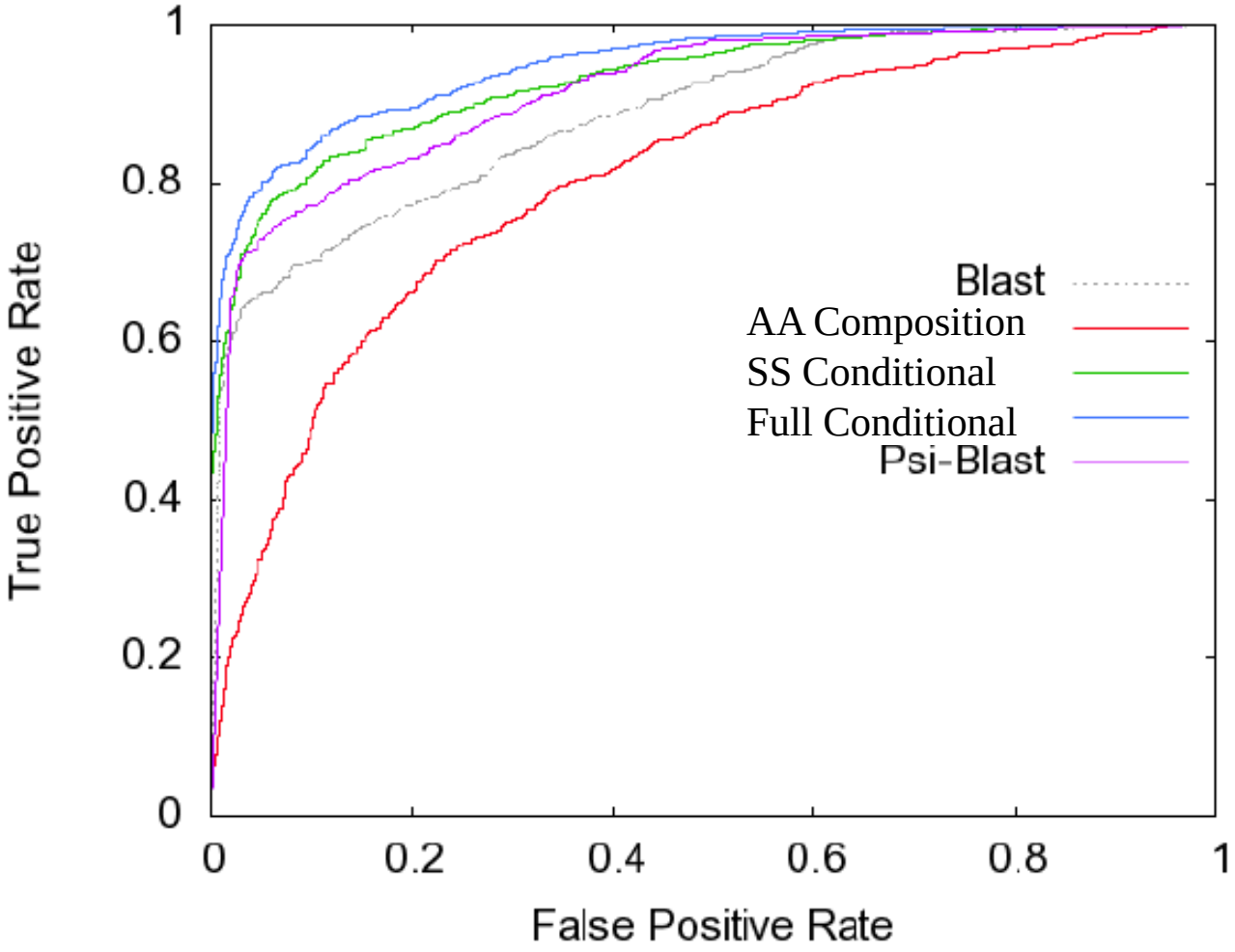
- Structure features should encode more about function
- However, less information with sequence - competitive

Robert E Langlois, Hui Lu (2010) Boosting the prediction and understanding of DNA-binding domains from sequence, 3149-3158. In Nucleic Acids Research 38 (10).

What rules do the machine learning algorithms use?

