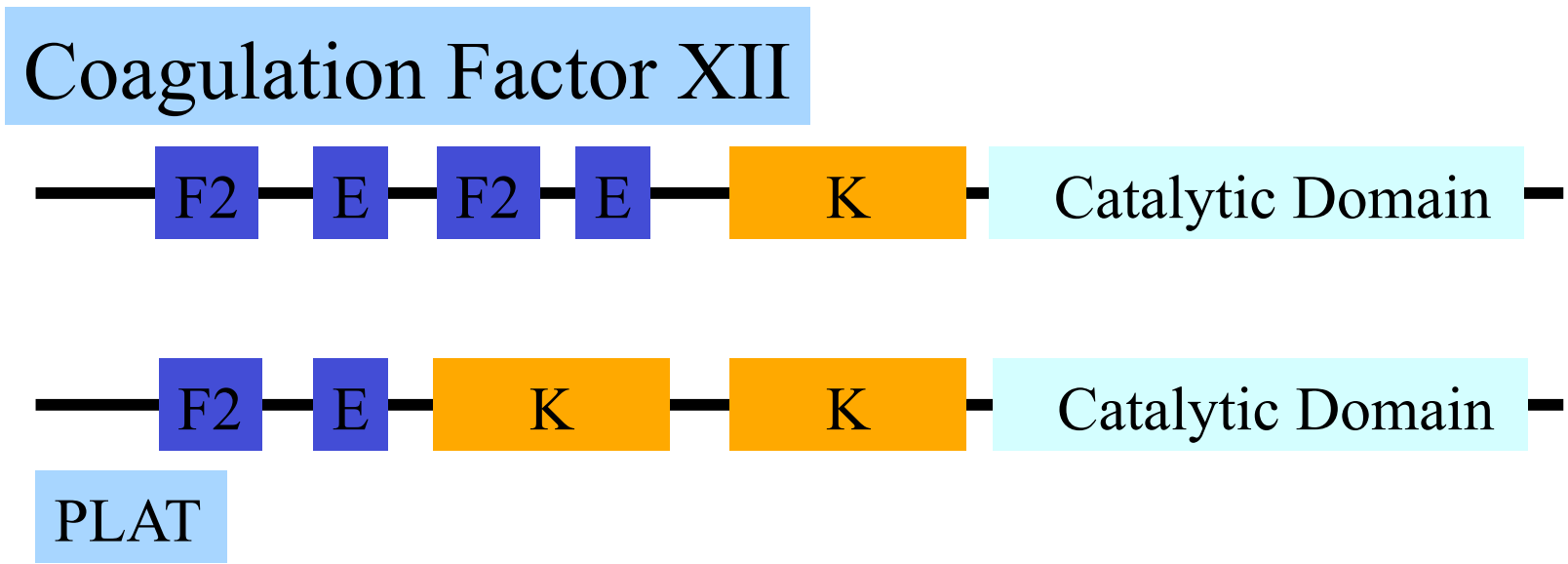
BSC 4934: Q'BIC Capstone Workshop

Giri Narasimhan
ECS 254A; Phone: x3748
giri@cs.fiu.edu
http://www.cs.fiu.edu/~giri/teach/BSC4934_Su11.html
July 2011

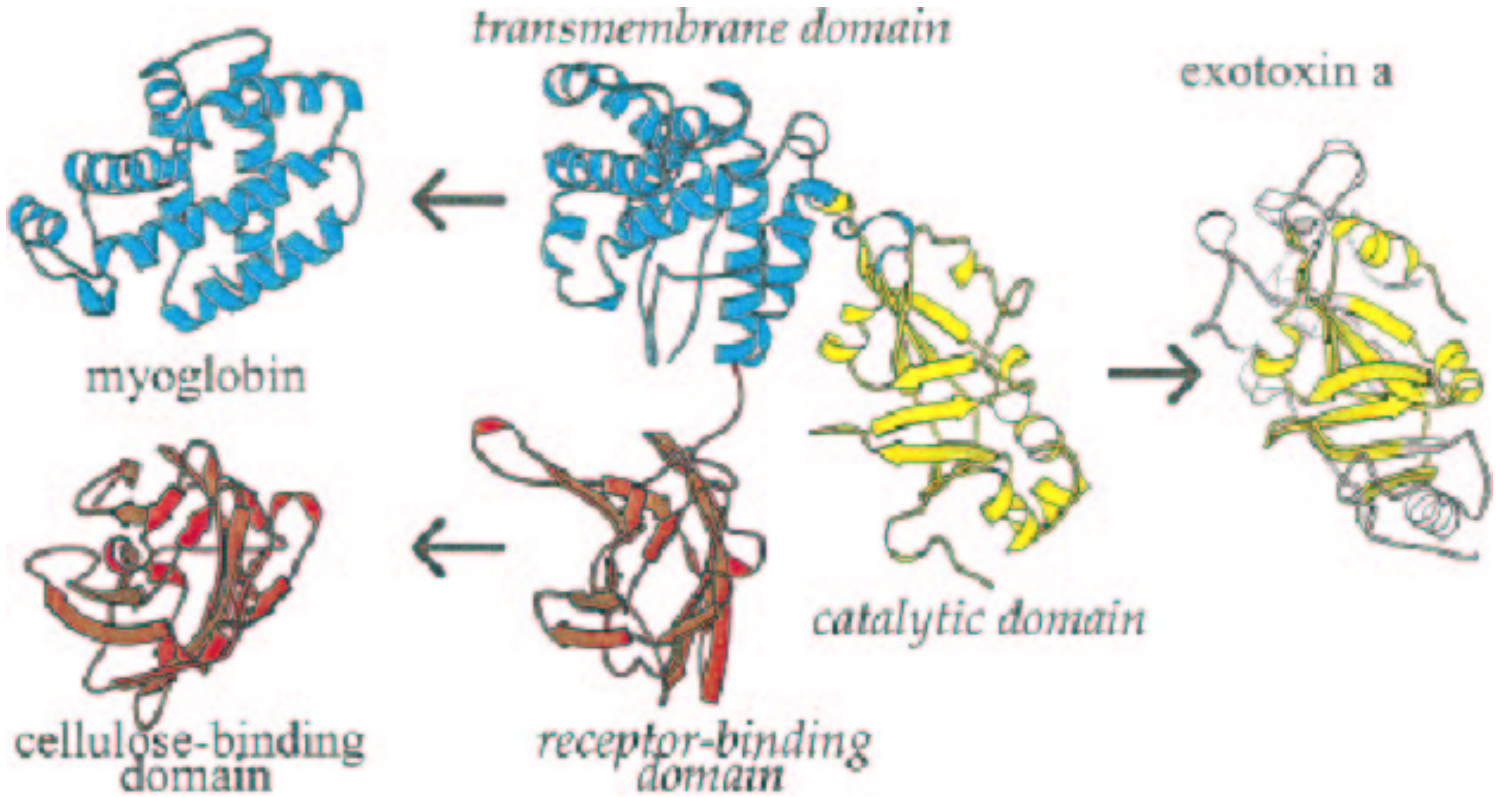
Modular Nature of Proteins

- Proteins are collections of "modular" domains. For example,



Modular Nature of Protein Structures

Example: Diphtheria Toxin



Domain Architecture Tools

- CDART
  - Protein AAH24495; Domain Architecture;
  - It's domain relatives;
  - Multiple alignment for 2nd domain
- SMART

Active Sites

Active sites in proteins are usually hydrophobic pockets/ crevices/troughs that involve sidechain atoms.

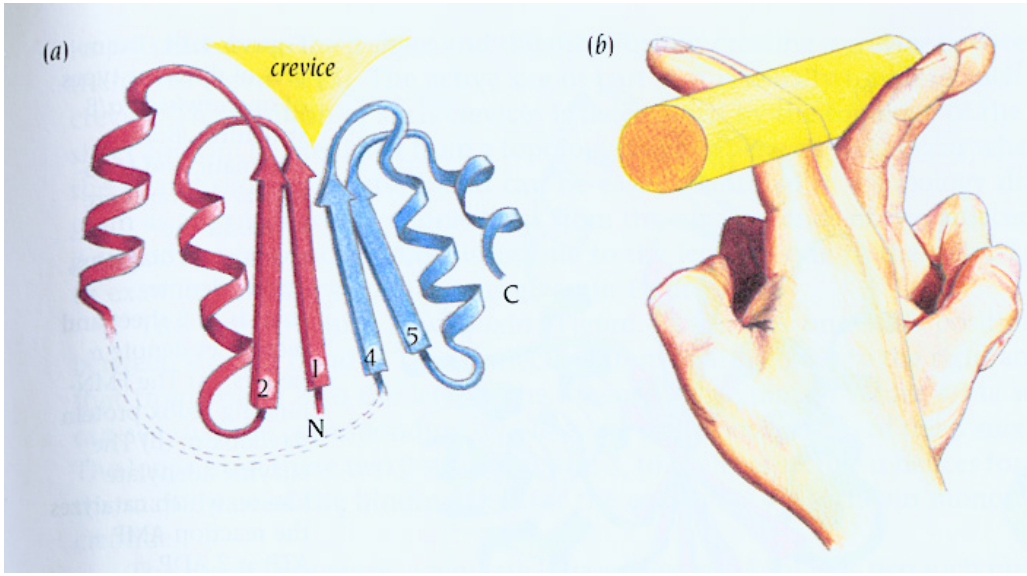
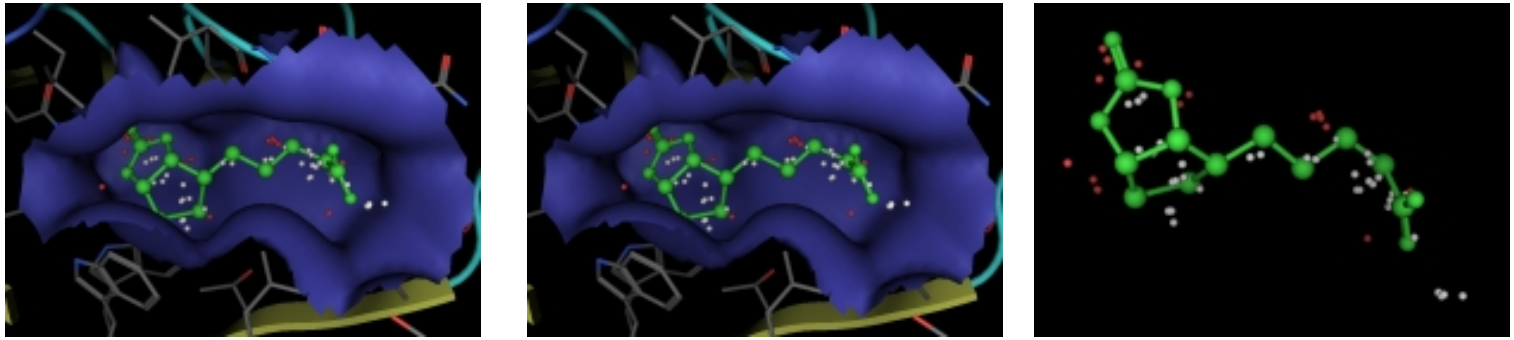


Figure 4.13 (a) The active site in open twisted α/β domains is in a crevice outside the carboxy ends of the β strands. This crevice is formed by two adjacent loop regions that connect the two strands with α helices on opposite sides of the β sheet. This is illustrated by the curled fingers of two hands (b), where the top halves of the fingers represent loop regions and the bottom halves represent the β strands. The rod represents a bound molecule in the binding crevice.

