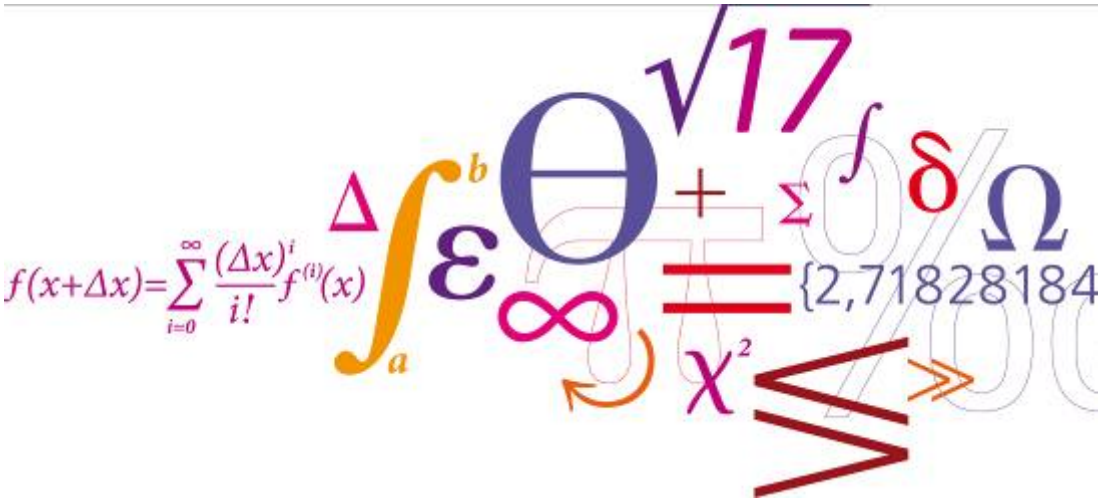
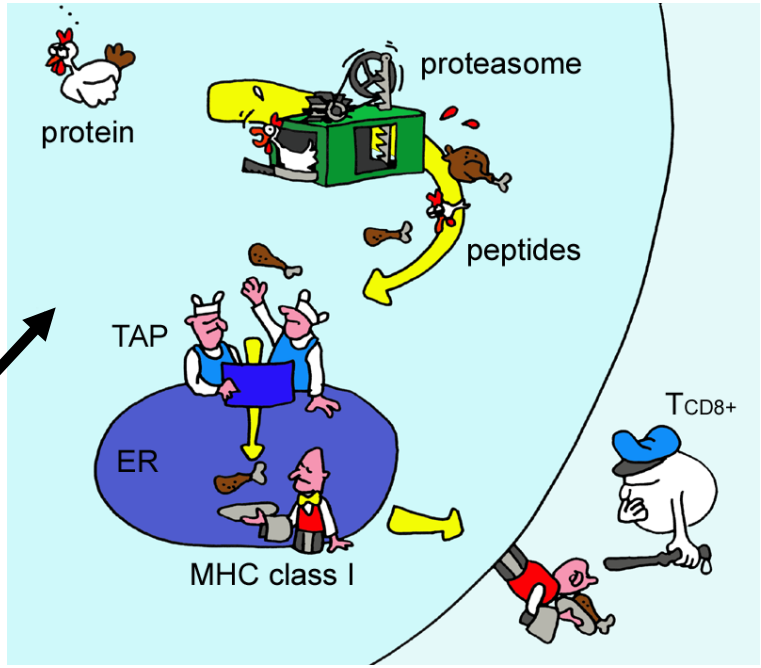
Workshop presentation: Development and application of Bioinformatics methods