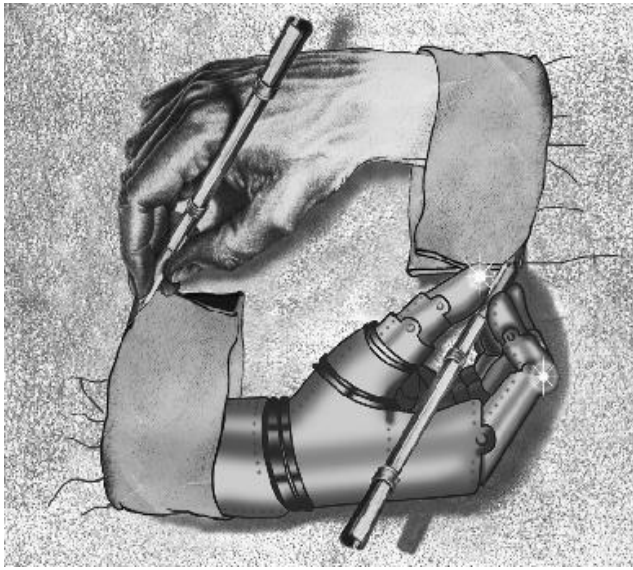


Genome-wide Results: 2000 Sequence < 25% Similarity



Outline

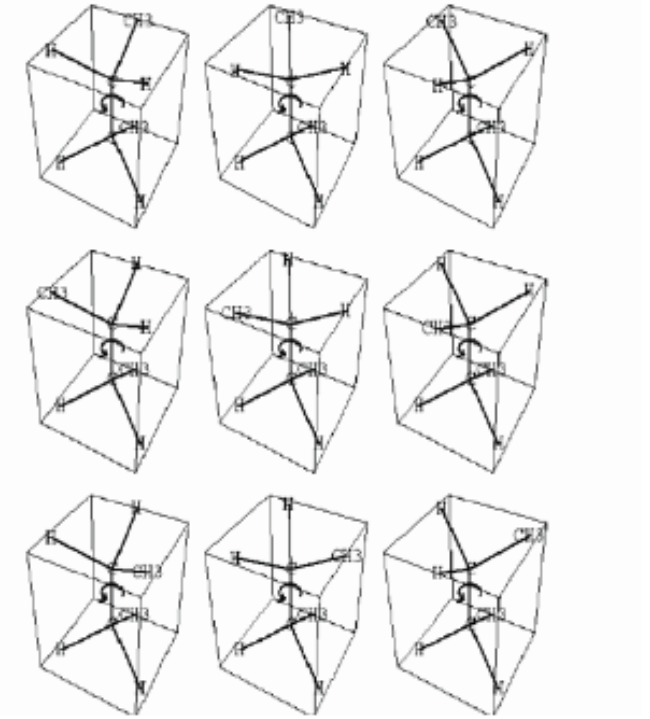
- Can a positive patch solve the problem?
  - Binary Classification
  - Represent entire sequence or structure
- Is there a more generic representation?
  - Binary Classification
  - Represent by residue content
- Most generic representation
  - Multiple-instance Learning
  - Represent each residue



http://ce.sharif.ir/courses/87-88/1/ce717/syllabus/Logo

Multiple-instance Learning

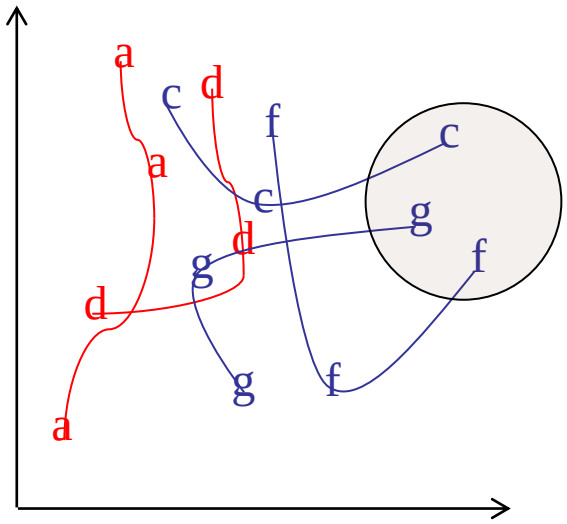
- Examples grouped into bags
- Label on group (bag) not example
- Bag is positive if at least one example is positive
- Otherwise bag is negative
- Only weak information about positive instances!



