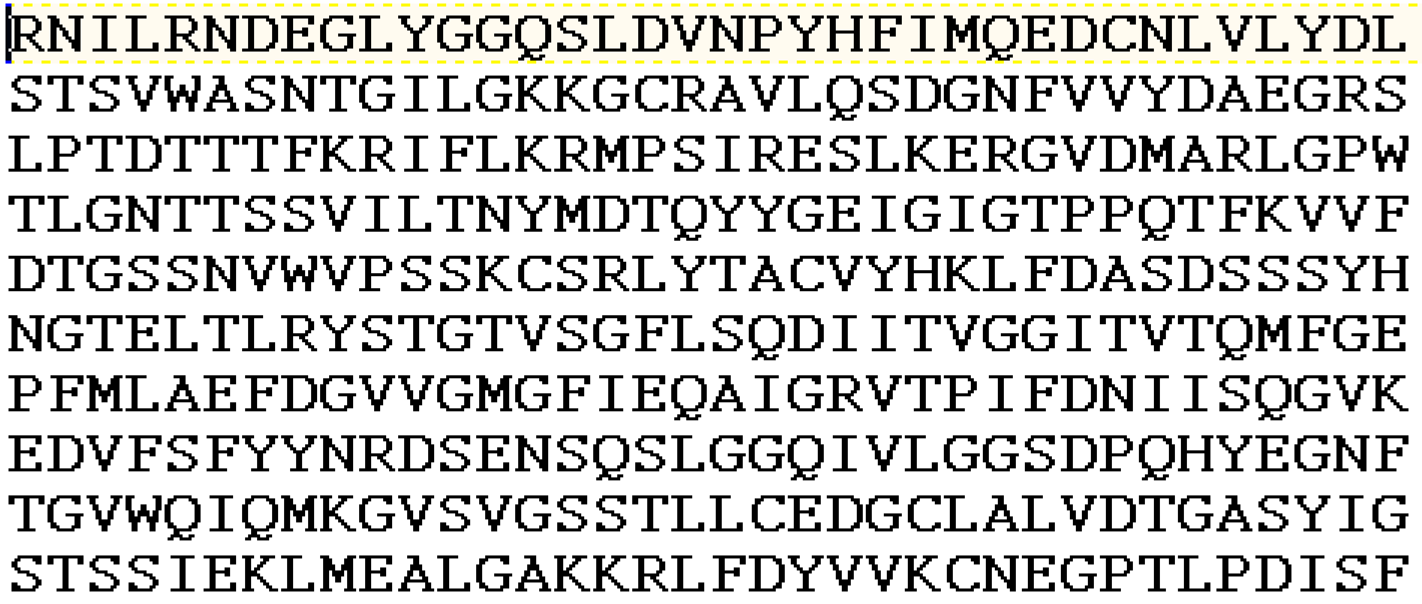
Application of Support Vector Machine In Bioinformatics

V. K. Jayaraman
Scientific and Engineering Computing Group
CDAC, Pune
jayaramanv@cdac.in

Arun Gupta
Computational Biology Group
AbhyudayaTech, Indore
arun@abhyudayatech.com

Introduction continue….

Given a set of sequences

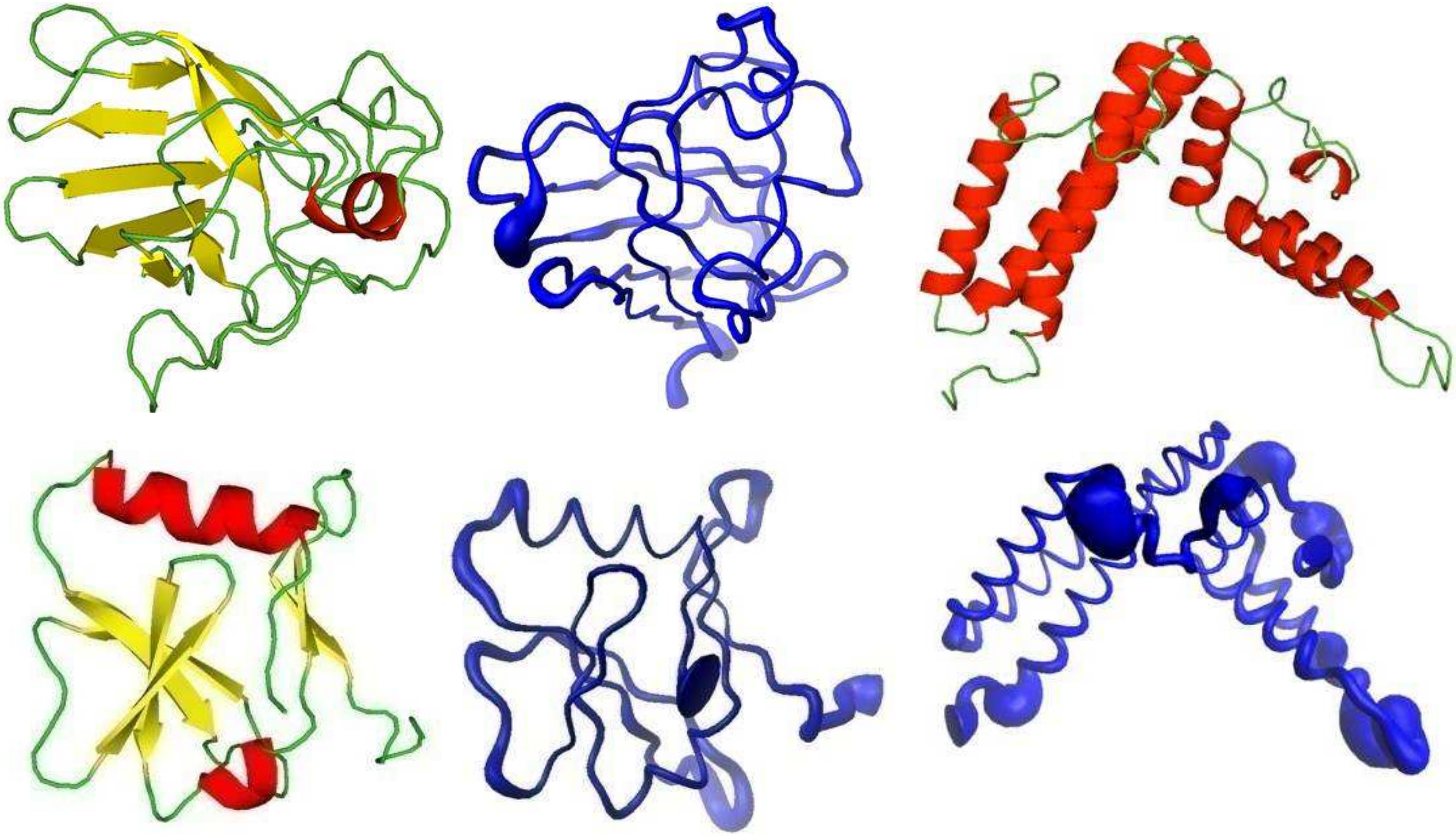


```
RNILRNDEGLYGGQSLDVNPYHFIMQEDCNLVLYDL
STSVWASNTGILGKKGCRAVLQSDGNFVVYDAEGRS
LPTDTTTFKRIFLKRMPSIRESLKERGVDMARLGPW
TLGNTTSSVILTNYMDTQYYGEIGIGTPPQTFKVVF
DTGSSNVWVPSSKCSRLYTACVYHKLFDASDSSSYH
NGTELTLRYSTGTVSGFLSQDIITVGGITVTQMFGE
PFMLAEFDGVVGMGFIEQAIGRVTPIFDNIISQGVK
EDVFSFYYNRDSENSQSLGGQIVLGGSDPQHYEGNF
TGVWQIQMKGVSVGSSTLLCEDGCLALVDTGASYIG
STSSIEKLMEALGAKKRLFDYVVKCNEGPTLPDISF
```

Can we Identify Allergen causing Sequences?

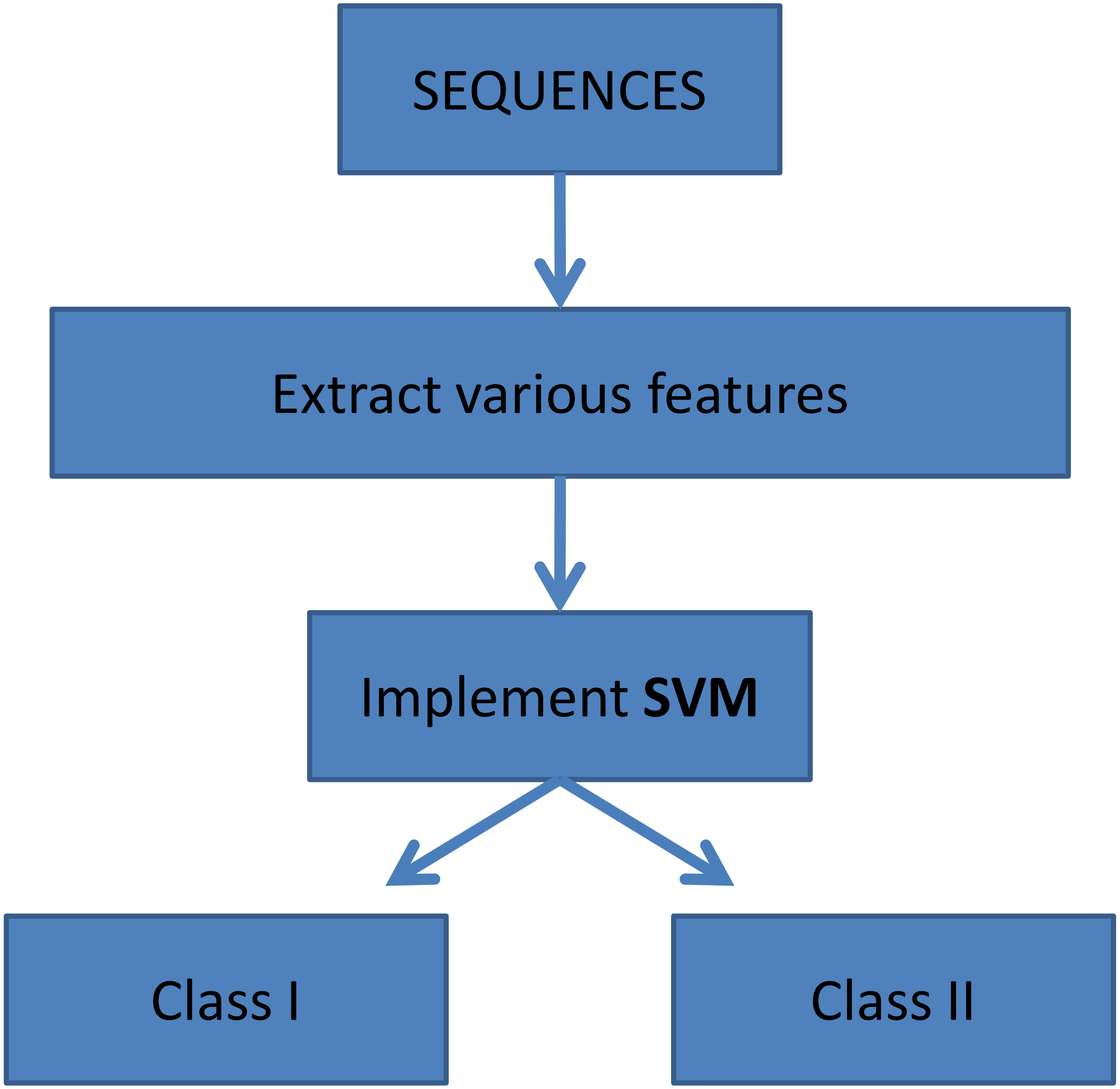
Introduction continue….

Given a set of proteins



Can we do the structure based classification?

Introduction continue….



