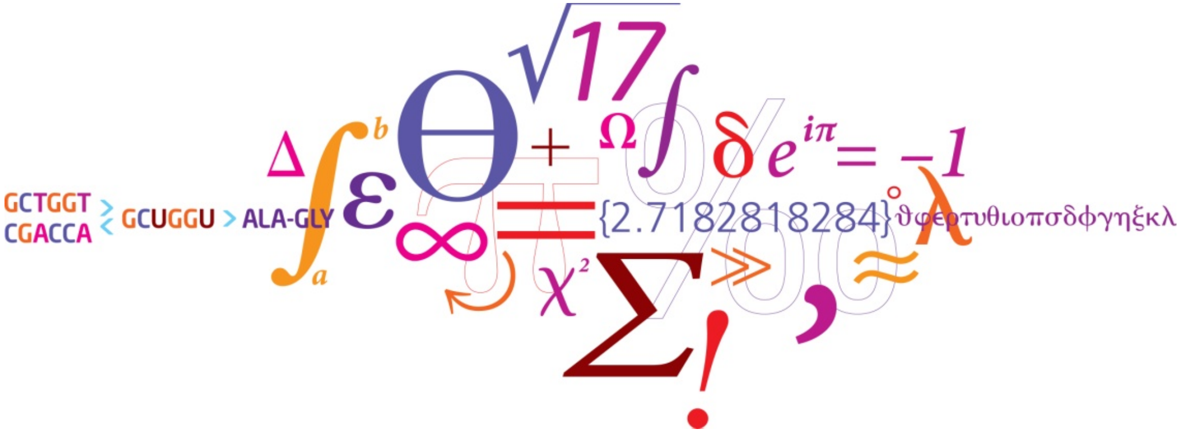
B-cell epitope Prediction

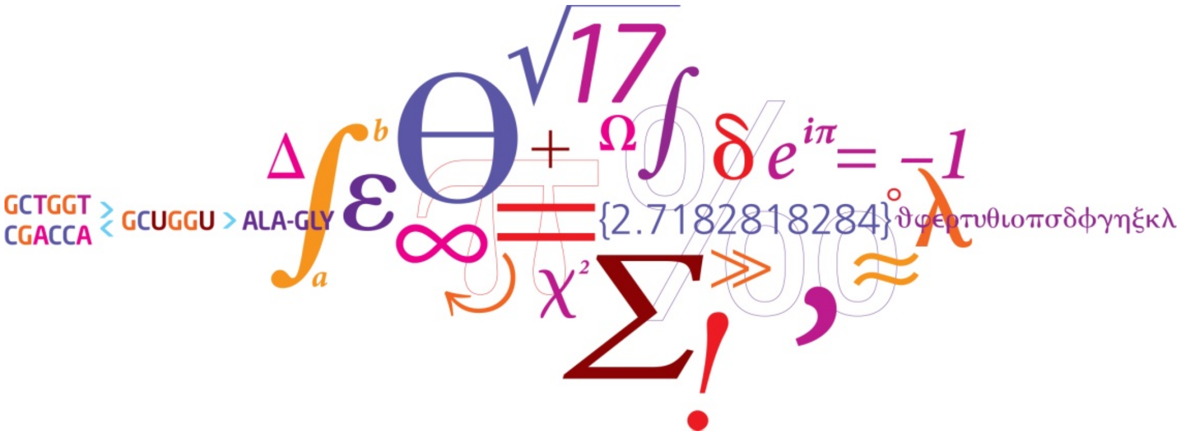
Paolo Marcatili



DTU Biosys
Department of Systems Biology

Linear B-cell epitope Prediction

Paolo Marcatili



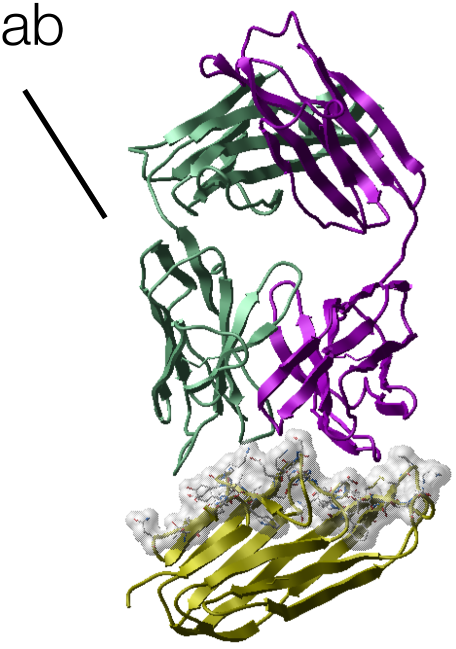
DTU Biosys
Department of Systems Biology

Outline

- What is a B-cell epitope?
- How can you predict B-cell epitopes?

What is a B-cell epitope?

- B-cell epitopes:
  - Accessible structural feature of a pathogen molecule.
  - Antibodies are developed to bind the epitope specifically using the complementary determining regions (CDRs).



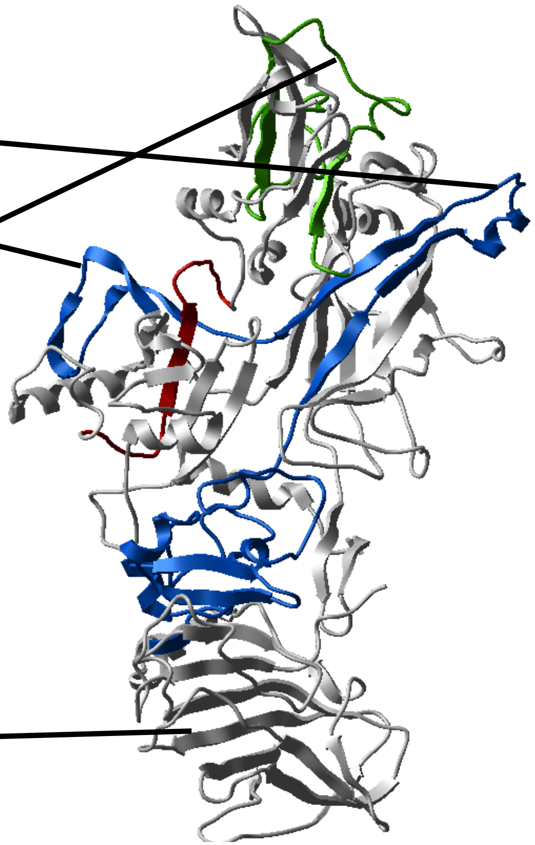
B-cell epitope classification

B-cell epitope: structural feature of a molecule or pathogen, accessible and recognizable by B-cell receptors and antibodies