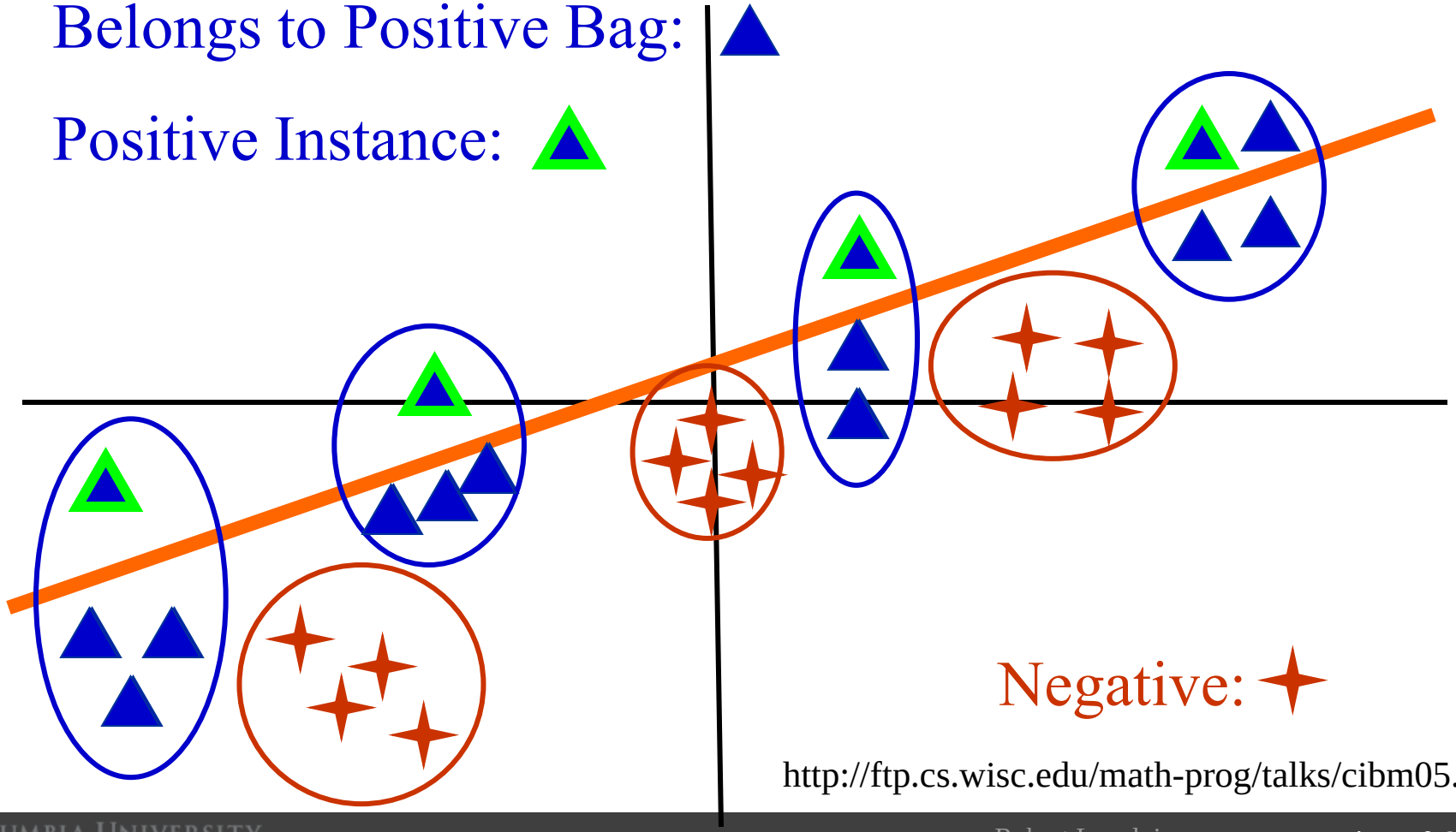
©1998 by Oded Maron

Intuitive Example

- Illustration
  - Each letter is an example
  - Each line indicates a bag (or group)
  - Blue: Examples in a positive bag
  - Red: Examples in a negative bag
- Classify bags by finding positive instances
  - Close to other instances in the positive bags
  - Far from instances in negative bags