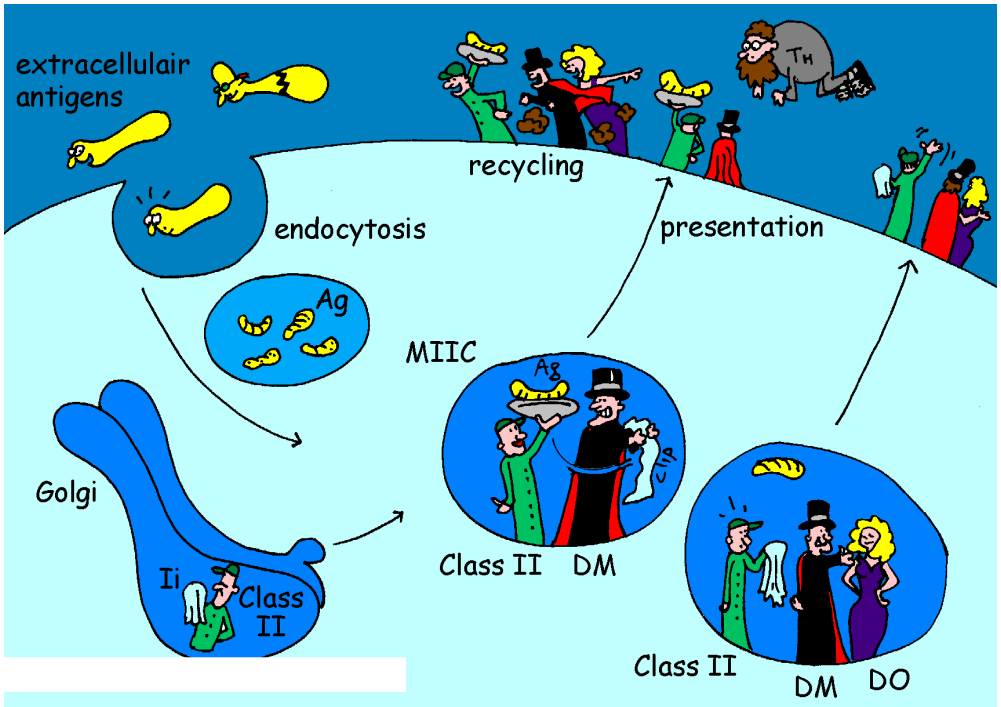


DTU Biosys
Department of Systems Biology

Immune system

- Innate – fast…
- Addaptive – remembers…
  - Cellular
    - Cytotoxic T lymphocytes (CTL)
    - Helper T lymphocytes (HTL)
  - Humoral
    - B lymphocytes





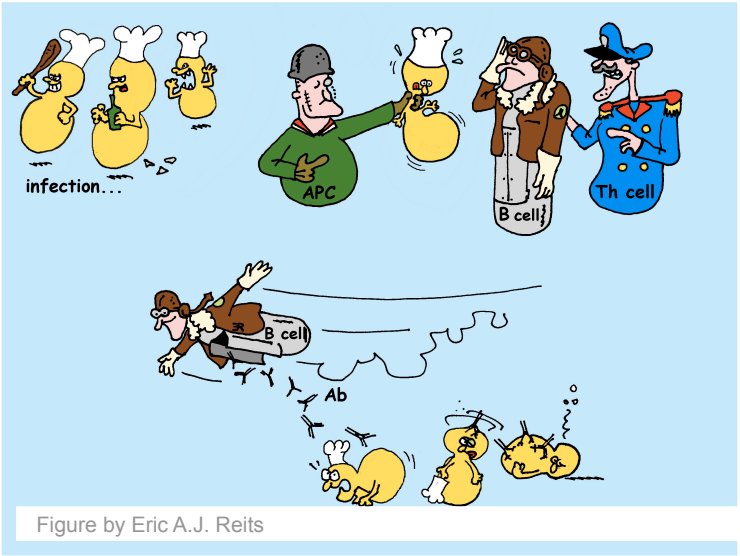
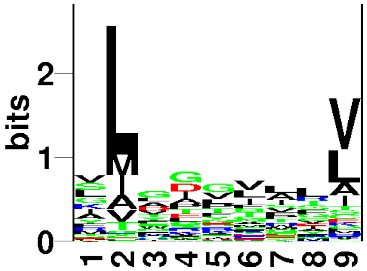


Figure by Eric A.J. Reits

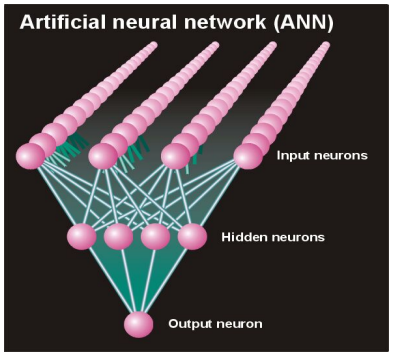
Data driven predictions

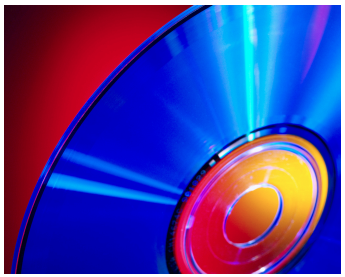
List of peptides that have a given biological feature