Linear epitopes
One segment of the amino acid chain

Discontinuous epitope
(with linear determinant)

Discontinuous epitope
Several small segments brought into proximity by the protein fold



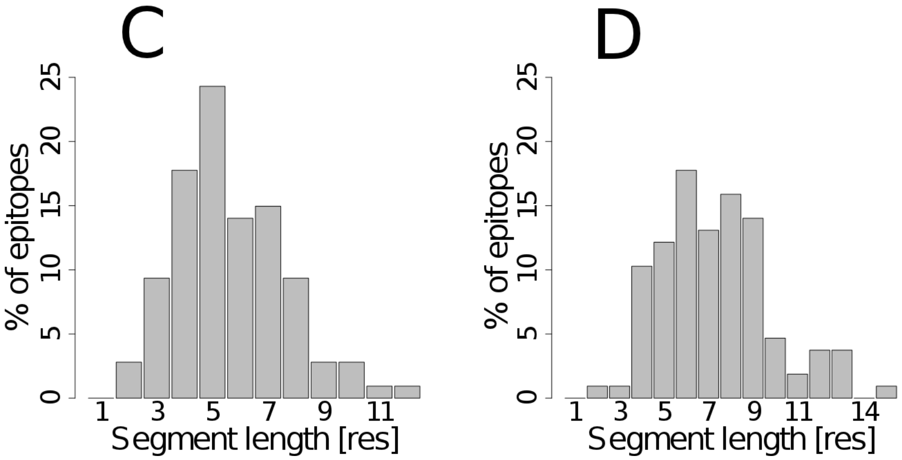
B-cell epitope annotation

- Linear epitopes:
  - Chop sequence into small pieces and measure binding to antibody
- Discontinuous epitopes:
  - Measure binding of whole protein to antibody
- The best annotation method : X-ray crystal structure of the antibody-epitope complex

B-cell epitope annotation

- Linear epitopes: 10%
  - Chop sequence into small pieces and measure binding to antibody
- Discontinuous epitopes: 90%
  - Measure binding of whole protein to antibody
- The best annotation method : X-ray crystal structure of the antibody-epitope complex

B-cell epitope annotation



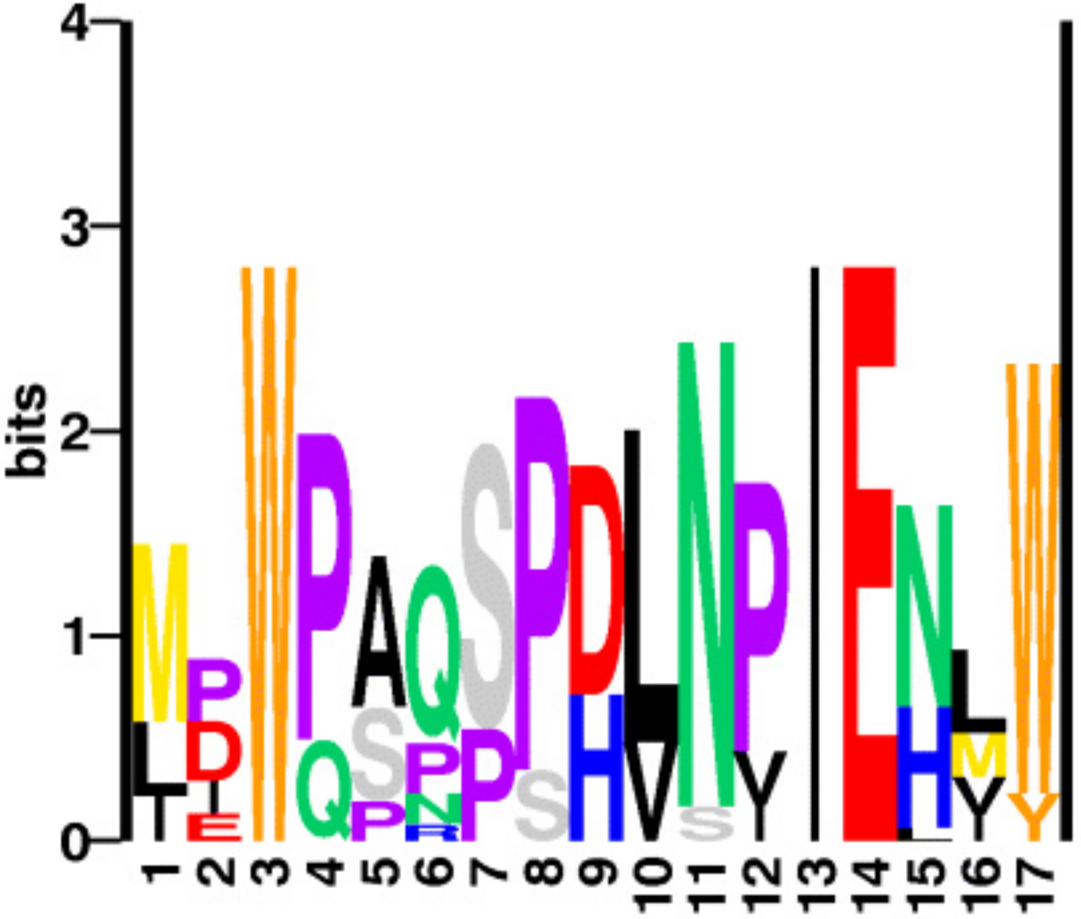
Longest linear stretch in epitope

- No epitope is purely linear
    - Epitopes contains linear determinants of 5 or more residues

B-cell epitope data bases

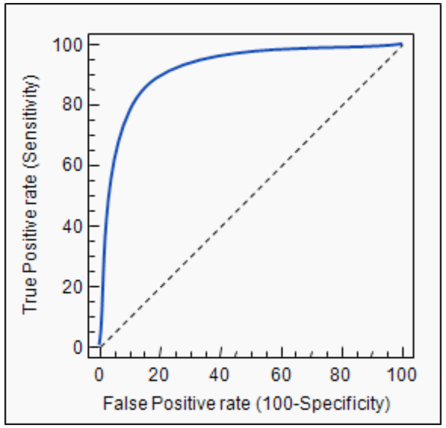
- Databases:
  - IEDB, Los Alamos HIV database, Protein Data Bank, AntiJen, BciPep
- Large amount of data available for linear epitopes
- Few data available for discontinuous

B cell epitope prediction



prediction tools

- Cytotoxic T cell epitope: (A_{ROC} ~ 0.9)
  - Will a given peptide bind to a given MHC class I molecule
- Helper T cell Epitope (A_{ROC} ~ 0.85)
  - Will a *part of* a peptide bind to a given MHC II molecule
- B cell epitope (A_{ROC} ~ 0.74)
  - Will a given part of a protein bind to one of the *billions of different* B Cell receptors