SEQUENCES

Extract various features

Implement **SVM**

Class I

Class II

SUPPORT VECTOR CLASSIFICATION ALGORITHM

SVM classification algorithm

- Introduced by Vapnik and co-workers in 1992
- Rigorously based on
  - Statistical Learning Theory
  - Perceptron

- SVM Classifier can Recognize Patterns efficiently
- Solve Real Life Problems in :
  - ❖ Chemo/Bio Informatics
  - ❖ Many many other fields

IDENTIFICATION OF ORDERED DISORDERED PROTEINS

- What is the importance of identifying disordered regions in proteins..?
- Dunker *et al.* have predicted the disordered regions in proteins only on the basis of hydrophobicity and charge.
- We shall now study how this can be done be supervised and unsupervised classification.

Supervised or Unsupervised

Unsupervised

Supervised

Supervised or Unsupervised

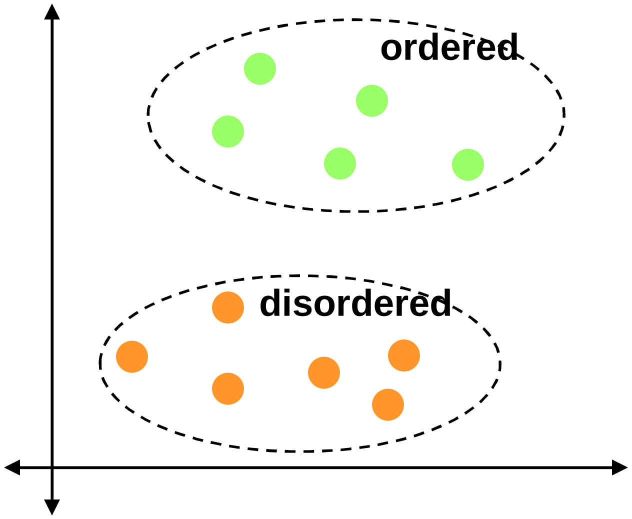
- ***Unsupervised*** : Provided a set of features algorithm groups data into clusters (does not require class information).
- ***Supervised*** : Employs class label information of some instances to build a model. Model is validated employing unseen data

Unsupervised Classification

- Algorithm will cluster the data points into groups

- Input

Mean Hydrophobicity	Net Charge
-1.68	-7
-1.315	1
-1.464	0
-1.961	2
-2.003	3
-1.594	-7

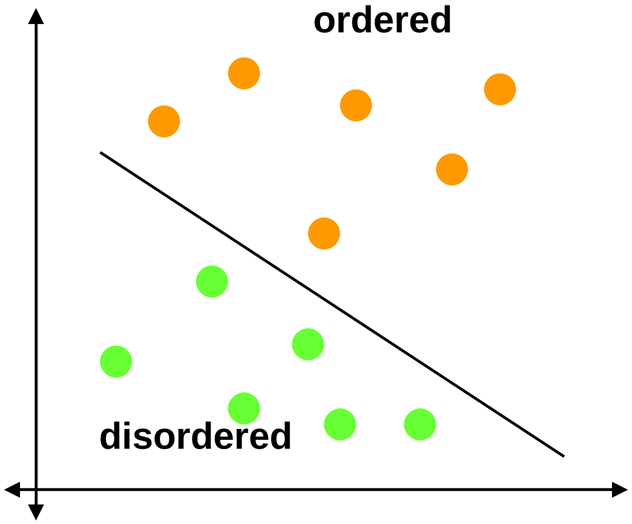


- Class information need not be provided

Supervised Classification

- Input

Mean Hydrophobicity	Net Charge	Class
-1.68	-7	1
-1.315	1	1
-1.464	0	1
-1.961	2	-1
-2.003	3	-1
-1.594	-7	-1



- Algorithm will be trained on a set of given data points.
- This algorithm will classify unseen data points into classes
- Class information is need for training

SVM Classification

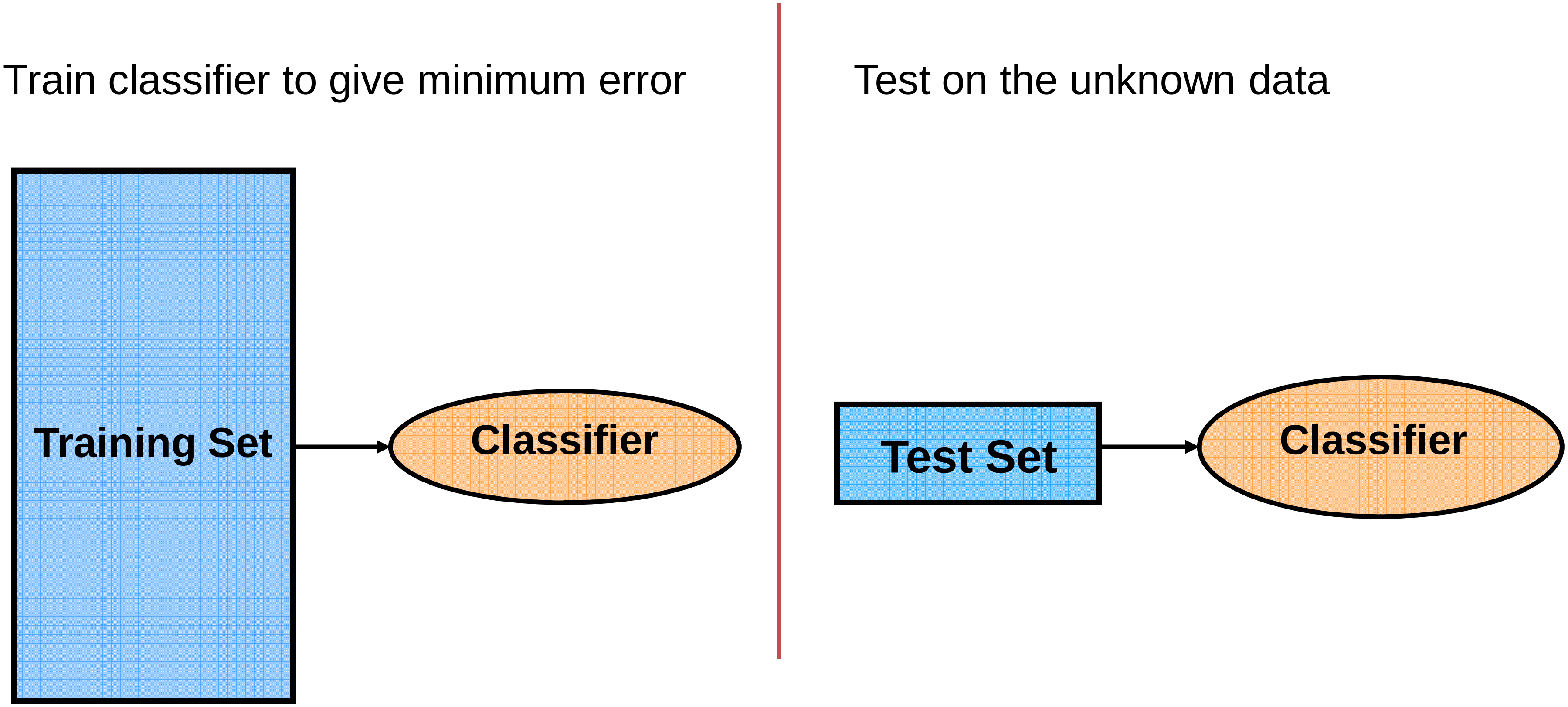
- SVM can be employed for both supervised and unsupervised classification
- Supervised classifications is more popular
- Entire lecture Is based on “supervised methodology”

Algorithm contd..

General Approach

Train classifier to give minimum error

Test on the unknown data



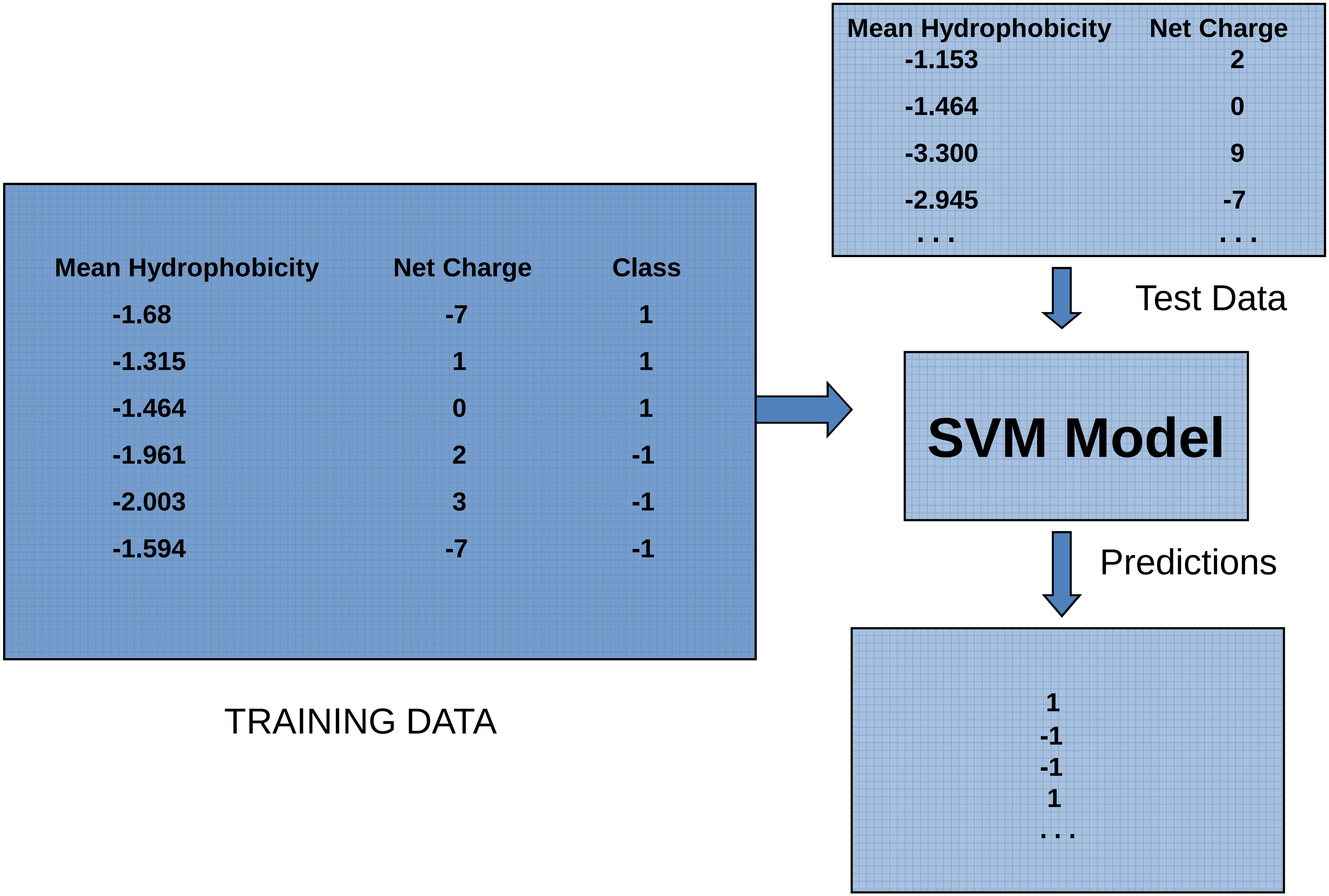
Training Set

Classifier

Test Set

Classifier

Algorithm contd…



Mean Hydrophobicity	Net Charge	Class
-1.68	-7	1
-1.315	1	1
-1.464	0	1
-1.961	2	-1
-2.003	3	-1
-1.594	-7	-1

TRAINING DATA

Mean Hydrophobicity	Net Charge
-1.153	2
-1.464	0
-3.300	9
-2.945	-7
. . .	. . .

