IMMUNE EPITOPE DATABASE AND ANALYSIS RESOURCE

Bjoern Peters
bpeters@lji.org
La Jolla Institute for Allergy and Immunology

October 23, 2018

Acknowledgments

La Jolla Institute for Allergy & Immunology and Leidos

Denmark

Consultants

- Ralph Kubo
- Chemical Entities of Biological Interest (ChEBI)

Prediction vs. Analysis Tools

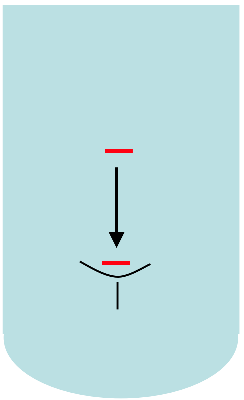
- Epitope ***prediction tools***
  - Machine learning algorithms that generalize the data contained in the IEDB to predict new epitopes
    - MHC class I & II binding
    - MHC class I processing & immunogenicity
    - B cell epitope predictions
- Epitope ***analysis tools***
  - Conservancy analysis
  - Population coverage
  - Cluster analysis
  - Validated reference epitope sets
  - Restrictor Analysis Tool for Epitopes (RATE)
  - Deimmunization

Epitope prediction tools: Machine learning explained using MHC-I binding as an example

Measuring and predicting MHC:peptide binding

Experimental Basis: MHC Binding Assay

List of peptides with allele specific binding affinity



Sequence	IC_{50}
QIVTMFEAL	3.6
LKGPDIYKG	308
NFCNLTSAF	50,000
AQSQCRTFR	38,000
CTYAGPFGM	143
CFGNTAVAK	50,000
. . .	

$\log(IC_{50})$ ~ Binding free Energy

low IC_{50} → high affinity

Impossible to measure all peptides

→ Predict binding peptides using machine learning

$\boxed{\text{Find}}$ function F_i in $\boxed{F_1, F_2, F_3, \ldots}$

F_i (Sequence) $\boxed{\approx}$ Affinity

Many different approaches (ANN, SVM, HMM, LP, ...)

T cell epitope mapping

ORF 1	M G Q I V T M F E A L P H I I D E V I N I V I I V L I V I T G I K A V Y N ...
ORF 2	M G L K G P D I Y K G V Y Q F K S V E F D M S H L N L T M P N A C S A N N ...
ORF 3	M H N F C N L T S A F N K K T F D H T L M S I V S S L H L S I D G N S N Y ...
ORF 4	M S A Q S Q C R T F R G R V L D M F R T A F G G K Y M R S G W G W T G S D ...
ORF 5	M H C T Y A G P F G M S R I L L S Q E K T K F F T R R L A G T F T W T L S ...
ORF 6	M K C F G N T A V A K C N V N H D A E F C D M L R L I D Y N K A A L S K F ...
ORF 7	M L M R N H L L D L M G V P Y C N Y S K F W Y L E H A K T G E T S V P K C ...

Calculate scoring matrix from affinities

Machine learning PSSM = Minimize the difference between predicted and measured binding affinities by varying the matrix values

