Computer-aided Vaccine and Drug Discovery

G.P.S. Raghava, Scientist and Head Bioinformatics Centre
Institute of Microbial Technology, Chandigarh

Vaccine Informatics

- **Understanding immune system**
- **Breaking complex problem**
- **Adaptive immunity**
- **Innate Immunity**
- **Vaccine delivery system**
- **ADMET of peptides**

Drug Informatics

- **Annotation of genomes**
- **Searching drug targets**
- **Properties of drug molecules**
- **Protein-chemical interaction**
- **Prediction of drug-like molecules**

Web Site: http://www.imtech.res.in/raghava/

Adaptive Immunity

Innate Immunity

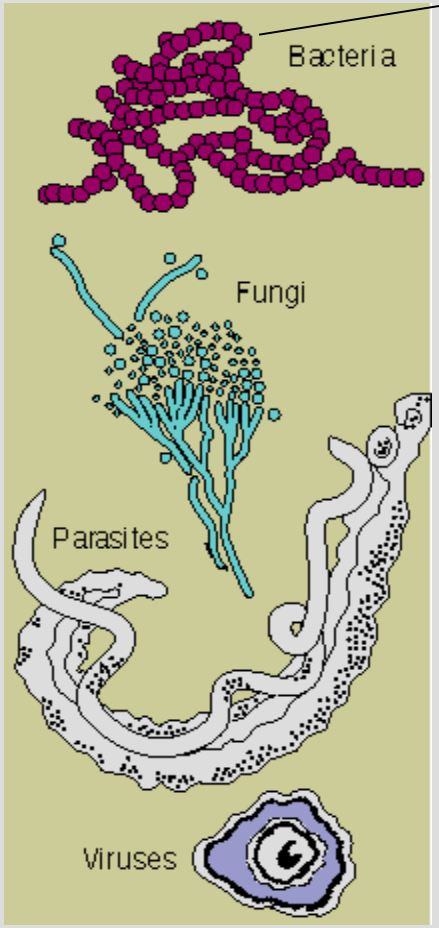
Vaccine Informatics

Bioinformatics Centre
IMTECH, Chandigarh

Protective Antigens

Vaccine Delivery

Disease Causing Agents



Pathogens/Invaders

Host Immune Defense System
For Microbial (Pathogens) Attack

Non-Specific Defense
(Innate Immunity)

Specific Defense
Adaptive Immunity
Responsible for Generating
Memory B & T Cells

TOLL LIKE RECEPTORS
PATHOGEN ASSOCIATED
MOLECULAR PATTERNS

Humoral Immunity
(B-Lymphocytes)

Cell-Mediated Immunity
(T-Lymphocytes)

B-Cell Epitopes
Antibody Prod.
Neutralization

Endogenous Antigen Processing

Exogenous Antigen Processing

CTL Epitopes
MHC I binders
TAP Binders
P. Cleavage Sites

T-Helper Epitopes
MHC II binders

Adaptive Immunity

Innate Immunity

Vaccine Informatics

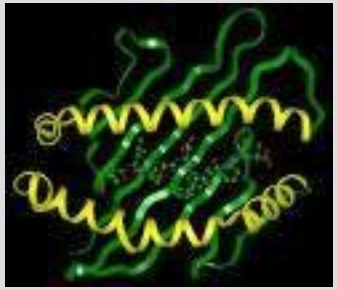
Bioinformatics Centre
IMTECH, Chandigarh

Protective Antigens

Vaccine Delivery

MHCBN: A database of MHC/TAP binders and T-cell epitopes

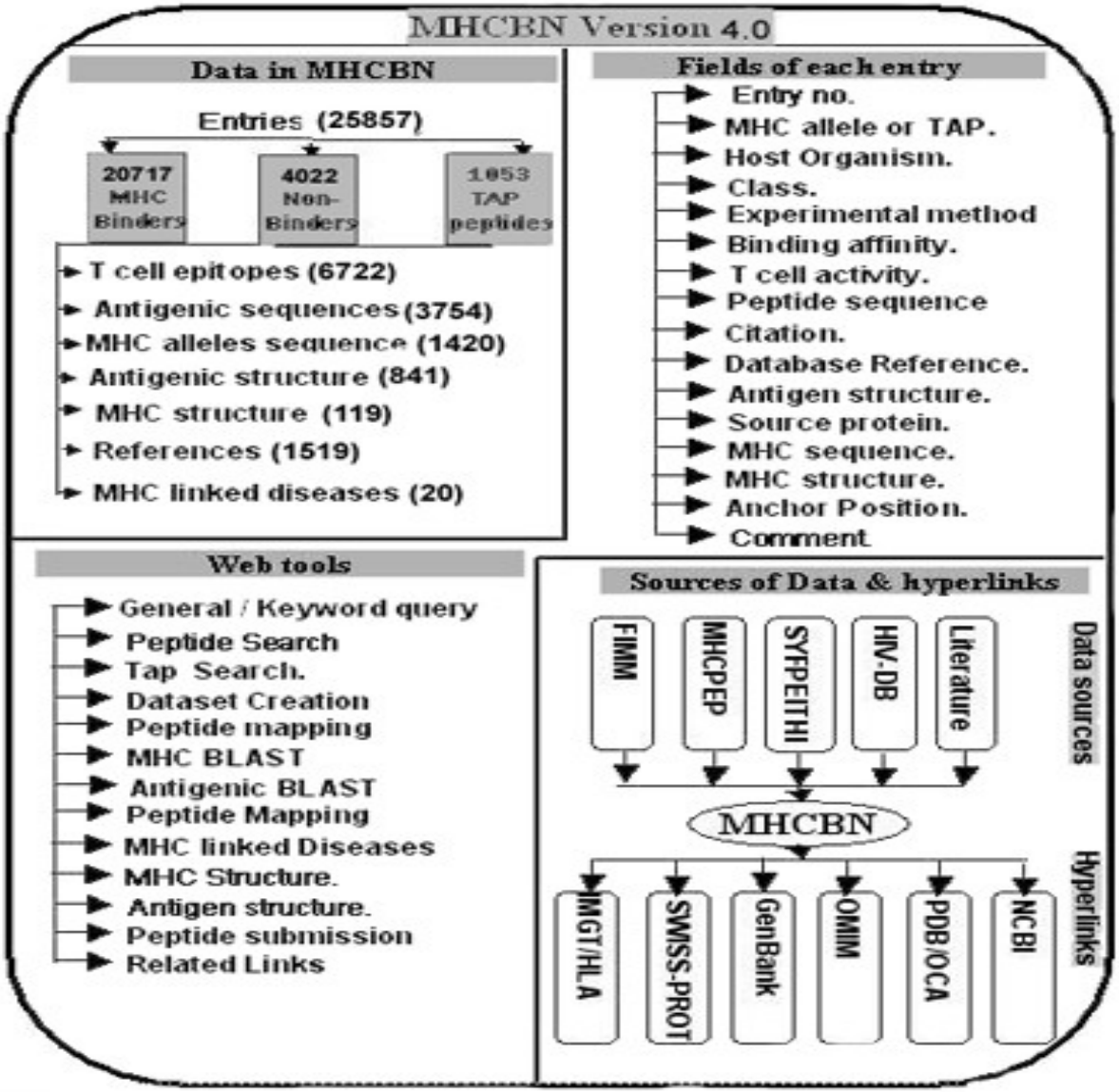
Distributed by EBI, UK



Reference database in T-cell epitopes
Highly Cited (~ 70 citations)

Bhasin et al. (2003) Bioinformatics 19: 665

Bhasin et al. (2004) NAR (Online)