YMNGTMSQV
GILGFVFTL
ALWGFFPVV
ILKEPVHGV
ILGFVFTLT
LLFGYPVYV
GLSPTVWLS
WLSLLVPFV
FLPSDFFPS
CVGGLLTMV
FIAGNSAYE



Mathematical model (neural network, hidden Markov model)





Search databases for other biological sequences with the same feature/property

>polymerase"
MERIKELRDLMSQSRTREILTKTTVDHMAIIKKYTSGRQEKNPALRMKWMMAMKYPITAD
KRIMEMIPERNEQGQTLWSKTNDAGSDRVMVSPLAVTWWNRNGPTTSTVHYPKVYKTYFE
KVERLKHGTFGPVHFRNQVKIRRRVDINPGHADLSAKEAQDVIMEVVFPNEVGARILTSE
SQLTITKEKKEELQDCKIAPLMVAYMLERELVRKTRFLPVAGGTSSVYIEVLHLTQGTCW
EQMYTPGGEVRNDDVDQSLIIAARNIVRRATVSADPLASLLEMCHSTQIGGIRMVDILRQ
NPTEEQAVDICKAAMGLRISSSFSFGGFTFKRTNGSSVKKEEEVLTGNLQTLKIKVHEGY
EEFTMVGRRATAILRKATRRLIQLIVSGRDEQSIAEAIIVAMVFSQEDCMIKAVRGDLNF
...

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

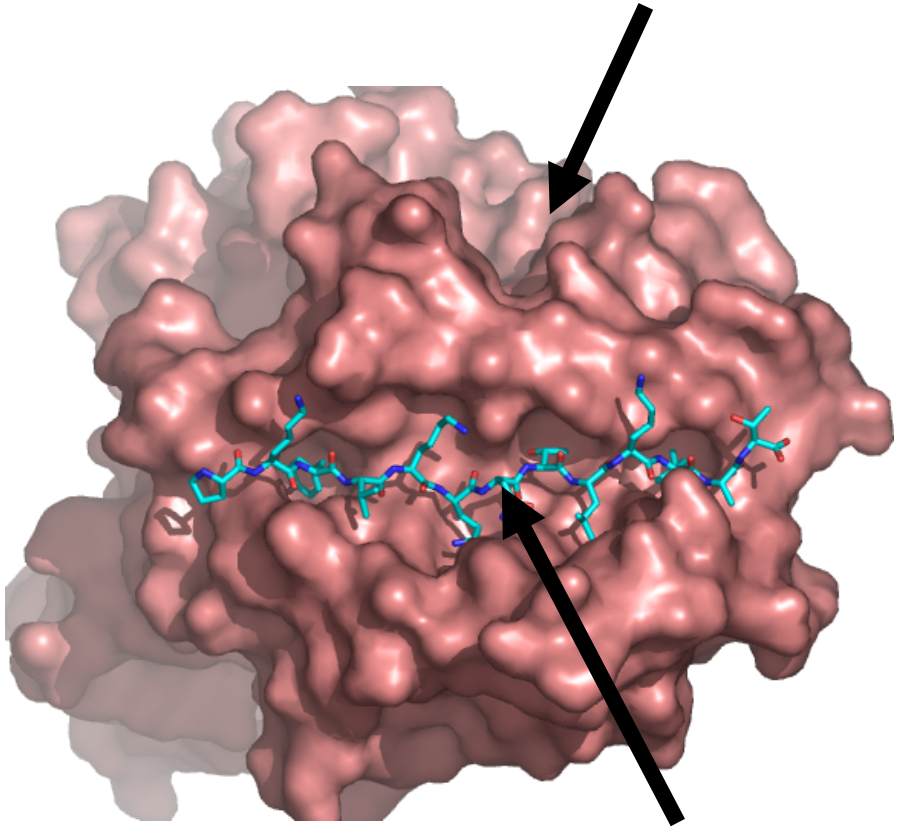
Prediction of HLA binding specificity

- Simple Motifs
  - Allowed/non allowed amino acids
- Extended motifs
  - Amino acid preferences (SYFPEITHI)
  - Anchor/Preferred/other amino acids
- Hidden Markov models
- Neural Networks

