

Weight matrices, Sequence motifs, information content, and sequence logos

Morten Nielsen,

Instituto de Investigaciones
Biotecnológicas, Universidad de San
Martín, Argentina,
and

Department of Health Technology, DTU,
Denmark

Why weight matrices?

- The vast majority of biological motifs are characterized by a linear motif
 - Post translational modifications
 - Signal peptides
 - T cell epitopes
 - Transcription binding sites
 - SH2/SH3 domain binding
 - MHC binding
 -
 - Predict impact of sequence variation (SNP)
 - ...
-

Identifying binding motifs (SH3)

Peptide

LMLSLFEQSLSCQAQ

QGTDATKSIIFEALER

RLEEAQAYLAAGQHD

EISELRTKVQEQQKQ

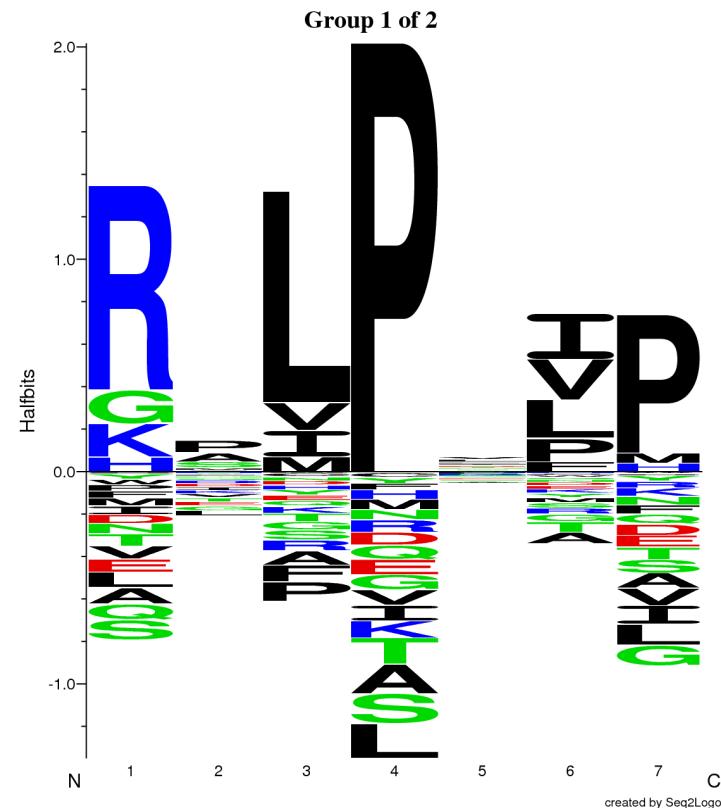
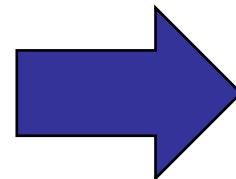
FAGAKKIFGSLAFLP

VRASSRVSGSFPEDS

CKAFFKRSIQGHNDY

CEGCKAFFKRSIQGH

RLSEADIRGFVAAVV

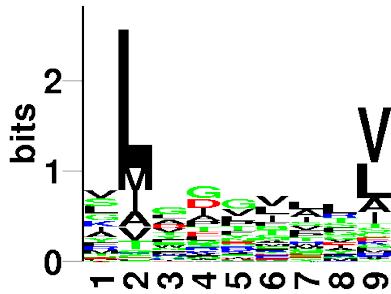


Objectives

- Understand the concepts of weight matrix construction
 - One of the most important methods of bioinformatics
- Visualization of binding motifs
 - Construction of sequence logos
- How to construct a weight matrix
- How to use weight matrices to characterize receptor-ligand interactions
- Case story from the MHC-peptide interactions guiding immune system reactions

Bioinformatics in a nutshell

List of peptides that have a given biological feature

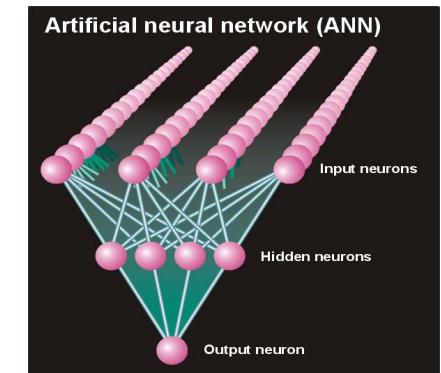


Mathematical model (neural network, hidden Markov model)



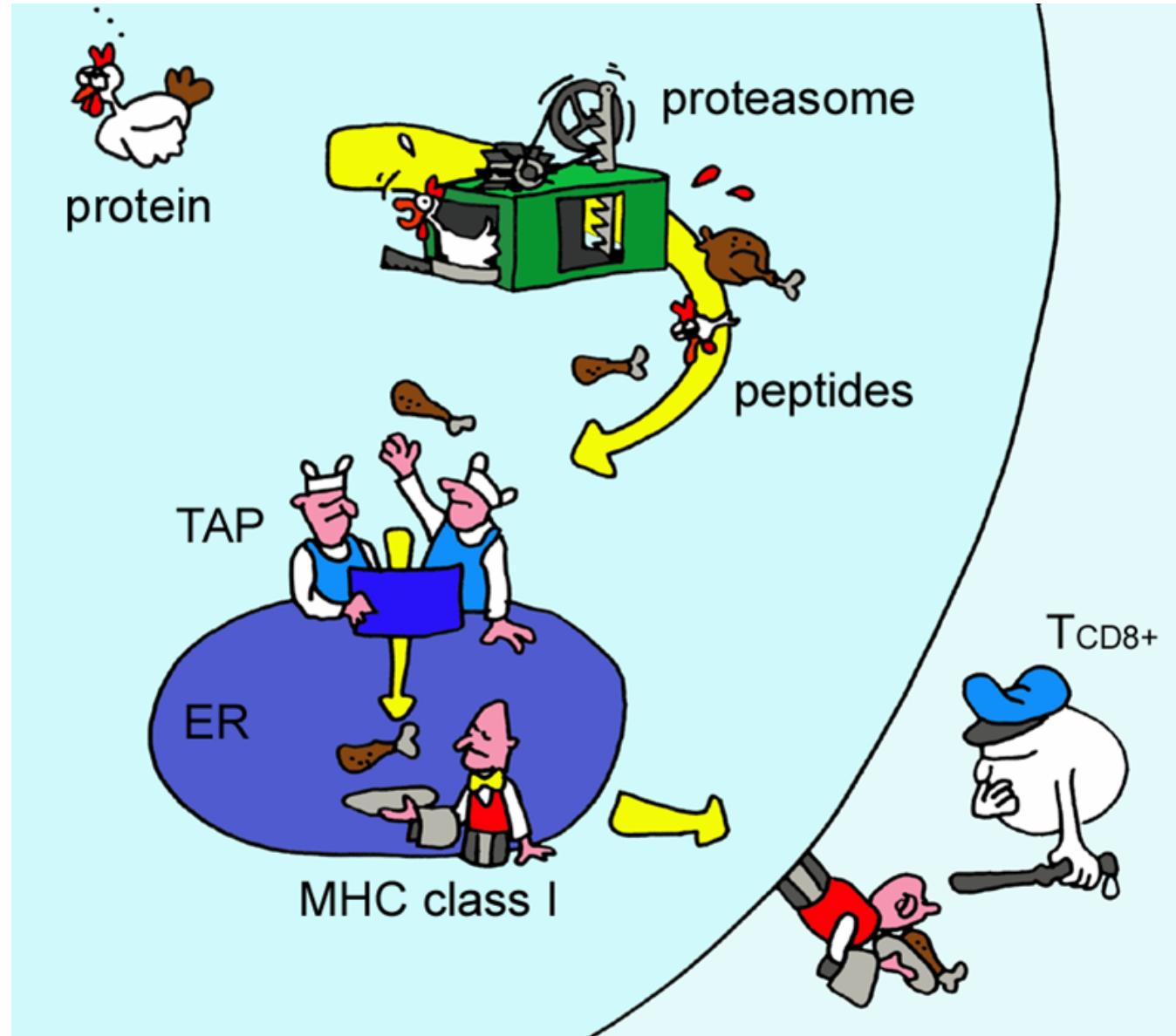
Search databases for other biological sequences with the same feature/property

YMNNGTMSQV
GILGFVFTL
ALWGFFPVV
IILKEPVHGV
IILGFVFTLT
LLFGYPVYV
GLSPTVWLS
WLSSLVPFV
FLPSPDFFPS
CVGGLLTMV
FIAGNSAYE



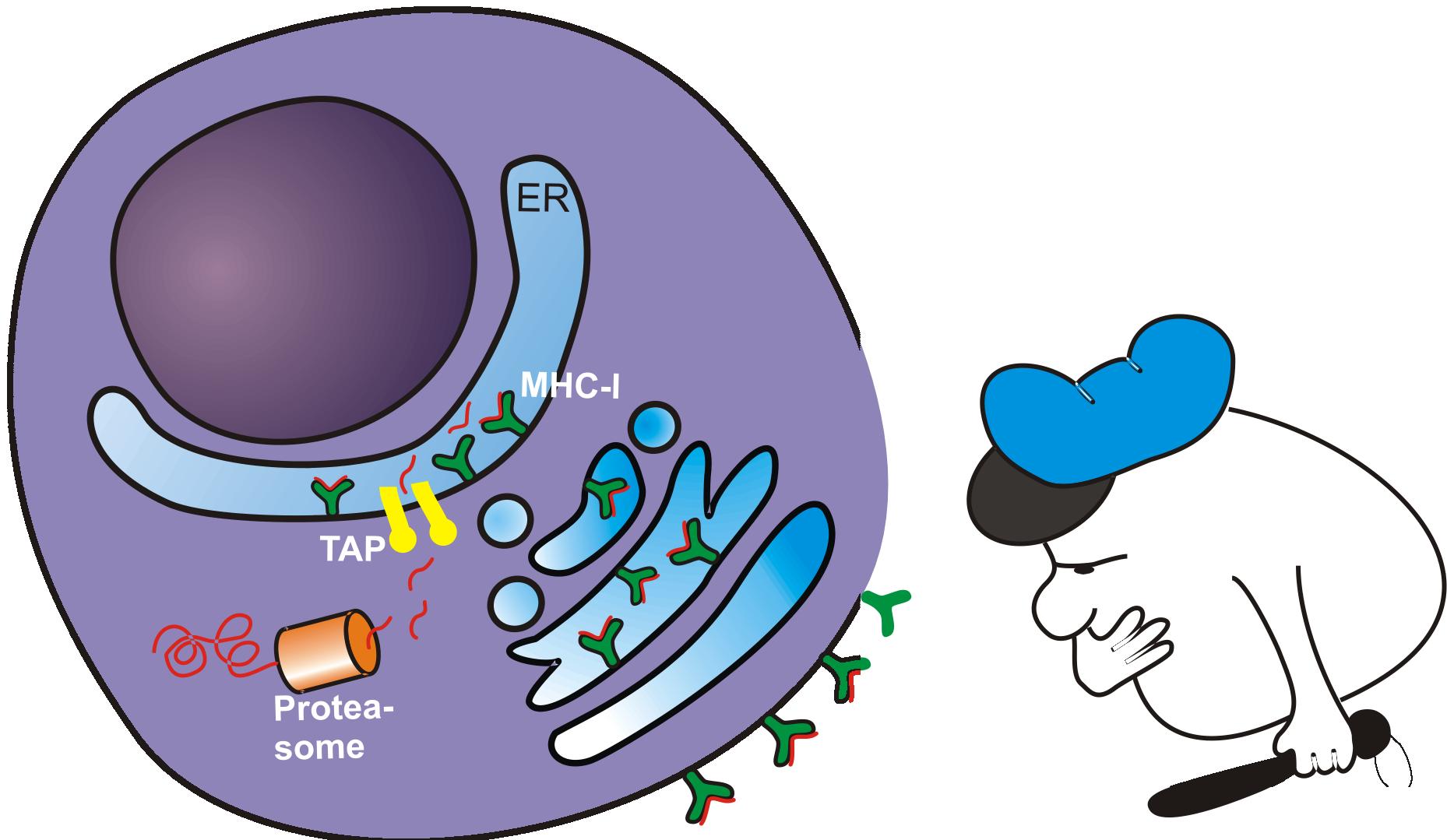
>polymerase"
MERIKELRDLMSQSRTREILTAKTTVD HMAII KKYT SGRQE KNPAL RMKWM MAMKY PITA D
KRIMEMIPERNEQQGTLWSKTNDAKS DRVMV SPLA VTWMN RNGPT TSTVH YPKVY KTYF E
KVERLKKGHTGPVHFRNQVKIRRRVD INPGH ADLS AKEAQ DVIME VVFPN EVGAR ILTS E
SQLTTITKEKKKEELQDKCIAPLMVAYNLEREL VRKT RFLPV AGGTS SVYIE VLHLT QOTC W
EQMYTPGGEVNRDDVDQSLIIAARNI VRRAT VSAD PLASL LEMCH STQIG GIRMV DILR Q
NPTEEQAVDICKAAMGLRRISSSFSGF GFTFK RTNG SSVKK EEEVL TGNLQ TLKIK VHEG Y
EEPTMVGRRAILRKATRRLIQIQLIV SGRDE QSIA EAIIV AMVFS QEDCM IKAVR GDLN F
...