Test Data

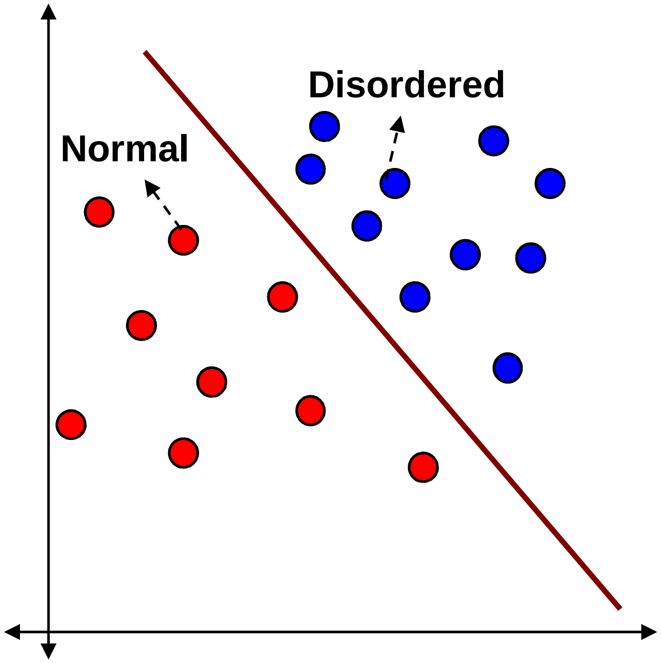
SVM Model

Predictions

1
-1
-1
1
. . .

Algorithm contd..

SVM employs a linear hyperplane



Mean Hydrophobicity	Net Charge	…	Class
-1.68	-7		1
-1.315	1		1
-1.464	0		1
-1.961	2		-1
-2.003	3		-1
-1.594	-7		-1

$$[\,w^T x\,] + b=0$$

W - weight vector

b - bias

W - vector : having dimensions equivalent to number of features

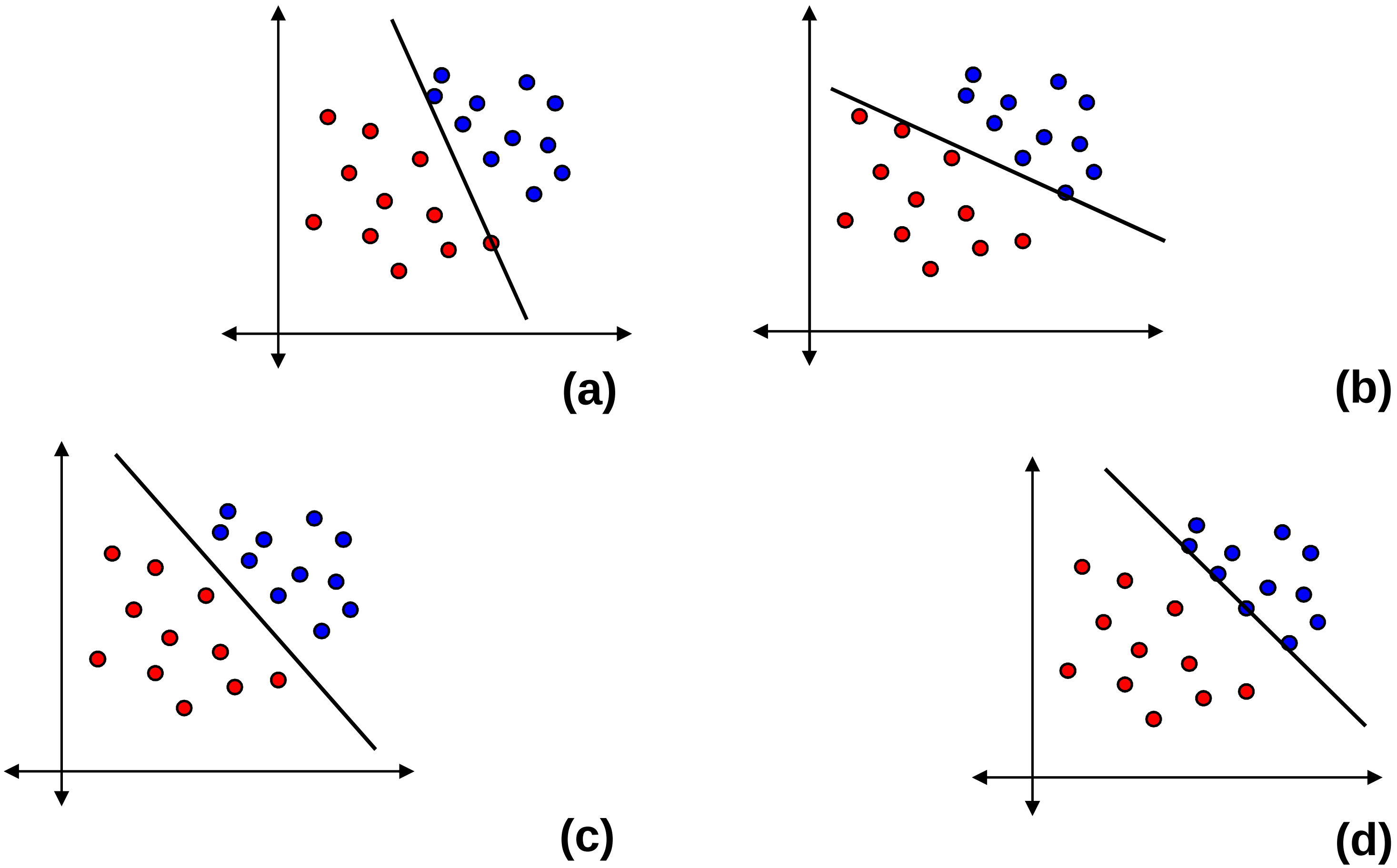
X : x_1, x_2, x_3, ……. Input Vector

y : +1 class 1

y : -1 class 2

Algorithm contd….

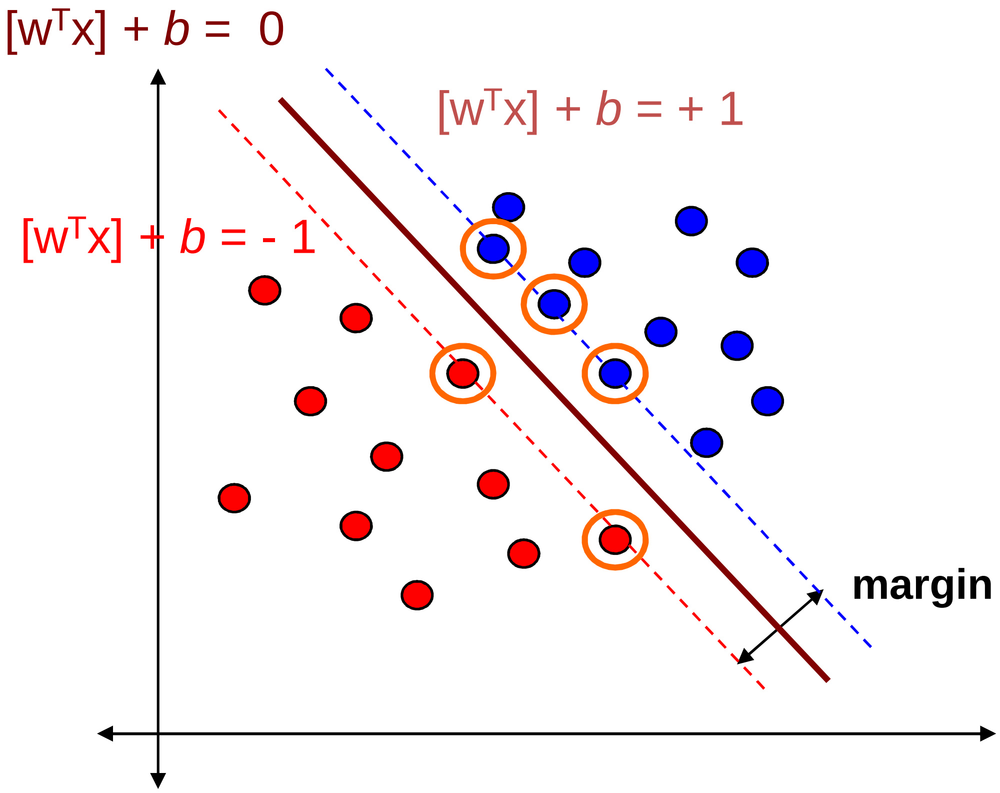
Which Hyperplane is Better ?



Algorithm contd..

Maximum Margin Classifier

- Constrain the data belonging to two different classes to be at least distance ‘1’ from the separating hyperplane
- Minimize the risk of overfitting by choosing the maximal margin hyperplane in feature space



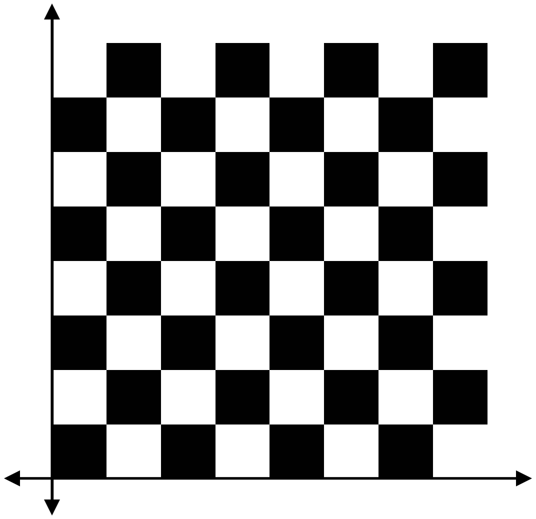
$$\text{Margin} = \frac{1}{2}\|w\|^{2}$$

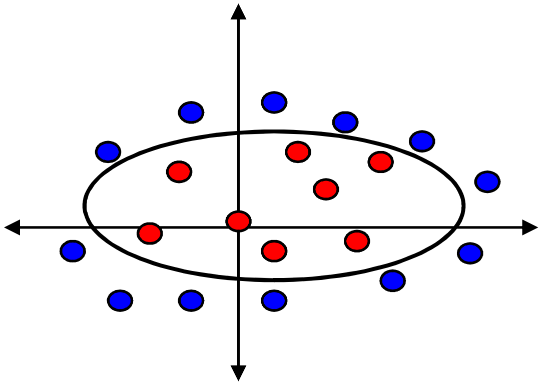
Maximize Margin

Algorithm contd…

Limitations of Linear Classifier

- Linear classifier is not always the winner.



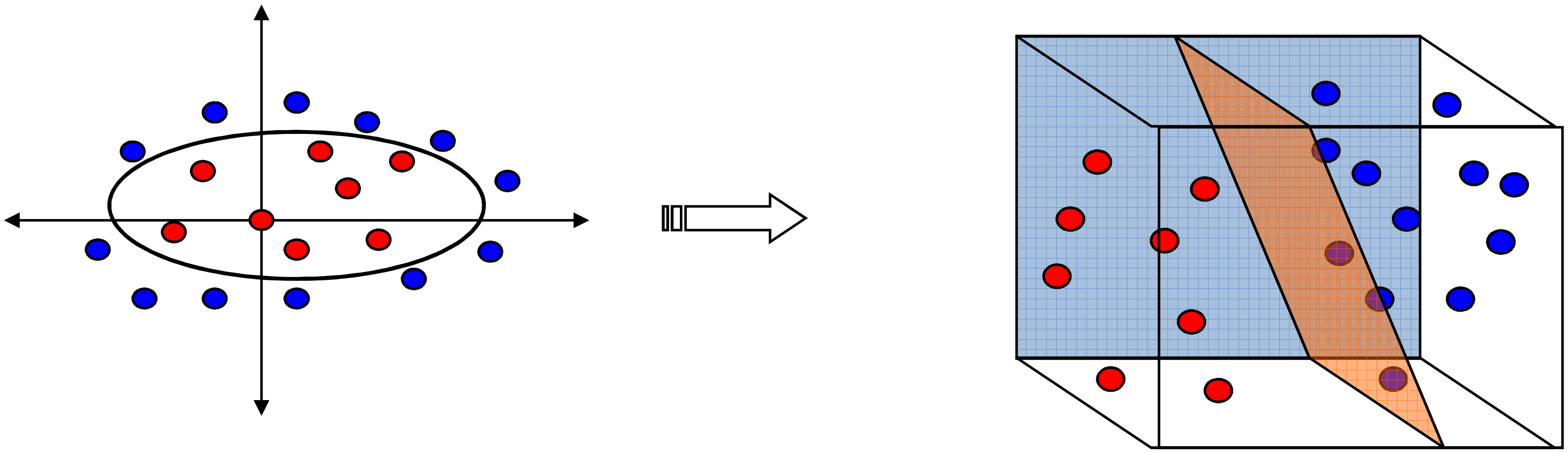


Algorithm contd..