Class II MHC binding

Human MHC II:
~1000 variants

- MHC class II binds peptides in the class II antigen presentation pathway
- Binds peptides of length 9-18 (even whole proteins can bind!)
- Binding cleft is open
- Binding core is 9 aa



Peptide:
up to 20^9
variants

MHC class II prediction

- Complexity of problem
  - Peptides of different length
  - Weak motif signal
- Alignment crucial
- Gibbs Monte Carlo sampler

- RFFGGDRGAPKRG
- YLDPLIRGLLARPAKLQV
- KPGQPPRLLIYDASNRATGIPA
- GSLFVYNITTNKYKAFLDKQ
- SALLSSDITASVNCAK
- PKYVHQNTLKLAT
- GFKGEQGPKGEP
- DVFKELKVHHANENI
- SRYWAIRTRSGGI
- TYSTNEIDLQLSQEDGQTIE

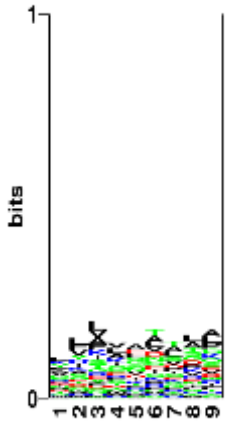
M Nielsen et al., Bioinformatics. 2004 20:1388-97

Class II binding motif

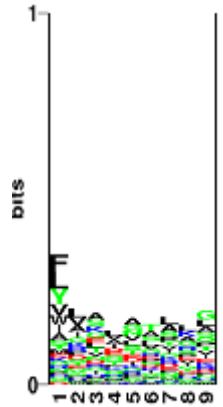
Alignment by Gibbs sampler

```
      RFFGGDRGAPKRG
   YLDPLIRGLLARPAKLQV
KPGQPPRLLIYDASNRATGIPA
     GSLFVYNITTNKYKAFLDKQ
      SALLSSDITASVNCAK
      PKYVHQNTLKLAT
       GFKGEQGPKGEP
      DVFKELKVHHANENI
      SRYWAIRTRSGGI
       TYSTNEIDLQLSQEDGQTI
```