Active Sites



Left PDB 3RTD (streptavidin) and the first site located by the MOE Site Finder. **Middle** 3RTD with complexed ligand (biotin). **Right** Biotin ligand overlaid with calculated alpha spheres of the first site.

Secondary Structure Prediction Software

Figure 11.3 Comparison of secondary structure predictions by various methods. The sequence of flavodoxin, an α/β protein, was used as the query and is shown on the first line of the alignment. For each prediction, H denotes an α helix, E a β strand, T a β turn; all other positions are assumed to be random coil. Correctly assigned residues are shown in inverse type. The methods used are listed along the left side of the alignment and are described in the text. At the bottom of the figure is the secondary structure assignment given in the PDB file for flavodoxin (1OFV, Smith et al., 1983).

PDB: Protein Data Bank

- Database of protein tertiary and quaternary structures and protein complexes. http://www.rcsb.org/pdb/
- Over 29,000 structures as of Feb 1, 2005.
- Structures determined by
  - NMR Spectroscopy
  - X-ray crystallography
  - Computational prediction methods
- Sample PDB file: Click here [ ]

Protein Folding

Unfolded

⇕ Rapid (< 1s)

Molten Globule State

⇕ Slow (1 – 1000 s)

Folded Native State

- How to find minimum energy configuration?

Protein Structures

- Most proteins have a hydrophobic core.
- Within the core, specific interactions take place between amino acid side chains.
- Can an amino acid be replaced by some other amino acid?
  - Limited by space and available contacts with nearby amino acids
- Outside the core, proteins are composed of loops and structural elements in contact with water, solvent, other proteins and other structures.

Viewing Protein Structures

- SPDBV
- RASMOL
- CHIME

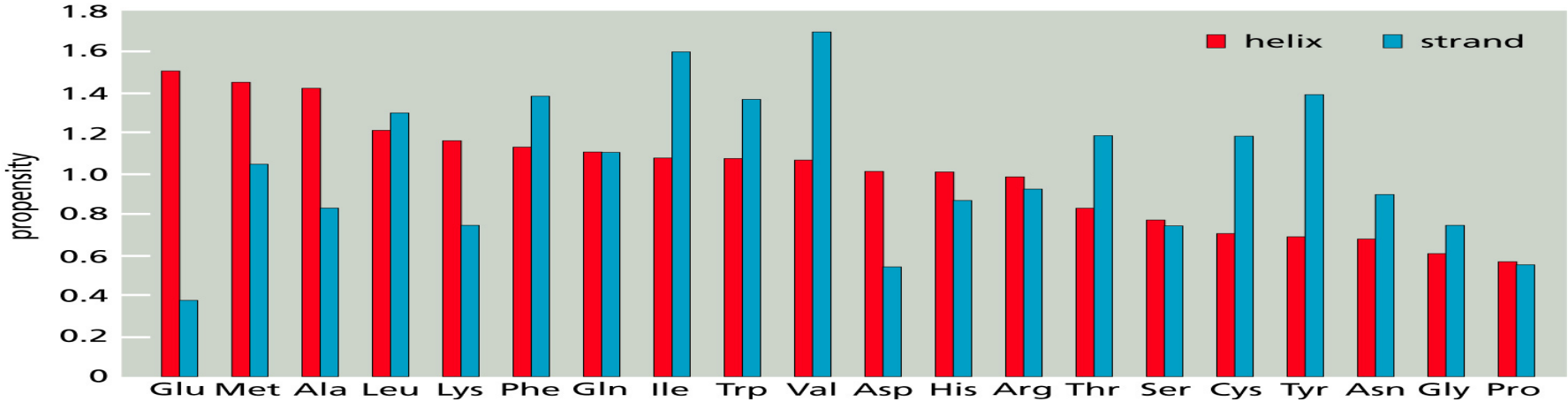
Secondary Structure Prediction Software

Figure 11.3 Comparison of secondary structure predictions by various methods. The sequence of flavodoxin, an α/β protein, was used as the query and is shown on the first line of the alignment. For each prediction, H denotes an α helix, E a β strand, T a β turn; all other positions are assumed to be random coil. Correctly assigned residues are shown in inverse type. The methods used are listed along the left side of the alignment and are described in the text. At the bottom of the figure is the secondary structure assignment given in the PDB file for flavodoxin (1OFV, Smith et al., 1983).

Recent Ones:
GOR V
PREDATOR
Zpred
PROF
NNSSP
PHD
PSIPRED
Jnet

Chou & Fasman Propensities

Amino Acid	helix			
	Designation	*P*	Designation	*P*
Ala	F	1.42	b	0.83
Cys	I	0.70	f	1.19
Asp	I	1.01	B	0.54
Glu	F	1.51	B	0.37
Phe	f	1.13	f	1.38
Gly	B	0.61	b	0.75
His	f	1.00	f	0.87
Ile	f	1.08	F	1.60
Lys	f	1.16	b	0.74
Leu	F	1.21	f	1.30
Met	F	1.45	f	1.05
Asn	b	0.67	b	0.89
Pro	**B**	**0.57**	**B**	**0.55**
Gln	f	1.11	h	1.10
Arg	I	0.98	I	0.93
Ser	I	0.77	b	0.75
Thr	I	0.83	f	1.19
Val	f	1.06	F	1.70
Trp	f	1.08	f	1.37
Tyr	b	0.69	F	1.4



GOR IV prediction for 1bbc

```
AFAGVLNDADIAAALEACKAADSFNHKAFFAKVGLTSKSADDVKKAFAII
CCCCCCCHHHHHHHHHHHHHCCCCCHHHHEEECCCCCCHHHHHHHHHHH
AQDKSGFIEEDELKLFLQNFKADARALTDGETKTFLKAGDSDGDGKIGVD
HHCCCCCHHHHHHHHHHHHHHHHHHHCCCCCEEEEEECCCCCCCCEEECC
DVTALVKA
CEEEEEEC
```

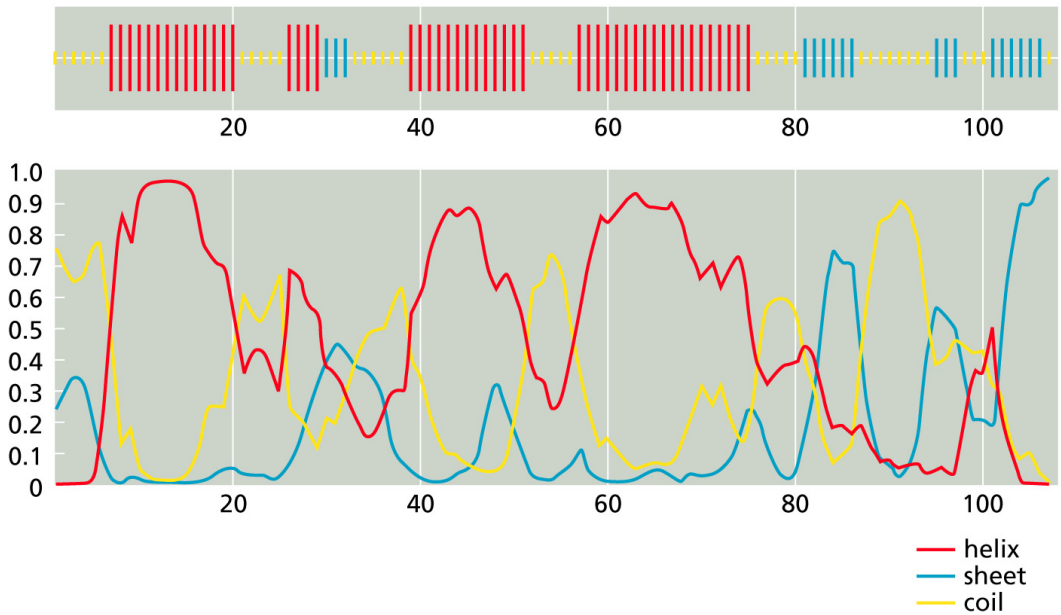
sequence length: 108

GOR IV:

alpha helix (Hh) : 50 is 46.30%

beta sheet (Ee) : 18 is 16.67%

random coil (Cc) : 40 is 37.04%



PDB: Protein Data Bank

- Database of protein tertiary and quaternary structures and protein complexes. http://www.rcsb.org/pdb/
- Over 29,000 structures as of Feb 1, 2005.
- Structures determined by
  - NMR Spectroscopy
  - X-ray crystallography
  - Computational prediction methods
- Sample PDB file: Click here [ ]

PDB Search Results

RCSB PDB PROTEIN DATA BANK

A MEMBER OF THE PDB

An Information Portal to Biological Macromolecular Structures

| PDB Statistics

Contact Us | Help | Print Page

PDB ID or keyword | Author | SEARCH | Advanced Search

Home | Search | Results | Queries

- Results (1-10 of 91)
- Results ID List
- Refine this Search
- 1 Structures Awaiting Release
- Select All
- Deselect All
- Download Selected
- Tabulate
- Narrow Query
- Sort Results
- Results per Page
- Show Query Details
- Results Help

91 Structure Hits | 127 Web Page Hits | 1 Unreleased Structure

1 2 3 4 5 .. 10

1X62 — Solution structure of the LIM domain of carboxyl terminal LIM domain protein 1

Characteristics: Release Date: 17-Nov-2005 Exp. Method: NMR 20 Structures

Classification: Structural Protein

Compound: Mol. Id: 1 Molecule: C Terminal Lim Domain Protein 1 Fragment: Lim Domain

Authors: Qin, X.R., Nagashima, T., Hayashi, F., Yokoyama, S.