Learning in the Feature Space

- Map the data into a feature space where they are linearly separable

Input Space $x \xrightarrow{\phi} \phi(x)$ **Feature Space**



Algorithm contd..

Learning in the Feature Space

Mean Hydrophobicity	Net Charge	Class
-1.68	-7	1
-1.464	0	1
-1.961	2	-1
-2.003	3	-1

Can we have a function that transforms lower dimensional non-Separable data into higher dimensional separable data

$$x \xrightarrow{\phi} \phi(x)$$

-1.68	-7	1
-1.464	0	1
-1.961	2	-1
-2.003	3	-1

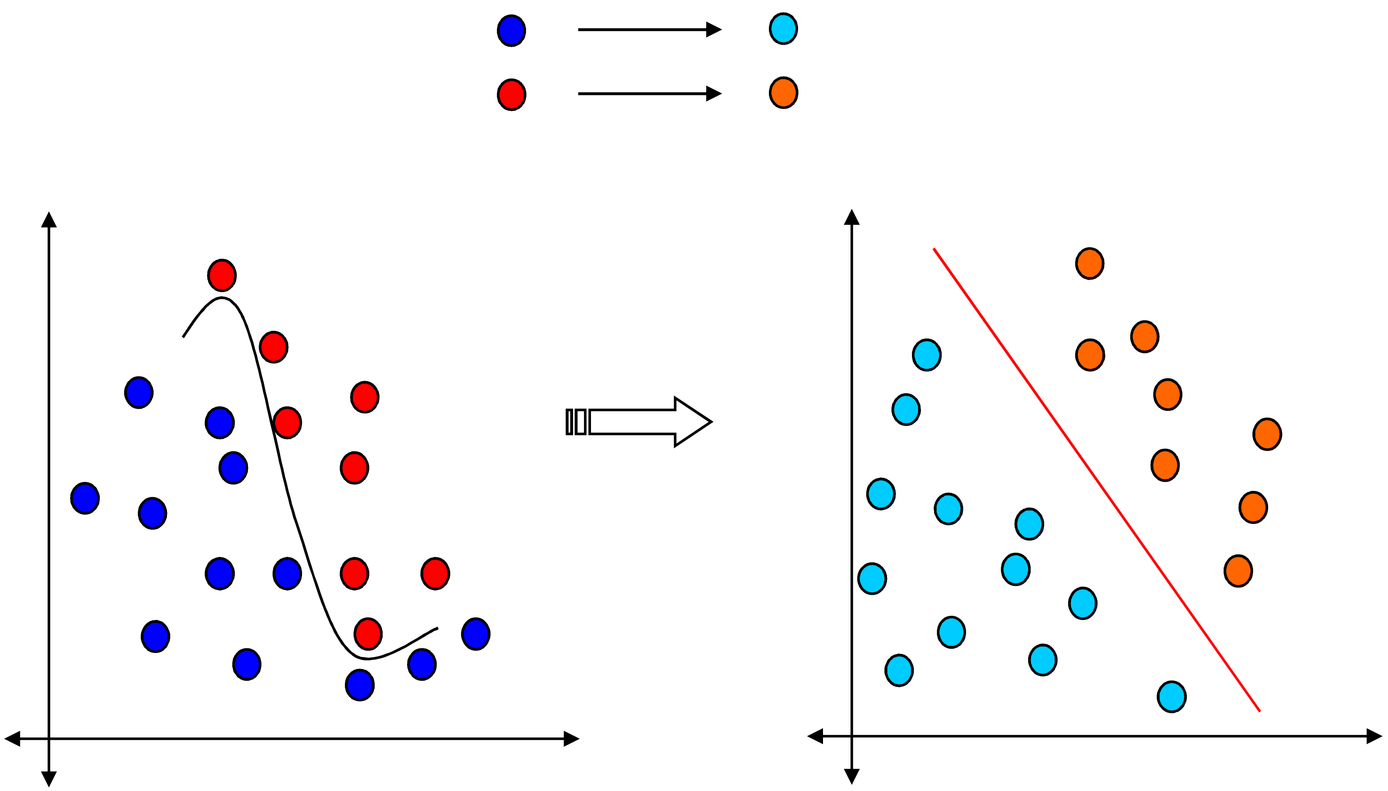
$$\Longrightarrow$$

6.01	3.9	2.5	1
0.17	2.0	8.0	1
3.10	5.6	3.6	-1
5.0	4.3	4.7	-1

Algorithm contd..

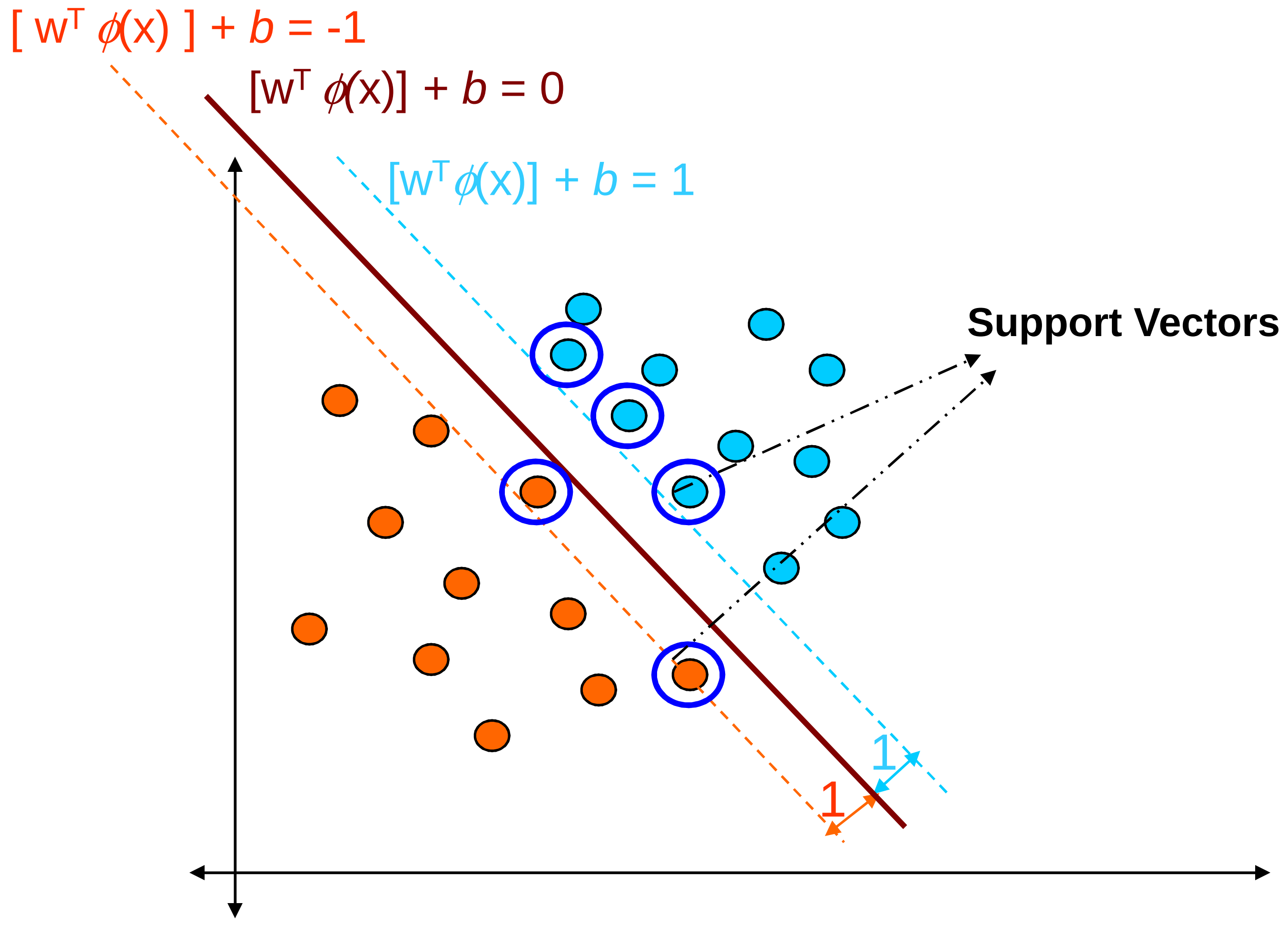
Learning in the Feature Space

$$x \xrightarrow{\phi} \phi(x)$$



Algorithm contd..

SVM Linear Classifier in High Dimensional Space


[$w^T \phi(x)$] + b = -1
[$w^T \phi(x)$] + b = 0
[$w^T \phi(x)$] + b = 1
Support Vectors
1
1

H Dim. feature space

- Real life problems have several input features
  - Gene Expression Profiles → Thousands of Genes
  - Drug Discovery → More than 100 thousand Descriptors
- Eg. 256 dimensional data, polynomial of degree 5 gives the feature space   dimension ≈ 10^{10}

Introduction to Kernel functions

- ➢ Working in high dimensional feature spaces solves the problem of expressing complex functions

  BUT….

  - ❖ Computationally intractable
  - ❖ Data Dimensionality increases exponentially in the feature space

SOLUTION…….

- ➢ Introduce kernel functions for simplification

The “Kernel Trick”

- The linear classifier relies on dot product between vectors

$$K(x_i,x_j)=x_i^T x_j$$

- If every data point is mapped into high-dimensional space via some transformation Φ: x → φ(x), the dot product becomes:

$$K(x_i,x_j)= \varphi(x_i)^T \varphi(x_j)$$

- A *kernel function* is some function that corresponds to an *inner product* in some expanded feature space.

Commonly-used kernel functions

- Linear kernel: $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$
- Polynomial kernel: $K(\mathbf{x}_i, \mathbf{x}_j) = (1 + \mathbf{x}_i^T \mathbf{x}_j)^p$
- Gaussian (Radial-Basis Function (RBF)) kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{2\sigma^2})$$

- Sigmoid:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\beta_0 \mathbf{x}_i^T \mathbf{x}_j + \beta_1)$$

Choosing the Kernel Function

- Probably the most tricky part of using SVM.
- The kernel function is important because it creates the kernel matrix, which summarizes all the data
- In practice, a ***low degree polynomial kernel*** or ***RBF kernel*** with a reasonable width is a good initial try
- Note that SVM with RBF kernel is closely related to RBF neural networks, with the centers of the radial basis functions automatically chosen for SVM

Applications of SVM in BI

Identification of protein functions

Gene functions

Micro array Classification

Identification of protein function

- Secondary structure prediction
- Identification of binding sites
- Sub nuclear localization of proteins
- Sub cellular localization
- Protein-protein interaction prediction
- Prediction of protein disorder

Identification of gene functions

- Promoter prediction
- Prediction of tissue specific localization of genes
- Prediction of DNA methylation sites
- DNA hot spots prediction

Microarray Classification

- Lukemia prediction
- Colon cancer prediction
- Prediction of several genetic disorders
- No of examples less & No Of Features very Large.
- Employ Feature Selection

Domain features extraction

Protein function Identification :

- Numerical representation of the sequence.
- Amino acid frequencies
- Dipeptide frequencies
- Tripeptide frequencies
- K-mer frequencies
- Homology information in terms of Blast and Psi-blast profiles.
- Remote Homology using PSSM
- Motive information
- Secondary Structure information
- Physical propertied like hydrophobic, hydrophilic, charge, etc.
- Structural features like surface accesibility, co-ordinates of atoms, contact order.