Prediction of MHC class II binders (exogenous pathway)

Available algorithms

Motifs

- Rammensee *et al.*, 1995
- Sette *et al.*, 1994
- Hammer *et al.*, 1994
- Max *et al.*, 1994
- Chicz *et al.*, 1994

SYFPEITHI
MOTIFSCAN

Matrices

- Marshal *et al.*, 1994
- Sturniolo *et al.*, 1999
- Brusic *et al.*, 1998
- Southwood *et al.*, 1998
- Borras-cuesta *et al.*,2000

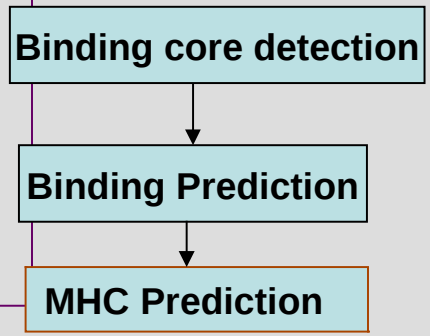
TEPITOPE
PROPRED

ANN

- Brusic *et al.*, 1998

Difficulties with MHC prediction

- Quality and quantity of data.
- Variable length of reported binders. So, a method is required additionally for detecting binding core.
- Poorly defined anchor residues.



Our approach:

Development of method for HLA-DRB1*0401

- Large dataset
- Computational approaches (ANN, SVM)
- Evaluation: 5-fold cross-validation
- Performance measures:
  - sensitivity, specificity
  - NPV, PPV, accuracy.
- Testing on independent dataset
- Extension of best approach to large number of alleles

Prediction of MHC class I binders

Existing methods:

- **Motifs:** consider occurrence of few residues.

  Low accuracy (Only 35% have motifs)

- Quantitative matrices: Consider independent contribution of each residue.

  Non-adaptive, non-linear.

- Machine learning techniques: ANN

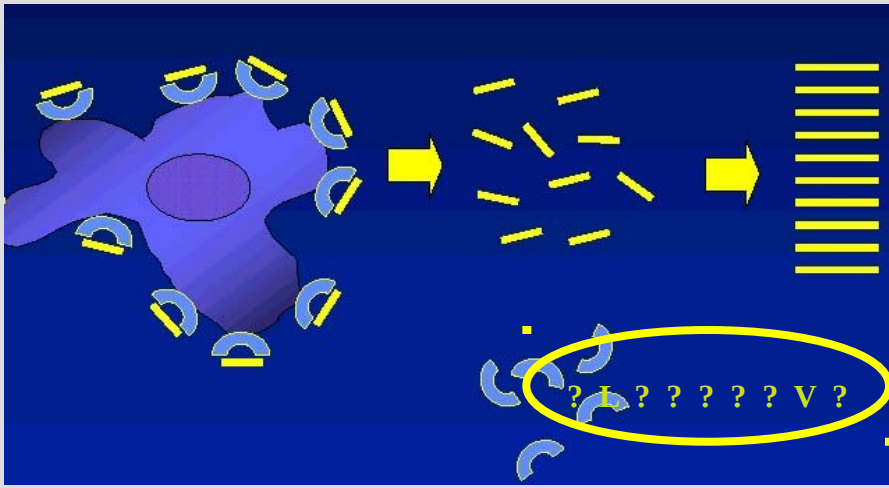
  Available for 1 or 2 alleles, require large data.

- MHC-peptide structure: available for 1 or two alleles.

  Very Slow, less amount of 3-D data.

Our Approach:

- Large & high quality dataset
- Computational techniques (ANN & QM).
- Combination of ANN & QM.
- Filtering of Potential CTL epitopes.
- Display favorable for locating promiscuous T cell epitopes.

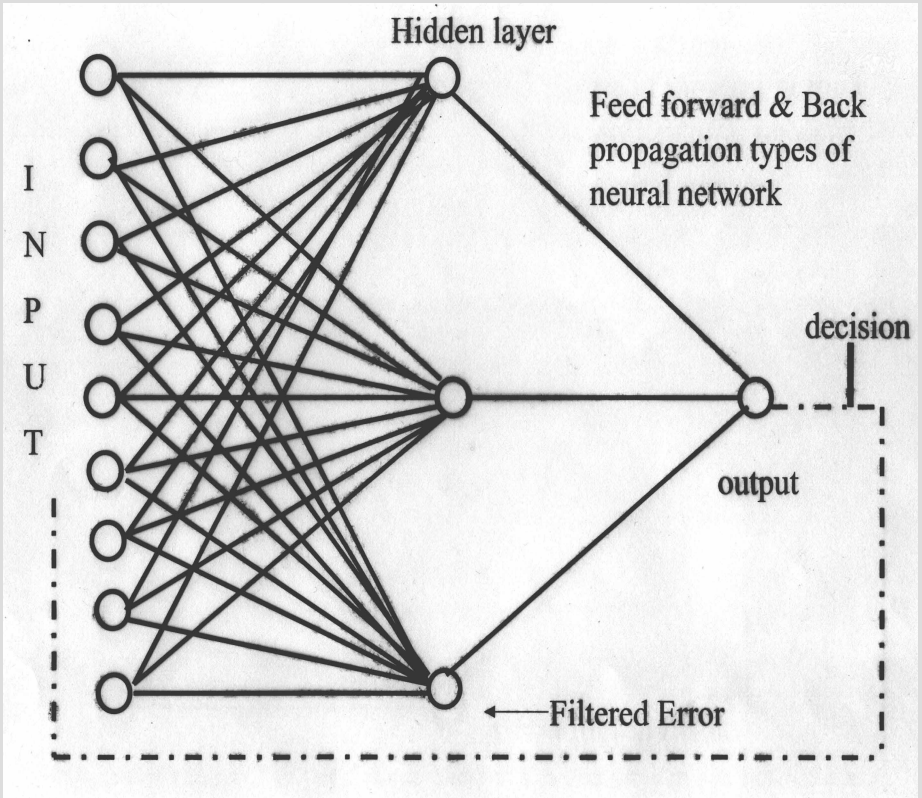


Amino acid/Position	P1	P2	P3	P4	P5	P6	P7	P8	P9
A	0.69	-0.99	0.52	-0.27	0.29	0.31	0.62	0.27	-0.09
C	-0.62	-1.47	0.08	0.71	-0.75	0.32	0.22	0.61	-0.40
D	-1.05	-1.75	0.18	0.19	-0.14	-0.53	-1.05	-1.06	-1.54
E	-1.52	-1.73	-1.14	0.56	-0.72	-0.86	-0.84	0.00	-1.79
F	1.08	-1.80	-0.15	-0.47	0.70	0.32	0.92	0.36	-1.67
G	-0.51	-1.74	-0.26	-0.22	-0.26	-0.91	-1.19	-0.58	-1.86
H	0.38	-1.80	-0.24	-0.40	0.00	-0.34	0.40	0.27	-2.00
I	0.00	0.12	-0.47	-0.47	-0.21	0.14	0.20	-0.18	0.38
K	0.14	-1.75	-1.14	-0.01	-1.17	-1.12	-1.69	-0.65	-1.94
L	0.09	6.31	0.54	-0.06	0.26	0.45	0.50	0.62	6.03
M	0.31	6.22	0.63	-0.86	0.08	0.33	-0.29	-0.43	-0.37
N	-0.67	-2.00	0.75	-0.04	-0.04	-0.20	0.10	-0.35	-2.00
P	-0.93	-1.88	-0.29	0.76	0.57	0.87	0.42	0.28	-2.00
Q	-0.69	-1.33	-0.61	-0.02	-0.38	-0.58	-0.76	-0.47	-1.50
R	0.31	-2.00	-0.67	0.20	0.00	-0.70	-0.31	-0.14	-1.82
S	0.72	-1.87	0.60	0.33	0.04	0.45	0.24	0.92	-1.62
T	-0.48	-0.16	-0.73	-0.35	-0.03	0.00	-0.11	0.27	-0.29
V	-0.22	-0.07	0.09	-0.26	0.49	0.58	0.43	-0.39	6.28
W	-0.67	-2.00	0.80	-0.74	0.50	-0.80	-0.40	-0.40	-2.00
X	0.00	2.00	2.00	0.00	2.00	2.00	2.00	2.00	2.00
Y	1.25	-1.58	0.88	-0.67	0.67	-0.34	0.22	-0.29	-2.00