N peptides with measured binding affinities

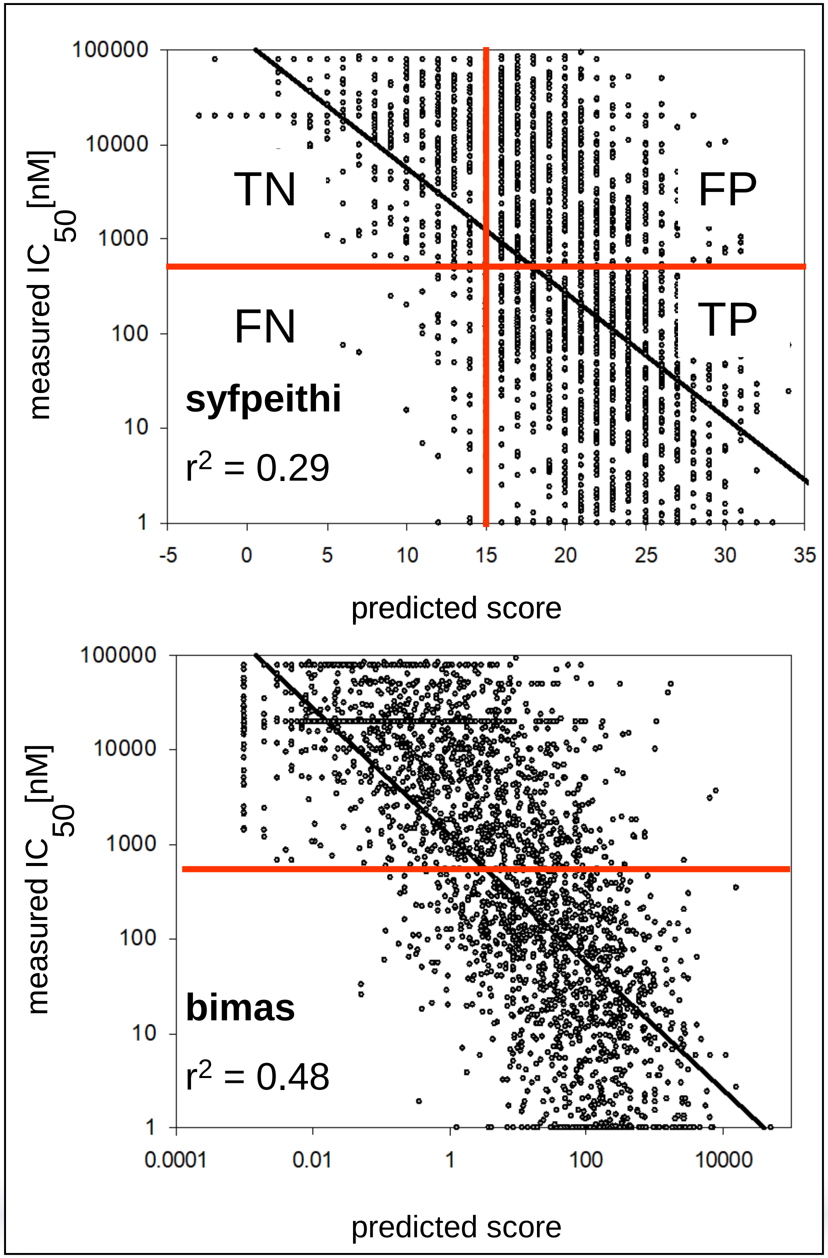
log (IC50)	Peptide
0.50	FQPQNGSFI
0.72	ISVANKIYM
2.37	RVYEALYYV
3.42	FQPQSGQFI
3.46	LYEKVKSQL
4.07	FKSVEFDMS
4.18	FQPQNGQFH
4.24	VLMLPVWFL
4.39	YMTLGQVVF
4.40	EDVKNAVGV
4.90	VFYEQMKRF
...	

	HLA A*0201								
	1	2	3	4	5	6	7	8	9
A	-0.3	0.8	-0.3	-0.3	-0.2	-0.3	0.0	0.0	-0.9
C	0.2	0.9	0.0	0.3	-0.5	-0.1	0.1	0.2	0.4
D	0.8	0.9	-0.4	-0.3	0.3	0.2	0.4	0.3	0.6
E	0.6	-0.4	0.7	-0.2	0.1	-0.4	-0.2	-0.2	-0.5
F	-1.3	0.5	-0.5	0.1	-0.1	0.0	-0.3	-0.4	-0.8
G	-0.2	0.1	0.3	-0.1	0.0	0.4	0.3	-0.1	0.2
H	1.1	0.9	-0.1	0.4	0.1	0.2	0.0	0.2	0.8
I	-0.4	-0.7	-0.4	0.1	-0.1	-0.4	-0.5	0.5	-1.4
K	-0.3	0.0	1.1	0.1	0.1	0.6	0.9	0.2	0.9
L	0.0	-1.9	-0.4	-0.2	0.0	-0.2	0.0	-0.1	-1.1
M	-0.7	-1.2	-0.7	0.2	-0.6	0.0	0.0	0.0	-0.8
N	-0.1	0.3	0.1	-0.3	-0.1	-0.3	0.0	0.2	0.7
P	1.2	0.5	0.6	-0.3	0.4	0.0	-0.4	-0.5	0.7
Q	0.4	-1.1	0.0	-0.1	0.4	-0.2	-0.3	0.2	0.7
R	-0.2	0.9	1.0	0.3	0.1	0.4	0.7	0.0	0.9
S	-0.3	0.1	0.1	-0.4	0.1	0.3	-0.2	-0.1	0.2
T	-0.2	-0.5	0.1	0.4	0.1	-0.5	0.2	0.0	-0.1
V	-0.1	-0.9	-0.1	0.2	0.0	-0.3	0.1	0.1	-1.9
W	0.0	0.7	-0.5	-0.2	-0.1	0.2	-0.3	-0.1	0.4
Y	-0.3	0.2	-0.6	0.2	0.0	0.4	-0.4	-0.3	0.8