MHC Class I pathway



MHC-I molecules present peptides on the surface of most cells

CENTERFO
R BIOLOGI
CAL SEQU
ENCEANA
LYSIS CBS



CTL response

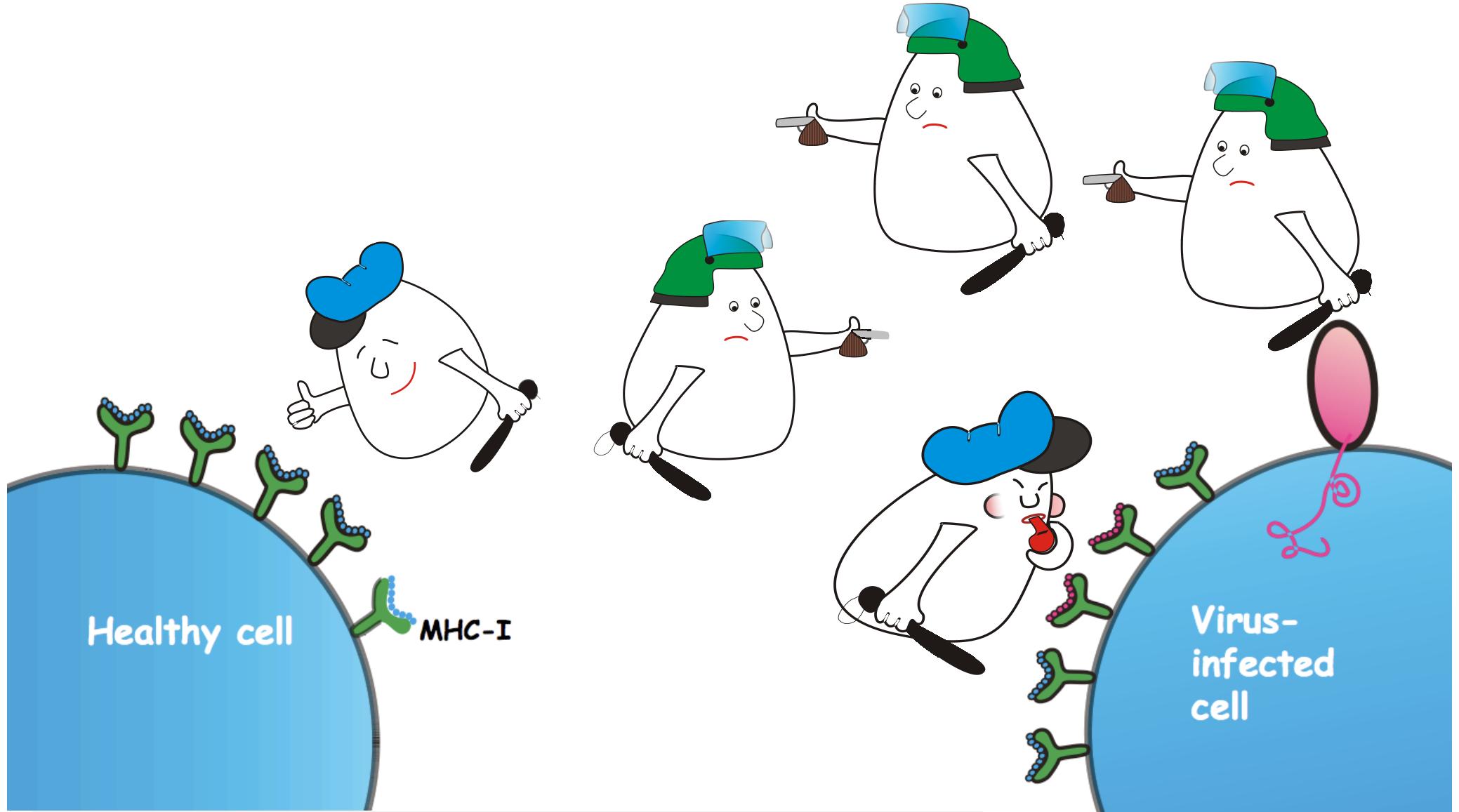
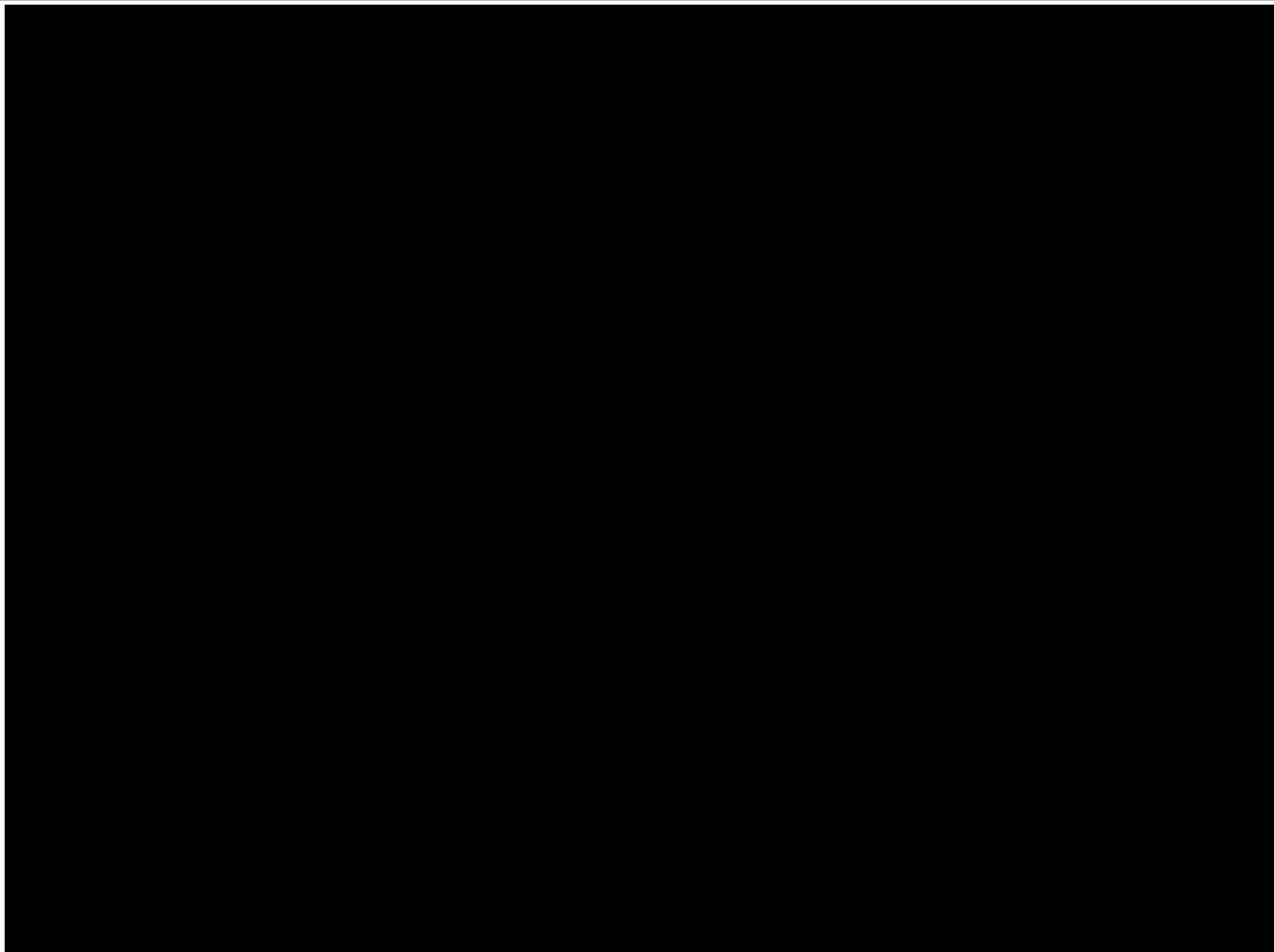


Figure courtesy Mette Voldby Larsen

Encounter with death



Sequence Information

- Say that a peptide must have L at P₂ in order to bind, and that A,F,W, and Y are found at P₁. Which position has most information?
- How many questions do I need to ask to tell if a peptide binds looking at only P₁ or P₂?