ANN implementation: For 30 alleles out of total 48 alleles having more than >15 binders

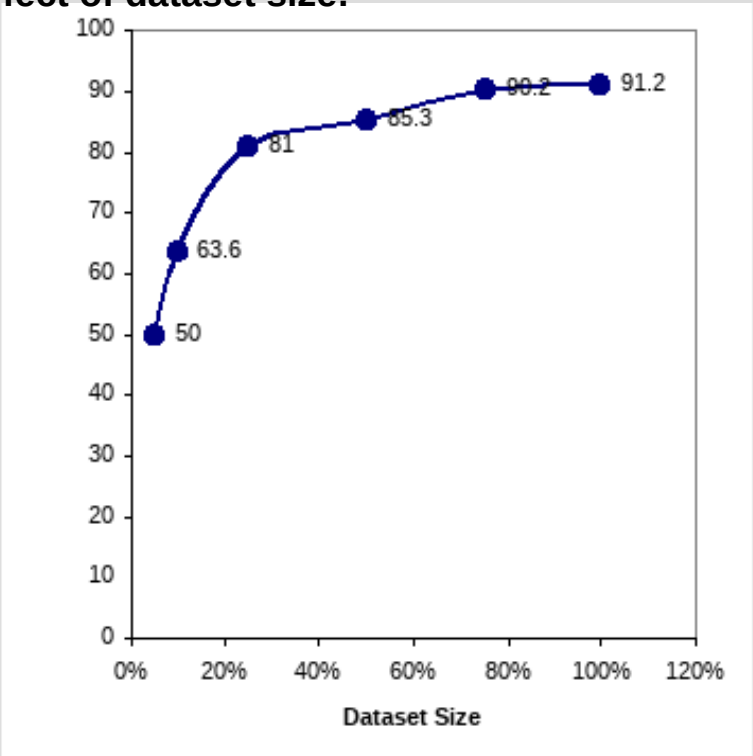
Implementation

Amino acid	Binary Encoding
Alanine (A)	1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Cysteine (C)	0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Aspartate (D)	0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Glutamate (E)	0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Phenylalanine (F)	0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0



Results

Effect of dataset size:



MHC Allele	Sensitivity	Specificity	PPV	Accuracy
HLA-A1	95.4	95.4	95.4	95.4
HLA-A2	83.0	88.5	64.9	87.5
Mean ±STDEV	**87.3±5.8**	**87.3±6.0**	**86.6±6.6**	**87.3±5.9**

Prediction of mutated MHC binders

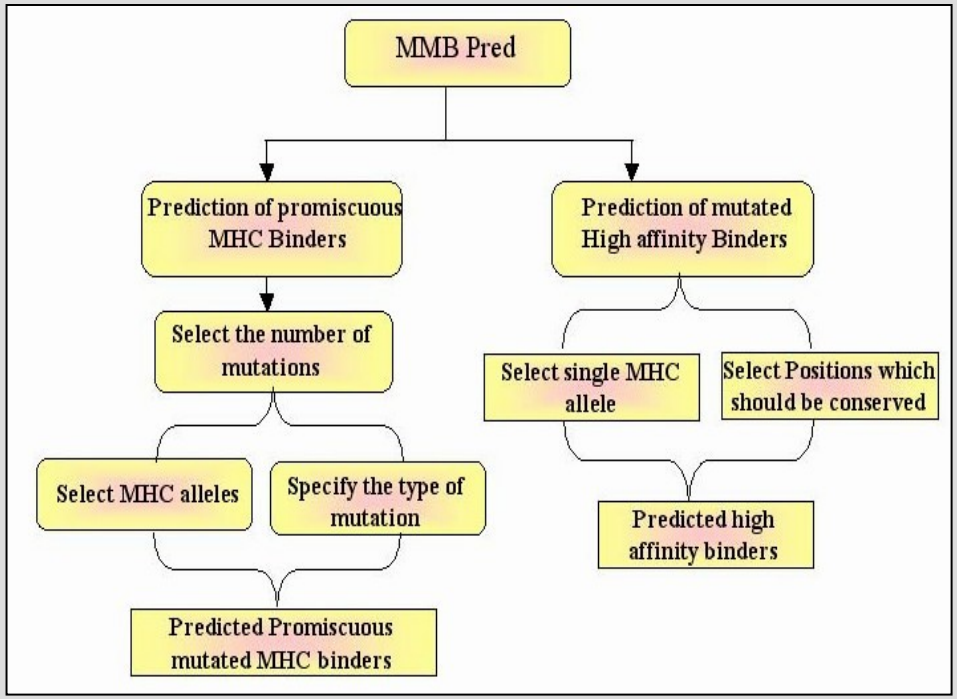
Role of mutations: Can increase the MHC binding affinity (Mucha *et al.*, 2002 .)
Enhance the promiscuousity of binders (Berzofsky *et al* ,1993) .

Available method

No

Our appraoch:

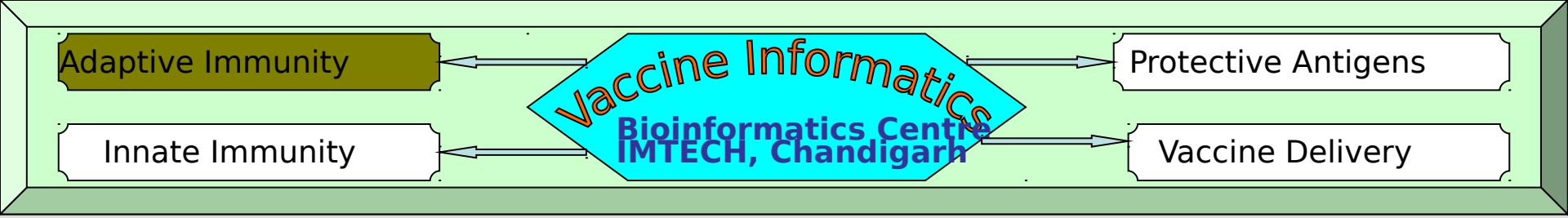
- Quantitative matrices (nHLAPred).
- Testing on independent data.
- Random or position specific mutations.
- Allows up to three mutations.