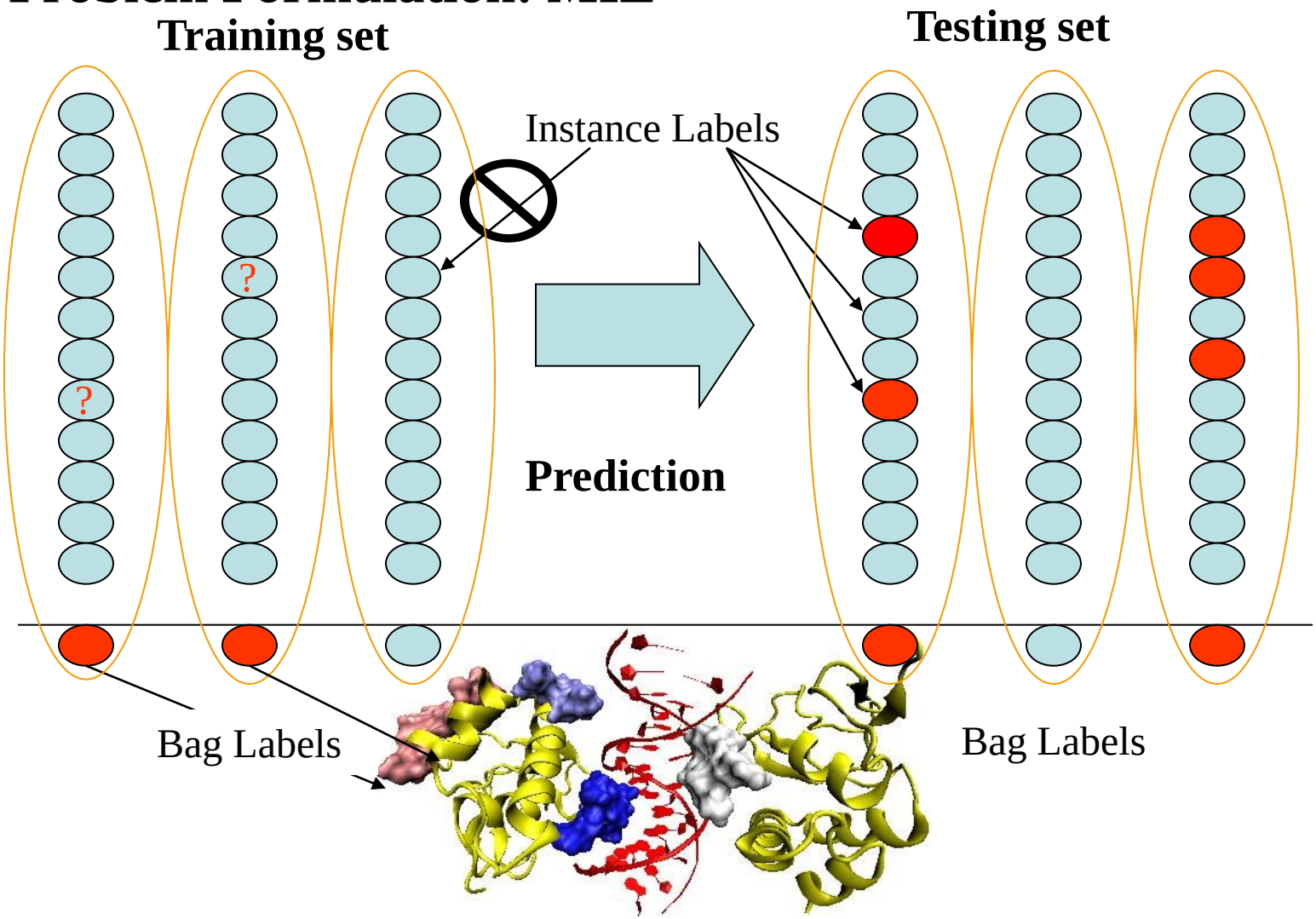


Multiple Instance Learning Example



http://ftp.cs.wisc.edu/math-prog/talks/cibm05.ppt

Problem Formulation: MIL



Amino Acid Representation

- Residue Identity (20)
- Secondary Structure (3)
- Structure Neighbors (20)
- PSSM for residue at position (20)
- Blosum for positions -3 … 3 (140)
- Properties: Charge, Surface Area (2)