Offset: 4.3

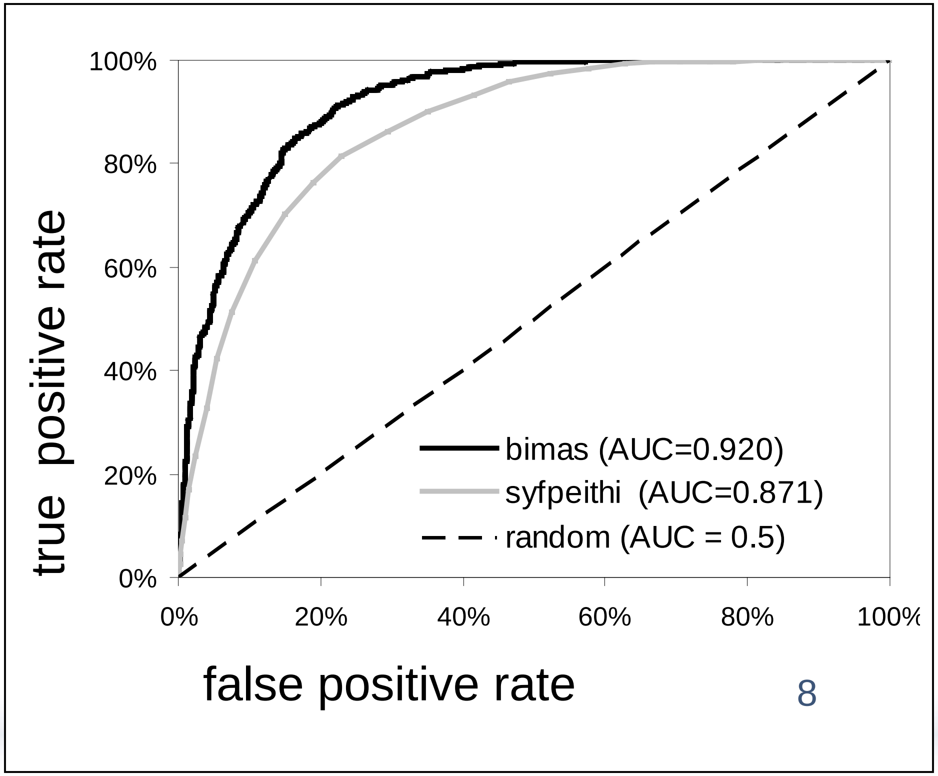
Evaluating prediction quality

Performance evaluation, external methods



Retrieved web-server predictions for each peptide in dataset
→ Scatterplot

ROC:
Predict Binders with IC_{50} < 500 nM



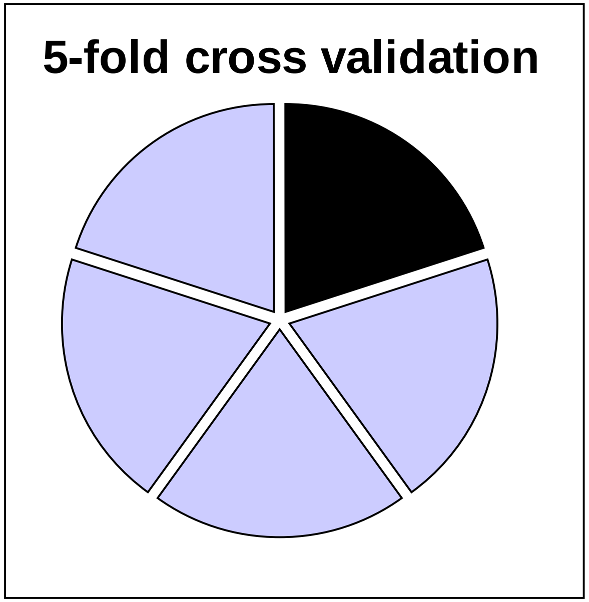
Performance evaluation, internal methods

- Split dataset in N subsets
- Train on N-1, predict left out subset
- Repeat N times

→Identical training and testing data for all methods

→Upcoming talks will show results

5-fold cross validation



Prediction tools available in the IEDB-AR

(detailed information in subsequent presentations)

Epitope Prediction Tools (B cell epitopes)

- Approaches
  - Amino acid property based predictions (e.g. hydrophilicity scale; historically first approaches)
  - Machine learning predictions + structure based approaches
- Breadth
  - applicable to all organisms
- Accuracy
  - Poor (AUC ~0.6-0.7), but evaluation is tricky

Epitope Prediction Tools (MHC Class I binding)

- Breadth
  - Humans: >80 alleles with specific predicitons
  - Non-human primates:
    - Chimpanzee (8 alleles)
    - Macaque (18 alleles)
  - Non-primates:
    - Pig (3 alleles)
    - Rat (1 alleles)
    - Mouse (6 alleles)
    - Cow (6 alleles)
  - ‘Pan’ predictions (NetMHCPan) covering all known human MHC alleles by extrapolating based on their sequence
- Accuracy
  - High, and highest in comparison to class II and antibodies (Average predictive performance with AUC greater than 0.9)
  - Incremental increases for many alleles since 2006

Epitope Prediction Tools (MHC Class II binding)

- Breadth
  - Humans (24)
  - Non humans (mouse, 3 alleles)

  → Pan predictions covering all known human class II allelels
- Accuracy
  - Still lower in comparison to class I, but higher than antibodies (average AUC=0.87 ±.05)