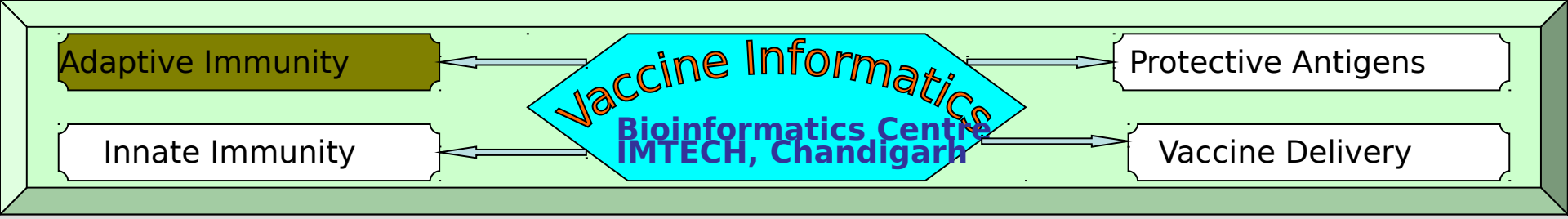
Results: high affinity binders from peptides "ITDQVPFSV" and "YLEPGPVTA"

Peptide sequence	Positions Of mutations	Experimentally proven [a]	Similarity [b]
ITDQVPFSV	1	2(F,Y)	2 (2)
ITDQVPFSV	2	3(L,M,I)	3 (3)
ITDQVPFSV	3	5(W,F,Y,A,M)	4 (7)
ITDQVPFSV	2 and 3	6(L&W, L&F, L&Y, L&A, L&M, L&S)	5 (12)
ITDQVPFSV	1 and 2	3(W&L,F&L,Y&L)	2 (2)
YLEPGPVTA	1	2(F, W)	-
YLEPGPVTA	3	6(Y, W, F, M, S, A)	6(11)
YLEPGPVTA	9	3(V, L, I)	3(3)
YLEPGPVTA	3 and 9	6(M&V, S&V, A&V, Y&V, F&V, W&V)	6 (17)

The method has been implemented online as **MMBPred.**




```
                ---------10--------20--------30--------40--------50--------60--
DRB1_0101:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
DRB1_0102:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
DRB1_0301:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
DRB1_0305:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
DRB1_0306:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
DRB1_0307:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
DRB1_0308:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
DRB1_0309:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
DRB1_0311:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
DRB1_0401:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
DRB1_0402:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
DRB1_0404:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
DRB1_0405:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
DRB1_0408:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
DRB1_0410:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
DRB1_0421:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
DRB1_0423:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
DRB1_0426:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
DRB1_0701:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
DRB1_0703:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
DRB1_0801:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
DRB1_0802:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
DRB1_0804:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
DRB1_0806:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
DRB1_0813:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
DRB1_0817:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
DRB1_1101:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
DRB1_1102:  MKVKYALLSAGALQLLVVGCGSSHHETHYGYATLSYADYWAGELGQSRDVLLAGNAEADRAGI
```



Prediction of MHC II Epitopes (T_{helper} Epitopes)

- **Propred: Promiscuous of binders for 51 MHC Class II binders**
  - **Virtual matrices**
  - **Singh and Raghava (2001) Bioinformatics 17:1236**
- **HLADR4pred: Prediction of HLA-DRB1*0401 binding peptides**
  - **Dominating MHC class II allele**
  - **ANN and SVM techniques**
  - Bhasin and Raghava (2004) Bioinformatics 12:421.
- **MHC2Pred: Prediction of MHC class II binders for 41 alleles**
  - **Human and mouse**
  - **Support vector machine (SVM) technique**
  - **Extension of HLADR4pred**
- **MMBpred: Prediction pf Mutated MHC Binder**
  - Mutations required to increase affinity
  - Mutation required for make a binder promiscuous
  - **Bhasin and Raghava (2003) Hybrid Hybridomics, 22:229**
- **MOT : Matrix optimization technique for binding core**
- **MHCBench: Benchmarting of methods for MHC binders**