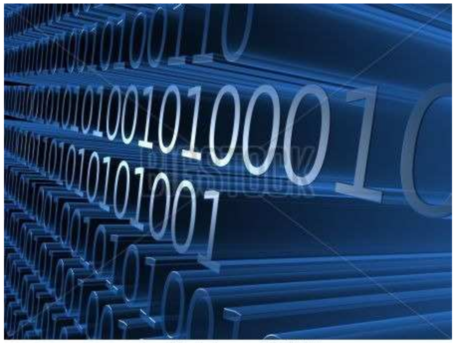
Features extraction

Protein function Identification :

Numerical representation of the sequence.

Alanine

Gylsine



www.bigstock.com · 2547744

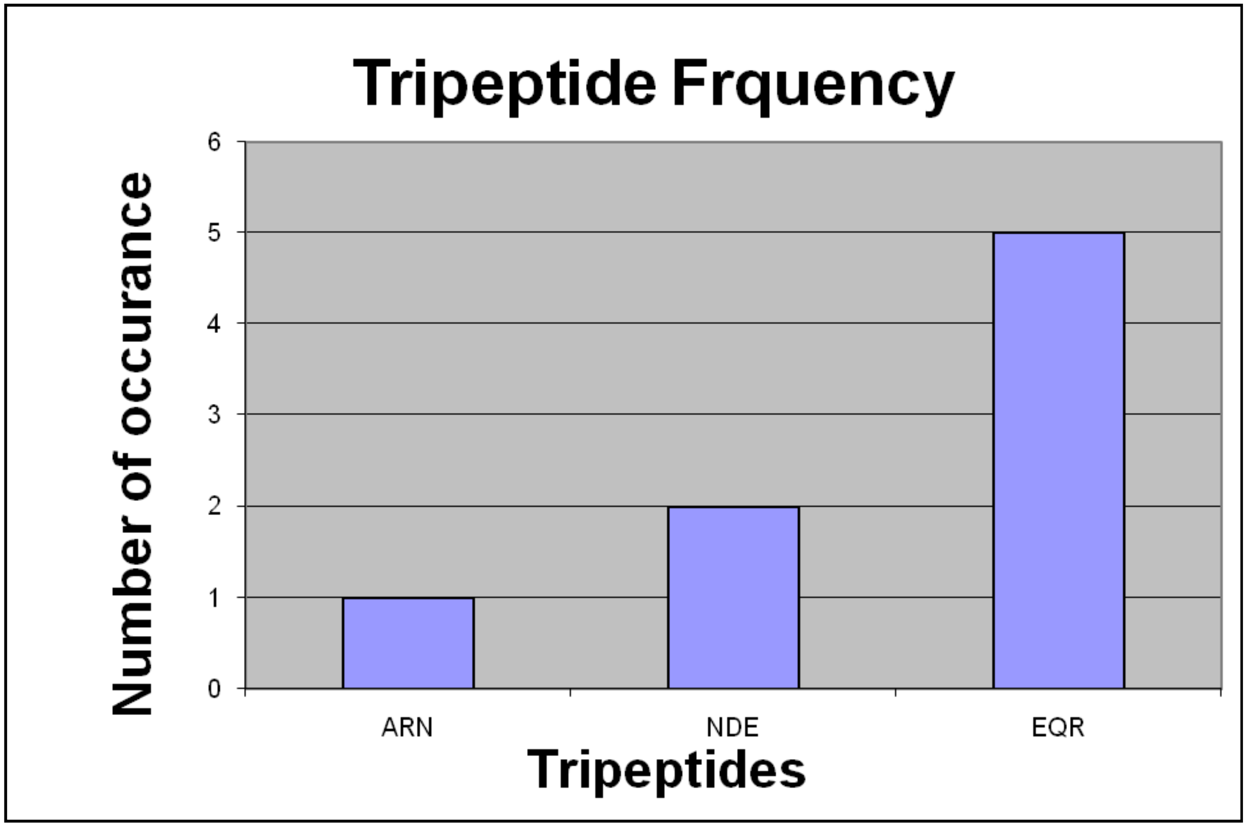
1	0	0	0	0	0	.	.	.	.
0	0	1	0	0	0	.	.	.	.

can be done for 20 amino acids

Features extraction

Protein function Identification :

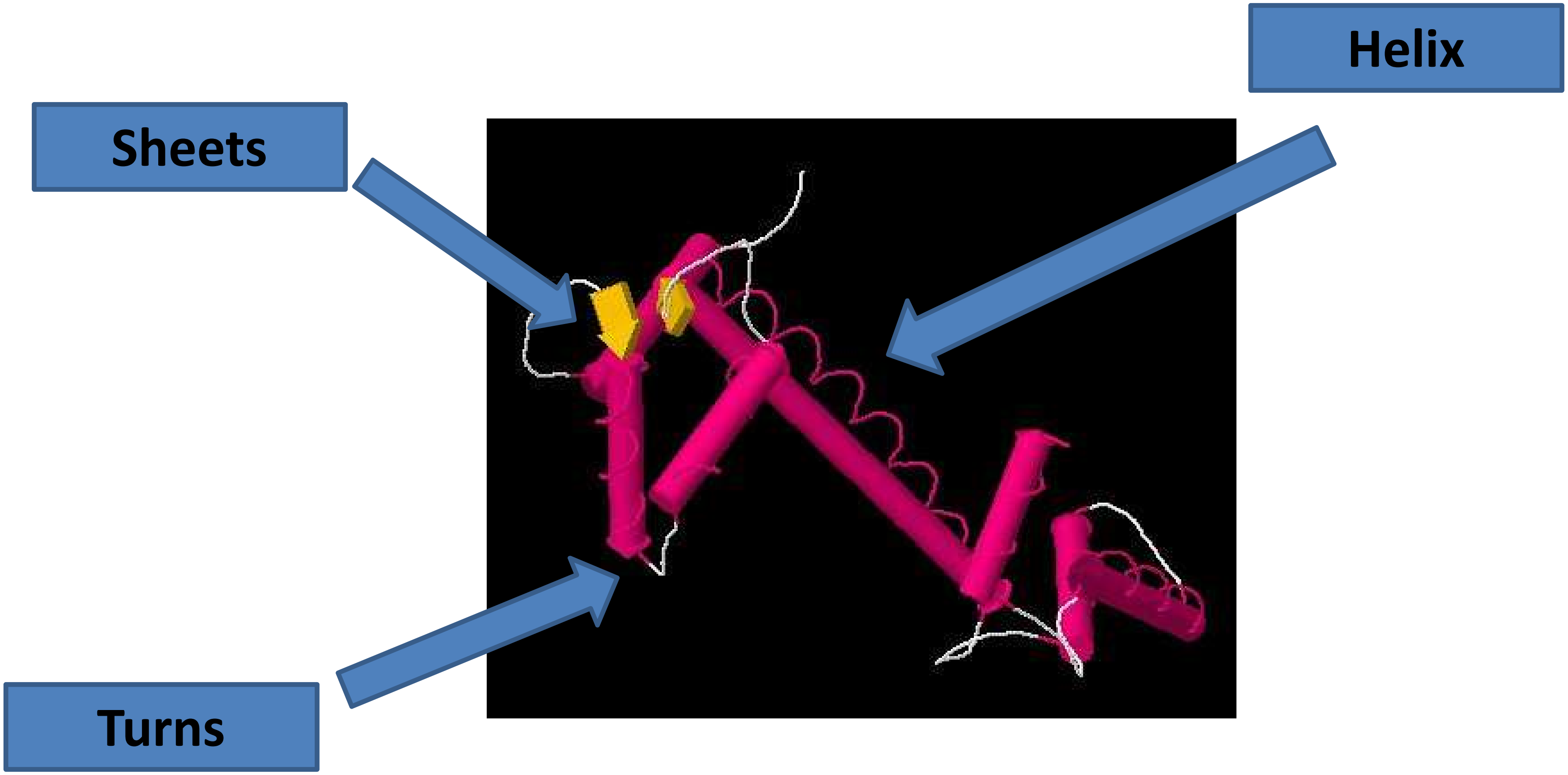
Amino acid frequencies



Features extraction

Protein function Identification :

Secondary structure information:



Helix

Sheets

Turns

Feature Calculation

>seq1
AASQRLAQS
>seq2
ASSNRADSQ
>seq3
ADSQRADSQ
….

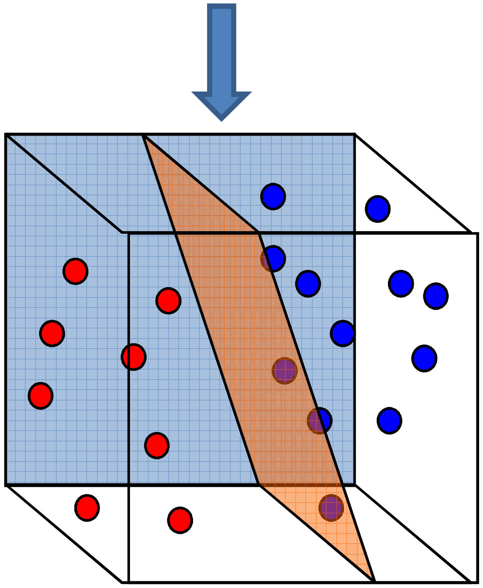
Training sequences

A_freq	C_freq	…	AA_freq	AC_freq	…	PSSM1	PSSM2	…
0.33	0.00	…	0.125	0.000	…	0.754	0.000	…
0.22	0.00	…	0.00	0.00	…	0.754	0.012	…
…	…	…	…	…	…	…	…	…

Input features (*X*)

Class
1
-1
…

Class labels (*Y*)