ALAKAAAAM
FLAKAAAAN
WLAKAAAAR
ALAKAAAAT
YLAKAAAAV
FLNERPILT
WLLGFVFTM
YLNAWVKVV
ALNEPVLLL
...
...
WLVPFIVSV

Sequence Information

- Say that a peptide must have L at P_2 in order to bind, and that A,F,W, and Y are found at P_1 . Which position has most information?
- How many questions do I need to ask to tell if a peptide binds looking at only P_1 or P_2 ?
 - P_1 : 4 questions (at most)
 - P_2 : 1 question (L or not)
 - P_2 has the most information

Sequence Information

- Say that a peptide must have L at P_2 in order to bind, and that A,F,W, and Y are found at P_1 . Which position has most information?
- How many questions do I need to ask to tell if a peptide binds looking at only P_1 or P_2 ?
- P_1 : 4 questions (at most)
- P_2 : 1 question (L or not)
- P_2 has the most information

- Calculate p_a at each position
- Entropy

$$S = - \sum_a p_a \log(p_a)$$

- Information content

$$I = \log(20) + \sum_a p_a \cdot \log(p_a)$$

or

$$I = \sum_a p_a \cdot \log\left(\frac{p_a}{q_a}\right)$$

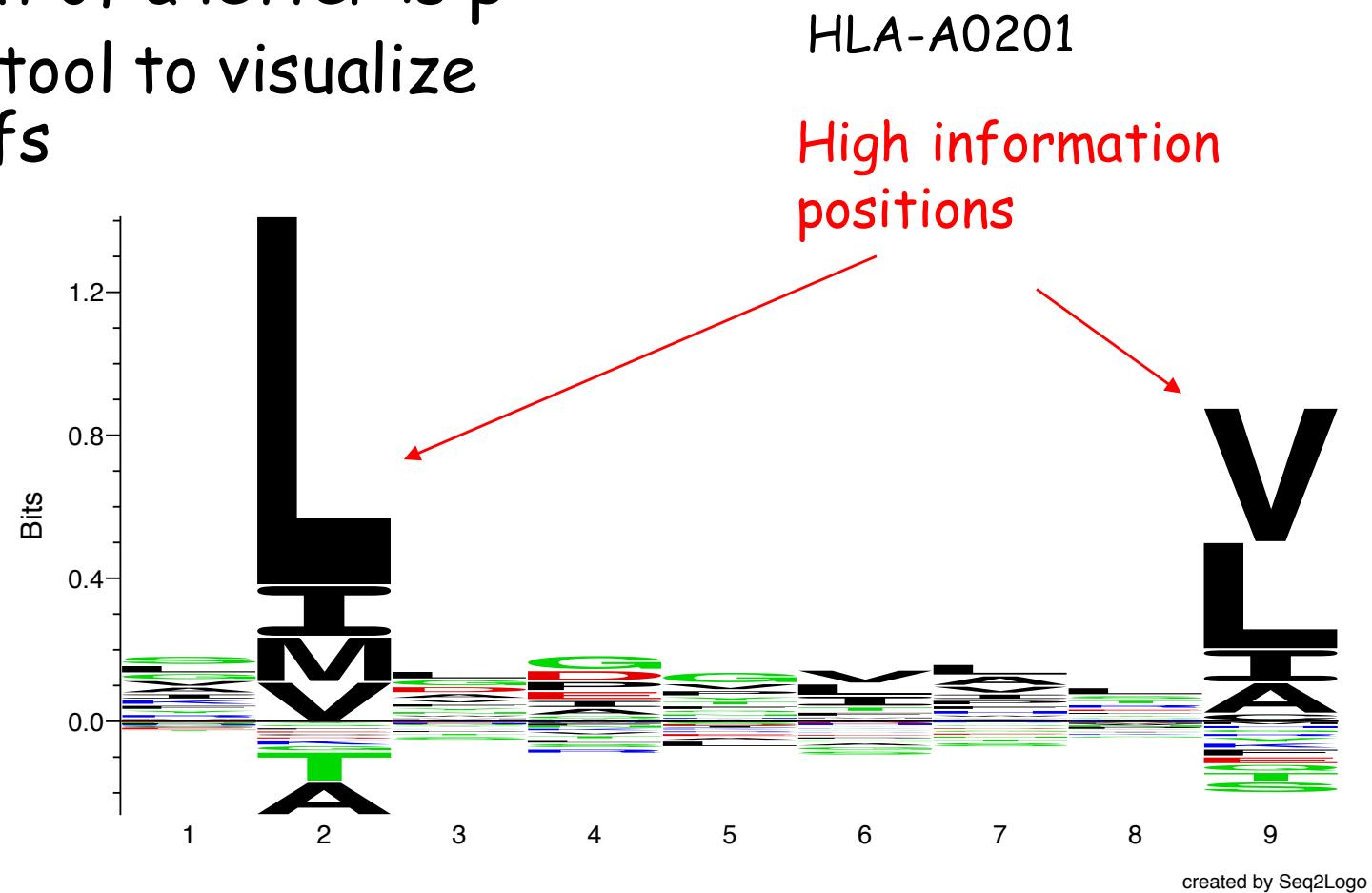
- Conserved positions
 - $P_L=1, P_{\neq L}=0 \Rightarrow S=0, I=\log(20)$
- Mutable positions
 - $P_{aa}=1/20 \Rightarrow S=\log(20), I=0$

Sequence information

SLLPAIVEL YLLPAIVHI TLWVDPYEV GLVPFLVSV KLLEPVLLL LLDVPTAAV LLDVPTAAV LLDVPTAAV
LLDVPTAAV VLFRGGPRG MVDGTLLLL YMNGTMSQV MLLSVPLLL SLLGLLVEV ALLPPINIL TLIKIQHTL
HLIDYLVTS ILAPPVVKL ALFPQLVIL GILGFVFTL STNRQSGRQ GLDVLTAKV RILGAVAKV QVCERIPTI
ILFGHENRV ILMEMHIHKL ILDQKINEV SLAGGIIGV LLIENVASL FLLWATAEA SLPDFGISY KKREEAPSL
LERPGGNEI ALSNLEVKL ALNELLQHV DLERKVESL FLGENISNF ALSDHHIYL GLSEFTEYL STAPPAHGV
PLDGEYFTL GVLVGVALI RTLDKVLEV HLSTAFARV RLDSYVRSL YMNGTMSQV GILGFVFTL ILKEPVHGV
ILGFVFTLT LLFGYPVYV GLSPTVWLS WLSLLVPFV FLPSDFFPS CLGGLLTMV FIAGNSAYE KLGEFYNQM
KLVALGINA DLMGYIPLV RLVTLKDIV MLLAVLYCL AAGIGILTV YLEPGPVTA LLDGTATLR ITDQVPFSV
KTWGQYWQV TITDQVPFS AFHHVAREL YLNKIQNSL MMRKLAILS AIMDKNIIL IMDKNIILK SMVGNWAKV
SLLAPGAKQ KIFGSLAFL ELVSEFSRM KLTPLCVTL VLYRYGSFS YIGEVLVSV CINGVCWTW VMNILLQYV
ILTVELGVL KVLEYVIKV FLWGPRALV GLSRYVARL FLLTRILTI HLGNVKYLV GIAGGLALL GLQDCTMLV
TGAPVTYST VIYQYMDDL VLPDVFIRC VLPDVFIRC AVGIGIAVV LVVLGLLAV ALGLGLLPV GIGIGVLA
GAGIGVAVL IAGIGILAI LIVIGILIL LAGIGLIAA VDGIGILTI GAGIGVLTA AAGIGIIQI QAGIGILLA
KARDPHSGH KACDPHSGH ACDPHSGHF SLYNTVATL RGPGRAFVT NLVPMVATV GLHCYEQLV PLKQHFQIV
AVFDRKSDA LLDFVRFMG VLVKSPNHV GLAPPQHLLI LLGRNSFEV PLTFGWCYK VLEWRFDSR TLNAWKVV
GLCTLVAML FIDSYICQV IISAVVGIL VMAGVGSPY LLWTLVVLL SVRDRLARL LLMDCSGSI CLTSTVQLV
VLHDDLLEA LMWITQCFL SLLMWITQC QLSLLMWIT LLGATCMFV RLTRFLSRV YMDGTMSQV FLTPKKLQC
ISNDVCAQV VKTDGNPPE SVYDFFFWL FLYGALLA VLFSSDFRI LMWAKIGPV SLLLELEEV SLSRFWSWGA
YTAFTIPSI RLMKQDFSV RLPRIFCSC FLWGPRAYA RLLQETELV SLFEGIDFY SLDQSVVEL RLNMFTPYI
NMFTPYIGV LMIPIPLINV TLFIGSHVV SLVIVTTFV VLQWASLAV ILAKFLHWL STAPPHVNV LLLLTVLTV
VVLGVVFGI ILHNGAYSL MIMVKCWMI MLGHTMEV MLGHTMEV SLADTNSLA LLWAARPRL GVALQTMKQ
GLYDGMEHL KMVELVHFL YLQLVFGIE MLMAQEALA LMAQEALAF VYDGREHTV YLSGANLNL RMFPNAPYL
EAAGIGILT TLDSQVMSL STPPPGRTRV KVAELVHFL IMIGVLVGV ALCRWGLLL LLFAGVQCQ VLLCESTAV
YLSTAFARV YLLEMWLRL SLDDYNHLV RTLDKVLEV GLPVEYLQV KLIANNTRV FIYAGSLSA KLVANNTRL
FLDEFMEGV ALQPGBTALL VLDGLDVLL SLYSFPEPE ALYVDSLFF SLLQHLLIGL ELTLGEFLK MINAYLDKL
AAGIGILTV FLPSDFFPS SVRDRLARL SLREWLLRI LLSAWILTA AAGIGILTV AVPDEIPPL FAYDGKDYL
AAGIGILTV FLPSDFFPS AAGIGILTV FLPSDFFPS AAGIGILTV FLWGPRALV ETVSEQSNV ITLWQRPLV

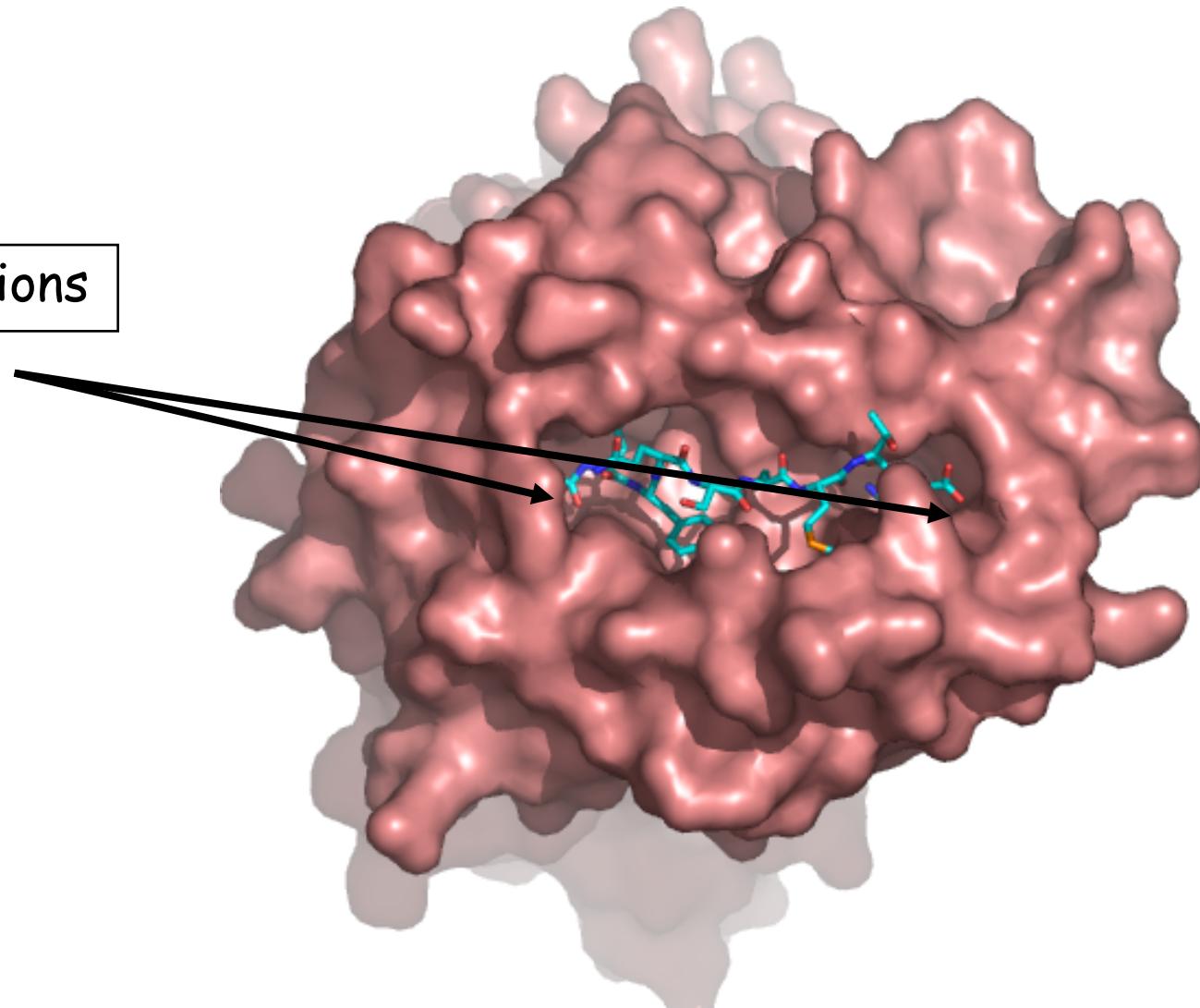
Sequence logos

- Height of a column equal to I
- Relative height of a letter is p
- Highly useful tool to visualize sequence motifs



Binding Motif. MHC class I with peptide

Anchor positions



An example!!
(See handout)

Cost of a motif characterization

- 200 peptides needed
 - 50-200 \$ per peptide = 10,000 - 40,000 \$
 - 1 PhD student manpower
- 2000 MHC class I molecules
 - So do the math your self ...