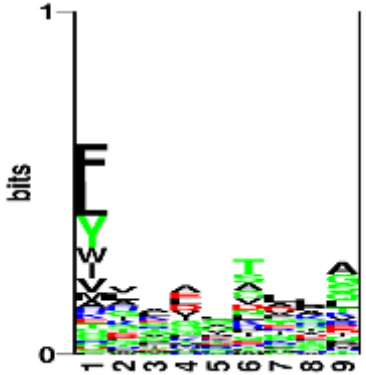
Random



ClustalW



Gibbs sampler

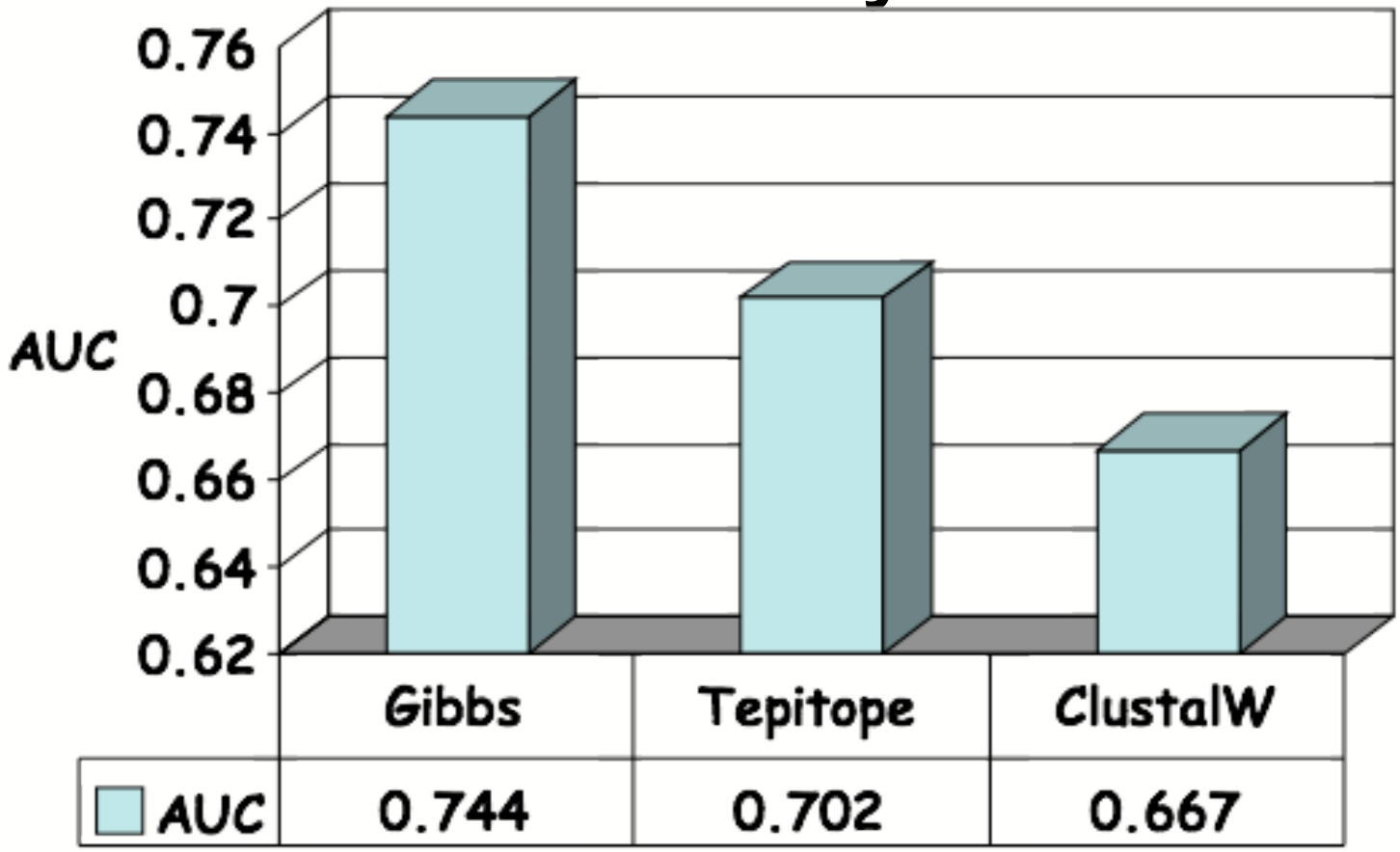


M Nielsen et al., Bioinformatics. 2004 20:1388-97

Earlier MHC Class II binding prediction methods

- Virtual matrices
  - TEPITOPE: Hammer, J., Current Opinion in Immunology 7, 263-269, 1995,
  - PROPRED: Singh H, Raghava GP Bioinformatics 2001 Dec;17(12):1236-7