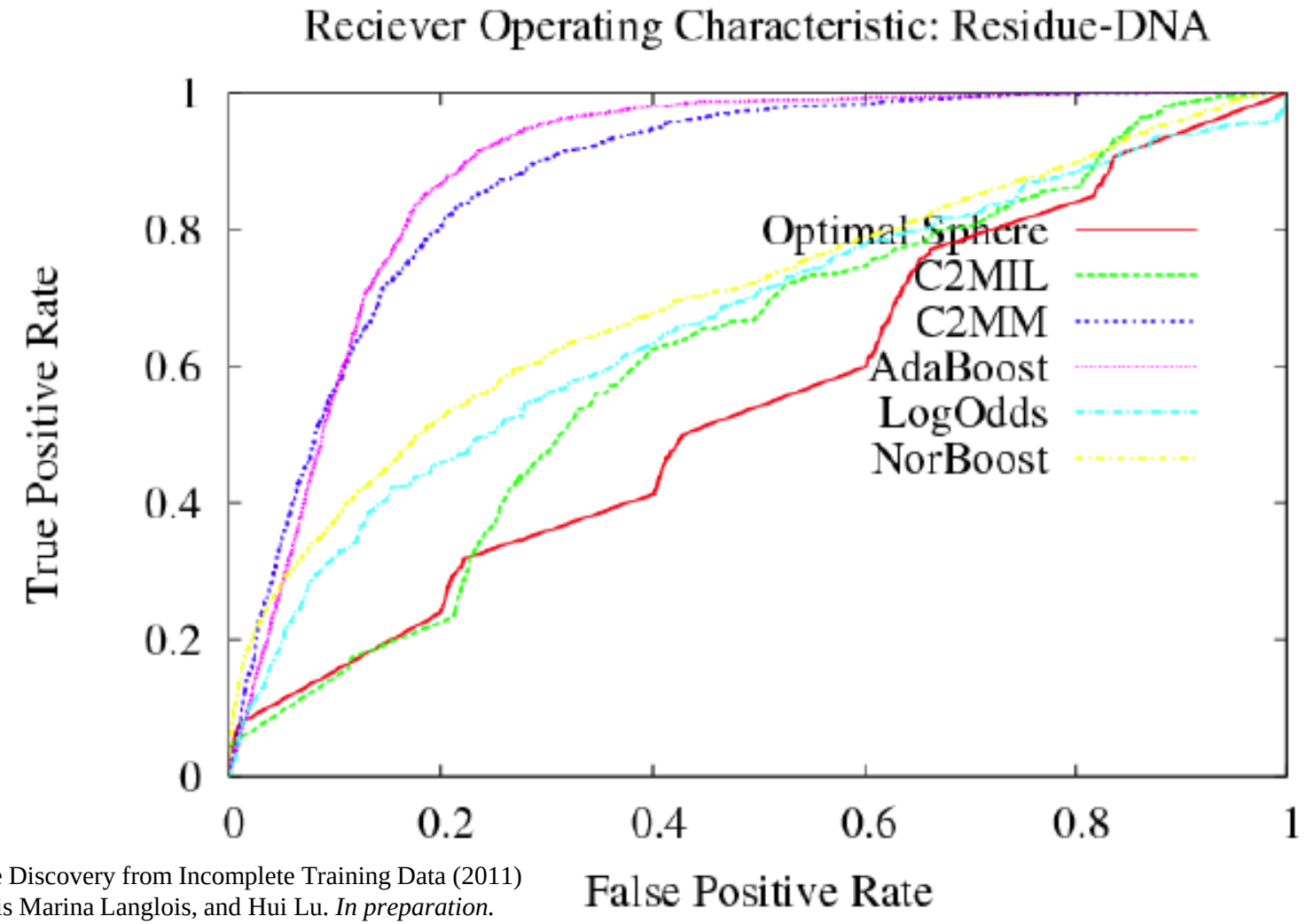
Experimental Results

- Similar Structure based features
- Achieves over 98 AUR!
- Change in representation effective in improving accuracy.