Characterizing a binding motif from small data sets

CENTER FOR BIOLOGICAL
SEQUENCE ANALYSIS CBS

10 MHC restricted peptides

ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV

What can we learn?

1. A at P1 favors binding?
2. I is not allowed at P9?
3. K at P4 favors binding?
4. Which positions are important for binding?

Simple motifs

Yes/No rules

10 MHC restricted peptides

ALAKAAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV

$[AGTK]_1 [LMIV]_2 [ANLV]_3 \dots [MNRTVL]_9$

- Only 11 of 212 peptides identified!
- Need more flexible rules
 - If not fit P1 but fit P2 then ok
- Not all positions are equally important
 - We know that P2 and P9 determines binding more than other positions
- Cannot discriminate between good and very good binders

Simple motifs

Yes/No rules

10 MHC restricted peptides

ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV

[AGTK]₁[LMIV]₂[ANLV]₃...[AIFKLV]₇...[MNRTVL]₉

- Example

RLLDDTPEV 84 nM

GLLGNVSTV 23 nM

ALAKAAAAL 309 nM

- Two first peptides will not fit the motif.
They are all good binders (aff< 500nM)

Extended motifs

- Fitness of aa at each position given by $P(aa)$

- Example P1

$$P_A = 6/10$$

$$P_G = 2/10$$

$$P_T = P_K = 1/10$$

$$P_C = P_D = \dots P_V = 0$$

- Problems

- Few data
- Data redundancy/duplication

ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWKVW
KLNEPVLLL
AVVPFIVSV

RLLDDDTPEV 84 nM
GLLGNVSTV 23 nM
ALAKAAAAL 309 nM

Sequence information

Raw sequence counting



Sequence weighting

- Poor or biased sampling of sequence space

- Example P1

$$P_A = 2/6$$

$$P_G = 2/6$$

$$P_T = P_K = 1/6$$

$$P_C = P_D = \dots P_V = 0$$

ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV

} Similar sequences
Weight 1/5

Sequence weighting



Pseudo counts

- I is not found at position P9.
Does this mean that I is forbidden ($P(I)=0$)?
- No! Use Blosum substitution matrix to estimate pseudo frequency of I at P9

ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV

Pseudo count estimation

- Calculate observed amino acids frequencies f_a
- Pseudo frequency for amino acid b

$$g_b = \sum_a f_a \cdot q_{b|a}$$

- Example

$$g_I = 0.2 \cdot q_{I|M} + 0.1 \cdot q_{I|R} + \dots + 0.3 \cdot q_{I|V} + 0.1 \cdot q_{I|L}$$

$$g_I = 0.2 \cdot 0.1 + 0.1 \cdot 0.02 + \dots + 0.3 \cdot 0.16 + 0.1 \cdot 0.12 = 0.094$$

ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV

Weight on pseudo count

- Pseudo counts are important when only limited data is available
- With large data sets only “true” observation should count

$$p_a = \frac{\alpha \cdot f_a + \beta \cdot g_a}{\alpha + \beta}$$

- α is the effective number of sequences ($N-1$), β is the weight on prior or weight on pseudo count

ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWKVV
KLNEPVLLL
AVVPFIVSV

Weight on pseudo count

- Example

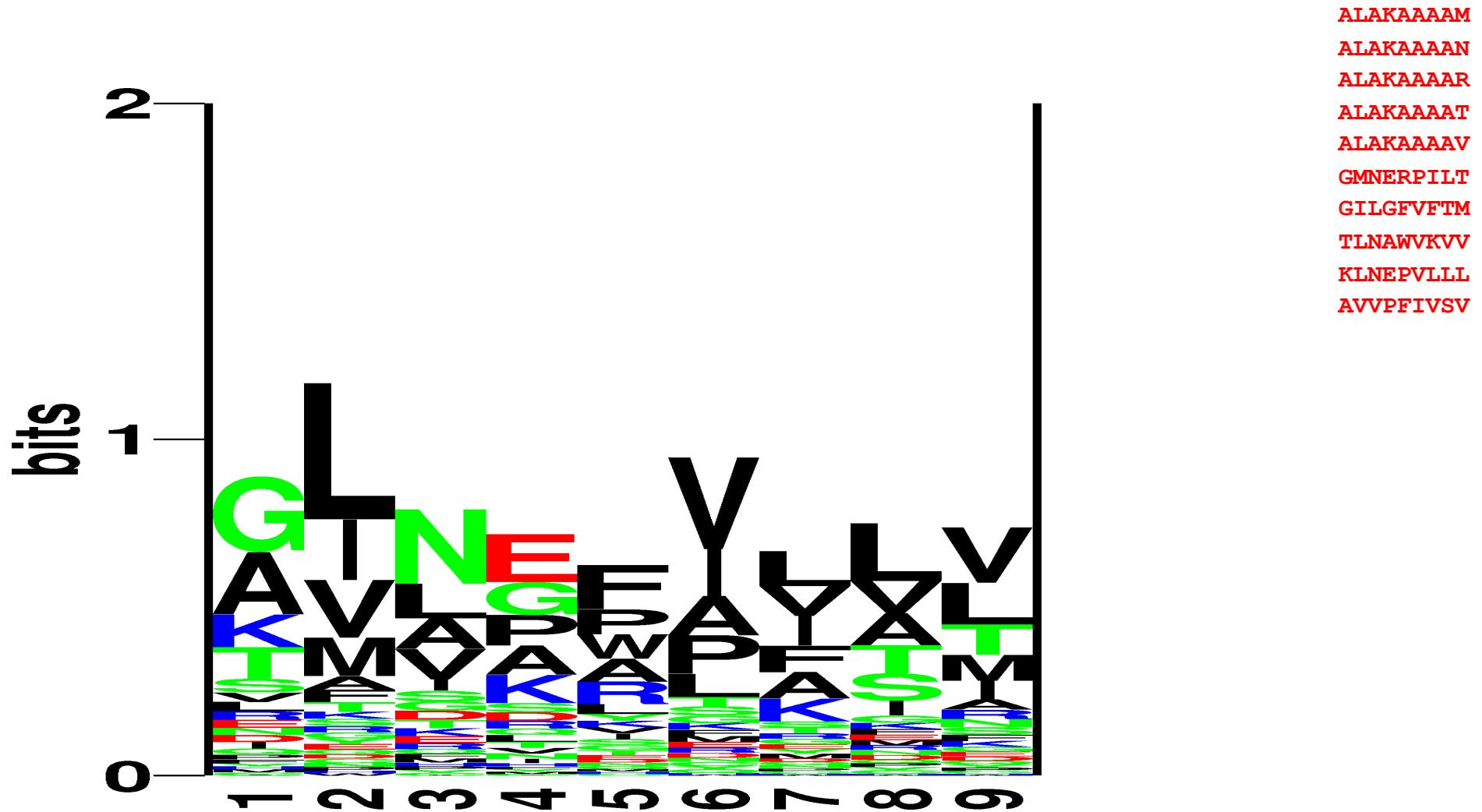
$$p_a = \frac{\alpha \cdot f_a + \beta \cdot g_a}{\alpha + \beta}$$

- If α large, $p \approx f$ and only the observed data defines the motif
- If α small, $p \approx g$ and the pseudo counts (or prior) defines the motif
- β is [50-200] normally (see how β is estimated later)

ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWKVV
KLNEPVLLL
AVVPFIVSV

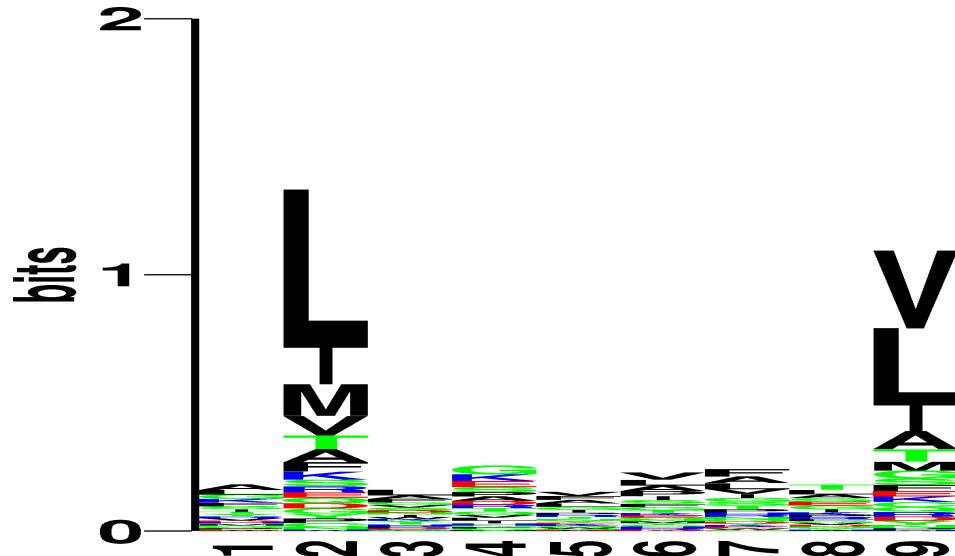
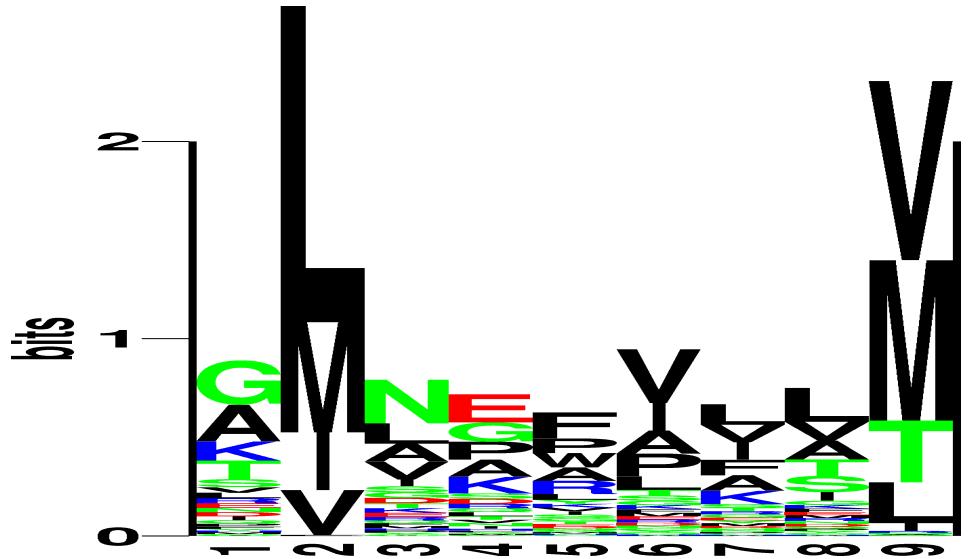
Sequence weighting and pseudo counts