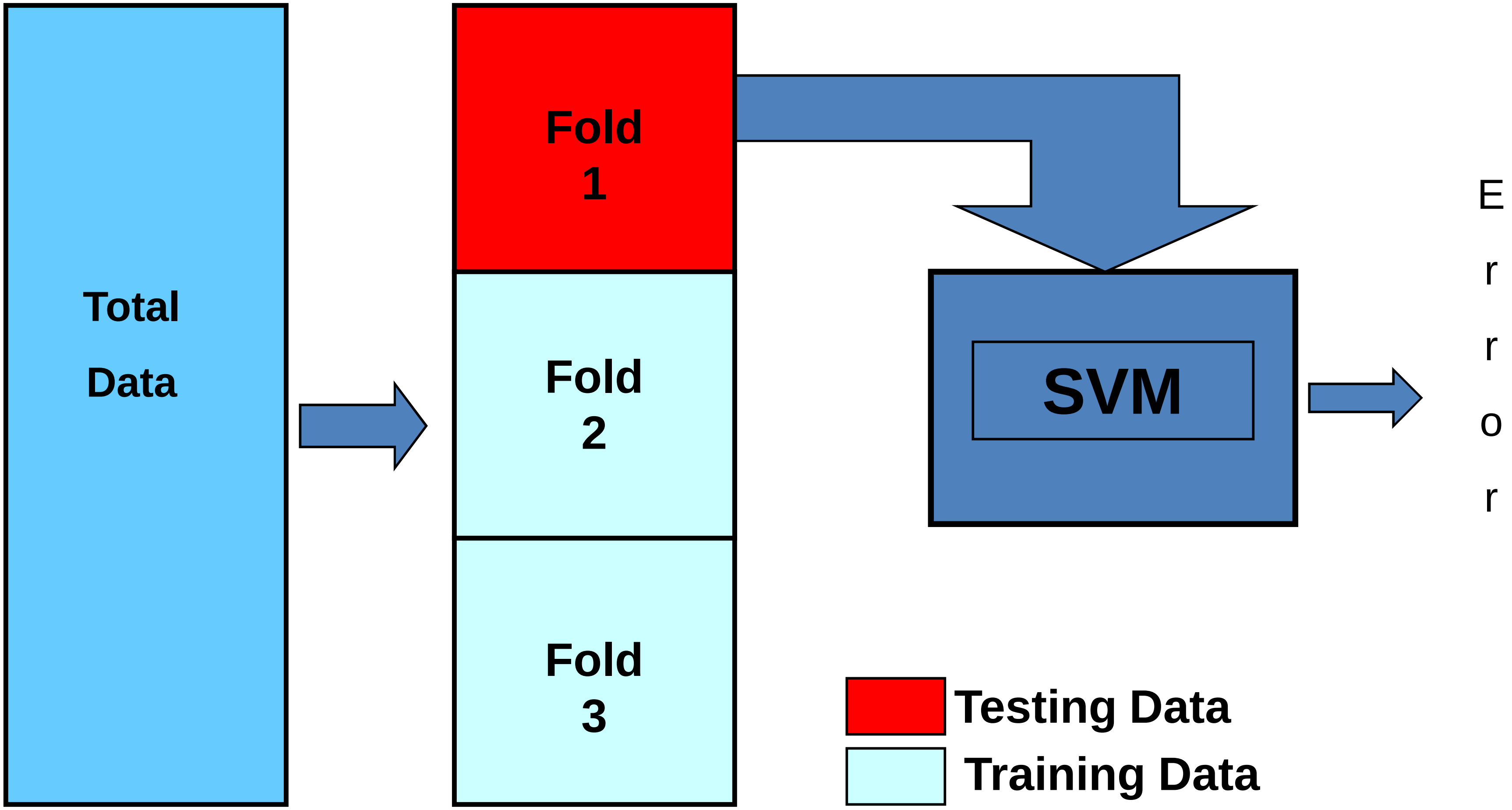
Commonly used Methods for Parameter Tuning

- k-fold cross validation
- Leave-one-out error estimation

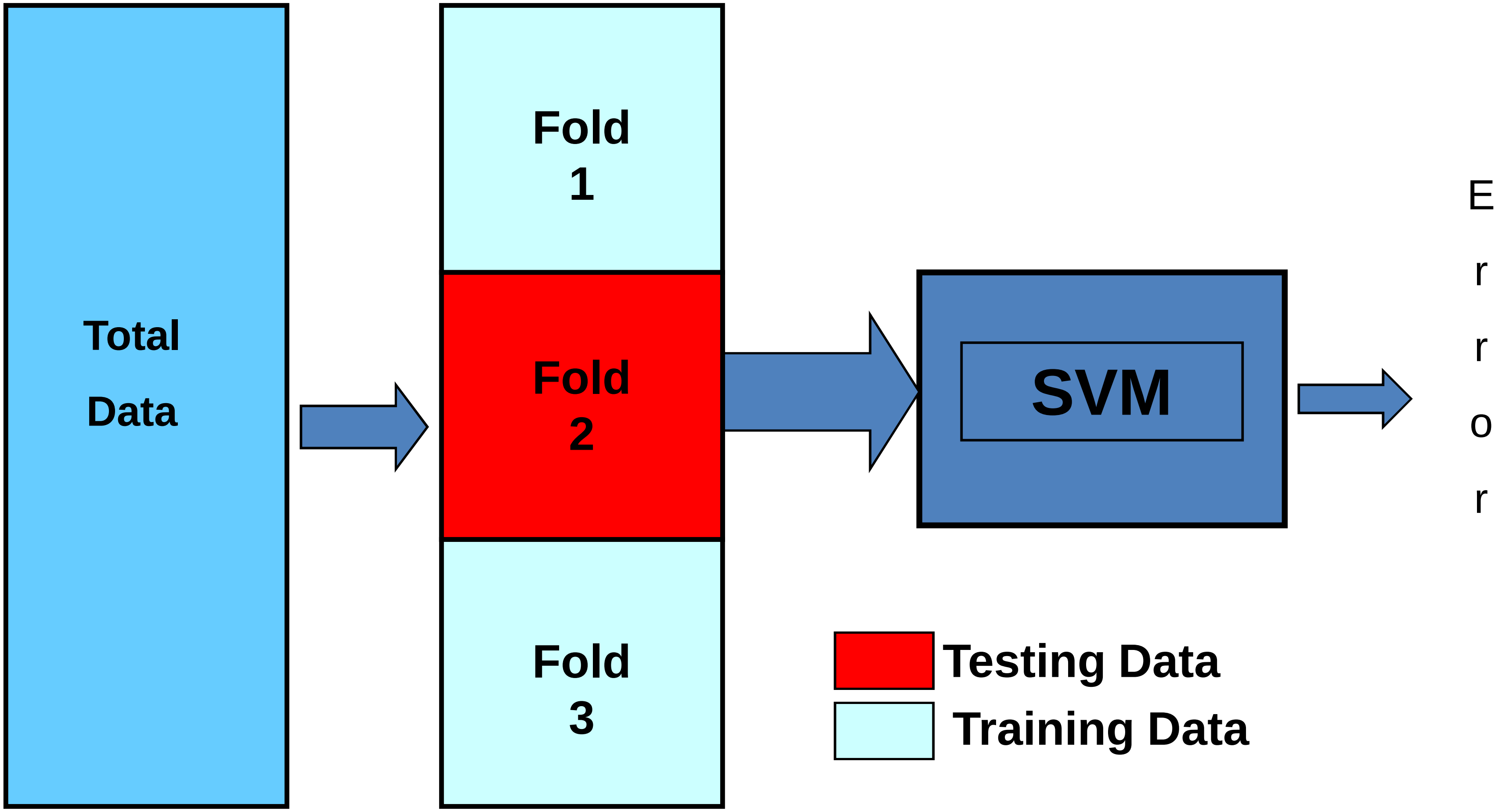
k-fold Cross Validation

- The training data is randomly split into k mutually exclusive subsets (or the folds) of approximately equal size
- For k times
  - SVM decision rule is obtained using k -1 of the subsets
  - Then tested on the subset left out