Adaptive Immunity

Innate Immunity

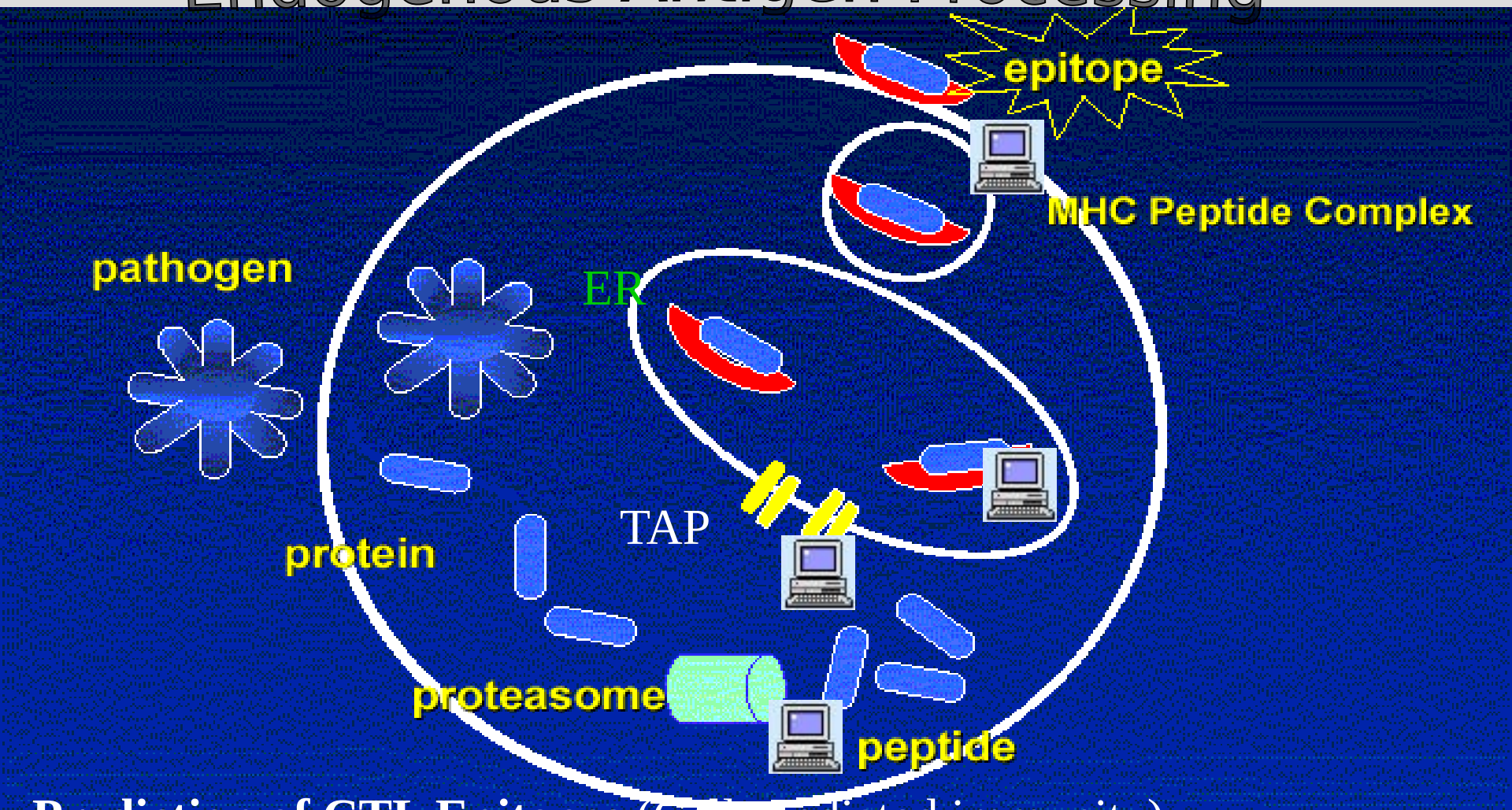
Vaccine Informatics

Bioinformatics Centre
IMTECH, Chandigarh

Protective Antigens

Vaccine Delivery

Endogenous Antigen Processing

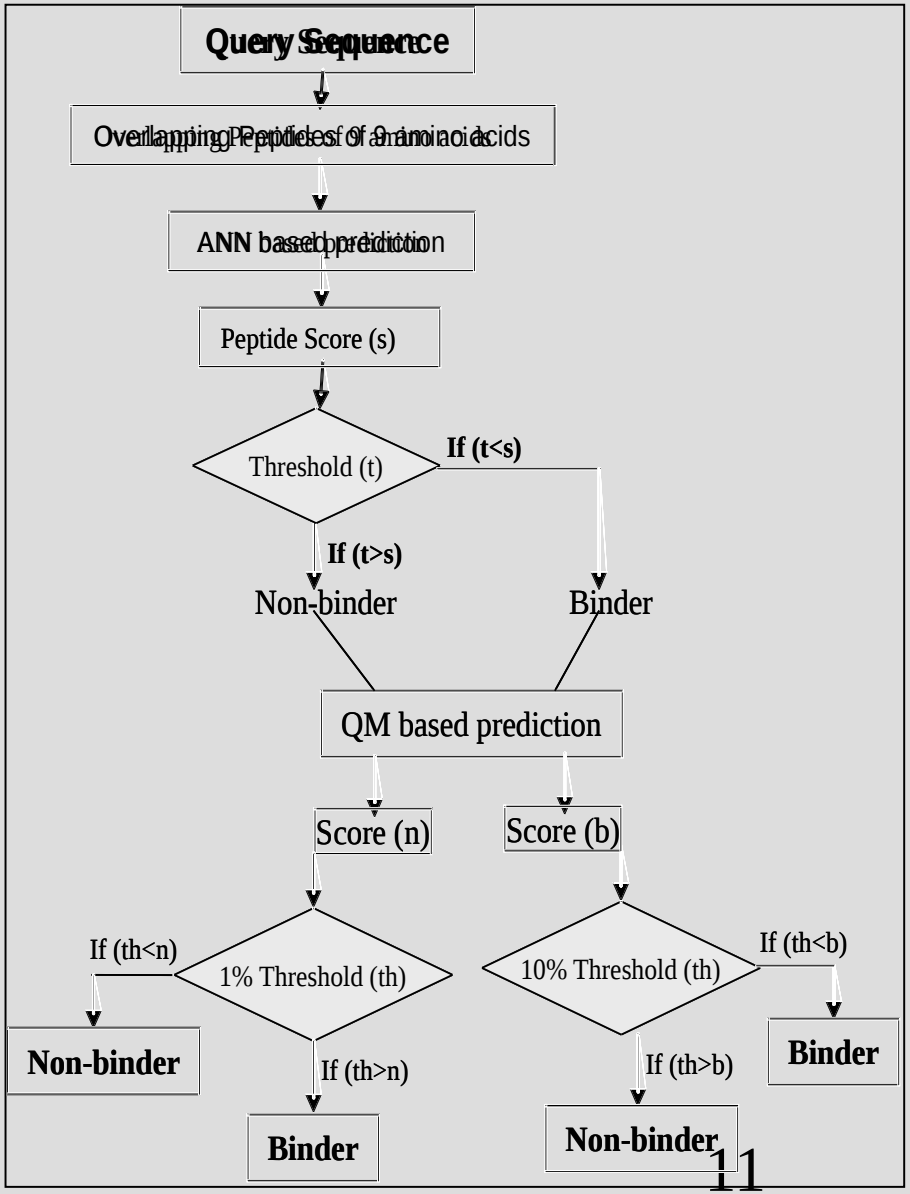


Prediction of CTL Epitopes (Cell-mediated immunity)

Hybrid approach: to further improve the accuracy

Result

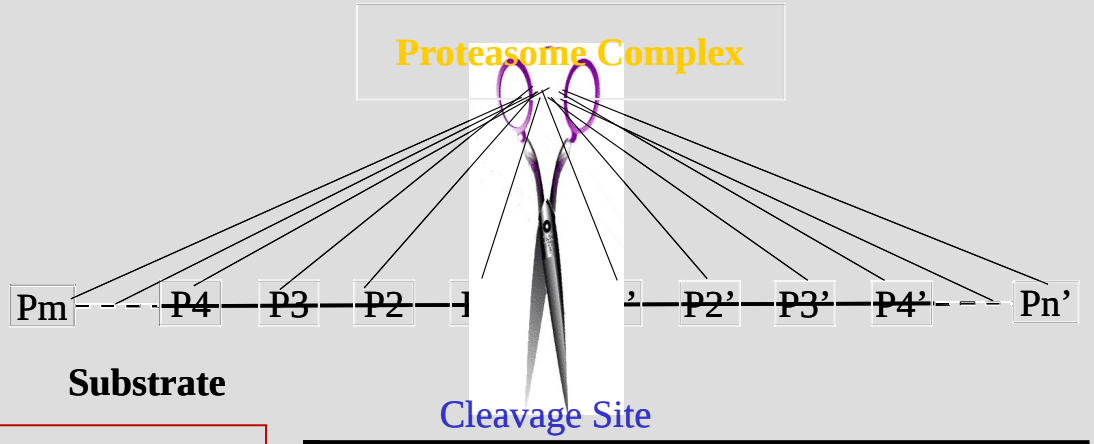
	ANN + QM
	91.8±5.4
	94.9±3.4
	94.0±4.9
	93.2±4.1
	93.6±2.9



Prediction of proteasome Cleavages

Proteasome cleavage specificity

Proteasome is a multienzyme complex whose cleavage specificity is not only effected by residues at cleavage site but also by the neighboring residues.



Existing methods:

FragPredict: 20S proteasome

Based on cleavage motifs and time dependent degradation of peptides.

PAProC: Human and Yeast Proteasome

Cleavage is determined using stochastic hill-climbing algorithm.

NetChop: For Proteasome and immunoproteasome.

Based on MHC class I ligand data using ANN

Recently, these three methods has been evaluated on independent dataset. The results showed that NetChop (MCC=0.32) is better then rest of methods. (Saxova et al., 2003). Low accuracy of all methods motivated us to develop better method

Our approach:

- **Generation of *in vitro* and *in vivo* data (MHC class I ligands).**
- **Application of machine learning techniques.**
  - **Support Vector Machine**
  - **Parallel Exemplar Based Learning (PEBLS)**
  - **Waikato Environment for Knowledge analysis (Weka)**
- **Evaluation**
  - **Leave One out cross validation and calculating Matthews Correlation Coefficient (MCC)**

Analysis & prediction of TAP binders

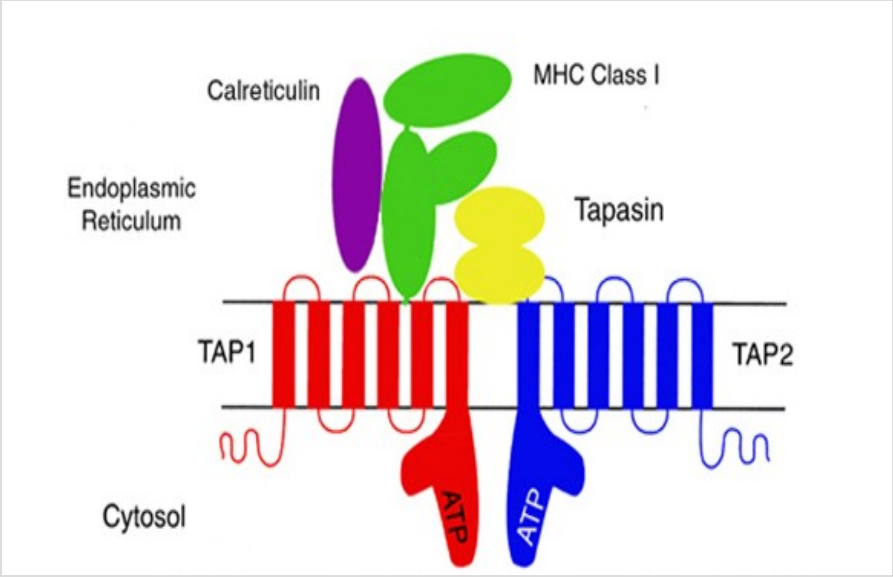
General about TAP:

❖The binding of the peptides to TAP is crucial for its translocation from cytoplasm to ER.