CENTER FOR BIOLOGICAL
SEQUENCE ANALYSIS CBS



Position specific weighting

- We know that positions 2 and 9 are anchor positions for most MHC binding motifs
 - Increase weight on high information positions
- Motif found on large data set



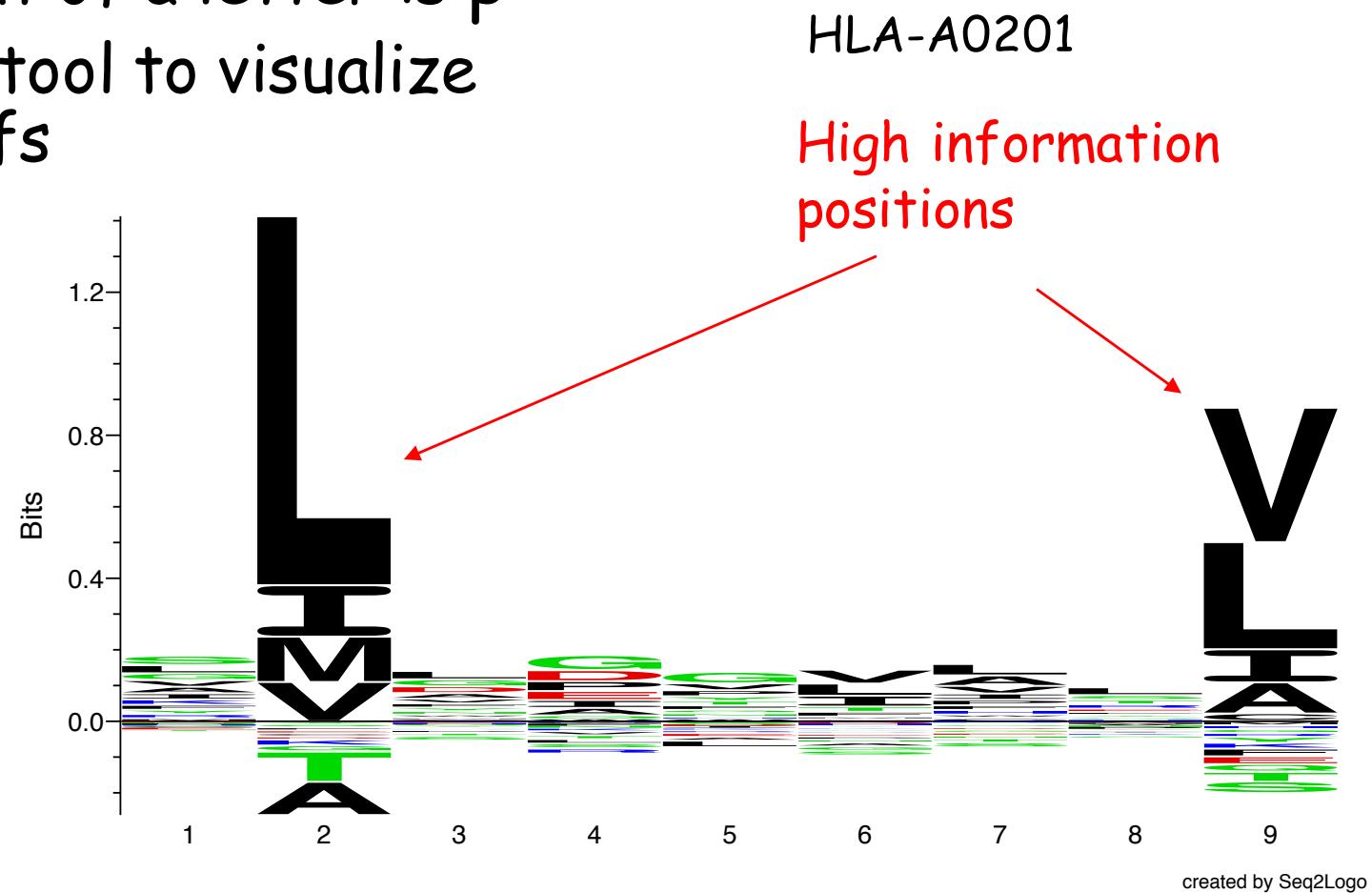
An example!!
(See handout)

Estimation of pseudo counts

	f_a	g_a	p_a	w_a
A	0	0.06	0.03	-2.61
R	0	0.053	0.027	-1.93
N	0	0.04	0.02	-2.33
D	0	0.083	0.042	-0.75
C	0	0.01	0.005	-4.64
Q	0.167	0.085	0.126	3.78
E	0.833	0.267	0.550	6.70
G	0	0.04	0.02	-3.78
H	0	0.03	0.015	-1.59
I	0	0.022	0.011	-5.30
L	0	0.042	0.021	-4.50
K	0	0.082	0.041	-1.01
M	0	0.012	0.006	-4.19
F	0	0.018	0.009	-4.72
P	0	0.028	0.014	-2.92
S	0	0.06	0.03	-1.85
T	0	0.04	0.02	-2.70
W	0	0.01	0.005	-2.76
Y	0	0.02	0.01	-3.36
V	0	0.032	0.016	-4.41

Sequence logos

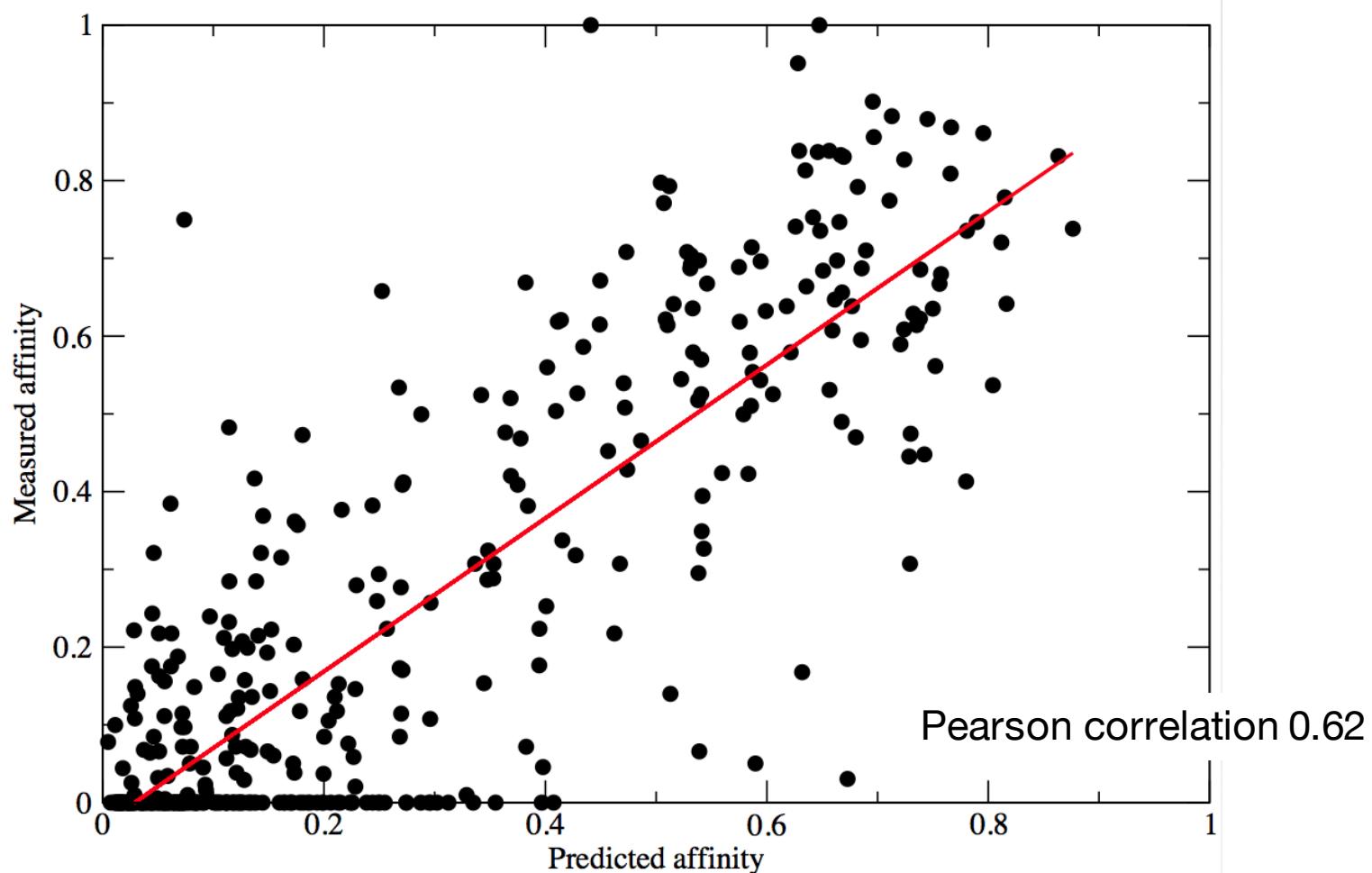
- Height of a column equal to I
- Relative height of a letter is p
- Highly useful tool to visualize sequence motifs