B-cell – prediction tools

- Sequence based prediction tools
  - Predominantly predicts linear epitopes
- Structure based epitopes
  - Predicts Conformational epitopes

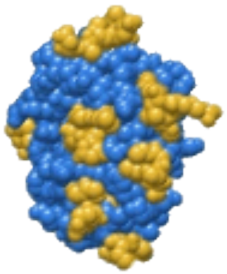
Sequence-based methods for prediction of linear epitopes

Input

TSQDLSVFPLASCCKDNIASTSVTLGCLVTGYLP
MSTTVTWDTGSLNKNVTTFPTTFHETYGLHSIV
SQVTASGKWAKQRFTCSVAHAESTAINKTFSAC
ALNFIPPTVKLFHSSCNPVGDTHTTIQLLCLISGY
VPGDMEVIWLVDGQKATNIFPYTAPGTKEGNVT
STHSELNITQGEWVSQKTYTCQVTYQGFTFKDE
ARKCSESDPRGVTSYLSPPSPL

Output

TSQDLSVFPLASCCKDNIASTSVTLGCLVTGYLP
MSTTVTWDTGSLNKNVTTFPTTFHETYGLHSIV
SQVTASGKWAKQRFTCSVAHAESTAINKTFSAC
ALNFIPPTVKLFHSSCNPVGDTHTTIQLLCLISGY
VPGDMEVIWLVDGQKATNIFPYTAPGTKEGNVT
STHSELNITQGEWVSQKTYTCQVTYQGFTFKDE
ARKCSESDPRGVTSYLSPPSPL



linear epitopes

- Protein hydrophobicity – hydrophilicity algorithms
  - Parker, Fauchere, Janin, Kyte and Doolittle, Manavalan
  - Sweet and Eisenberg, Goldman, Engelman and Steitz (GES), von Heijne
- Protein flexibility prediction algorithm
  - Karplus and Schulz
- Protein secondary structure prediction algorithms
  - PsiPred (D. Jones)

Idea

Epitopes are exposed regions

+

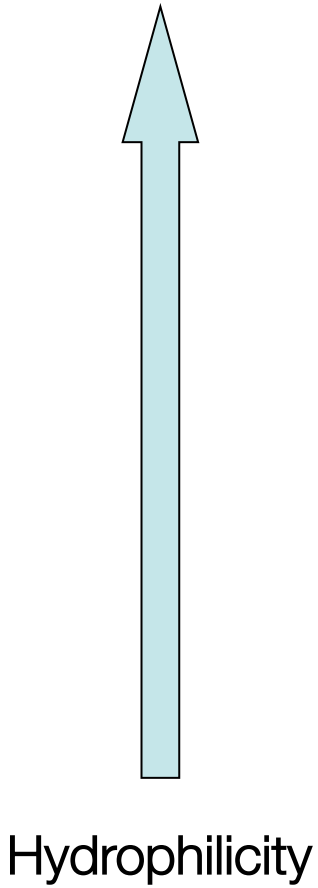
Hydrophilic residues are usually exposed

Propensity scales: The principle

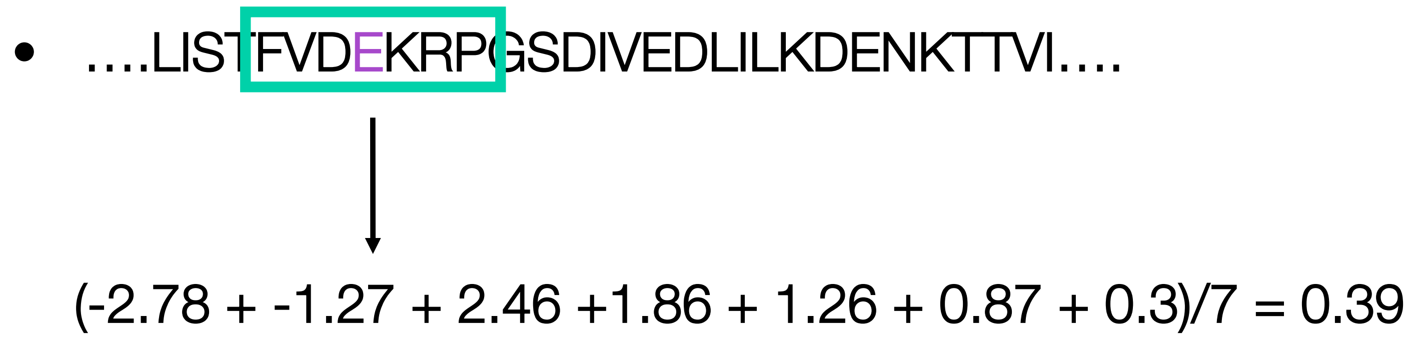
- The Parker hydrophilicity scale

- Derived from experimental data

D	2.46
E	1.86
N	1.64
S	1.50
Q	1.37
G	1.28
K	1.26
T	1.15
R	0.87
P	0.30
H	0.30
C	0.11
A	0.03
Y	-0.78
V	-1.27
M	-1.41
I	-2.45
F	-2.78
L	-2.87
W	-3.00



Propensity scales: The principle



- ....LISTFVDEKRPGSDIVEDLILKDENKTTVI....

(-2.78 + -1.27 + 2.46 +1.86 + 1.26 + 0.87 + 0.3)/7 = 0.39

Prediction scores:

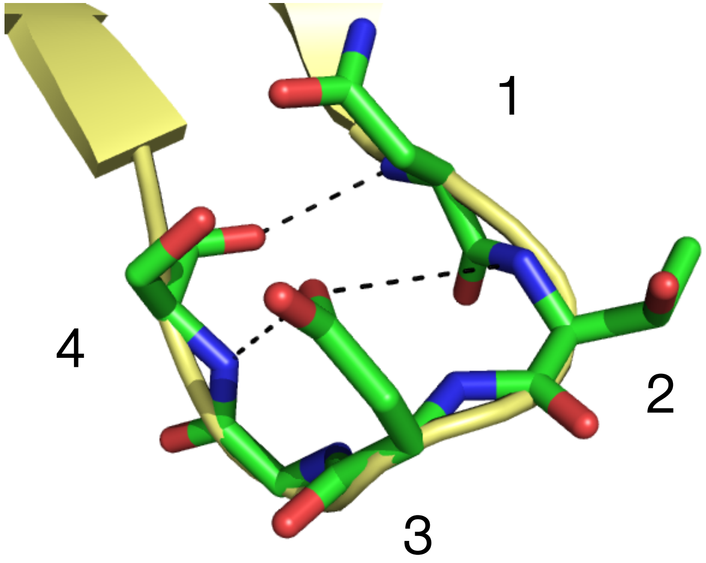
0.39 0.1 0.6 0.9 1.0 1.2 2.6 1.0 0.9 0.5 -0.5

Epitope

Turns and epitopes

- Pellequer found that 50% of the epitopes in a data set of 11 proteins were located in turns

Turn propensity scales for each position in the turn were used for epitope prediction.