❖Three N terminal residues and C terminal residue is important for TAP selectivity

❖Charged amino acids are preferred whereas aromatics, acidic (in P1) and proline residue (in P2, P3) are disfavored.

❖Hydrophobic residues at C terminal



Available methods:

Daniel *et al*., 1998 has developed an ANN Based Methods & achieved an correlation of 0.73.

Brusic *et al.,* 1999 has developed another ANN based method to predict TAP binders with fair accuracy.

Limitations: Low correlation.
Not available as software or webtool.

Our approach:

High quality dataset

Analysis of TAP binding peptides.

Prediction using QM
Prediction using SVM:

I. Binary sequence.
II. Sequence features.
III. Combination of both.

Evaluation: Leave One Out Cross-validation (LOOCV) using Correlation

Analysis of TAP binders:

Dataset :

Dataset of 431 peptides of 9 aa whose TAP binding affinity is experimentally known

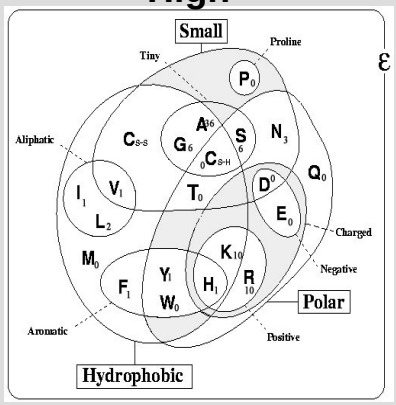
↓

Binding affinity of peptides were normalized to 0-10.

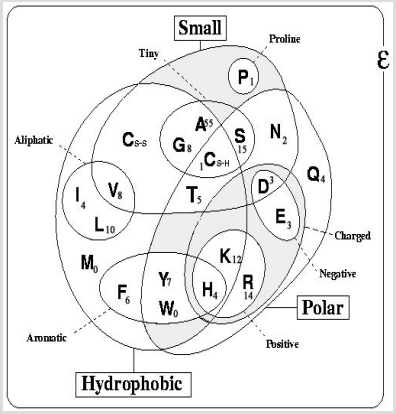
409	22
Binder	Negligible binder

Analysis by generating venn diagrams

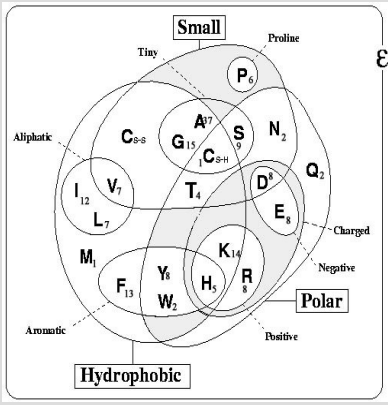
High



Intermediate



Low



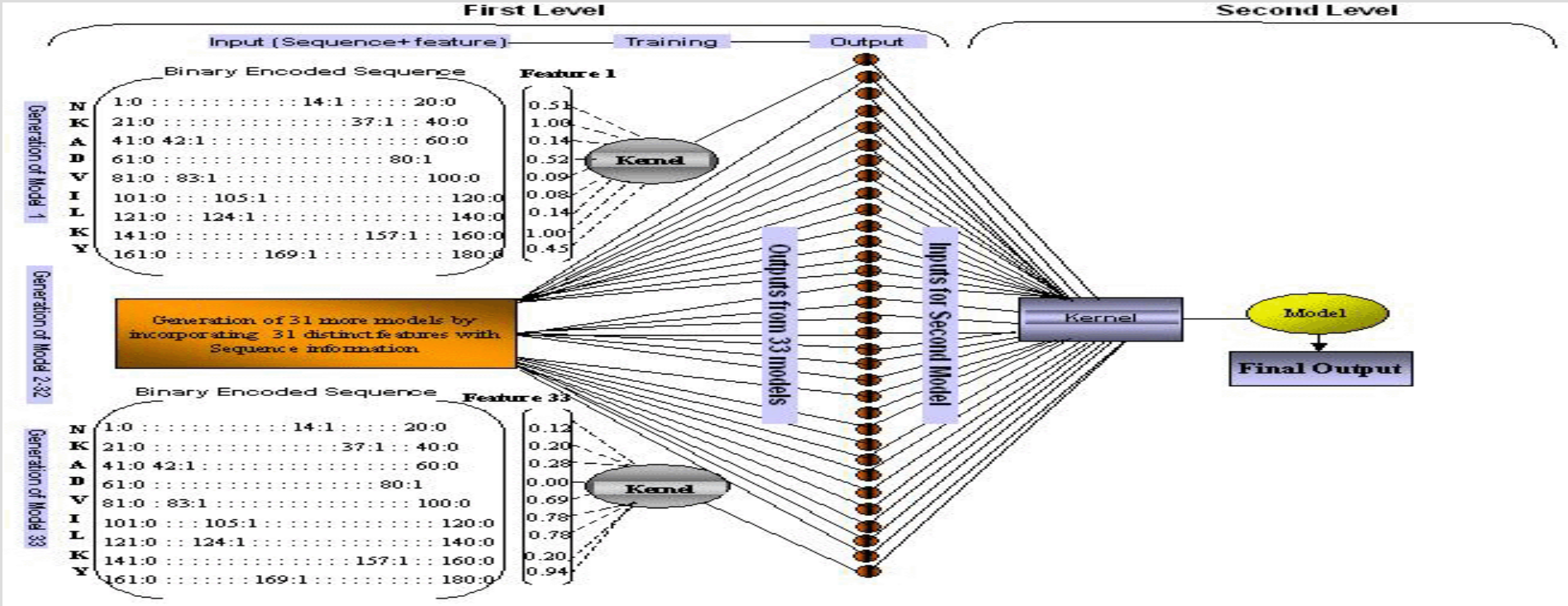
Correlation between features & position

Results

Features	P1	P2	P3	P4	P5	P6	P7	P8	P9
Volume	○	●	●				●		●
Charge	●	●					●		●
Aromatic	○		●				●		●
Hydrophobic	○	○	●				●		●
Hydrophilic	●	●	○				○		○
Hydropathy		○					●		
Accessibility		●	●						●
Flexibility		●	○				○		
% buried Residues		○					●		○

● liked ○ Disliked

Prediction approaches: cascade SVM (a novel approach)



Results:

Method	Correlation
Quantitative matrix	0.65
SVM (Binary)[1]	0.81
SVM (Feature based)[2]	0.80
SVM(1+2)	0.82
Cascade SVM (1st model)	0.80
Cascade SVM (2nd model)	0.88

Cascade SVM is able to outperform the other SVMs and existing methods.

The method based on cascade SVM has been implemented online as TAPPred

Direct prediction of CTL epitope

Why direct CTL epitope prediction methods required?

1. MHC binder prediction methods are not able to discriminate between T cell epitopes and non-epitopes (MHC binders).

Available algorithms:

AMPHI: T cell epitope form amphipathic helix.

SOHHA: Strip of Helix Hydrophobicity.