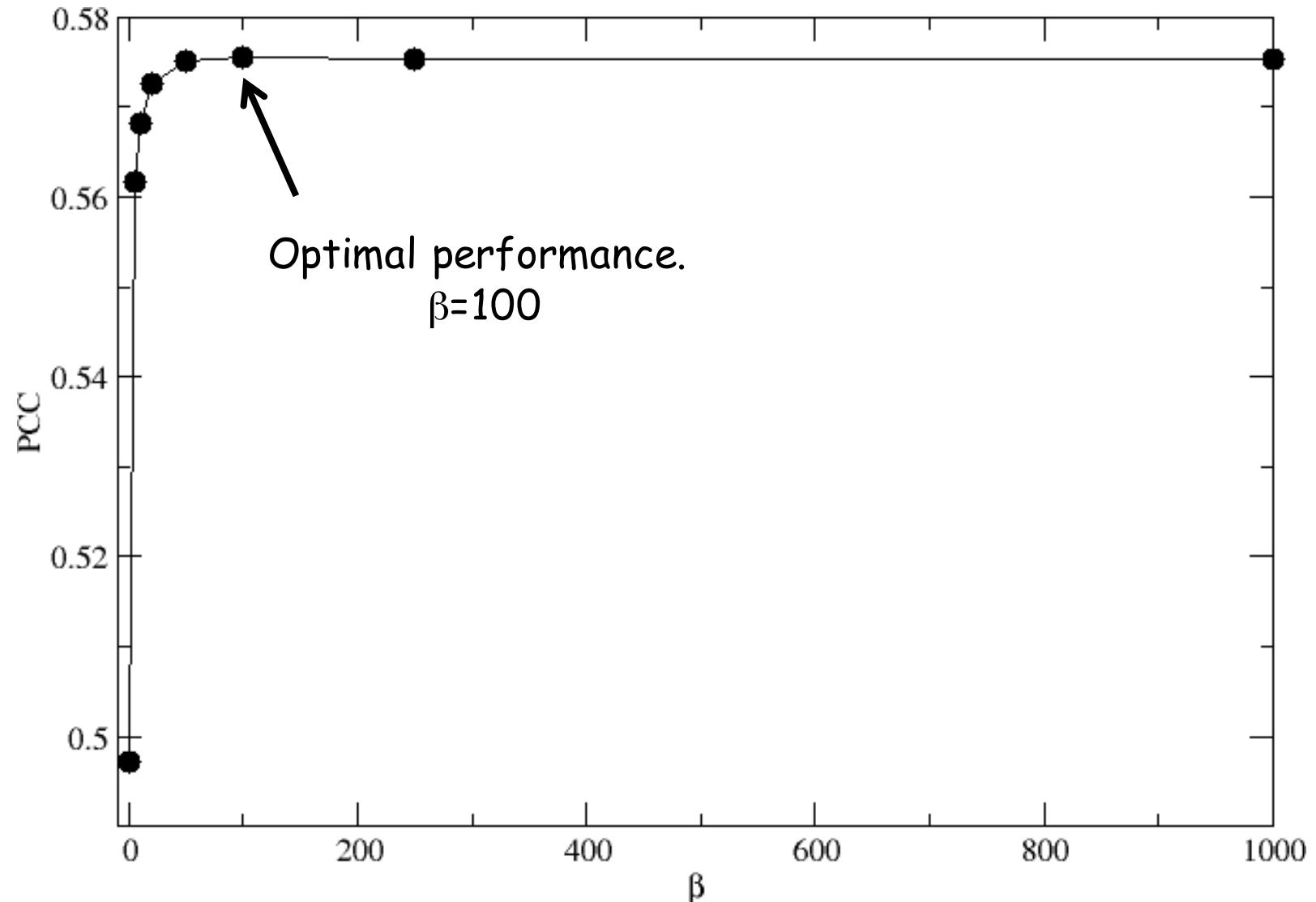
Example from real life

- 10 peptides from MHCpep database
 - ALAKAAAAM**
 - ALAKAAAAN**
 - ALAKAAAAR**
 - ALAKAAAAT**
 - ALAKAAAAV**
 - GMNERPILT**
 - GILGFVFTM**
 - TLNAWVKVV**
 - KLNEPVLLL**
 - AVVPFIVSV**
- Bind to the MHC complex
- Relevant for immune system recognition
- Estimate sequence motif and weight matrix
- Evaluate motif “correctness” on 528 peptides

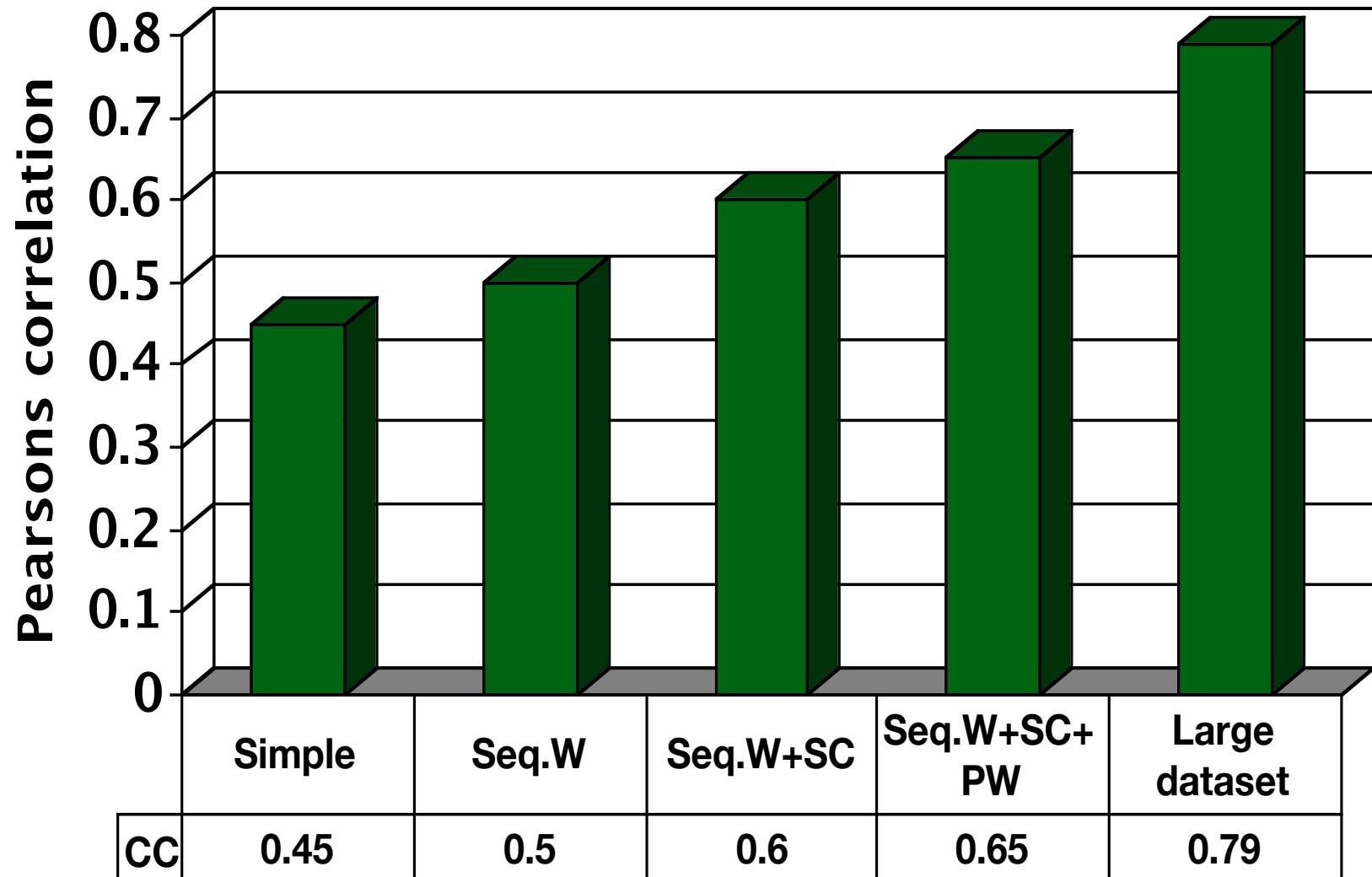
Prediction accuracy



How to define β ?