Pellequer et al.,
Immunology letters, 1993

Blythe and Flower 2005

- Extensive evaluation of propensity scales for epitope prediction
- Conclusion:
  - Basically all the classical scales perform close to random!
  - Other methods must be used for epitope prediction

Blythe and Flower 2005

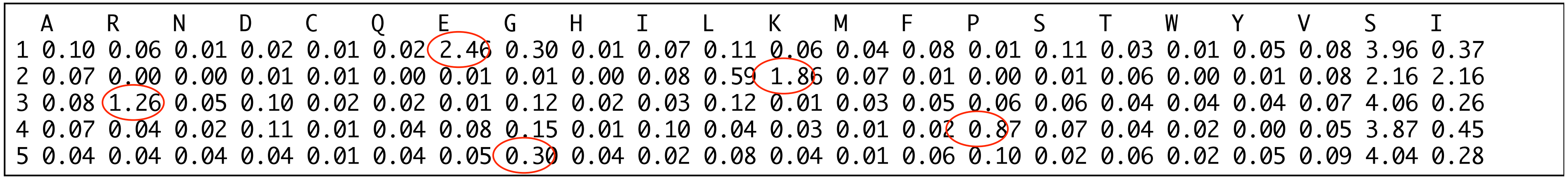
- Extensive evaluation of propensity scales for epitope prediction
- Conclusion:
  - Basically all the classical scales perform close to random!
  - Other methods must be used for epitope prediction

WHY?

BepiPred 1.0

- Parker hydrophilicity scale
- PSSM
- PSSM based on linear epitopes extracted from the AntiJen database
- Combination of the Parker prediction scores and PSSM leads to prediction score
- Tested on the Pellequer dataset and epitopes in the HIV Los Alamos database

PSSM



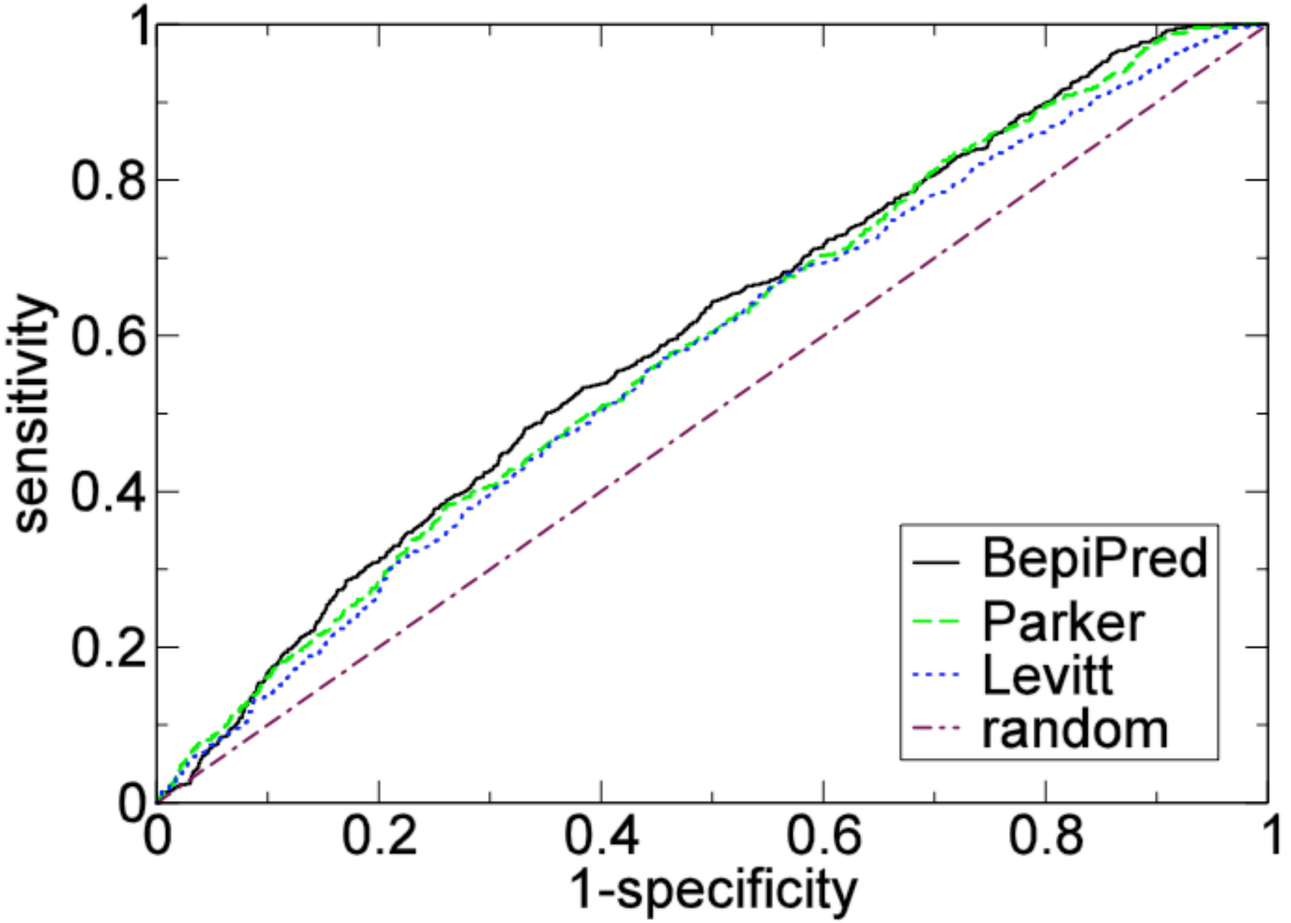
```
   A    R    N    D    C    Q    E    G    H    I    L    K    M    F    P    S    T    W    Y    V    S    I
1 0.10 0.06 0.01 0.02 0.01 0.02 2.46 0.30 0.01 0.07 0.11 0.06 0.04 0.08 0.01 0.11 0.03 0.01 0.05 0.08 3.96 0.37
2 0.07 0.00 0.00 0.01 0.01 0.00 0.01 0.01 0.00 0.08 0.59 1.86 0.07 0.01 0.00 0.01 0.06 0.00 0.01 0.08 2.16 2.16
3 0.08 1.26 0.05 0.10 0.02 0.02 0.01 0.12 0.02 0.03 0.12 0.01 0.03 0.05 0.06 0.06 0.04 0.04 0.04 0.07 4.06 0.26
4 0.07 0.04 0.02 0.11 0.01 0.04 0.08 0.15 0.01 0.10 0.04 0.03 0.01 0.02 0.87 0.07 0.04 0.02 0.00 0.05 3.87 0.45
5 0.04 0.04 0.04 0.04 0.01 0.04 0.05 0.30 0.04 0.02 0.08 0.04 0.01 0.06 0.10 0.02 0.06 0.02 0.05 0.09 4.04 0.28
```

- ….LISTFVDIEKRPGSDIVEDLILKDENKTTVI….