1X4K — Solution structure of LIM domain in LIM-protein 3

Characteristics: Release Date: 14-Nov-2005 Exp. Method: NMR 20 Structures

Classification: Metal Binding Protein

Compound: Mol. Id: 1 Molecule: Skeletal Muscle Lim Protein 3 Fragment: Lim Domain

Authors: He, F., Muto, Y., Inoue, M., Kigawa, T., Shirouzu, M., Terada, T., Yokoyama,

1X4L — Solution structure of LIM domain in Four and a half LIM domains protein 2

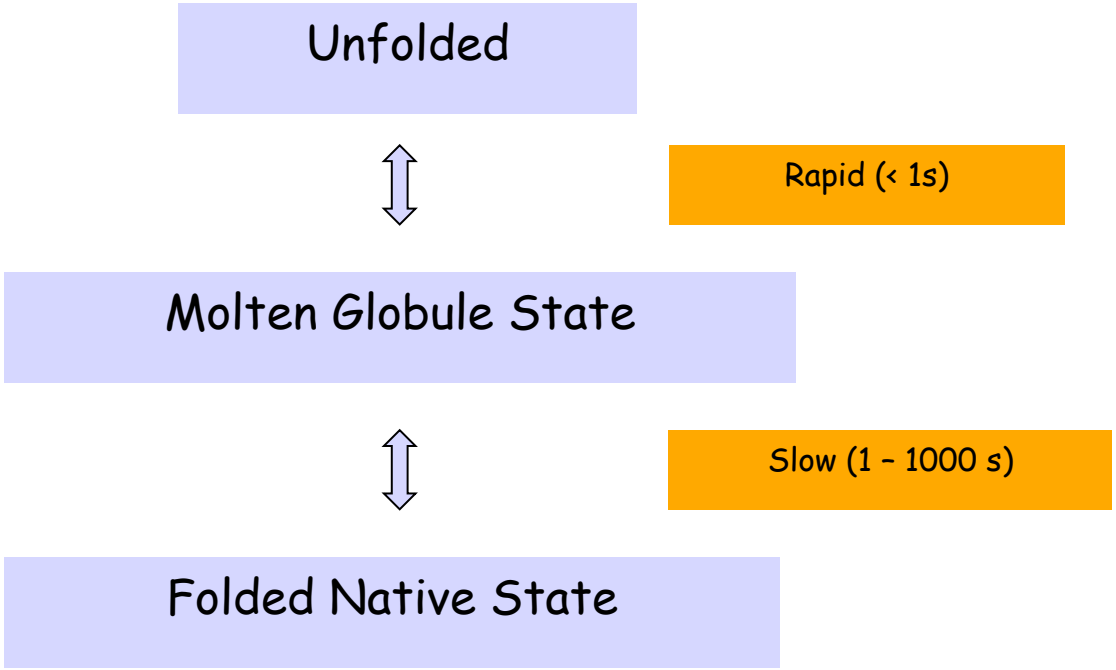
Characteristics: Release Date: 14-Nov-2005 Exp. Method: NMR 20 Structures

Classification: Metal Binding Protein

Compound: Mol. Id: 1 Molecule: Skeletal Muscle Lim Protein 3 Fragment: Lim Domain

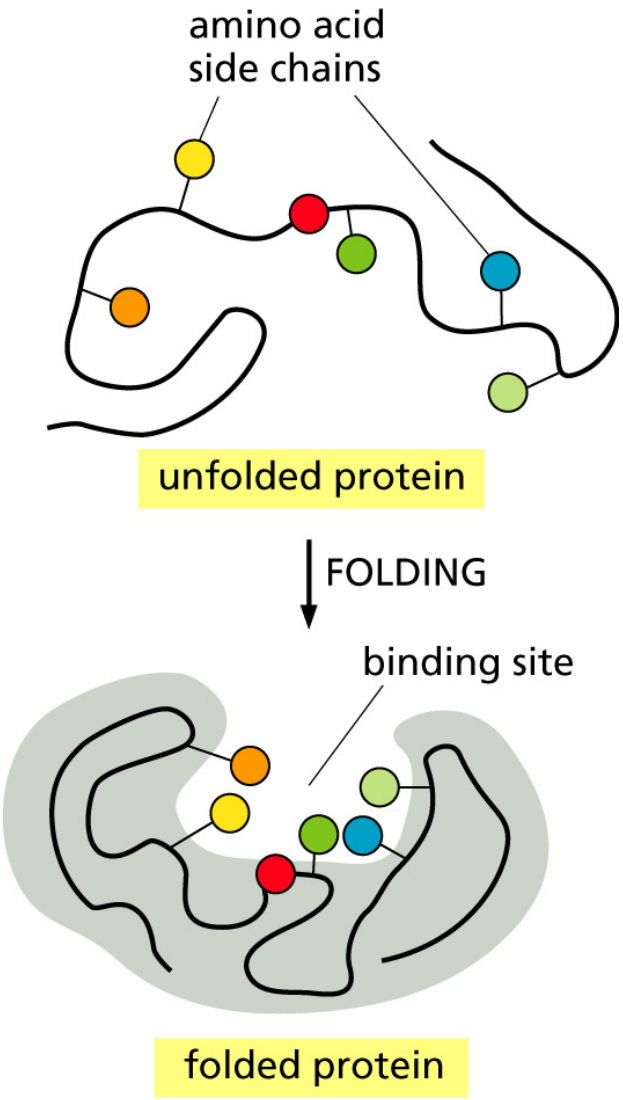
Authors: He, F., Muto, Y., Inoue, M., Kigawa, T., Shirouzu, M., Terada, T., Yokoyama,

Protein Folding

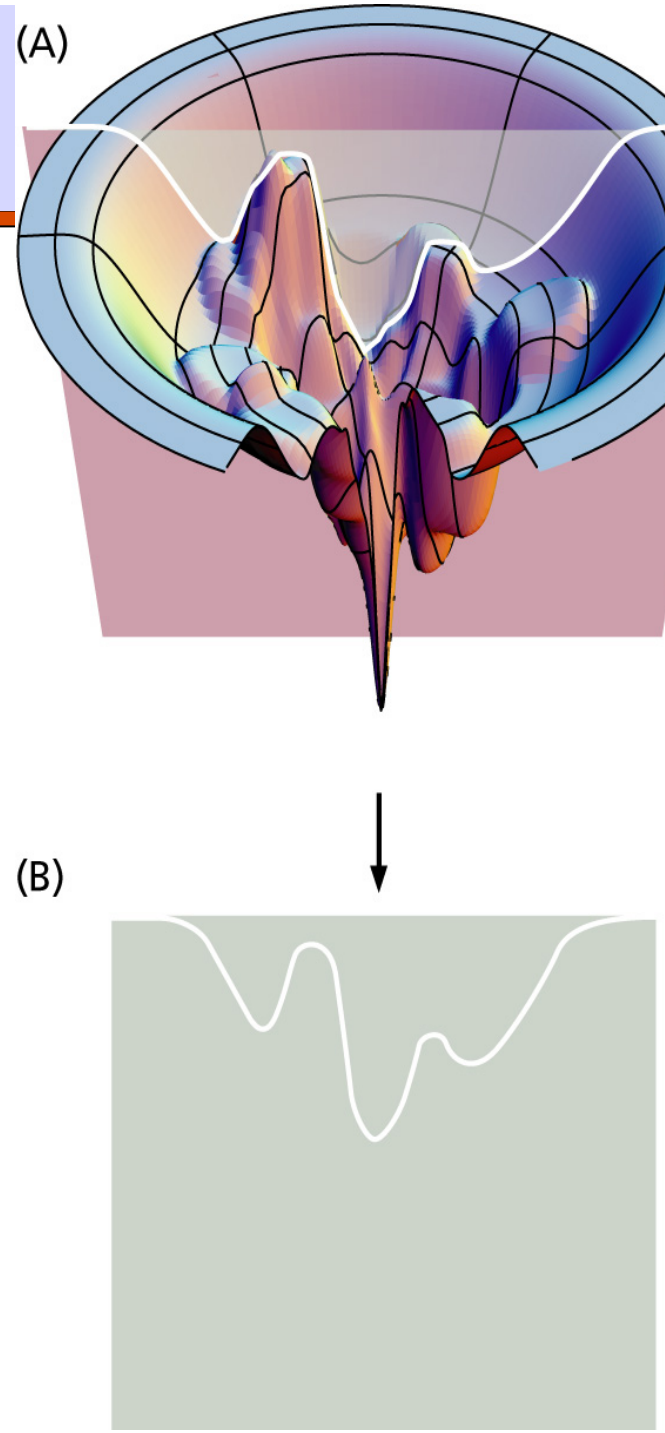


- How to find minimum energy configuration?

Protein Folding



Energy Landscape



Protein Structures

- Most proteins have a hydrophobic core.
- Within the core, specific interactions take place between amino acid side chains.
- Can an amino acid be replaced by some other amino acid?
  - Limited by space and available contacts with nearby amino acids
- Outside the core, proteins are composed of loops and structural elements in contact with water, solvent, other proteins and other structures.

Viewing Protein Structures

- SPDBV
- RASMOL
- CHIME

Structural Alignment

- What is structural alignment of proteins?
  - 3-d superimposition of the atoms as "best as possible", i.e., to minimize RMSD (root mean square deviation).
  - Can be done using VAST and SARF
- Structural similarity is common, even among proteins that do not share sequence similarity or evolutionary relationship.

Other databases & tools

- MMDB contains groups of structurally related proteins
- SARF structurally similar proteins using secondary structure elements
- VAST Structure Neighbors
- SSAP uses double dynamic programming to structurally align proteins

Protein Structure Prediction

- Holy Grail of bioinformatics
- Protein Structure Initiative to determine a set of protein structures that span protein structure space sufficiently well. WHY?
  - Number of folds in natural proteins is limited. Thus a newly discovered proteins should be within modeling distance of some protein in set.
- CASP: Critical Assessment of techniques for structure prediction
  - To stimulate work in this difficult field

PSP Methods

- homology-based modeling
- methods based on fold recognition
  - Threading methods
- *ab initio* methods
  - From first principles
  - With the help of databases

ROSETTA

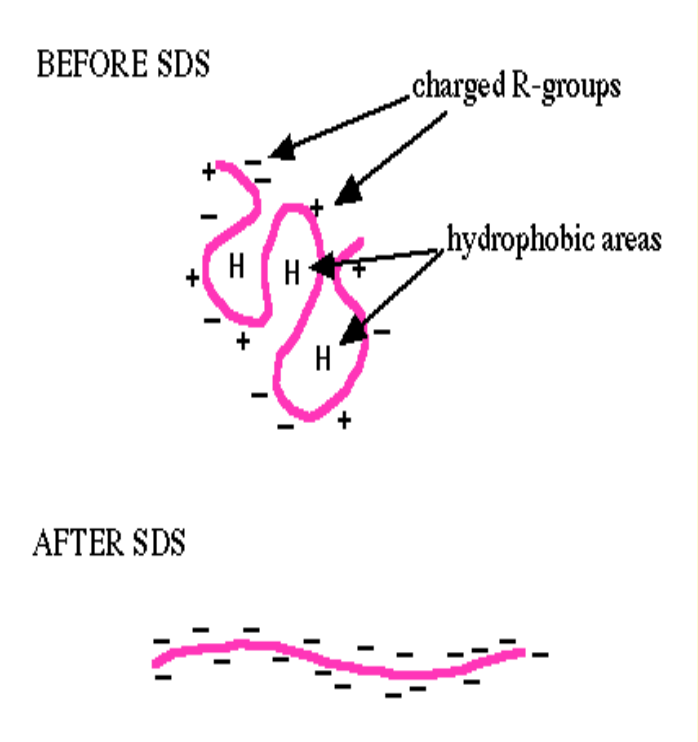
- Best method for PSP
- As proteins fold, a large number of partially folded, low-energy conformations are formed, and that local structures combine to form more global structures with minimum energy.
- Build a database of known structures (I-sites) of short sequences (3-15 residues).
- Monte Carlo simulation assembling possible substructures and computing energy