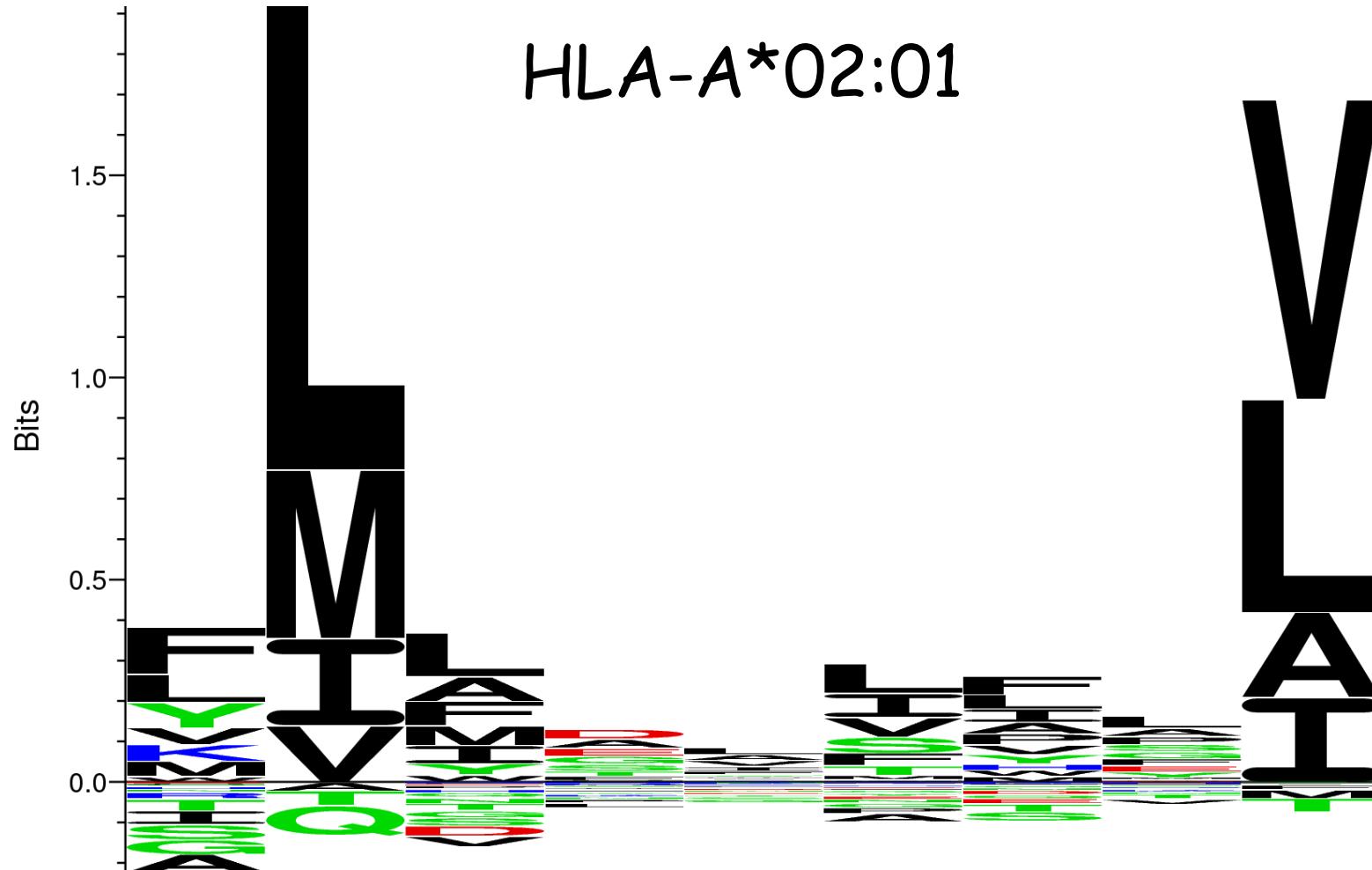
Predictive performance



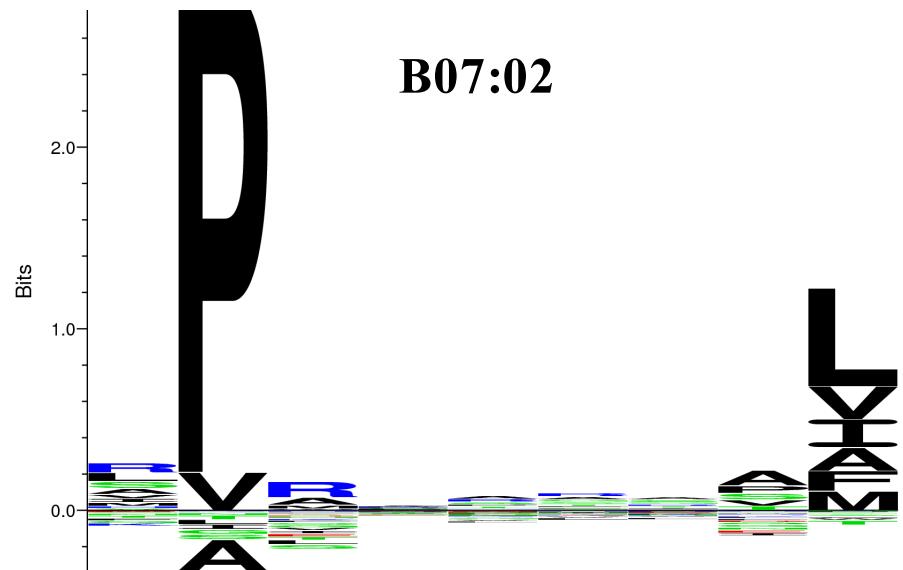
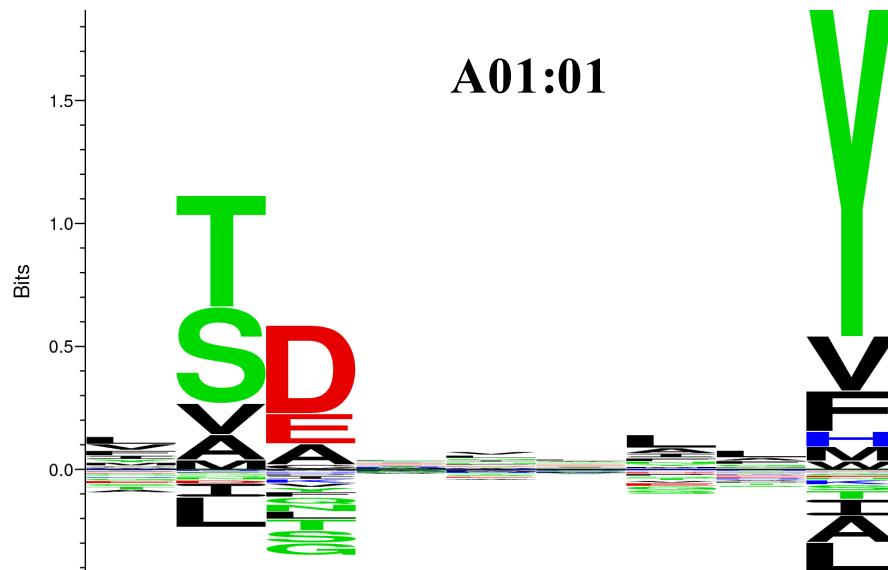
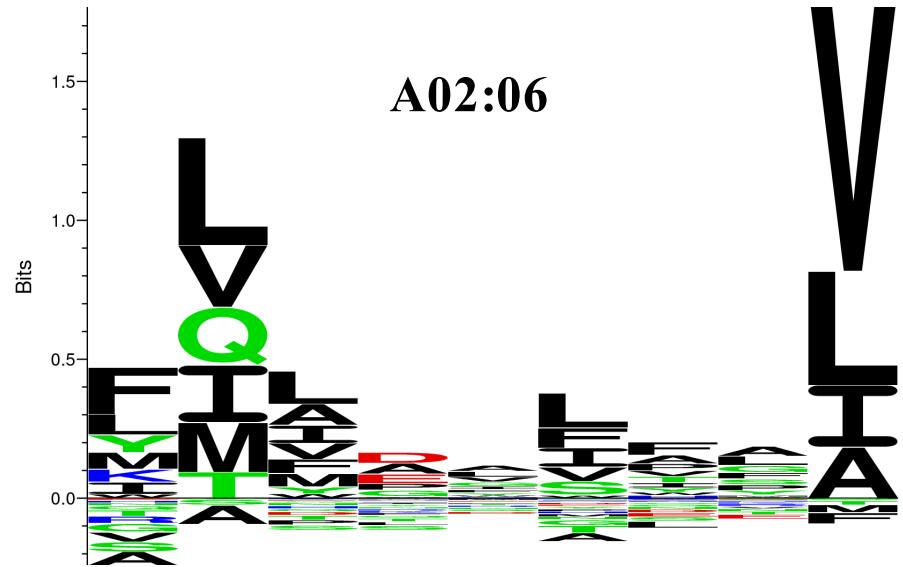
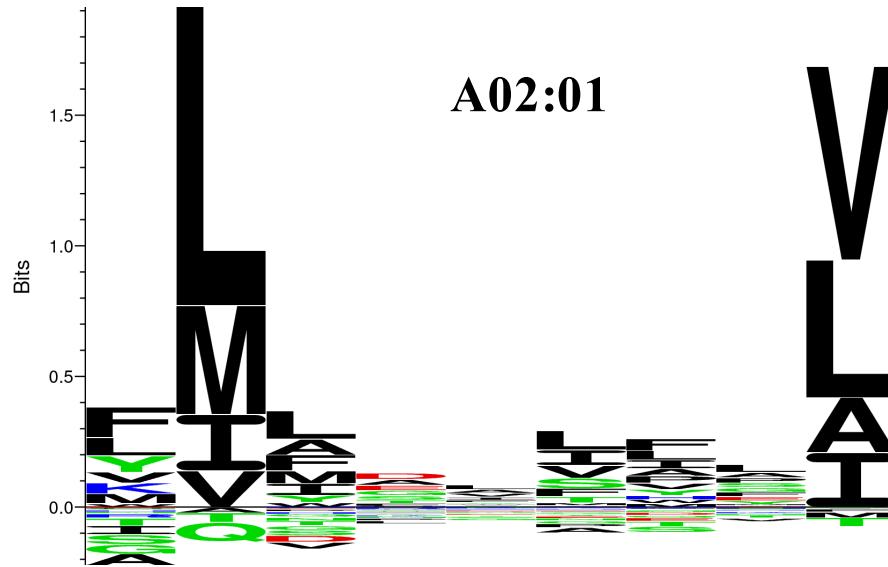
MHC binding motifs

SLLPAIVEL YLLPAIVHI TLWVDPYEV GLVPFLVSV KLLEPVLLL LLDVPTAAV LLDVPTAAV LLDVPTAAV
LLDVPTAAV VLFRGGPRG MVDGTLLLL YMNGTMSQV MLLSVPLLL SLLGLLVEV ALLPPINIL TLIKIQHTL
HLIDYLVTS ILAPPVVKL ALFPQLVIL GILGFVFTL STNRQSGRQ GLDVLTAKV RILGAVAKV QVCERIPTI
ILFGHENRV ILMEMHIHKL ILDQKINEV SLAGGIIGV LLIENVASL FLLWATAEA SLPDFGISY KKREEAPSL
LERPGGNEI ALSNLEVKL ALNELLOQHV DLERKVESL FLGENISNF ALSDHHIYL GLSEFTEYL STAPPAHGV
PLDGEYFTL GVLVGVALI RTLDKVLEV HLSTAFARV RLDSYVRSL YMNGTMSQV GILGFVFTL ILKEPVHGV
ILGFVFTLT LLFGYPVYV GLSPTVWLS WLSLLVPFV FLPSDFFPS CLGGLLTMV FIAGNSAYE KLGEFYNQM
KLVALGINA DLMGYIPLV RLVTLKDIV MLLAVLYCL AAGIGILTV YLEPGPVTA LLDGTATLR ITDQVPFSV
KTWGQYWQV TITDQVPFS AFHHVAREL YLNKIQNSL MMRKLAILS AIMDKNIIL IMDKNIILK SMVGNWAKV
SLLAPGAKQ KIFGSLAFL ELVSEFSRM KLTPLCVTL VLYRYGSFS YIGEVLVSV CINGVCWTW VMNILLQYV
ILTVELGVL KVLEYVIKV FLWGPRALV GLSRYVARL FLLTRILTI HLGNVKYLV GIAGGLALL GLQDCTMLV
TGAPVTYST VIYQYMDDL VLPDVFIRC VLPDVFIRC AVGIGIAVV LVVLGLLAV ALGLGLLPV GIGIGVLA
GAGIGVAVL IAGIGILAI LIVIGILIL LAGIGLIAA VDGIGILTI GAGIGVLTA AAGIGIIQI QAGIGILLA
KARDPHSGH KACDPHSGH ACDPHSGHF SLYNTVATL RGPGRAFVT NLVPMVATV GLHCYEQLV PLKQHFQIV
AVFDRKSDA LLDFVRFMG VLVKSPNHV GLAPPQHLLI LLGRNSFEV PLTFGWCYK VLEWRFDSR TLNAWKVV
GLCTLVAML FIDSYICQV IISAVVGIL VMAGVGSPY LLWTLVVLL SVRDRLARL LLMDCSGSI CLTSTVQLV
VLHDDLLEA LMWITQCFL SLLMWITQC QLSLLMWIT LLGATCMFV RLTRFLSRV YMDGTMSQV FLTPKKLQC
ISNDVCAQV VKTDGNPPE SVYDFFFVWL FLYGALLA VLFSSDFRI LMWAKIGPV SLLLELEEV SLSRFWSWGA
YTAFTIPSI RLMKQDFSV RLPRIFCSC FLWGPRAYA RLLQETELV SLFEGIDFY SLDQSVVEL RLNMFTPYI
NMFTPYIGV LMIPIPLINV TLFIGSHVV SLVIVTTFV VLQWASLAV ILAKFLHWL STAPPHVNV LLLLTVLTV
VVLGVVFGI ILHNGAYSL MIMVKCWMI MLGHTMEV MLGHTMEV SLADTNSLA LLWAARPRL GVALQTMKQ
GLYDGMEHL KMVELVHFL YLQLVFGIE MLMAQEALA LMAQEALAF VYDGREHTV YLSGANLNL RMFPNAPYL
EAAGIGILT TLDSQVMSL STPPPGRTRV KVAELVHFL IMIGVLVGV ALCRWGLLL LLFAGVQCQ VLLCESTAV
YLSTAFARV YLLEMWLRL SLDDYNHLV RTLDKVLEV GLPVEYLQV KLIANNTRV FIYAGSLSA KLVANNTRL
FLDEFMEGV ALQPGBTALL VLDGLDVLL SLYSFPEPE ALYVDSLFF SLLQHLLIGL ELTLGEFLK MINAYLDKL
AAGIGILTV FLPSDFFPS SVRDRLARL SLREWLLRI LLSAWILTA AAGIGILTV AVPDEIPPL FAYDGKDYL
AAGIGILTV FLPSDFFPS AAGIGILTV FLPSDFFPS AAGIGILTV FLWGPRALV ETVSEQSNV ITLWQRPLV