2.46+1.86+1.26+0.87+0.3 = 6.75 Prediction value

ROC evaluation

Evaluation on HIV Los Alamos data set



BepiPred performance

- Pellequer data set:
  - Levitt AROC = 0.66
  - Parker AROC = 0.65
  - BepiPred AROC = 0.68
- HIV Los Alamos data set
  - Levitt AROC = 0.57
  - Parker AROC = 0.59
  - BepiPred AROC = 0.60

Improving BepiPred

BepiPred conclusion:

- On both of the evaluation data sets, Bepipred was shown to perform better
- Still the AROC value is low compared to T-cell epitope prediction tools!

Dataset

675 Ag-Ab complexes from PDB (Ab specific hmm)

- resolution <3Å
- antigen > 60 residues
- no unnatural aa

antigen redundacy reduction: 70% seq id

170 cluster (165 training + cross fold, 5 final evaluation)

Training variables

For each antigen residue:

sequence +-4 aas

(encoded as AA volume, polarity and hydrophobicity)

Secondary structure (3 classes, sparse encoded)

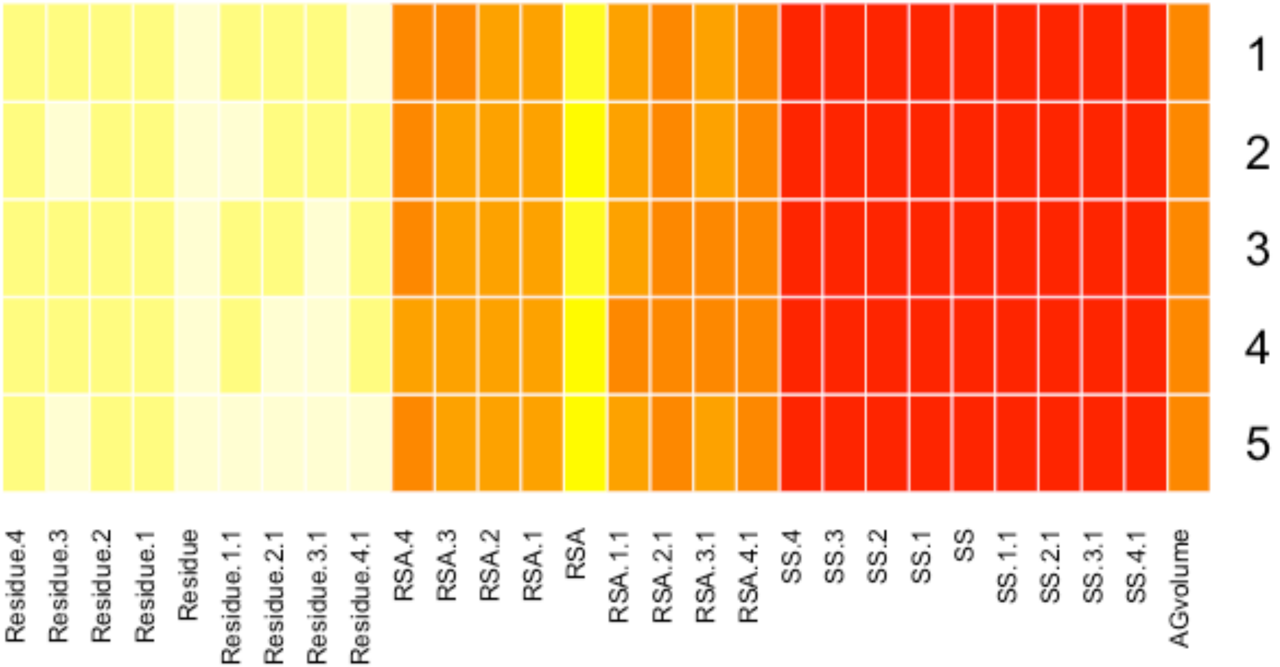
RSA ([0..1] values)

overall antigen volume

Variable Importance

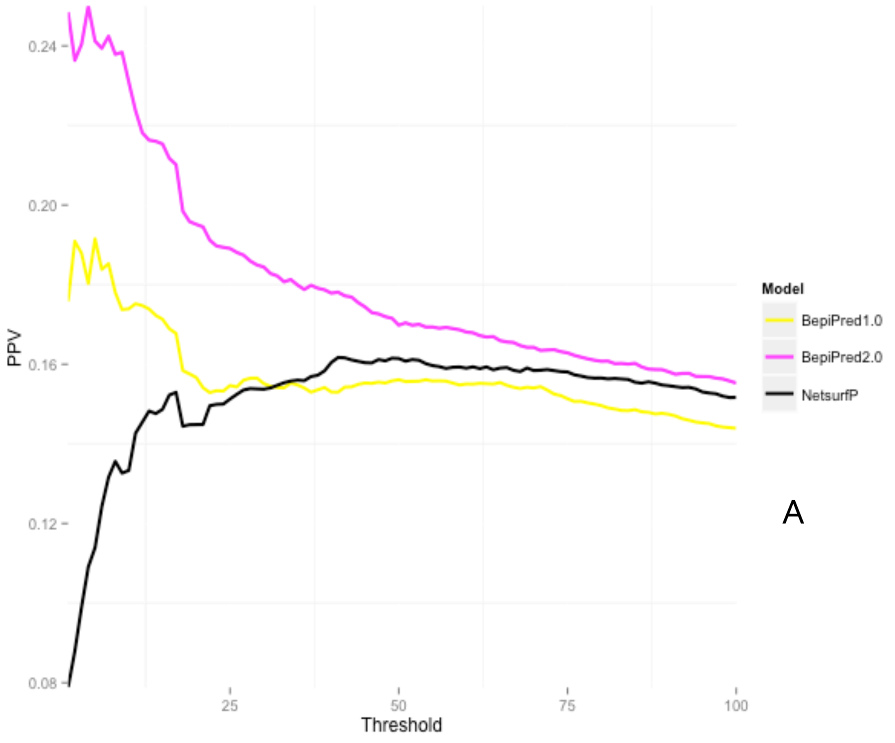
Gini Importance:

central residue > window > RSA of central residue
NB: **no threshold** on residue accessibility in negative dataset