Modeling Servers

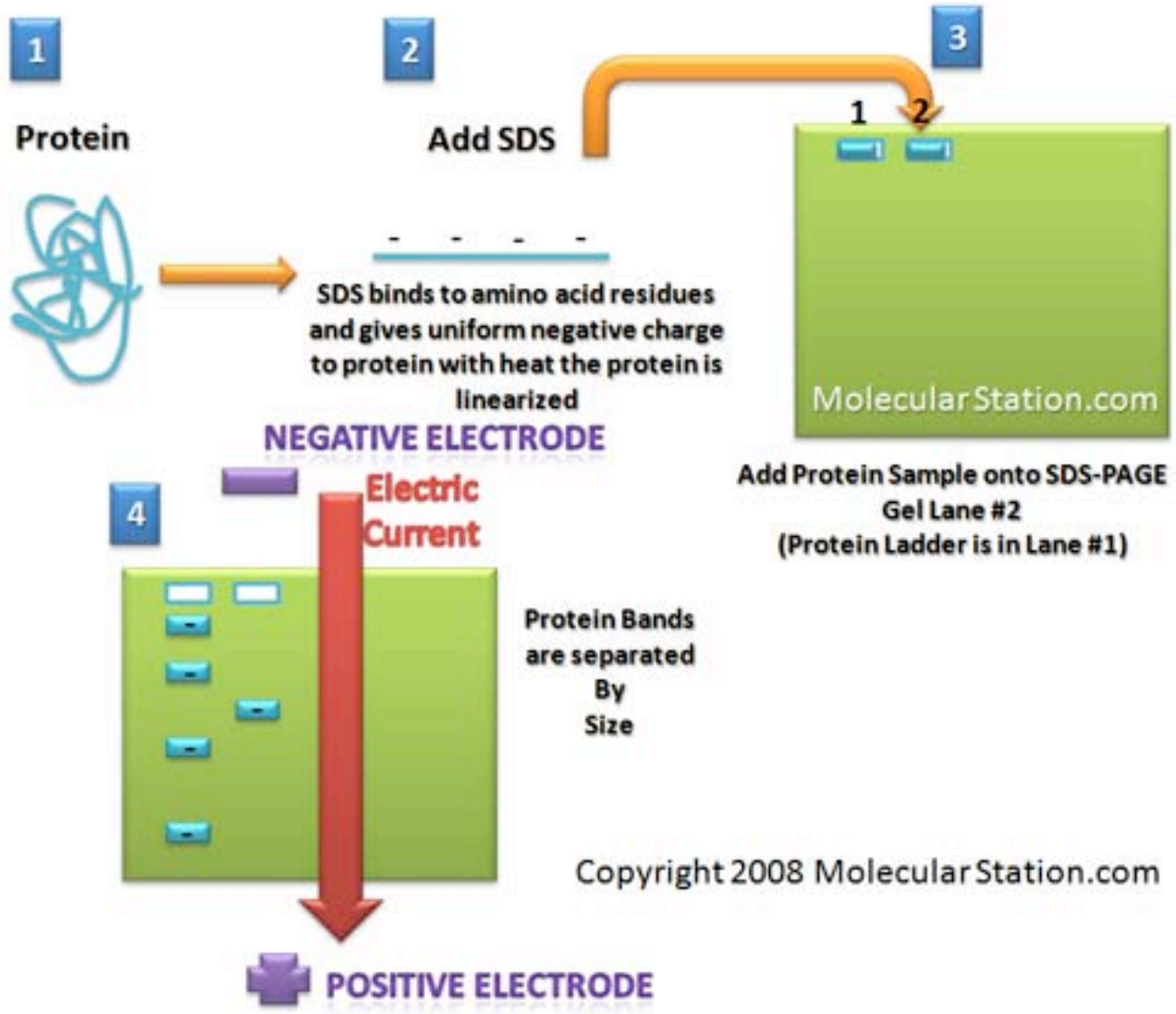
- SwissMODEL
- 3DJigsaw
- CPHModel
- ESyPred3D
- Geno3D
- SDSC1
- Rosetta
- MolIDE
- SCWRL
- PSIPred
- MODELLER
- LOOPY

Gel Electrophoresis for Protein

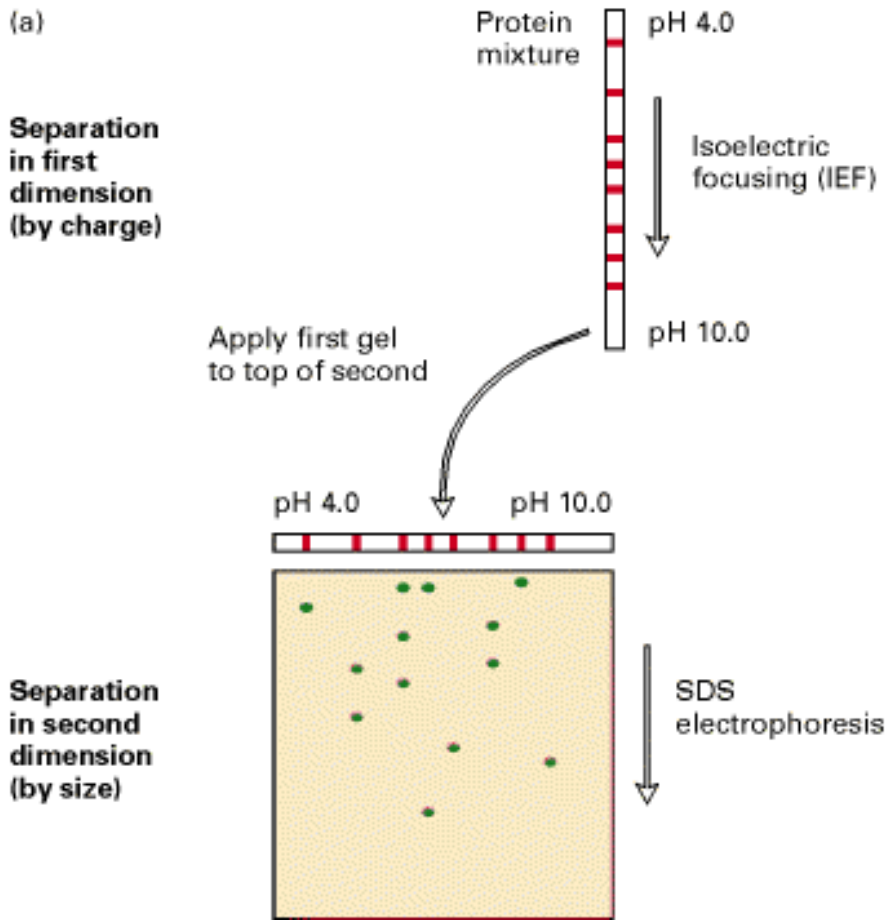
- Protein is also charged
- Has to be denatured - WHY
- Gel: SDS-Polyacrylamide gels
- Add sample to well
- Apply voltage
- Size determines speed
- Add dye to assess the speed
- Stain to see the protein bands



Protein Gel

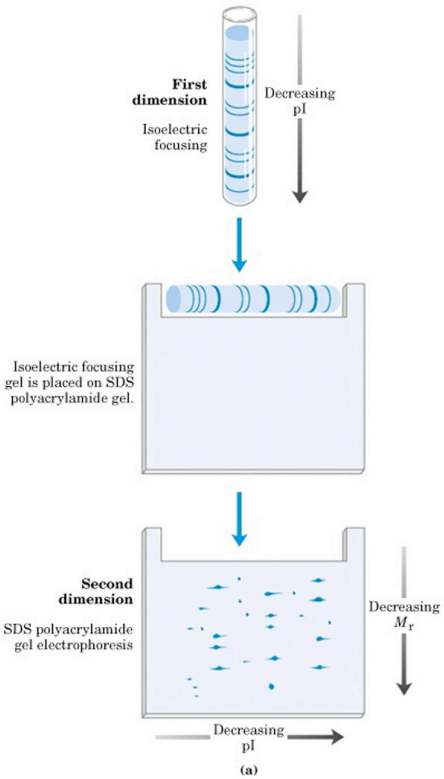

1
Protein
2
Add SDS
SDS binds to amino acid residues and gives uniform negative charge to protein with heat the protein is linearized
3
1
2
MolecularStation.com
Add Protein Sample onto SDS-PAGE Gel Lane #2 (Protein Ladder is in Lane #1)
NEGATIVE ELECTRODE
4
Electric Current
Protein Bands are separated By Size
Copyright 2008 MolecularStation.com
POSITIVE ELECTRODE

2D-Gels


(a)
Protein mixture
pH 4.0
Separation in first dimension (by charge)
Isoelectric focusing (IEF)
pH 10.0
Apply first gel to top of second
pH 4.0
pH 10.0
Separation in second dimension (by size)
SDS electrophoresis

2D Gel Electrophoresis

Two-D Gels



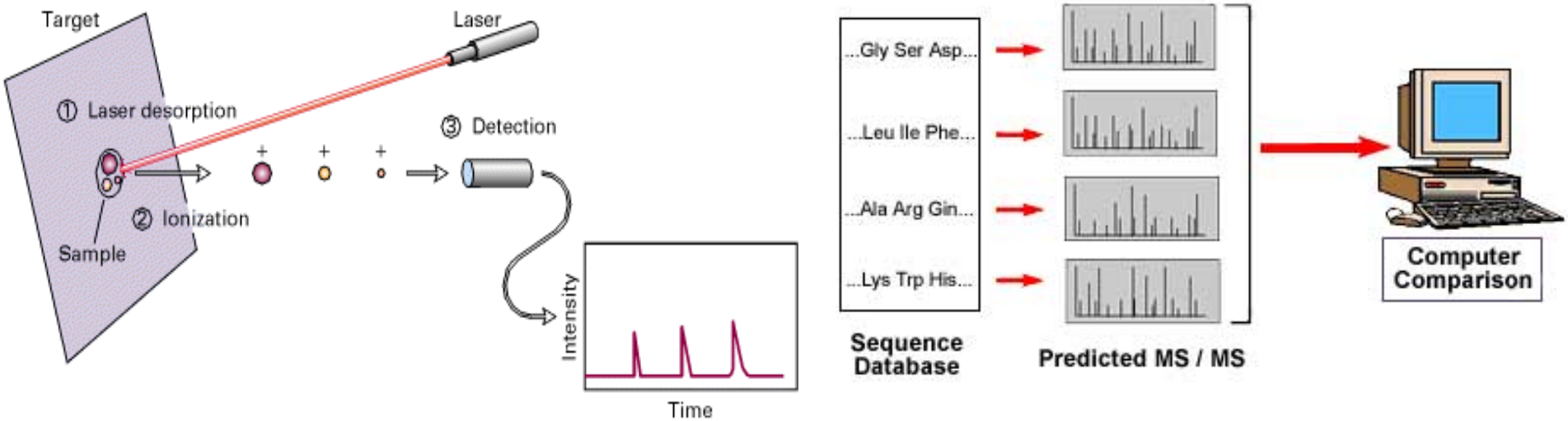
(b)

Mass Spectrometry

- **Mass measurements By Time-of-Flight**
  Pulses of light from laser ionizes protein that is absorbed on metal target. Electric field accelerates molecules in sample towards detector. The time to the detector is inversely proportional to the mass of the molecule. Simple conversion to mass gives the molecular weights of proteins and peptides.