X-ray crystallography proved that T cell epitopes have extended conformation.

Optimer: Based on T cell epitope conformation & motif density.

Limitations:

Based only on T cell epitopes.
Based on small dataset.
No one is available online.
Performance is nearly random (Deavin *et al*., 1996)

Our Approach:

- Larger & high quality dataset.

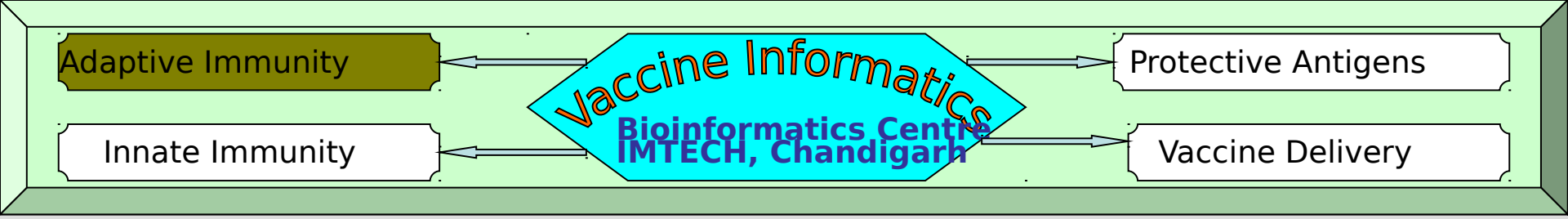
Computational techniques

- QM
- ANN
- SVM

Consensus & Combined prediction

- Evaluation 5-fold cross-validation
- Testing on independent dataset

Dataset: From MHCBN database.
1,137 CTL epitopes
1,134 non-epitopes



Prediction of MHC I binders and CTL Epitopes

Propred1: Promiscuous binders for 47 MHC class I alleles

- **Cleavage site at C-terminal**
- ***Singh and Raghava (2003) Bioinformatics 19:1109***

nHLApred: Promiscuous binders for 67 alleles using ANN and QM

- **Bhasin and Raghava (2007) J. Biosci. 32:31-42**

TAPpred: Analysis and prediction of TAP binders

- ***Bhasin and Raghava (2004) Protein Science 13:596***

Pcleavage: Proteasome and Immuno-proteasome cleavage site.

- **Trained and test on in vitro and in vivo data**
- ***Bhasin and Raghava (2005) Nucleic Acids Research 33: W202-7***

CTLpred: Direct method for Predicting CTL Epitopes

- ***Bhasin and Raghava (2004) Vaccine 22:3195***

Adaptive Immunity

Innate Immunity

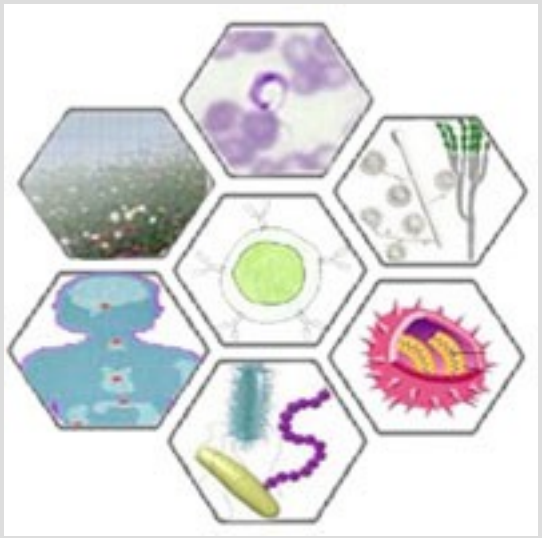
Vaccine Informatics

Bioinformatics Centre
IMTECH, Chandigarh

Protective Antigens

Vaccine Delivery

BCIPEP: A database of B-cell epitopes.



SWISS-PROT | PDB | MHCBN | PUBMED | GenBank

Structure of Each Entry
- Entry No
- Peptide Sequence
- Source Protein
- Experimental Method
- Immunogenicity
- Neutralization
- Model Organism
- Antibody
- Pathogen Group
- Publication reference
- Database Reference
- Antigen Structure
- Comment

Hyperlinks

Bcipep

Data structure

Interactive Tools
- Keyword Search
- Peptide Search
- Peptide Mapping
- Antigenic BLAST
- Data Submission

Web Resources
- Ab Structure
- Db. Information
- Related Links
- SRS @ IMTECH
- SRS @ EBI

Total Entries (3031)

Virus (2046) | Bacteria (539) | Others (157) | Fungi (53) | Protozoa (236)

Saha et al.(2005) BMC Genomics 6:79.

Saha et al. (2006) NAR (Online)

BCEpred: Benchmarking of physico-cemical properties used in existing B-cell epitope prediction methods

In 2003, we evaluate parmeters on 1029 non-redudant B-cell epitopes obtained from BCIpep and 1029 random peptide

Saha and Raghava (2004) ICARIS 197-204.

Physico-chemical Properties	Threshold	Sensitivity	Specificity	Accuracy% (Max)
Hydrophilicity [1]## (Parker et al., 1986)**	2.00	33	76	54.47
Accessibility[2] (Emini et al., 1985)	2.00	65	46	55.49
Flexibility [3] **(**Karplus and Schulz, 1985)	**1.90**	**47**	**68**	**57.53**
Surface [4] (Janin and Wodak, 1978)	2.40	37	74	55.73
Polarity [5] (Ponnuswamy et al., 1980)	2.30	2.8	81	54.08
Turns [6] (Pellequer et al., 199)	1.90	17	89	52.92
Antigenic Scale [7] (Kolaskar and Tongaonkar, 1990)	1.80	59	52	55.59
[3]+[1]+[5]+[4]	**2.38**	**56**	**61**	**58.70**

Challenge in Developing Method

1. Machine learnning technique needs fixed length pattern where epitope have variable length
2. Classification methods need positive and negative dataset
3. There are experimentally proved B-cell epitopes (positive) dataset but not Non-epitopes (negative)

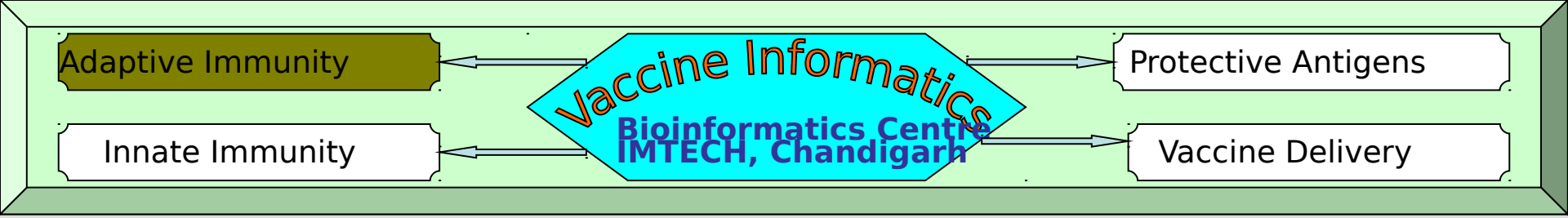
Assumptions to fix the Problem

1. More than 90% epitope have less than 20 residue so we fix maximum length 20
2. We added residues at both end of small epitopes from source protein to make length of epitope 20
3. We generate random peptides from proteins and used them as non-epitopes