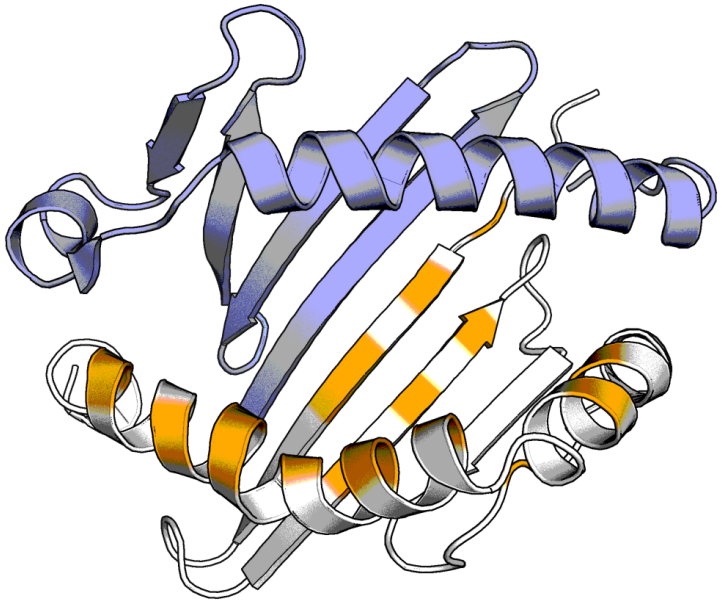
Gibbs sampler. Prediction accuracy



NetMHC-IIpan pseudo sequence

- Include polymorphic residues in potential contact with the bound peptide

- The contact residues are defined as being within 4.0 Å of the peptide in any of a representative set of HLA-DR, -DQ, and DP structures with peptides.
- Only polymorphic residues are included
- Pseudo-sequence consisting of *25 amino acid residues*.

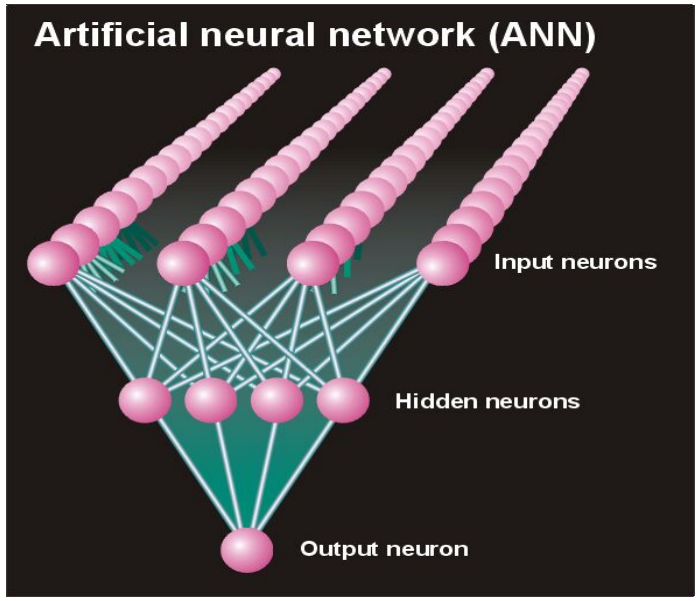


M Nielsen et al., Comput Biol. 2008

NetMHC-IIpan Method

```
XAAAAAVAAEAYXXX WLFECLECYPDYWLQRATYCHNVGF 0.119861 DRB1_0101 AAAAAVAAEAY
ALNVKRREGMFIDEX WLFECLECYPDYWLQRATYCHNVGF 0.238877 DRB1_0101 AALNVKRREGMFIDE
QPGLTSAVIEALPXX WLFECLECYPDYWLQRATYCHNVGF 0.169769 DRB1_0101 AAQPGLTSAVIEALP
XACVKDLVSKYLADN WLFECLECYPDYWLQRATYCHNVGF 0.577653 DRB1_0101 ACVKDLVSKYLADNE
KIGLHTEFQTVSFXX WLFECLECYPDYWLQRATYCHNVGF 0.982712 DRB1_0101 AFKIGLHTEFQTVSF
AGDLGRDELMELASD WLFECLECYPDYWLQRATYCHNVGF 0.061007 DRB1_0101 AGDLGRDELMELASD
XAGLIAIVMVTILLC WLFECLECYPDYWLQRATYCHNVGF 0.104993 DRB1_0101 AGLIAIVMVTILLCC
XAGYAATNDDNILSH WLFECLECYPDYWLQRATYCHNVGF 0.364429 DRB1_0101 AGYAATNDDNILSHV
XXAKCNLDHSSEFCD WLFECLECYPDYWLQRATYCHNVGF 0.156760 DRB1_0101 AKCNLDHSSEFCDML
XAKMKCFGNTAVAKC WLFECLECYPDYWLQRATYCHNVGF 0.734955 DRB1_0101 AKMKCFGNTAVAKCN
```

+ peptide length
+ PFR length




```
AAAAAVAAEAY      0.11986 0.33183
AALNVKRREGMFIDE  0.23888 0.44053
AAQPGLTSAVIEALP  0.16977 0.56296
ACVKDLVSKYLADNE  0.57765 0.59562
AFKIGLHTEFQTVSF  0.98271 0.46325
AGDLGRDELMELASD  0.06101 0.18933
AGLIAIVMVTILLCC  0.10499 0.36359
AGYAATNDDNILSHV  0.36443 0.29837
AKCNLDHSSEFCDML  0.15676 0.17179
AKMKCFGNTAVAKCN  0.73496 0.69730
```