- **Using Peptide Masses to Identify Proteins:**
  One powerful use of mass spectrometers is to identify a protein from its peptide mass fingerprint. A peptide mass fingerprint is a compilation of the molecular weights of peptides generated by a specific protease. The molecular weights of the parent protein prior to protease treatment and the subsequent proteolytic fragments are used to search genome databases for any similarly sized protein with identical or similar peptide mass maps. The increasing availability of genome sequences combined with this approach has almost eliminated the need to chemically sequence a protein to determine its amino acid sequence.

Mass Spectrometry



Protein Sequence

- 20 amino acids
- How is it ordered?
- Basis: Edman Degradation (Pehr Edman)
  - Limited ~30 residues
  - React with Phenylisothiocyanate
  - Cleave and chromatography
- First separate the proteins – Use 2D gels
- Then digest to get pieces
- Then sequence the smaller pieces
- Tedious
- Mass spectrometry

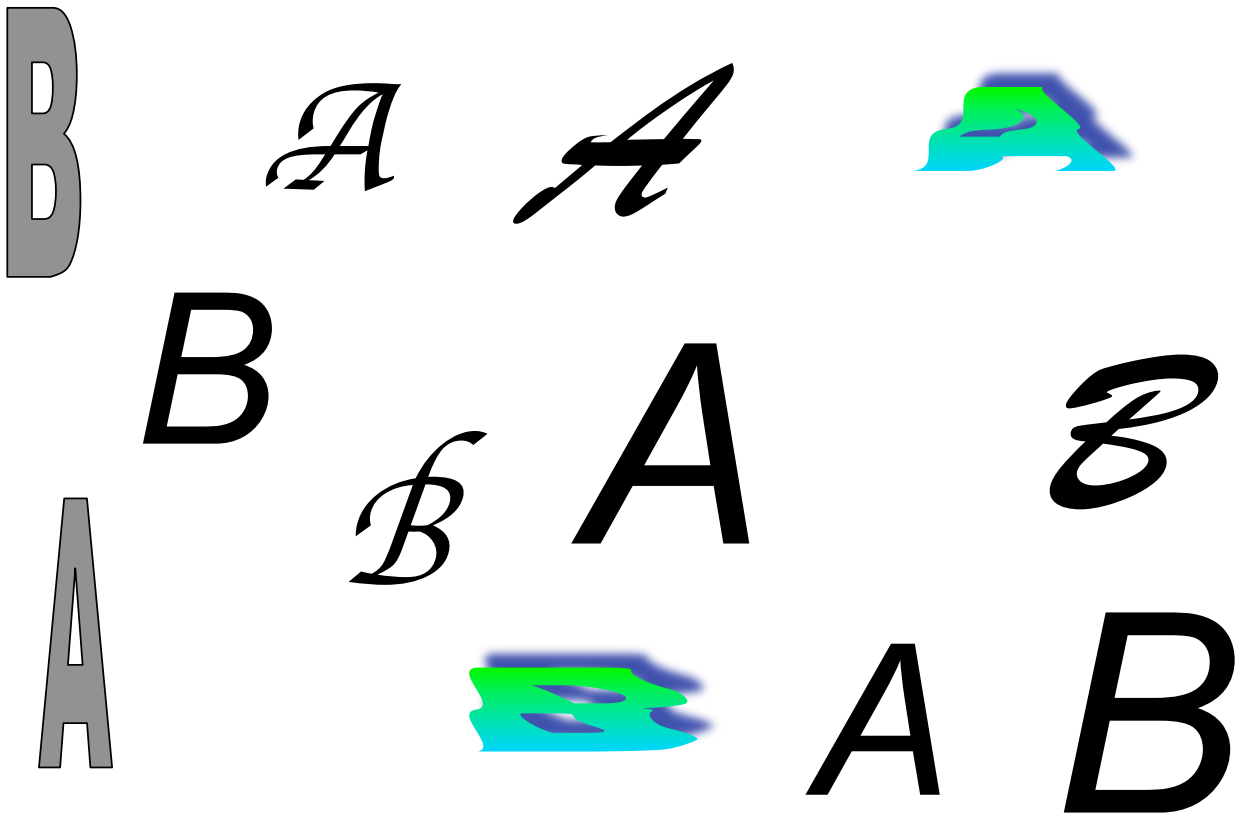
Machine Learning

- Human Endeavor
  - Data → Information → Knowledge
- Machine Learning
  - Automatically extracting information from data
- Types of Machine Learning
  - Unsupervised
    - Clustering
    - Pattern Discovery
  - Supervised
    - Learning
    - Classification

Support Vector Machines

- Supervised Statistical Learning Method for:
  - Classification
  - Regression
- Simplest Version:
  - Training: Present series of labeled examples (e.g., gene expressions of tumor vs. normal cells)
  - Prediction: Predict labels of new examples.