  - Substantial improvements over the last few years (in 2006, average AUC=0.76 ±.05)

Epitope Prediction Tools (MHC Class I - Processing)

- Two main approaches
  - 1) Separate predictors for proteasomal cleavage, TAP transport from independent experimental datasets
  - 2) Predictors trained on eluted MHC ligands
- Breadth
  - Only available (and validated) for human
- Accuracy
  - The improvement on prediction performance of pure MHC binding predictions is small but significant.
  - Large scale validation is still outstanding

Epitope Analysis Tools:

Add value to epitope datasets

Analysis tools available in the IEDB-AR

- **Conservancy analysis →** Analyze if epitopes are found conserved across different protein sequences
- **Cluster analysis →** Analyze how many epitopes in a set have significant sequence homology
- **Population coverage →** Analyze how many T cell epitopes with known HLA restriction will be recognized in a human population based on HLA frequencies

Analysis tools available in the IEDB-AR (2)

- **EpiFilter →** Generate reference datasets of high quality epitopes for various disease indications based on original query input
- **Restrictor Analysis Tool for Epitopes (RATE) →** Infer HLA restriction for a set of given epitopes from large datasets of T cell responses in HLA typed subjects
- **Deimmunization →** Identify immunodominant regions in a given protein, and suggest amino-acid substitutions that create non-immunogenic versions of the protein

Summary

- IEDB Prediction tools extrapolate from existing data to identify new candidate epitopes
  - ‘Machine learning’ approaches identify patterns
  - ROC curves / AUC values as preferred metrics of prediction performance
- IEDB Analysis tools help to examine existing sets of epitopes and gain new knowledge
  - No single metric of performance
  - Broad array of applications