JMB06	Blast	82.5
	AdaBoost	93.9
	Szilagyi (2006)	93.0

Functional Site Discovery from Incomplete Training Data (2011)
Robert Langlois Marina Langlois, and Hui Lu. *In preparation.*

Instance-level (Binding Residue) Results



Functional Site Discovery from Incomplete Training Data (2011) Robert Langlois Marina Langlois, and Hui Lu. *In preparation.*

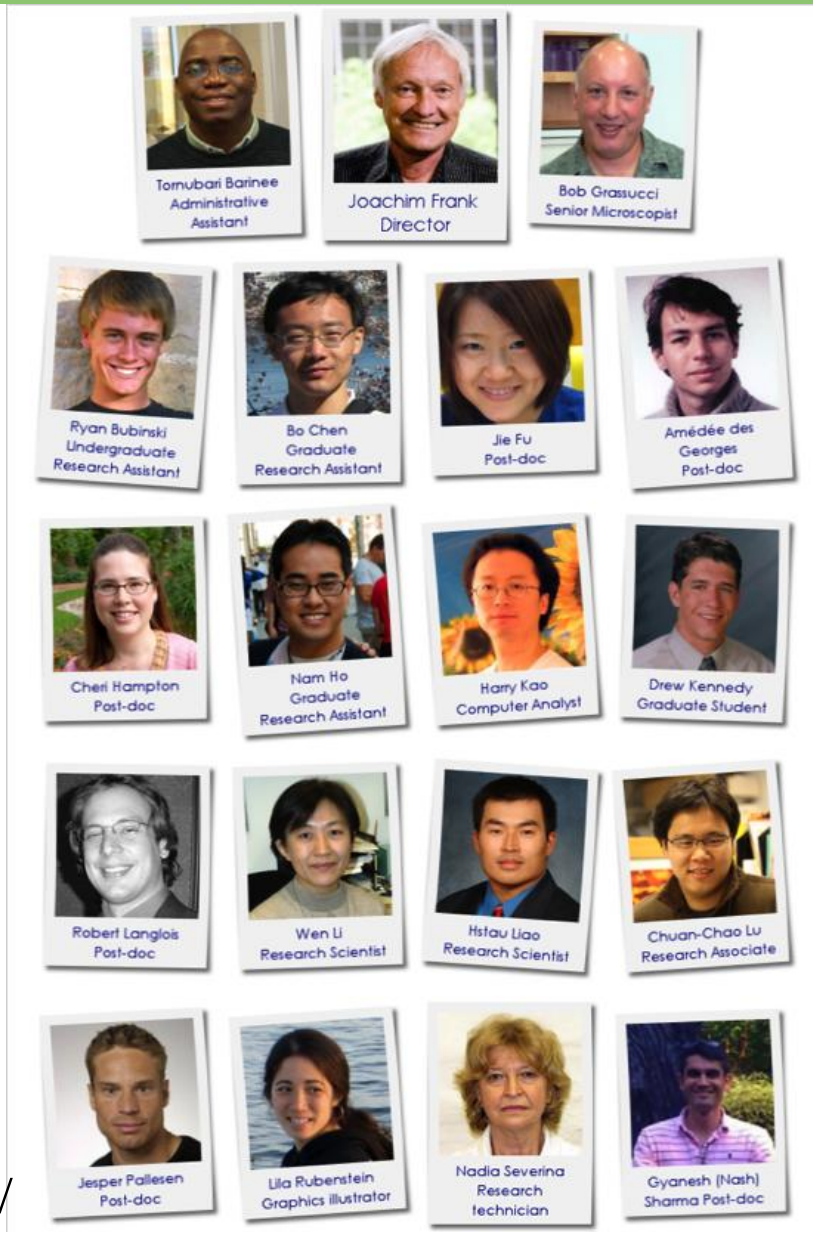
Open Problems

- How to incorporate prediction dependencies in MIL?
  - A protein (bag) is DNA-binding if **a set (not a single)** residue binds DNA
  - The prediction of one residues increases likelihood of one nearby!
  - However, both classification and MIL assume IID – assumption not valid

- Can we improve learning by knowing the label/function of certain proteins?
  - Multiple-instance Active Learning
  - Expensive to label all proteins – possible to label a subset
  - Learning algorithm to find the *optimal* subset

Acknowledgements

- Joachim Frank (PI) - Current
  - Everyone in the Frank Lab
- Hui Lu (PI) – Previous
  - Everyone in the Lu Lab



http://www.mendeley.com/profiles/robert-langlois/