Learning Problems



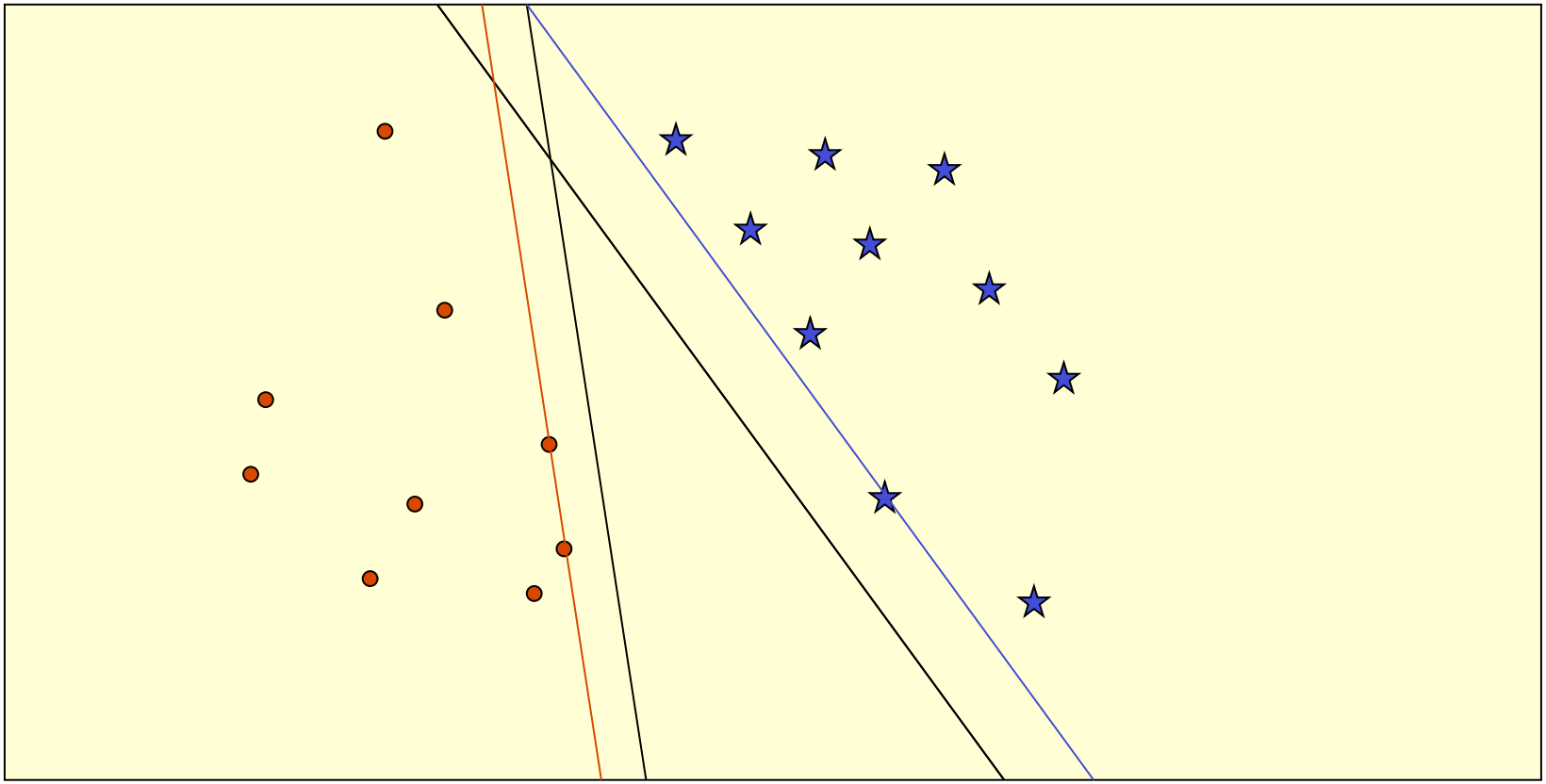
Learning Problems

- Binary Classification
- Multi-class classification
- Regression

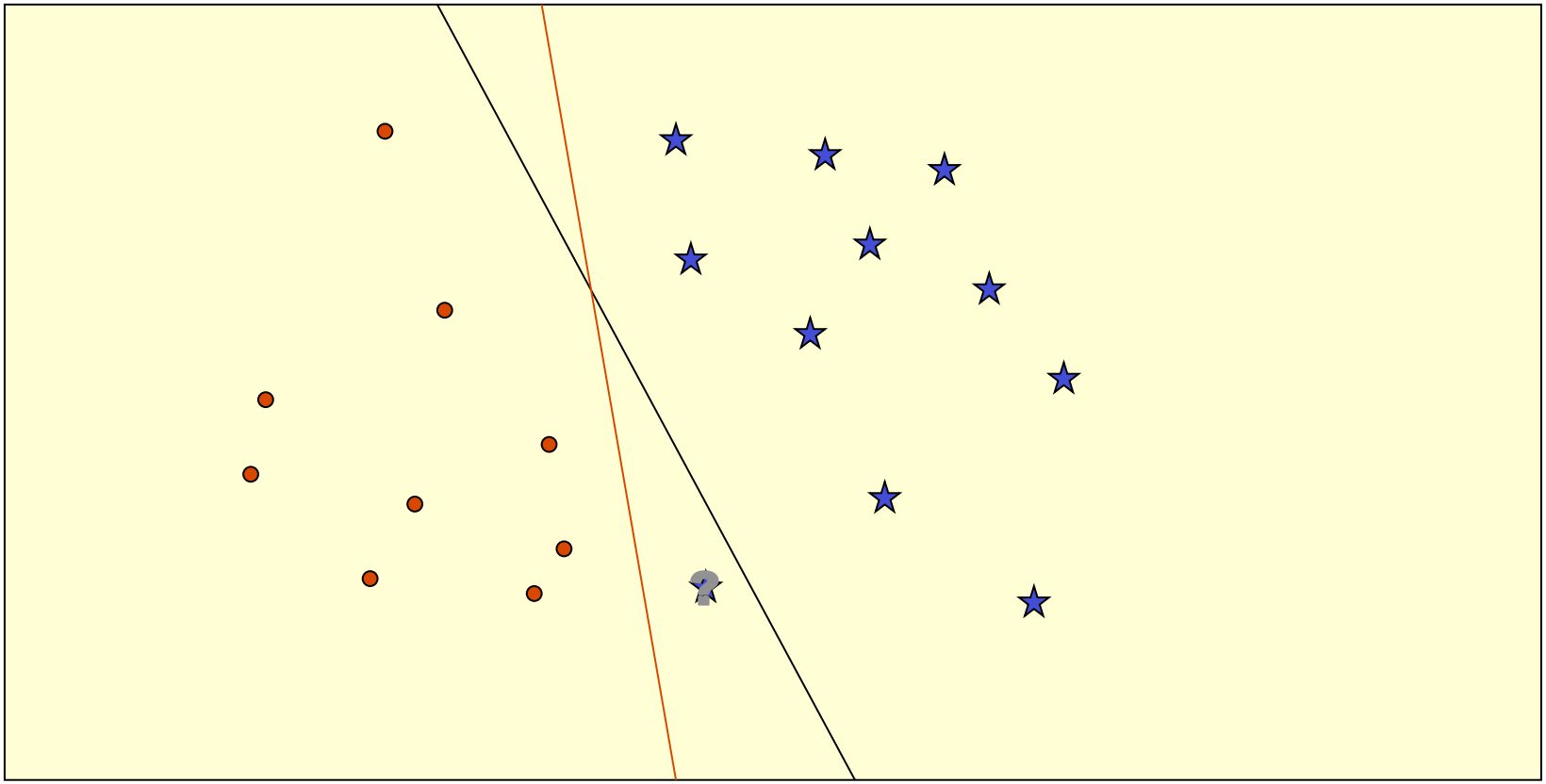
SVM – Binary Classification

- Partition feature space with a surface.
- Surface is implied by a subset of the training points (vectors) near it. These vectors are referred to as Support Vectors.
- Efficient with high-dimensional data.
- Solid statistical theory
- Subsume several other methods.

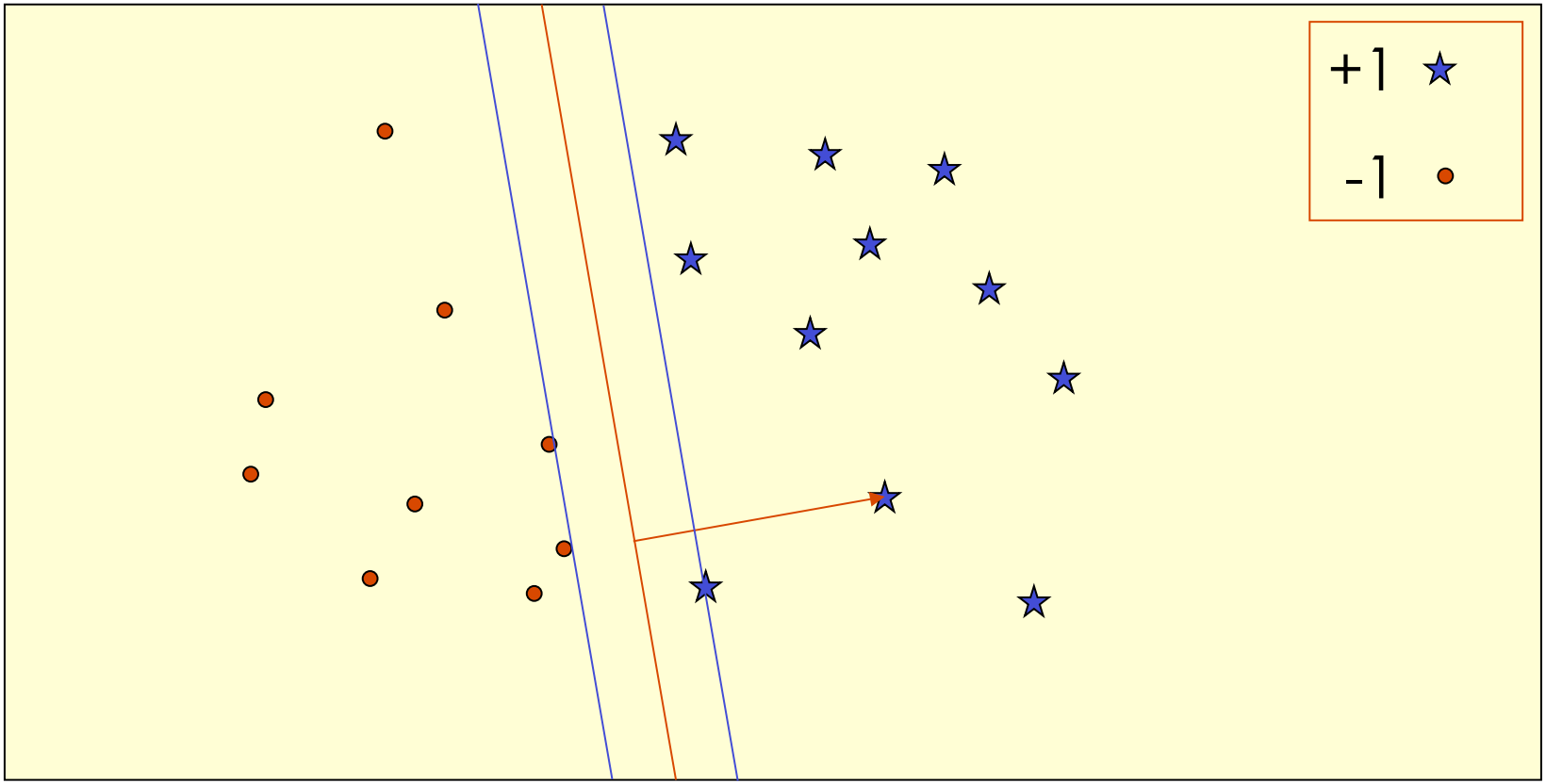
Classification of 2-D (Separable) data



Classification of (Separable) 2-D data

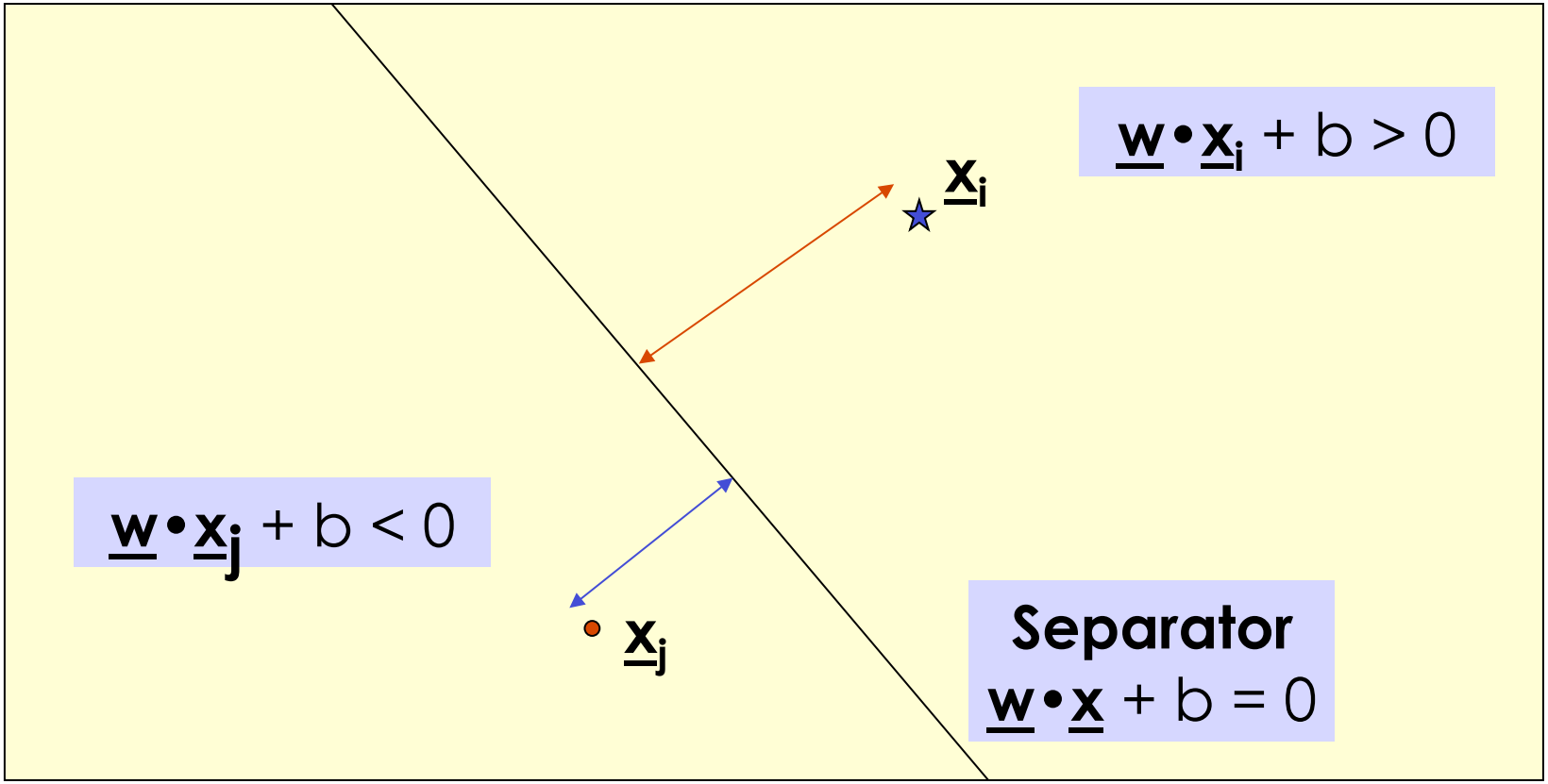


Classification of (Separable) 2-D data



- Margin of a point
- Margin of a point set

Classification using the Separator



Perceptron Algorithm (Primal)

Rosenblatt, 1956

Given separable training set S and learning rate $\eta>0$

$\underline{\mathbf{w}}_0 = \underline{0}$; // Weight

$b_0 = 0$; // Bias

$k = 0$; $R = \max |\underline{x}_i|$

$\underline{\mathbf{w}} = \Sigma\, a_i y_i \underline{\mathbf{x}}_i$

repeat

  for $i = 1$ to N

    if $y_i\,(\underline{\mathbf{w}}_k \bullet \underline{x}_i + b_k) \leq 0$ **then**