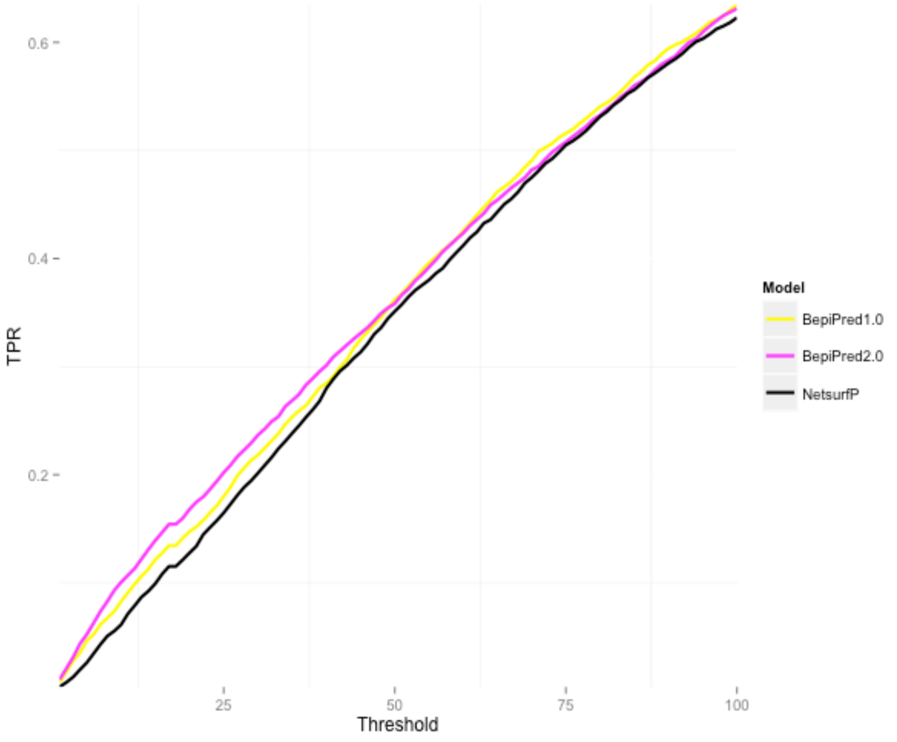
Evaluation on structural data

PPV



A

TPR



B

External structural validation

Results are comparable to the cross fold validation

	BEPIPRED 1.0	**BEPIPRED 2.0**	**BEPIPRED 1.0**	**BEPIPRED 2.0**
PDB ID	**AUC**	**AUC**	**AUC10%**	**AUC10%**
4WFF	0.678	**0.715**	0.169	**0.230**
4XAK	**0.739**	0.657	**0.183**	0.104
4Z5R	0.327	**0.576**	0.000	**0.038**
5BVP	0.525	**0.569**	0.082	**0.228**
5C0N	**0.596**	0.473	0.000	0.000
AVERAGE	0.573	**0.598**	0.088	**0.120**