Creation of fixed pattern of 20 from epitopes

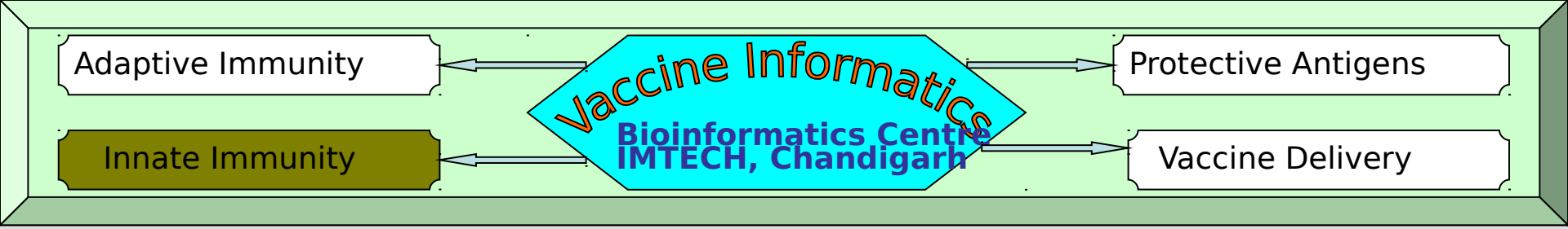
Window Length/ Peptide	**AEFPLDIT**	**ACVPTDPNPQEVVLVNVTEN** (20 amino acid length)
20	PKGYVG**AEFPLDIT**AGTEAA	**ACVPTDPNPQEVVLVNVTEN**
18	KGYVG**AEFPLDIT**AGTEA	**CVPTDPNPQEVVLVNVTE**
16	GYVG**AEFPLDIT**AGTE	**VPTDPNPQEVVLVNVT**
14	YVG**AEFPLDIT**AGT	**PTDPNPQEVVLVNV**
12	VG**AEFPLDIT**AG	**TDPNPQEVVLVN**
10	G**AEFPLDIT**A	**DPNPQEVVLV**

Results



Prediction of B-Cell Epitopes

- **BCEpred: Prediction of Continuous B-cell epitopes**
  - Benchmarking of existing methods
  - Evaluation of Physico-chemical properties
  - Poor performance slightly better than random
  - Combine all properties and achieve accuracy around 58%
  - Saha and Raghava (2004) ICARIS 197-204.
- **ABCpred: ANN based method for B-cell epitope prediction**
  - Extract all epitopes from BCIPEP (around 2400)
  - 700 non-redundant epitopes used for testing and training
  - Recurrent neural network
  - Accuracy 66% achieved
  - Saha and Raghava (2006) Proteins,65:40-48
- **ALGpred: Mapping and Prediction of Allergenic Epitopes**
  - Allergenic proteins
  - IgE epitope and mapping
  - Saha and Raghava (2006) Nucleic Acids Research 34:W202-W209



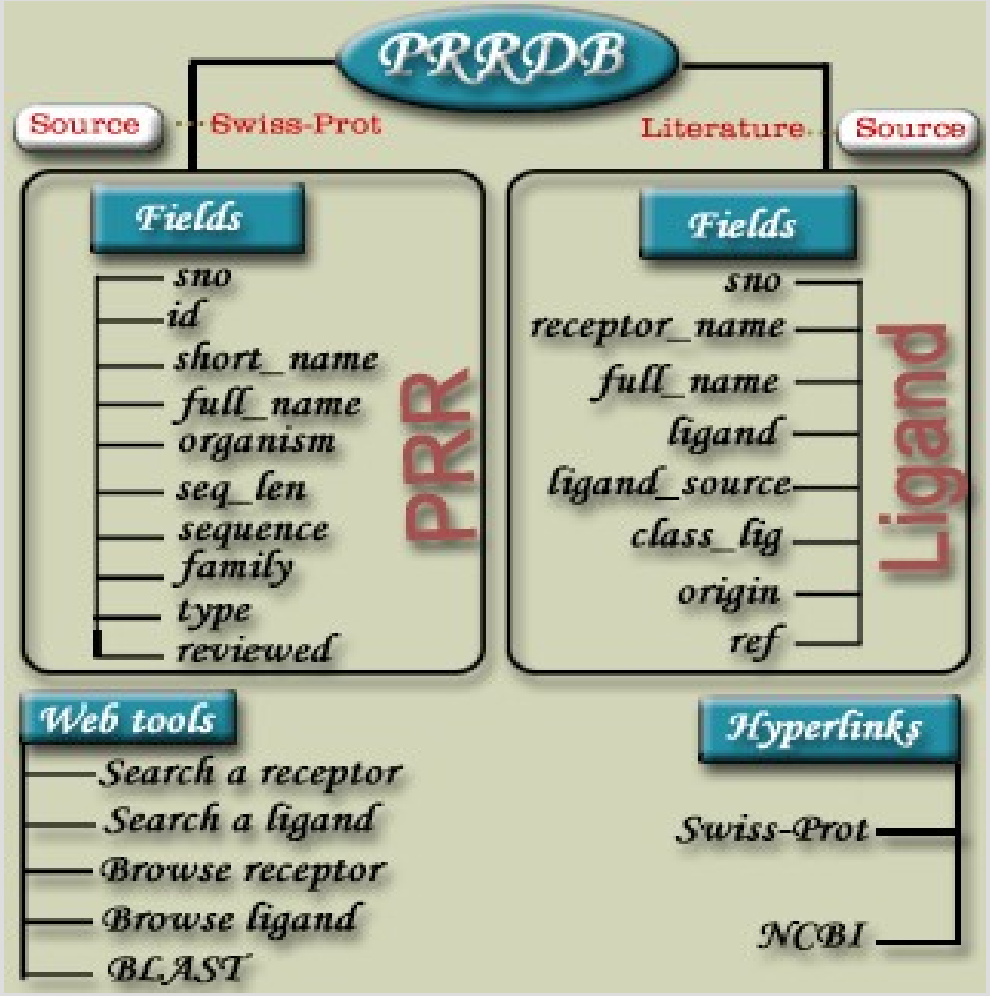
PRRDB is a database of pattern recognition receptors and their ligands

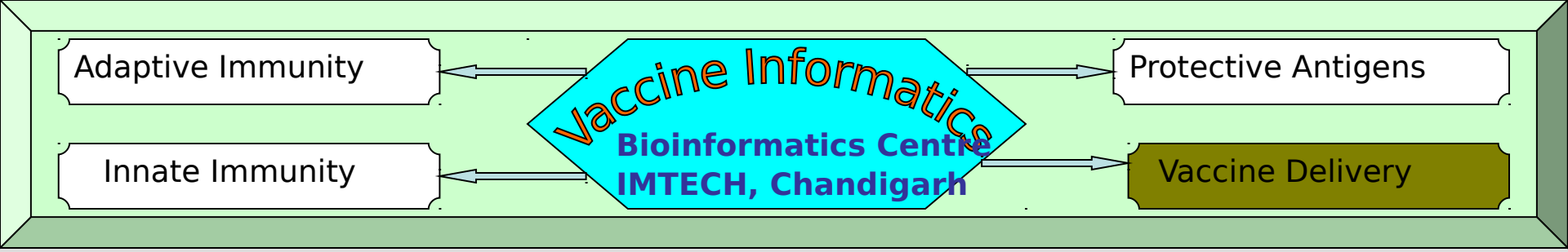
~500 Pattern-recognition Receptors

228 ligands (PAMPs)

77 distinct organisms

720 entries





Major Challenges in Vaccine Design

- **ADMET of peptides and proteins**
- **Activate innate and adaptive immunity**
- **Prediction of carrier molecules**
- **Avoid cross reactivity (autoimmunity)**
- **Prediction of allergic epitopes**
- **Solubility and degradability**
- **Absorption and distribution**
- **Glycocylated epitopes**