      $\underline{\mathbf{w}}_{k+1} = \underline{\mathbf{w}}_k + \eta y_i \underline{\mathbf{x}}_i$

      $b_{k+1} = b_k + \eta y_i R^2$

      $k = k + 1$

Until no mistakes made within loop

Return k, and $(\underline{\mathbf{w}}_k, b_k)$ where k = # of mistakes

Performance for Separable Data

Theorem:

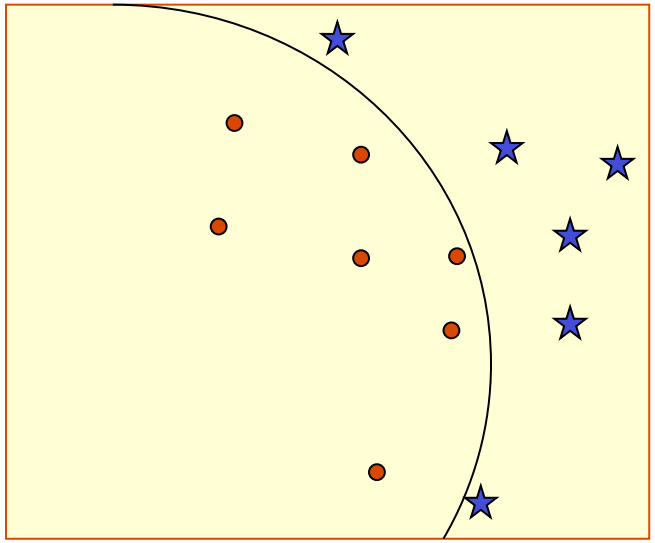
If **margin** **m** of S is positive, then

$$k \leq (2R/m)^2$$

i.e., the algorithm will always converge,

and will converge quickly.

Non-linear Separators



Main idea: Map into feature space

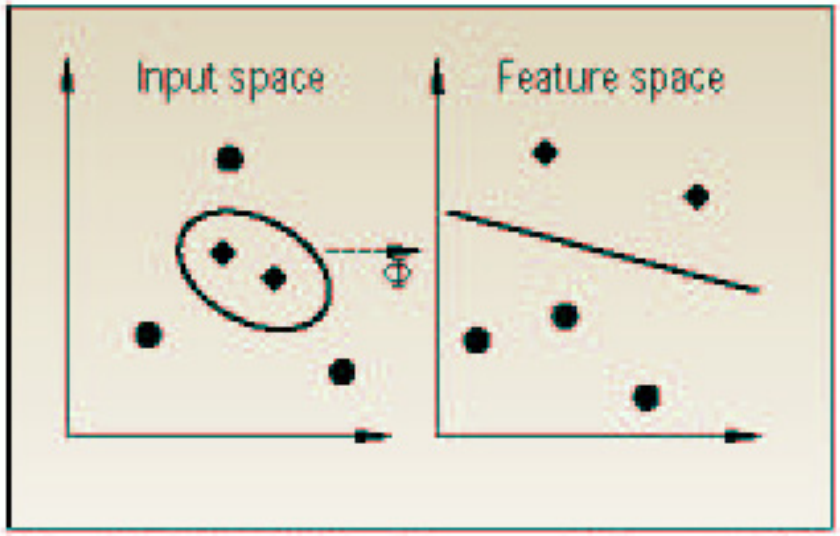
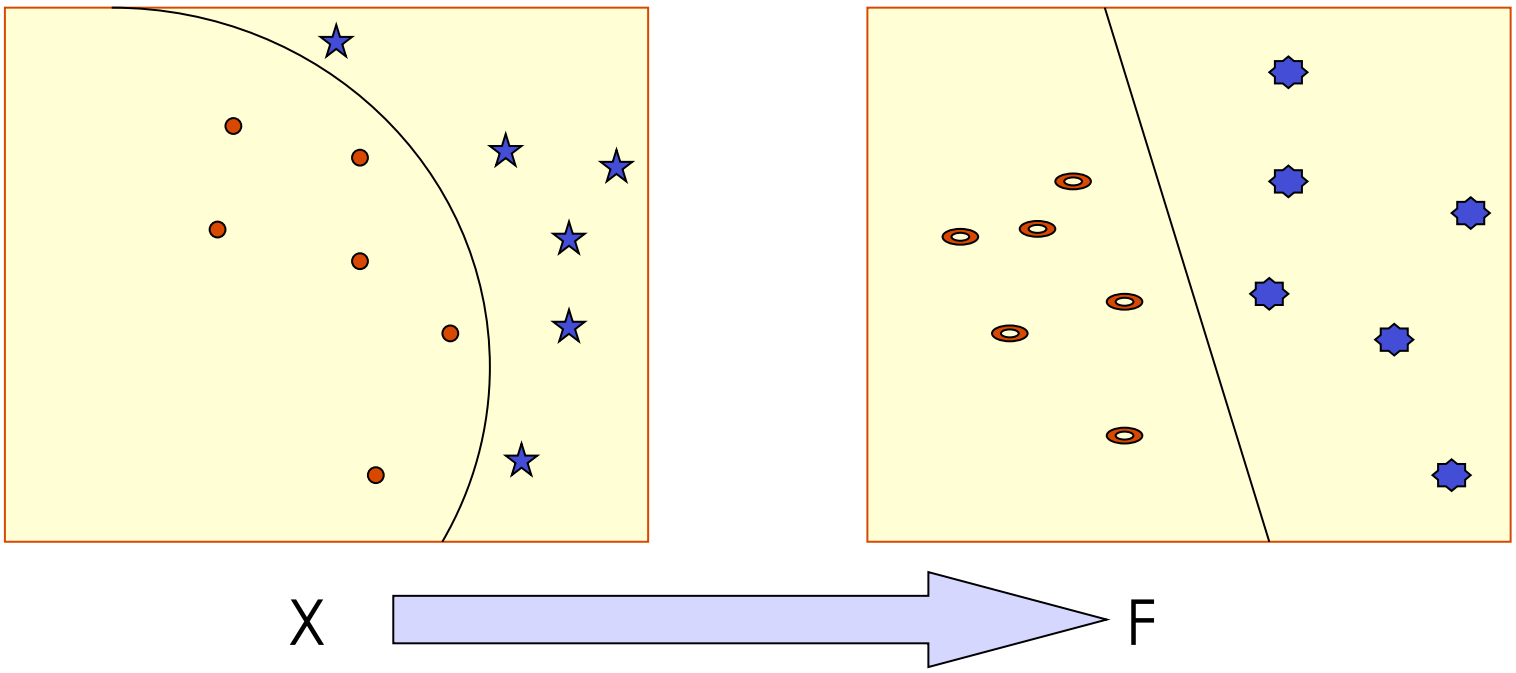


Figure 2. The idea of SV machines: map the training data nonlinearly into a higher-dimensional feature space via Φ, and construct a separating hyperplane with maximum margin there. This yields a nonlinear decision boundary in input space. By the use of a kernel function, it is possible to compute the separating hyperplane without explicitly carrying out the map into the feature space.

Non-linear Separators



Useful URLs

- http://www.support-vector.net

Perceptron Algorithm (Primal)

Rosenblatt, 1956

Given separable training set S and learning rate $\eta>0$

$\underline{\mathbf{w}}_0 = \underline{0}$; // Weight

$b_0 = 0$; // Bias

$k = 0$; $R = \max |\underline{x}_i|$

$\underline{\mathbf{w}} = \Sigma\, a_i y_i \underline{\mathbf{x}}_i$

repeat

  for $i = 1$ to N

    if $y_i\,(\underline{\mathbf{w}}_k \bullet \underline{x}_i + b_k) \leq 0$ **then**

      $\underline{\mathbf{w}}_{k+1} = \underline{\mathbf{w}}_k + \eta y_i \underline{\mathbf{x}}_i$

      $b_{k+1} = b_k + \eta y_i R^2$

      $k = k + 1$

Until no mistakes made within loop

Return k, and $(\underline{\mathbf{w}}_k, b_k)$ where k = # of mistakes

Perceptron Algorithm (Dual)

Given a separable training set S

$\underline{\mathbf{a}} = \underline{0}$; $b_0 = 0$;

$R = \max \; |\underline{x}_i|$

repeat

for $i = 1$ to N

if $y_i \left(\sum a_j y_j \, \underline{x}_i \bullet \underline{x}_j + b\right) \leq 0$ **then**

$a_i = a_i + 1$

$b = b + y_i R^2$

endif

Until no mistakes made within loop

Return $(\underline{\mathbf{a}}, b)$

Perceptron Algorithm (Dual)

Given a separable training set S

$\underline{\mathbf{a}} = \underline{0}$; $b_0 = 0$;

$R = \max \; |\underline{x}_i|$

repeat

  for $i = 1$ to N

    if $y_i \left(\sum a_j y_j \, W(\underline{x}_i, \underline{x}_j) + b\right) \leq 0$ **then**

      $a_i = a_i + 1$

      $b = b + y_i R^2$

Until no mistakes made within loop

Return $(\underline{\mathbf{a}}, b)$

$W(\underline{x}_i, \underline{x}_j) = \Phi(\underline{x}_i) \bullet \Phi(\underline{x}_j)$

Different Kernel Functions

- Polynomial kernel

$$\kappa(X,Y) = (X \bullet Y)^d$$

- Radial Basis Kernel

$$\kappa(X,Y) = \exp\left(\frac{-\|X-Y\|^2}{2\sigma^2}\right)$$

- Sigmoid Kernel

$$\kappa(X,Y) = \tanh(\omega(X \bullet Y) + \theta)$$

SVM Ingredients

- Support Vectors
- Mapping from Input Space to Feature Space
- Dot Product – Kernel function
- Weights

Generalizations

- How to deal with more than 2 classes?

  Idea: Associate weight and bias for each class.
- How to deal with non-linear separator?

  Idea: Support Vector Machines.
- How to deal with linear regression?
- How to deal with non-separable data?