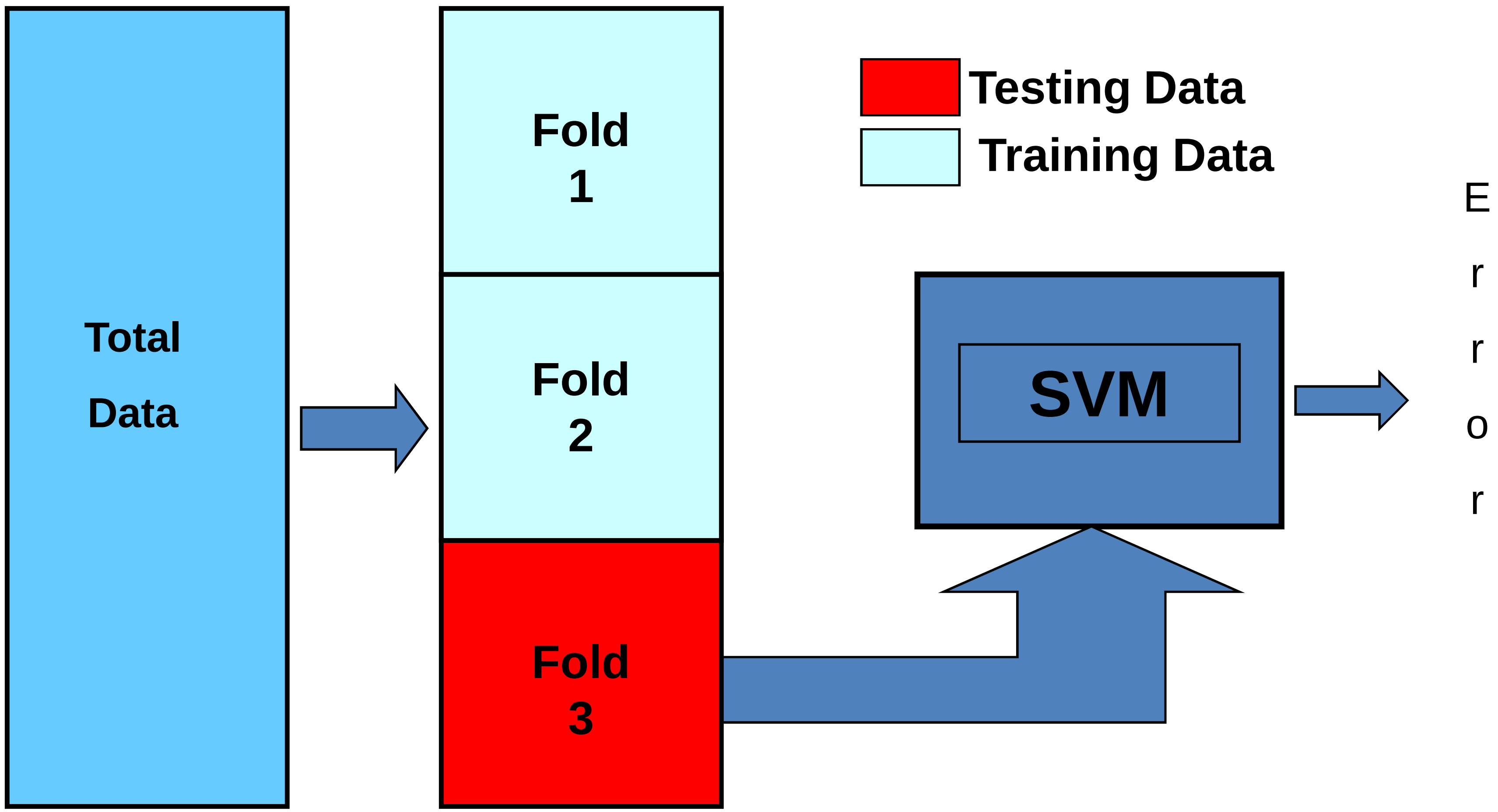
3-fold Cross Validation



3-fold Cross Validation



3-fold Cross Validation



Total Data

Fold 1

Fold 2

Fold 3

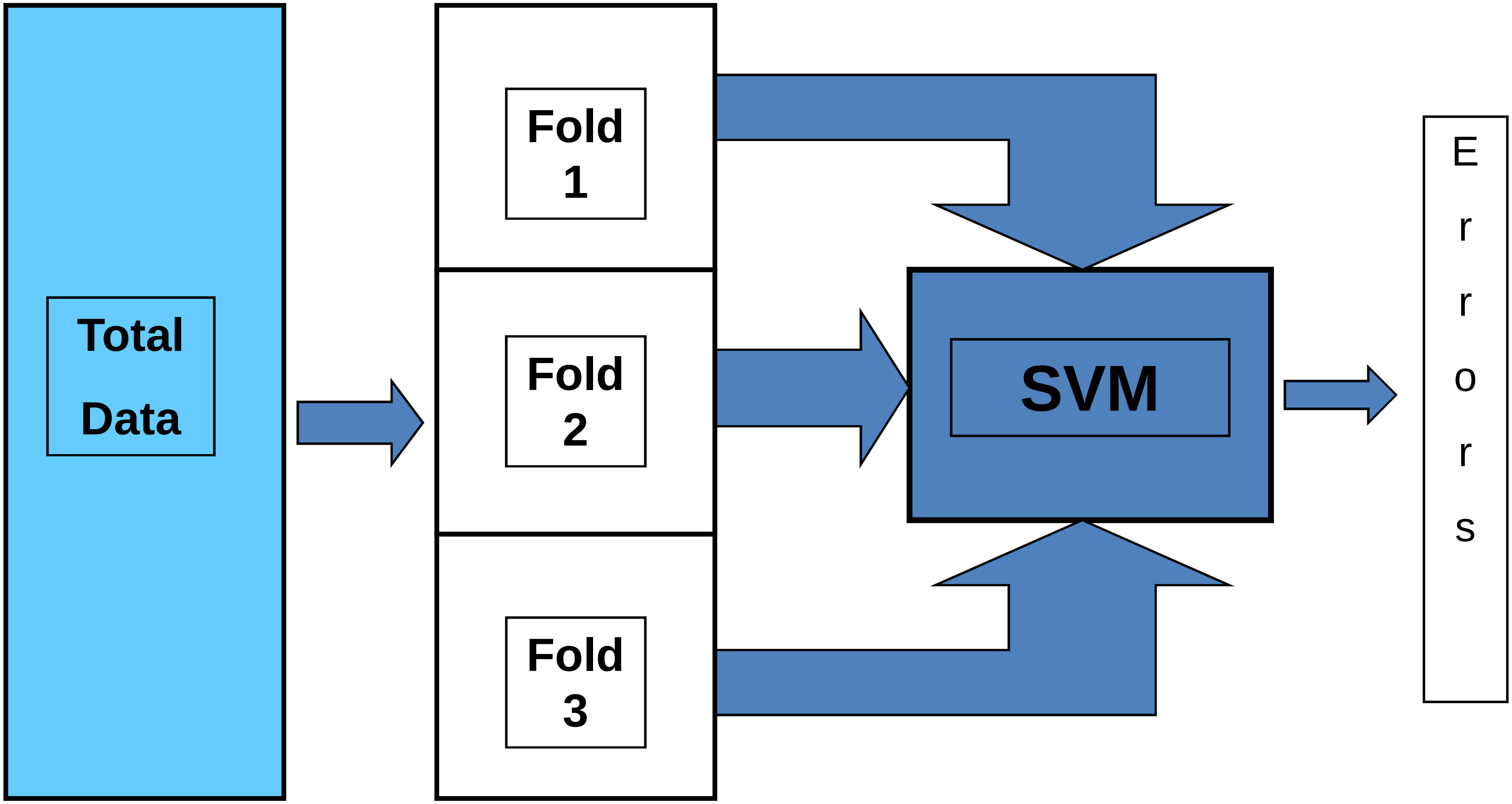
🟥 Testing Data

⬜ Training Data

SVM

Error

3-fold Cross Validation



Validation error = Average of three errors

Leave-one-out error (LOO) estimate

- Extreme form of k -fold cross-validation

  k is equal to the number of examples, ***l***
- For ***l*** times

  SVM decision rule is obtained using ***l*** -1 of the examples

  Then tested on the subset left out example
- Good thing about LOO

  Almost unbiased estimate of the expected generalization error
- Limitation

  Computationally expensive since computations require running the training algorithm ***l*** times.

Model Evaluation Measures

Sensitivity = $\dfrac{TP}{TP+FN}$

Specificity = $\dfrac{TN}{TN+FP}$

Precision = $\dfrac{TP}{TP+FP}$

TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

Matthew's Correlation Coefficient

$$MCC(X)=\frac{TPTN-FPFN}{\sqrt{(TN+FN)(TP+FN)(TN+FP)(TP+FP)}}$$

SVM based prediction servers

Cyclin Pred
http://bioinfo.icgeb.res.in/cyclinpred/ **CyclinPred** is a **SVM based prediction method** to identify novel cyclins using various features of proteins.
SS PRED
http://www.bioinformatics.org/sspred/html/sspred.html
Identification & Classification of proteins involved in bacterial secretion systems

Bayes Server
http://immunopred.org/bayesb/server/index.html
BayesB: Server for SVM Prediction of Linear B-cell Epitopes using Bayes Feature Extraction

NRpred
http://www.imtech.res.in/raghava/lgepred/
This server allows user to analyse the expression data (Microarray Data).

UbiPred
http://iclab.life.nctu.edu.tw/ubipred/
UbiPred is a SVM-based prediction server using that detects the presence /absence of ubiquitylation site in a protein sequence.

OUR WORK

HLAB27Pred

A machine learning HLA-B*2705 Binders Prediction Method

- HLA-B27 is found to be associated with the development of variety of autoimmune diseases including *Ankylosing spondylitis*.
- Several theories have been proposed to explain the association of HLA-B27 with *spondyloarthritis*.
- ***HLAB27Pred*** will be helpful in designing new peptide vaccines through the prediction of corresponding binding peptides.

UnPublished (under Review)

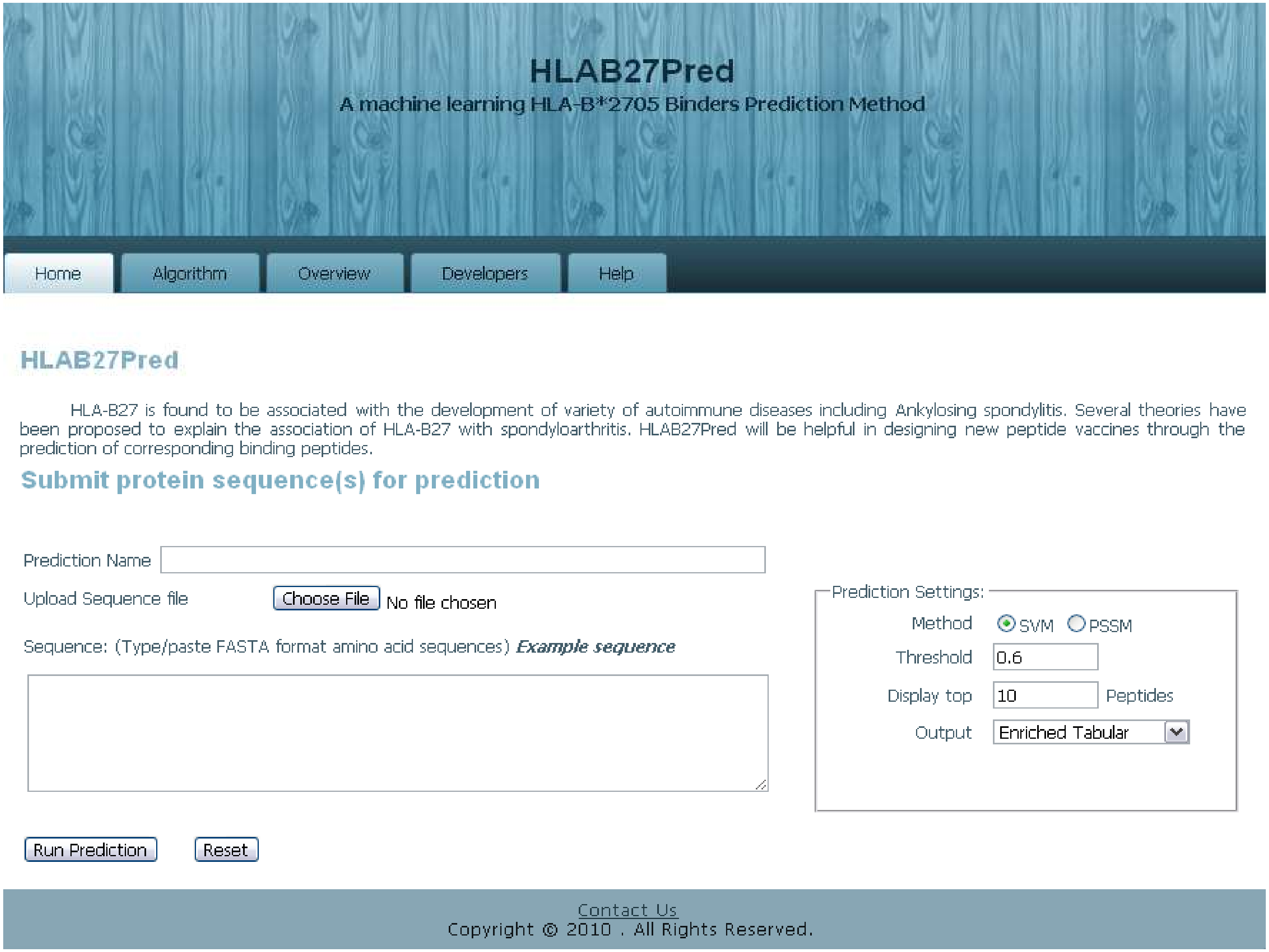
HLAb27Pred

UnPublished (under Review)

- HLAB27Pred is a server designed for the prediction of HLA-B*2705 (MHC class I allele) based nanomer epitopic binding peptides. Server implements 2 techniques for the purpose of prediction, viz. SVM and PSSM.
- SVM based prediction are deployed by training a set of experimentally validated nanomeric binding and non-binding peptides.
- The performance of the SVM predictions has been tested through **5 cross-validation**.
- The ***specificity*** and ***sensitivity*** obtained during the development of this server is ***84.54%*** and ***85.57%*** respectively.
- Whereas average ***precision*** and average ***MCC*** values were observed to be ***84.69%*** and ***0.8%*** respectively.

HLAb27Pred

UnPublished (under Review)



HLAB27Pred

HLA-B27 is found to be associated with the development of variety of autoimmune diseases including Ankylosing spondylitis. Several theories have been proposed to explain the association of HLA-B27 with spondyloarthritis. HLAB27Pred will be helpful in designing new peptide vaccines through the prediction of corresponding binding peptides.

Submit protein sequence(s) for prediction

Prediction Name

Upload Sequence file No file chosen

Sequence: (Type/paste FASTA format amino acid sequences) ***Example sequence***

Prediction Settings:

Method ● SVM ○ PSSM

Threshold 0.6

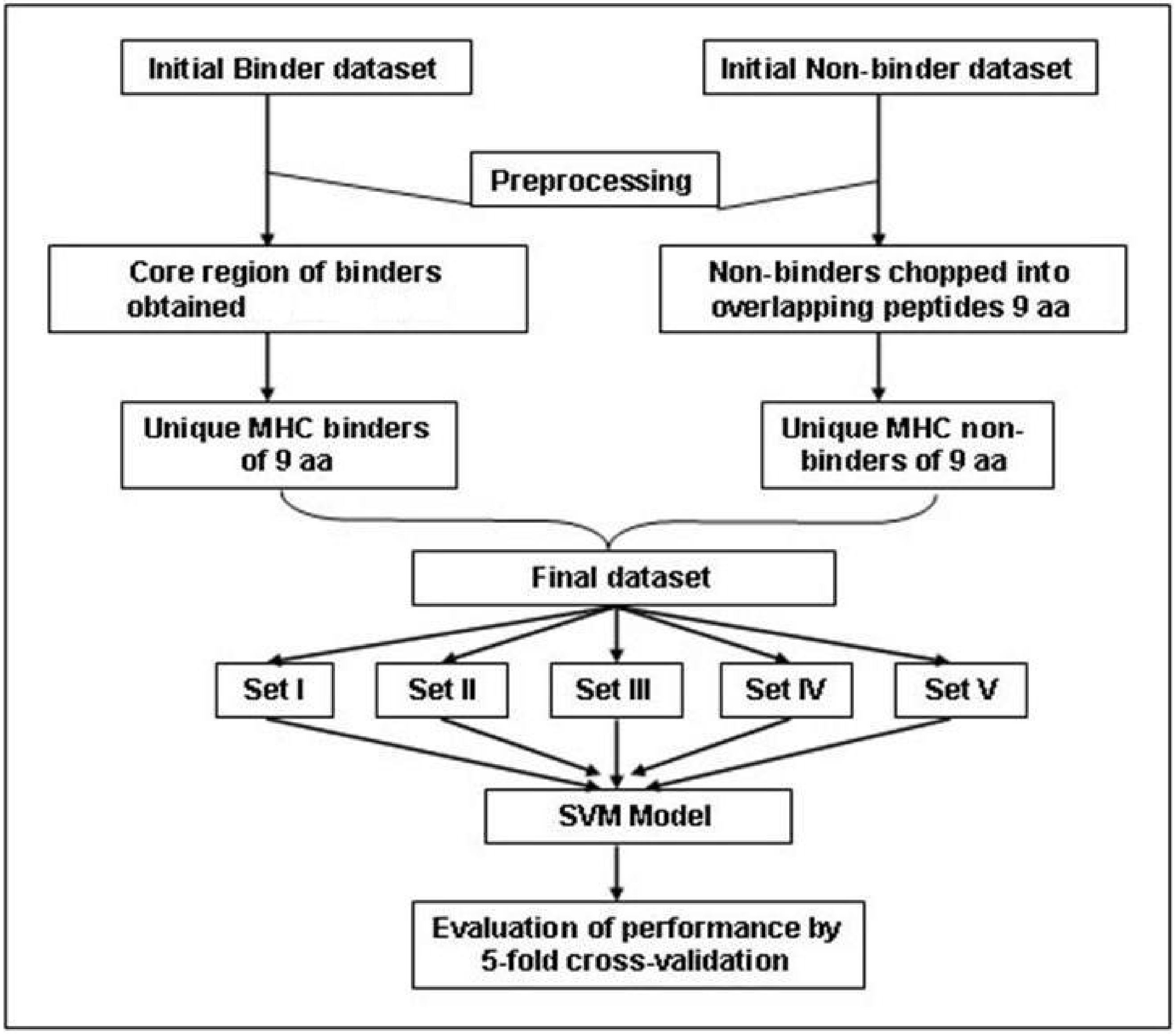
Display top 10 Peptides

Output Enriched Tabular

Copyright © 2010 . All Rights Reserved.

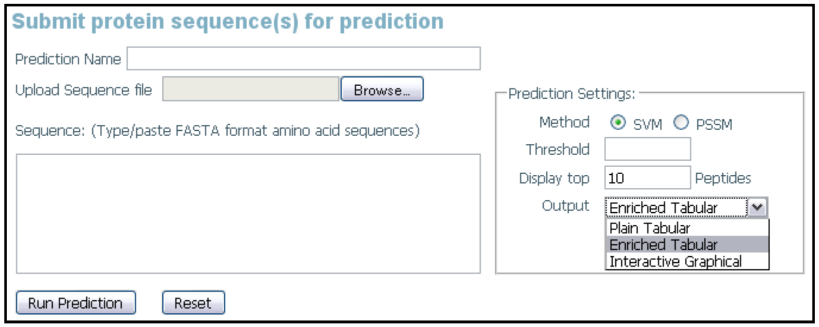
HLAb27Pred

UnPublished (under Review)


Initial Binder dataset
Initial Non-binder dataset
Preprocessing
Core region of binders obtained
Non-binders chopped into overlapping peptides 9 aa
Unique MHC binders of 9 aa
Unique MHC non-binders of 9 aa
Final dataset
Set I
Set II
Set III
Set IV
Set V
SVM Model
Evaluation of performance by 5-fold cross-validation

HLAb27Pred

UnPublished (under Review)



Submit protein sequence(s) for prediction

Prediction Name

Upload Sequence file [Browse...]

Sequence: (Type/paste FASTA format amino acid sequences)

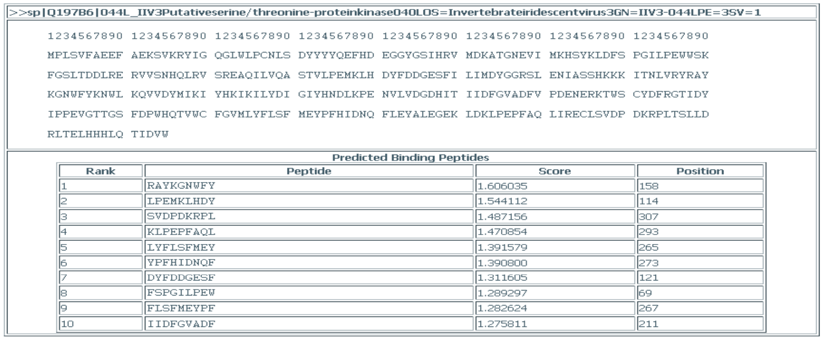
Prediction Settings:

Method ● SVM ○ PSSM

Threshold

Display top 10 Peptides

Output Enriched Tabular

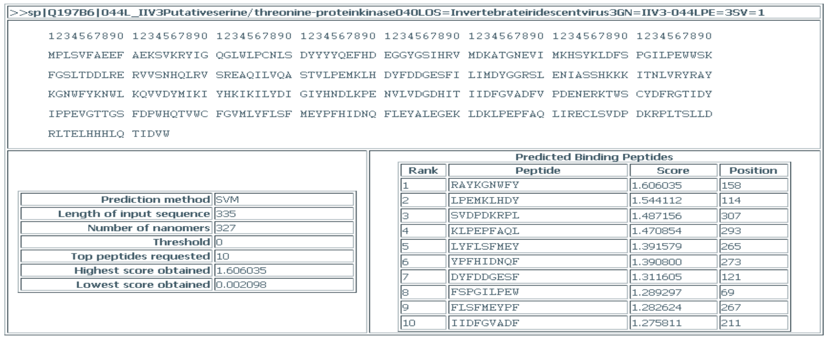


>>sp|Q197B6|044L_IIV3Putativeserine/threonine-proteinkinase040LOS=Invertebrateiridescentvirus3GN=IIV3-044LPE=3SV=1

```
1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890
MPLSVFAEEF AEKSVKRYIG QGLWLPCNLS DYYYYQEFHD EGGYGSIHRV MDKATGNEVI MKHSYKLDFS PGILPEWWSK
FGSLTDDLRE RVVSNHQLRV SREAQILVQA STVLPEMKLH DYFDDGESFI LIMDYGGRSL ENIASSHKKK ITNLVRYRAY
KGNWFYKNWL KQVVDYMIKI YHKIKILYDI GIYHNDLKPE NVLVDGDHIT IIDFGVADFV PDENERKTWS CYDFRGTIDY
IPPEVGTTGS FDPWHQTVWC FGVMLYFLSF MEYPFHIDNQ FLEYALEGEK LDKLPEPFAQ LIRECLSVDP DKRPLTSLLD
RLTELHHHLQ TIDVW
```

Predicted Binding Peptides

Rank	Peptide	Score	Position
1	RAYKGNWFY	1.606035	158
2	LPEMKLHDY	1.544112	114
3	SVDPDKRPL	1.487156	307
4	KLPEPFAQL	1.470854	293
5	LYFLSFMEY	1.391579	265
6	YPFHIDNQF	1.390800	273
7	DYFDDGESF	1.311605	121
8	FSPGILPEW	1.289297	69
9	FLSFMEYPF	1.282624	267
10	IIDFGVADF	1.275811	211



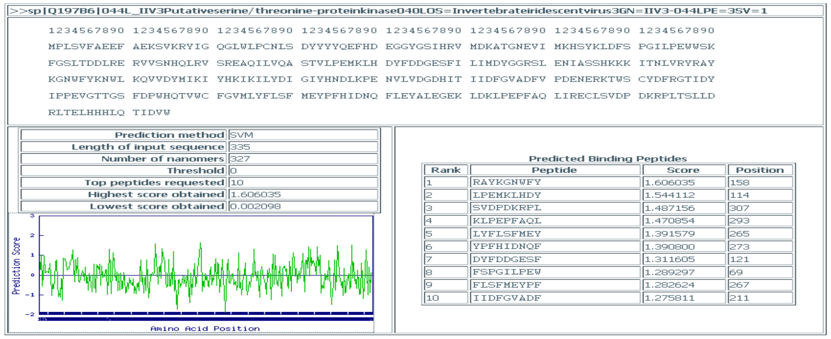
>>sp|Q197B6|044L_IIV3Putativeserine/threonine-proteinkinase040LOS=Invertebrateiridescentvirus3GN=IIV3-044LPE=3SV=1

```
1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890
MPLSVFAEEF AEKSVKRYIG QGLWLPCNLS DYYYYQEFHD EGGYGSIHRV MDKATGNEVI MKHSYKLDFS PGILPEWWSK
FGSLTDDLRE RVVSNHQLRV SREAQILVQA STVLPEMKLH DYFDDGESFI LIMDYGGRSL ENIASSHKKK ITNLVRYRAY
KGNWFYKNWL KQVVDYMIKI YHKIKILYDI GIYHNDLKPE NVLVDGDHIT IIDFGVADFV PDENERKTWS CYDFRGTIDY
IPPEVGTTGS FDPWHQTVWC FGVMLYFLSF MEYPFHIDNQ FLEYALEGEK LDKLPEPFAQ LIRECLSVDP DKRPLTSLLD
RLTELHHHLQ TIDVW
```

Prediction method	SVM
Length of input sequence	335
Number of nanomers	327
Threshold	0
Top peptides requested	10
Highest score obtained	1.606035
Lowest score obtained	0.002098

Predicted Binding Peptides

Rank	Peptide	Score	Position
1	RAYKGNWFY	1.606035	158
2	LPEMKLHDY	1.544112	114
3	SVDPDKRPL	1.487156	307
4	KLPEPFAQL	1.470854	293
5	LYFLSFMEY	1.391579	265
6	YPFHIDNQF	1.390800	273
7	DYFDDGESF	1.311605	121
8	FSPGILPEW	1.289297	69
9	FLSFMEYPF	1.282624	267
10	IIDFGVADF	1.275811	211



>>sp|Q197B6|044L_IIV3Putativeserine/threonine-proteinkinase040LOS=Invertebrateiridescentvirus3GN=IIV3-044LPE=3SV=1

```
1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890
MPLSVFAEEF AEKSVKRYIG QGLWLPCNLS DYYYYQEFHD EGGYGSIHRV MDKATGNEVI MKHSYKLDFS PGILPEWWSK
FGSLTDDLRE RVVSNHQLRV SREAQILVQA STVLPEMKLH DYFDDGESFI LIMDYGGRSL ENIASSHKKK ITNLVRYRAY
KGNWFYKNWL KQVVDYMIKI YHKIKILYDI GIYHNDLKPE NVLVDGDHIT IIDFGVADFV PDENERKTWS CYDFRGTIDY
IPPEVGTTGS FDPWHQTVWC FGVMLYFLSF MEYPFHIDNQ FLEYALEGEK LDKLPEPFAQ LIRECLSVDP DKRPLTSLLD
RLTELHHHLQ TIDVW
```

Prediction method	SVM
Length of input sequence	335
Number of nanomers	327
Threshold	0
Top peptides requested	10
Highest score obtained	1.606035
Lowest score obtained	0.002098

Predicted Binding Peptides

Rank	Peptide	Score	Position
1	RAYKGNWFY	1.606035	158
2	LPEMKLHDY	1.544112	114
3	SVDPDKRPL	1.487156	307
4	KLPEPFAQL	1.470854	293
5	LYFLSFMEY	1.391579	265
6	YPFHIDNQF	1.390800	273
7	DYFDDGESF	1.311605	121
8	FSPGILPEW	1.289297	69
9	FLSFMEYPF	1.282624	267
10	IIDFGVADF	1.275811	211

Software to try your hands on

- SVMLight
- LibSVM
- WEKA and Bio-Weka
- MATLAB

Questions and/or Comments…?

Thank You…