Evaluation on IEDB dataset

Epitopes mapped on proteins

BEPIPRED 1.0	**0.562**	**0.082**
BEPIPRED 2.0	0.573	0.084
P VALUE	$< 1 \cdot 10^{-6}$	0.052

Evaluation on IEDB dataset

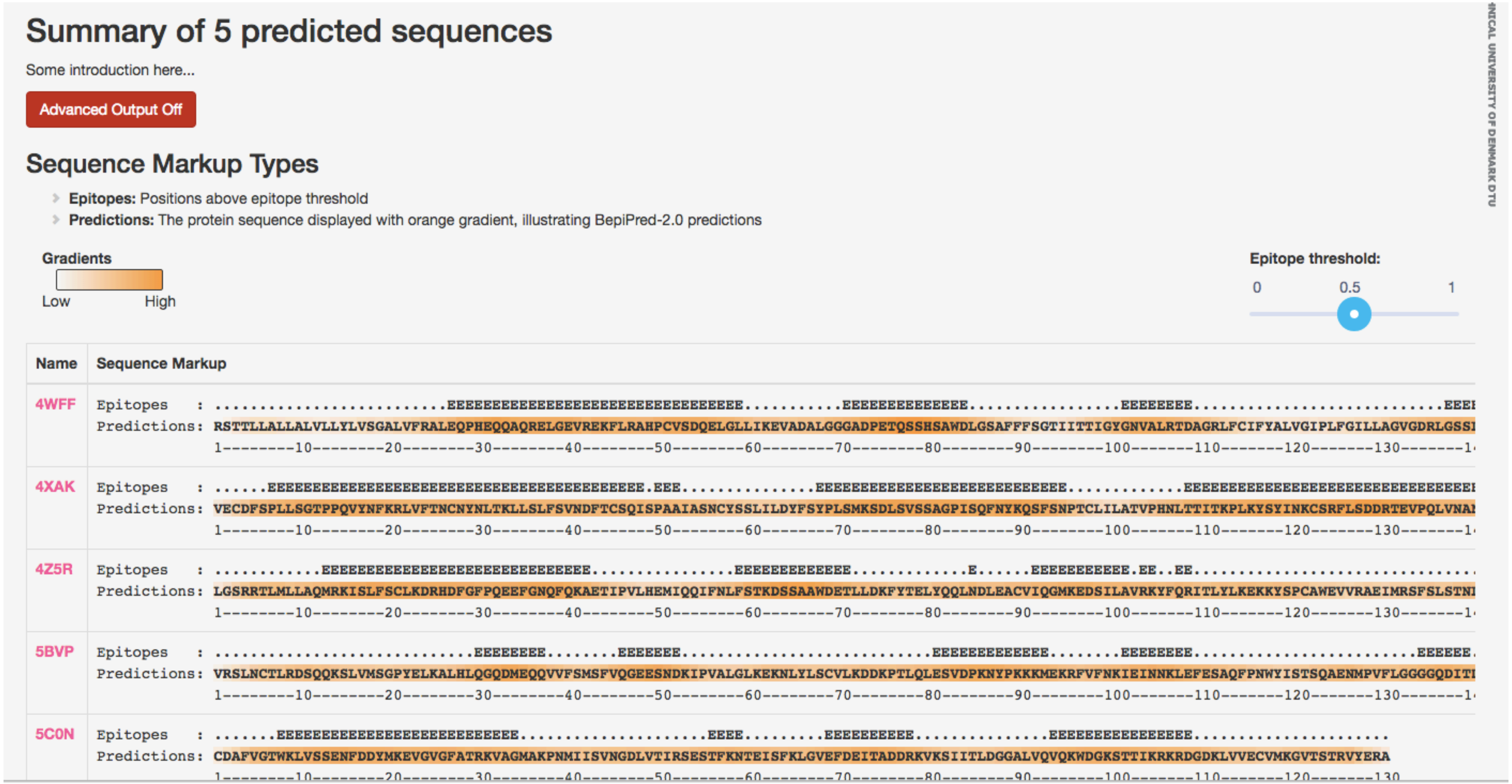
Epitopes mapped on proteins

BEPIPRED 1.0	**0.562**	**0.082**
BEPIPRED 2.0	0.573	0.084
P VALUE	$< 1 \cdot 10^{-6}$	0.052

100 aa window centered on the epitope

	AUC	**AUC10%**
BEPIPRED 1.0	0.540	0.104
BEPIPRED 2.0	0.547	0.114
P VALUE	$< 1 \cdot 10^{-6}$	$< 1 \cdot 10^{-6}$

Web server



Summary of 5 predicted sequences

Some introduction here...

Advanced Output Off

Sequence Markup Types

- **Epitopes:** Positions above epitope threshold
- **Predictions:** The protein sequence displayed with orange gradient, illustrating BepiPred-2.0 predictions

Gradients

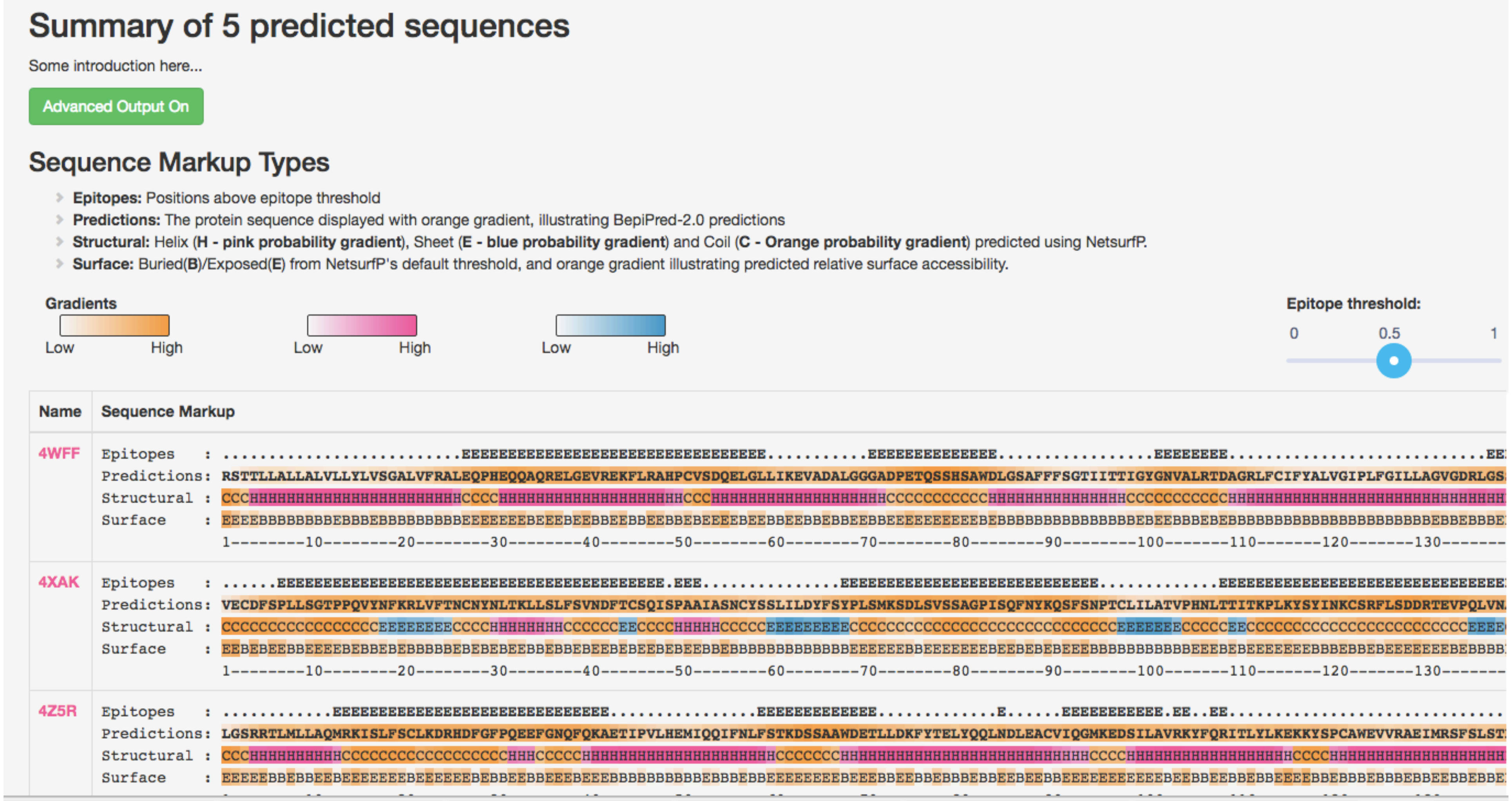
Low High

Epitope threshold:

0 0.5 1

Name	Sequence Markup
4WFF	Epitopes : ..........................EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE...........EEEEEEEEEEEEEE.................EEEEEEEE............................EEE Predictions: RSTTLLALLALVLLYLVSGALVFRALEQPHEQQAQRELGEVREKFLRAHPCVSDQELGLLIKEVADALGGGADPETQSSHSAWDLGSAFFFSGTIITTIGYGNVALRTDAGRLFCIFYALVGIPLFGILLAGVGDRLGSS 1--------10--------20--------30--------40--------50--------60--------70--------80--------90--------100-------110-------120-------130-------1
4XAK	Epitopes : ......EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE.EEE...............EEEEEEEEEEEEEEEEEEEEEEEEEEEE.............EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE Predictions: VECDFSPLLSGTPPQVYNFKRLVFTNCNYNLTKLLSLFSVNDFTCSQISPAAIASNCYSSLILDYFSYPLSMKSDLSVSSAGPISQFNYKQSFSNPTCLILATVPHNLTTITKPLKYSYINKCSRFLSDDRTEVPQLVNA 1--------10--------20--------30--------40--------50--------60--------70--------80--------90--------100-------110-------120-------130-------1
4Z5R	Epitopes : ............EEEEEEEEEEEEEEEEEEEEEEEEEEEEEE................EEEEEEEEEEEEE.............E......EEEEEEEEEEE.EE..EE............................... Predictions: LGSRRTLMLLAQMRKISLFSCLKDRHDFGFPQEEFGNQFQKAETIPVLHEMIQQIFNLFSTKDSSAAWDETLLDKFYTELYQQLNDLEACVIQGMKEDSILAVRKYFQRITLYLKEKKYSPCAWEVVRAEIMRSFSLSTN 1--------10--------20--------30--------40--------50--------60--------70--------80--------90--------100-------110-------120-------130-------1
5BVP	Epitopes : .............................EEEEEEEE........EEEEEEE............................EEEEEEEEEEEEE........EEEEEEEE.........................EEEEEE Predictions: VRSLNCTLRDSQQKSLVMSGPYELKALHLQGQDMEQQVVFSMSFVQGEESNDKIPVALGLKEKNLYLSCVLKDDKPTLQLESVDPKNYPKKKMEKRFVFNKIEINNKLEFESAQFPNWYISTSQAENMPVFLGGGGQDIT 1--------10--------20--------30--------40--------50--------60--------70--------80--------90--------100-------110-------120-------130-------1
5CON	Epitopes : .......EEEEEEEEEEEEEEEEEEEEEEEEEEE.....................EEEE.........EEEEEEEEEE...............EEEEEEEEEEEEEEEE..................... Predictions: CDAFVGTWKLVSSENFDDYMKEVGVGFATRKVAGMAKPNMIISVNGDLVTIRSESTFKNTEISFKLGVEFDEITADDRKVKSIITLDGGALVQVQKWDGKSTTIKRKRDGDKLVVECVMKGVTSTRVYERA 1--------10--------20--------30--------40--------50--------60--------70--------80--------90--------100-------110-------120-------130

Web server



Summary of 5 predicted sequences

Some introduction here...

Advanced Output On

Sequence Markup Types

- **Epitopes:** Positions above epitope threshold
- **Predictions:** The protein sequence displayed with orange gradient, illustrating BepiPred-2.0 predictions
- **Structural:** Helix (**H - pink probability gradient**), Sheet (**E - blue probability gradient**) and Coil (**C - Orange probability gradient**) predicted using NetsurfP.
- **Surface:** Buried(**B**)/Exposed(**E**) from NetsurfP's default threshold, and orange gradient illustrating predicted relative surface accessibility.

Gradients

Low High

Low High

Low High

Epitope threshold:

0 0.5 1

Name	Sequence Markup

4WFF	`Epitopes    : .........................EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE...........EEEEEEEEEEEEEE................EEEEEEEE...........................EE`
	`Predictions: RSTTLLALLALVLLYLVSGALVFRALEQPHEQQAQRELGEVREKFLRAHPCVSDQELGLLIKEVADALGGGADPETQSSHSAWDLGSAFFFSGTIITTIGYGNVALRTDAGRLFCIFYALVGIPLFGILLAGVGDRLGS`
	`Structural  : CCCHHHHHHHHHHHHHHHHHHHHHHHHCCCCHHHHHHHHHHHHHHHHHHHHHCCCHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCHHHHHHHHHHHHHHHHCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH`
	`Surface     : EEEBBBBBBBBEBBBBEBBBBBBBBBBBEEEEEEEEBEEEBEEBBEEBBEEBBEBEEEEBEEBBEEBBEBBEEBBEEBBEEBBEEEEEEEEEEBEBBBBBBBBBBBBBBBBEBEEBBBEBEBBBBBBBBBBBBBBBBBBBBBBBBBEBBEBBBE`
	`              1--------10--------20--------30--------40--------50--------60--------70--------80--------90--------100-------110-------120-------130-------`

4XAK	`Epitopes    : ......EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE.EEE...............EEEEEEEEEEEEEEEEEEEEEEEEEEEE............EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE`
	`Predictions: VECDFSPLLSGTPPQVYNFKRLVFTNCNYNLTKLLSLFSVNDFTCSQISPAAIASNCYSSLILDYFSYPLSMKSDLSVSSAGPISQFNYKQSFSNPTCLILATVPHNLTTITKPLKYSYINKCSRFLSDDRTEVPQLVN`
	`Structural  : CCCCCCCCCCCCCCCCCEEEEEEEECCCCHHHHHHHHCCCCCCEECCCCHHHHHCCCCCEEEEEEEEECCCCCCCCCCCCCCCCCCCCCCCCCCCCCEEEEEEECCCCCEECCCCCCCCCCCCCCCCCCCCCCCCEEEE`
	`Surface     : EEBEBEEBBEEEEBEBBEBEBBBBBEBEBEBEEBBEBBEBEEBEBEEBEBEEBBEBBBBBBBBBBBBEEEEEEBBEEEEEEEEBEEBEBEBEEEBBBBBBBBBBBBEEEEBEBEEEEEEEBBBEBBEBEEEEEEEEEBEBBBB`
	`              1--------10--------20--------30--------40--------50--------60--------70--------80--------90--------100-------110-------120-------130-------`

4Z5R	`Epitopes    : ............EEEEEEEEEEEEEEEEEEEEEEEEEEEEEE...............EEEEEEEEEEEEE.............E......EEEEEEEEEEE.EE..EE............................`
	`Predictions: LGSRRTLMLLAQMRKISLFSCLKDRHDFGFPQEEFGNQFQKAETIPVLHEMIQQIFNLFSTKDSSAAWDETLLDKFYTELYQQLNDLEACVIQGMKEDSILAVRKYFQRITLYLKEKKYSPCAWEVVRAEIMRSFSLST`
	`Structural  : CCCHHHHHHHHHHCCCCCCCCCCCCCCCCCCHHHCCCCCHHHHHHHHHHHHHHHHHHHHHCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCHHHHHHHHHHHHHHHHHHCCCCHHHHHHHHHHHHHHHHHHH`
	`Surface     : EEEEEBBEBBEEBEEEEEEEEBEEEEEEEBEBBEEBBEEEEBEEEBBBBBBBBBBBEBBBEBBEEEEEEEEEBEEEEBBEBEEEBBEEBBEBBBEBBEEBEEBBEEEEEEEEEEEEBEEBBEEBBEBBEEEEBBEBBBEBBBE`

Prediction of linear epitopes

- Pro
- easily predicted computationally
- easily identified experimentally
- immunodominant epitopes in many cases
- do not need 3D structural information
- easy to produce and check binding activity experimentally

Con

- only ~10% of epitopes can be classified as "linear"
- weakly immunogenic in most cases
- most epitope peptides do not provide antigen-neutralizing immunity
- in many cases represent hypervariable regions