Applications

- Text Categorization & Information Filtering
  - 12,902 Reuters Stories, 118 categories (91% !!)
- Image Recognition
  - Face Detection, tumor anomalies, defective parts in assembly line, etc.
- Gene Expression Analysis
- Protein Homology Detection

		Learned threshold					Optimized threshold				
Class	Method	FP	FN	TP	TN	Cost	FP	FN	TP	TN	Cost
Tricarboxylic acid	Radial SVM	8	8	9	2442	24	4	7	10	2446	18
	Dot-product-1 SVM	11	9	8	2439	29	3	6	11	2447	15
	Dot-product-2 SVM	5	10	7	2445	25	4	6	11	2446	16
	Dot-product-3 SVM	4	12	5	2446	28	4	6	11	2446	16
	Parzen	4	12	5	2446	28	0	12	5	2450	24
	FLD	9	10	7	2441	29	7	8	9	2443	23
	C4.5	7	17	0	2443	41	–	–	–	–	–
	MOC1	3	16	1	2446	35	–	–	–	–	–
Respiration	Radial SVM	9	6	24	2428	21	8	4	26	2429	16
	Dot-product-1 SVM	21	10	20	2416	41	6	9	21	2431	24
	Dot-product-2 SVM	7	14	16	2430	35	7	6	24	2430	19
	Dot-product-3 SVM	3	15	15	2434	33	7	6	24	2430	19
	Parzen	22	10	20	2415	42	7	12	18	2430	31
	FLD	10	10	20	2427	30	14	4	26	2423	22
	C4.5	18	17	13	2419	52	–	–	–	–	–
	MOC1	12	26	4	2425	64	–	–	–	–	–
Ribosome	Radial SVM	9	4	117	2337	17	6	1	120	2340	8
	Dot-product-1 SVM	13	6	115	2333	25	11	1	120	2335	13
	Dot-product-2 SVM	7	10	111	2339	27	9	1	120	2337	11
	Dot-product-3 SVM	3	18	103	2343	39	7	1	120	2339	9
	Parzen	6	8	113	2340	22	5	8	113	2341	21
	FLD	15	5	116	2331	25	8	3	118	2338	14
	C4.5	31	21	100	2315	73	–	–	–	–	–
	MOC1	26	26	95	2320	78	–	–	–	–	–

Table 2: **Comparison of error rates for various classification methods.** Classes are as described in Table 1. The methods are the radial basis function SVM, the SVMs using the scaled dot product kernel raised to the first, second and third power, Parzen windows, Fisher's linear discriminant, and the two decision tree learners, C4.5 and MOC1. The next five columns are the false positive, false negative, true positive and true negative rates summed over three cross-validation splits, followed by the cost, which is the number of false positives plus twice the number of false negatives. These five columns are repeated twice, first using the threshold learned from the training set, and then using the threshold that minimizes the cost on the test set. The threshold optimization is not possible for the decision tree methods, since they do not produce ranked results.

		Learned threshold					Optimized threshold				
Class	Method	FP	FN	TP	TN	Cost	FP	FN	TP	TN	Cost
Proteasome	Radial SVM	3	7	28	2429	17	4	5	30	2428	14
	Dot-product-1 SVM	14	11	24	2418	36	2	7	28	2430	16
	Dot-product-2 SVM	4	13	22	2428	30	4	6	29	2428	16
	Dot-product-3 SVM	3	18	17	2429	39	2	7	28	2430	16
	Parzen	21	5	30	2411	31	3	9	26	2429	21
	FLD	7	12	23	2425	31	12	7	28	2420	26
	C4.5	17	10	25	2415	37	–	–	–	–	–
	MOC1	10	17	18	2422	44	–	–	–	–	–
Histone	Radial SVM	0	2	9	2456	4	0	2	9	2456	4
	Dot-product-1 SVM	0	4	7	2456	8	0	2	9	2456	4
	Dot-product-2 SVM	0	5	6	2456	10	0	2	9	2456	4
	Dot-product-3 SVM	0	8	3	2456	16	0	2	9	2456	4
	Parzen	2	3	8	2454	8	1	3	8	2455	7
	FLD	0	3	8	2456	6	2	1	10	2454	4
	C4.5	2	2	9	2454	6	–	–	–	–	–
	MOC1	2	5	6	2454	12	–	–	–	–	–
Helix-turn-helix	Radial SVM	1	16	0	2450	33	0	16	0	2451	32
	Dot-product-1 SVM	20	16	0	2431	52	0	16	0	2451	32
	Dot-product-2 SVM	4	16	0	2447	36	0	16	0	2451	32
	Dot-product-3 SVM	1	16	0	2450	33	0	16	0	2451	32
	Parzen	14	16	0	2437	46	0	16	0	2451	32
	FLD	14	16	0	2437	46	0	16	0	2451	32
	C4.5	2	16	0	2449	34	–	–	–	–	–
	MOC1	6	16	0	2445	38	–	–	–	–	–

Table 3: **Comparison of error rates for various classification methods (continued).** See caption for Table 2.

Class	Kernel	Cost for each split					Total
Tricarboxylic acid	Radial	18	21	15	22	21	97
	Dot-product-1	15	22	18	23	22	100
	Dot-product-2	16	22	17	22	22	99
	Dot-product-3	16	22	17	23	22	100
Respiration	Radial	16	18	23	20	16	93
	Dot-product-1	24	24	29	27	23	127
	Dot-product-2	19	19	26	24	23	111
	Dot-product-3	19	19	26	22	21	107
Ribosome	Radial	8	12	15	11	13	59
	Dot-product-1	13	18	14	16	16	77
	Dot-product-2	11	16	14	16	15	72
	Dot-product-3	9	15	11	15	15	65
Proteasome	Radial	14	10	9	11	11	55
	Dot-product-1	16	12	12	17	19	76
	Dot-product-2	16	13	15	17	17	78
	Dot-product-3	16	13	16	16	17	79
Histone	Radial	4	4	4	4	4	20
	Dot-product-1	4	4	4	4	4	20
	Dot-product-2	4	4	4	4	4	20
	Dot-product-3	4	4	4	4	4	20

Table 4: **Comparison of SVM performance using various kernels.** For each of the MYGD classifications, SVMs were trained using four different kernel functions on five different random three-fold splits of the data, training on two-thirds and testing on the remaining third. The first column contains the class, as described in Table 1. The second column contains the kernel function, as described in Table 2. The next five columns contain the threshold-optimized cost (i.e., the number of false positives plus twice the number of false negatives) for each of the five random three-fold splits. The final column is the total cost across all five splits.

Family	Gene	Locus	Error	Description
TCA	YPR001W	CIT3	FN	mitochondrial citrate synthase
	YOR142W	LSC1	FN	α subunit of succinyl-CoA ligase
	YNR001C	CIT1	FN	mitochondrial citrate synthase
	YLR174W	IDP2	FN	isocitrate dehydrogenase
	YIL125W	KGD1	FN	α-ketoglutarate dehydrogenase
	YDR148C	KGD2	FN	component of α-ketoglutarate dehydrogenase complex in mitochondria
	YDL066W	IDP1	FN	mitochondrial form of isocitrate dehydrogenase
	YBL015W	ACH1	FP	acetyl CoA hydrolase
Resp	YPR191W	QCR2	FN	ubiquinol cytochrome-c reductase core protein 2
	YPL271W	ATP15	FN	ATP synthase epsilon subunit
	YPL262W	FUM1	FP	fumarase
	YML120C	NDI1	FP	mitochondrial NADH ubiquinone 6 oxidoreductase
	YKL085W	MDH1	FP	mitochondrial malate dehydrogenase
	YDL067C	COX9	FN	subunit VIIa of cytochrome c oxidase
Ribo	YPL037C	EGD1	FP	β subunit of the nascent-polypeptide-associated complex (NAC)
	YLR406C	RPL31B	FN	ribosomal protein L31B (L34B) (YL28)
	YLR075W	RPL10	FP	ribosomal protein L10
	YAL003W	EFB1	FP	translation elongation factor EF-1β
Prot	YHR027C	RPN1	FN	subunit of 26S proteasome (PA700 subunit)
	YGR270W	YTA7	FN	member of CDC48/PAS1/SEC18 family of ATPases
	YGR048W	UFD1	FP	ubiquitin fusion degradation protein
	YDR069C	DOA4	FN	ubiquitin isopeptidase
	YDL020C	RPN4	FN	involved in ubiquitin degradation pathway
Hist	YOL012C	HTA3	FN	histone-related protein
	YKL049C	CSE4	FN	required for proper kinetochore function

Table 6: **Consistently misclassified genes.** The table lists all 25 genes that are consistently misclassified by SVMs trained using the MYGD classifications listed in Table 1. Two types of errors are included: a false positive (FP) occurs when the SVM includes the gene in the given class but the MYGD classification does not; a false negative (FN) occurs when the SVM does not include the gene in the given class but the MYGD classification does.

Kernel DF	Feature	FP	FN	TP	TN
dot-product 0	25	5	4	10	12
dot-product 2	25	5	2	12	12
dot-product 5	25	4	2	12	13
dot-product 10	25	4	2	12	13
dot-product 0	50	4	2	12	13
dot-product 2	50	3	2	12	14
dot-product 5	50	3	2	12	14
dot-product 10	50	3	2	12	14
dot-product 0	100	4	3	11	13
dot-product 2	100	5	3	11	12
dot-product 5	100	5	3	11	12
dot-product 10	100	5	3	11	12
dot-product 0	500	5	3	11	12
dot-product 2	500	4	3	11	13
dot-product 5	500	4	3	11	13
dot-product 10	500	4	3	11	13
dot-product 0	1000	7	3	11	10
dot-product 2	1000	5	3	11	12
dot-product 5	1000	5	3	11	12
dot-product 10	1000	5	3	11	12
dot-product 0	97802	17	0	14	0
dot-product 2	97802	9	2	12	8
dot-product 5	97802	7	3	11	10
dot-product 10	97802	5	3	11	12

Table 1: **Error rates for ovarian cancer tissue experiments.** For each setting of the SVM consisting of a kernel and diagonal factor (DF), each tissue was classified. Column 2 is the number of features (clones) used. Reported are the number of normal tissues misclassified (FP), tumor tissues misclassified (FN), tumor tissues classified correctly (TP), and normal tissues classified correctly (TN).

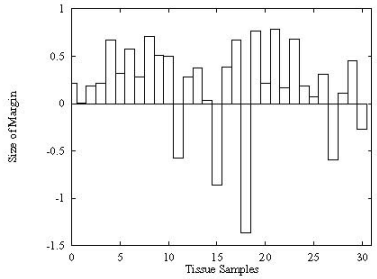


Figure 1: **SVM classification margins for ovarian tissues.** When classifying, the SVM calculates a margin which is the distance of an example from the decision boundary it has learned. In this graph, the margin for each tissue sample calculated using (10) is shown. A positive value indicates a correct classification, and a negative value indicates an incorrect classification. The most negative point corresponds to tissue N039. The second most negative point corresponds to tissue HWBC3.

Dataset	Features	FP	FN	SVM FP	SVM FN
Ovarian(original)	97802	4.6	4.8	5	3
Ovarian(modified)	97802	4.4	3.4	0	0
AML/ALL train	7129	0.6	2.8	0	0
AML treatment	7129	4.8	3.5	3	2
Colon	2000	3.8	3.7	3	3

Table 5: **Results for the perceptron on all data sets.** The results are averaged over 5 shufflings of the data as this algorithm is sensitive to the order in which it receives the data points. The first column is the dataset used and the second is number of features in the dataset. For the ovarian and colon datasets, the number of normal tissues misclassified (FP) and the number of tumor tissues misclassified (FN) is reported. For the AML/ALL training dataset, the number of AML samples misclassified (FP) and the number of ALL patients misclassified (FN) is reported. For the AML treatment dataset, the number of unsuccessfully treated patients misclassified (FP) and the number of successfully treated patients misclassified (FN) is reported. The last two columns report the best score obtained by the SVM on that dataset.