M Nielsen et al., Comput Biol. 2008

Final NetMHC-IIpan method

		Pan		SMM-align		TEPITOPE
Allele	**N**	**Pearson**	**AUC**	**Pearson**	**AUC**	**AUC**
DRB1*0101	5166	0.682	0.841	0.610	0.802	0.720
DRB1*0301	1020	0.659	0.846	0.563	0.795	0.664
DRB1*0401	1024	0.625	0.816	0.496	0.751	0.716
DRB1*0404	663	0.701	0.860	0.579	0.801	0.770
DRB1*0405	630	0.637	0.833	0.560	0.789	0.759
DRB1*0701	853	0.725	0.870	0.618	0.812	0.761
DRB1*0802	420	0.660	0.843	0.555	0.787	0.766
DRB1*0901	530	0.517	0.728	0.360	0.655	
DRB1*1101	950	0.724	0.871	0.581	0.796	0.721
DRB1*1302	498	0.663	0.819	0.558	0.785	0.652
DRB1*1501	934	0.644	0.800	0.528	0.727	0.686
DRB3*0101	549	0.619	0.841	0.585	0.836	
DRB4*0101	446	0.695	0.871	0.541	0.793	
DRB5*0101	924	0.703	0.856	0.529	0.761	0.680
Ave*	**14**	**0.661**	**0.835**	**0.547**	**0.778**	
Ave**	**11**	**0.675**	**0.841**	**0.562**	**0.782**	**0.718**

M Nielsen et al., Comput Biol. 2008

Prediction servers at CBS

Web servers

CTL epitopes

http://www.cbs.dtu.dk/services/NetCTL/

MHC binding

http://www.cbs.dtu.dk/services/NetMHC/
http://www.cbs.dtu.dk/services/NetMHCII/
http://www.cbs.dtu.dk/services/NetMHCpan/
http://www.cbs.dtu.dk/services/NetMHCIIpan/

MHC Motif viewer

http://www.cbs.dtu.dk/biotools/MHCMotifViewer/Home.html

Proteasome processing

http://www.cbs.dtu.dk/services/NetChop-3.0/

B-cell epitopes

http://www.cbs.dtu.dk/services/BepiPred/
http://www.cbs.dtu.dk/services/DiscoTope/

Plotting of epitopes relative to reference sequence

http://www.cbs.dtu.dk/services/EpiPlot-1.0/

Analysis of human immunoglobulin VDJ recombination

http://www.cbs.dtu.dk/services/VDJsolver/

Geno-pheno type association based mapping of binding sites

http://www.cbs.dtu.dk/services/SigniSite/

PhD/master course in Immunological Bioinformatics, June, 2009

http://www.cbs.dtu.dk/courses/27685.imm/

Proposed application in assessment of protein drugs

1. Compare amino acid sequence of drug with the human proteome
2. Predict epitopes in regions that differ from the human proteome
3. Select representative HLA alleles
4. Verify binding experimentally
5. Assess predicted immunogenecity using blood from treated patients/transgenic animals
6. Compare with clinical findings of immunogenecity/adverse effects/lack of effect

Work in progress: Pilot study based on DrugBank

- www.drugbank.ca
- Records corresponding to 123 FDA-approved biotech (protein/peptide) drugs were downloaded
- Sequences were compared to the human proteome (sequences from “Homo Sapiens” in NR (non redundant database from NCBI)) using blast.
- Sequences found in DrugBank and NR need to be manually validated/curated

Types of proteins

- Human/Human protein sequence Identical proteins
- Non human proteins
- Modified/allelic human proteins
- Antibodies
  - Non human
  - Human-murine chimaer
  - Humanized
  - Human

Immunonological Bioinformatics Group

- Ole Lund
- Claus Lundegaard
- Morten Nielsen
- Mette Voldby Larsen
- Sheila Tang
- Jorid Sørli
- Marlene Erup Larsen
- Bent Petersen
- Thomas Stranzl
- Massimo Andretta
- Edita Bartaseviciute