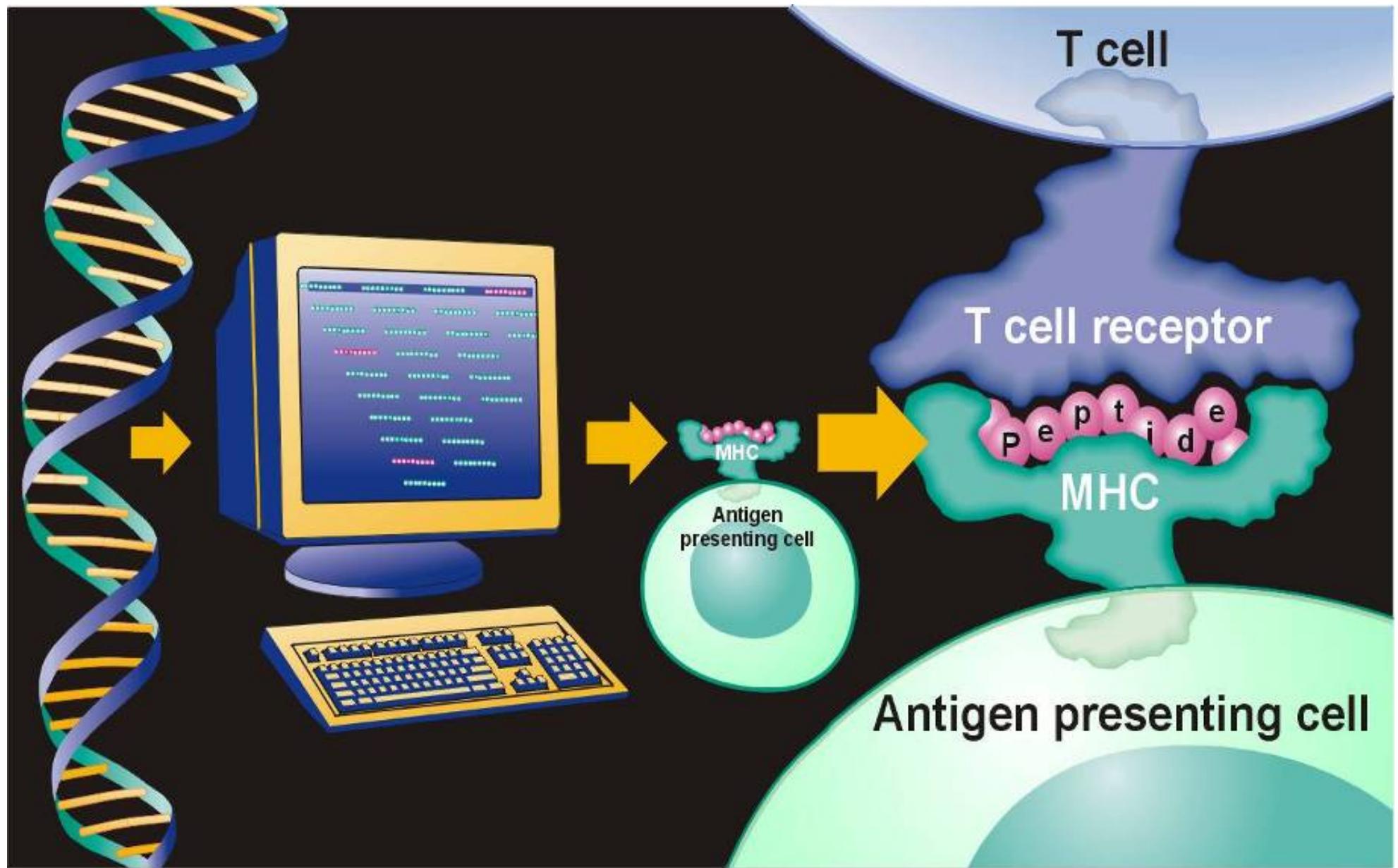
HLA binding motifs



HLA specificities



Antigen Discovery



Influenza A virus (A/Goose/Guangdong/1/96(H5N1))

Genome

>Segment 1

```
agcaaaaaggcaggtaattatattcaatatggaaagaataaaagaactaagagatctaatg  
tcgcagtcccgcaactcgcgagatactaacaaaaaccactgtggatcatatggccataatc  
aagaaatacacatcaggaagacaagagaagaaccctgctcagaatgaaatggatgatg  
gcaatgaaatatccaatcacagcagacaagagaataatggagatgattcctgaaaggaat
```

and 13350 other nucleotides on 8 segments

↓
Proteins

>polymerase"

```
MERIKELRDLMSQSRTREILTKTTVDHMAIIKKYTSRQEKNPALRMKWMAMKYPITAD  
KRIMEMIPERNEQGQTLWSKTNDAGSDRVMVSPLAVTWWNRNGPTTSTVHPKVYKTYFE  
KVERLKHGTFGPVHFRNQVKIRRVDINPGHADLSAKEAQDVIMEVFPNEVGARILTSE  
SQLTITKEKKEELQDCKIAPLMAYMLERELVRKTRFLPVAGGTSSVYIEVLHLTQGTCW  
EQMYTPGGEVRNDVDQSLIIAARNIVRRATVSADPLASLLEMCHSTQIGGIRMVDILRQ  
NPTEEQAVDICKAAMGLRISSSFSGGTFKRTNGSSVKKEEEVLTGNLQTLKIKVHEGY  
EEFTMVGRATAILRKATRRLIQLIVSGRDEQSIAEAIIVAMVFSQEDCMIKAVRGDLNF  
...
```

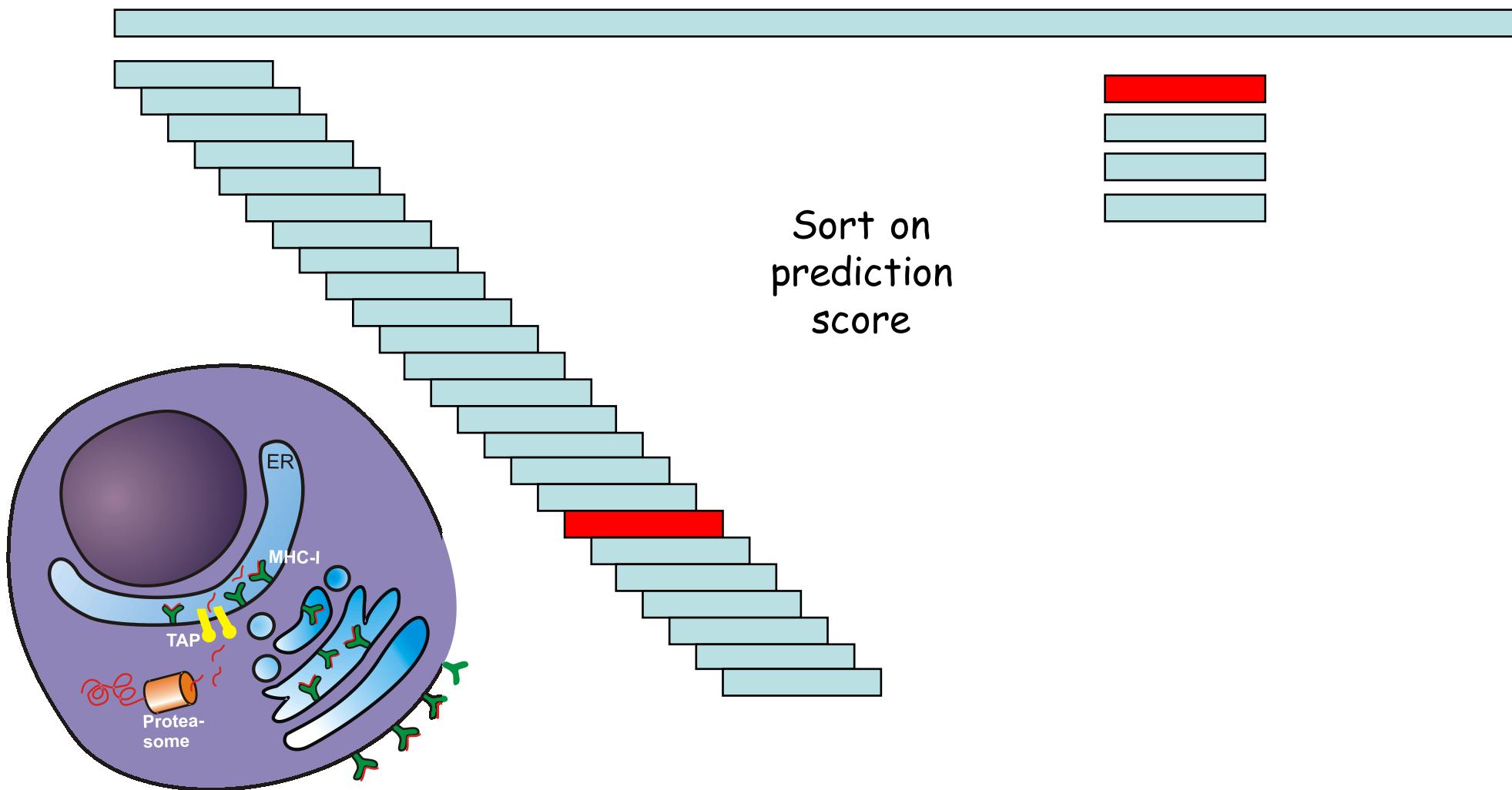
and 9 other proteins

**9mer
peptides**

MERIKELRD
ERIKELRDL
RIKELRDLM
IKELELDLMS
KELRDLMQS
EIRDLMSQS
LRDLMSQR
RDLMQSRT
DILMSQSRTR
LMSQSRTRE

and 4376 other 9mers

Predicting T cell epitopes?



Validation of binding predictions

1. Identification of MHC class II restricted T-cell-mediated reactivity against MHC class I binding *Mycobacterium tuberculosis* peptides. Wang M, Tang ST, Stryhn A, Justesen S, Larsen MV, Dziegial MH, Lewinsohn DM, Buus S, Lund O, Claesson MH. *Immunology*. 2011 Apr;132(4):482-91. doi: 10.1111/j.1365-2567.2010.03383.x. Epub 2011 Feb 7.
 2. Genome-Based In Silico Identification of New *Mycobacterium tuberculosis* Antigens Activating Polyfunctional CD8+ T Cells in Human Tuberculosis. Tang ST, van Meijgaarden KE, Caccamo N, Guggino G, Klein MR, van Weeren P, Kazi F, Stryhn A, Zaigler A, Sahin U, Buus S, Dieli F, Lund O, Ottenhoff TH. *J Immunol*. 2011 Jan 15;186(2):1068-80. Epub 2010 Dec 17.
 3. Identification of CD8+ T cell epitopes in the West Nile virus polyprotein by reverse-immunology using NetCTL. Larsen MV, Lelic A, Parsons R, Nielsen M, Hoof I, Lambeth K, Loeb MB, Buus S, Bramson J, Lund O. *PLoS One*. 2010 Sep 14;5(9):e12697.
 4. HLA class I binding 9mer peptides from influenza A virus induce CD4 T cell responses. Wang M, Larsen MV, Nielsen M, Harndahl M, Justesen S, Dziegial MH, Buus S, Tang ST, Lund O, Claesson MH. *PLoS One*. 2010 May 7;5(5):e10533.
 5. Interdisciplinary Analysis of HIV-Specific CD8+ T Cell Responses against Variant Epitopes Reveals Restricted TCR Promiscuity. Hoof I, Pérez CL, Buggert M, Gustafsson RK, Nielsen M, Lund O, Karlsson AC. *J Immunol*. 2010 May 1;184(9):5383-91. Epub 2010 Apr 2.
 6. High-affinity human leucocyte antigen class I binding variola-derived peptides induce CD4 T cell responses more than 30 years post-vaccinia virus vaccination. Wang M, Tang ST, Lund O, Dziegial MH, Buus S, Claesson MH. *Clin Exp Immunol*. 2009 Mar;155(3):441-446.
 7. MHC I restricted epitopes conserved among variola and other related orthopoxviruses are recognised by T cells 30 years after vaccination. Tang ST, Wang M, Lambeth K, Harndahl M, Dziegial MH, Claesson MH, Buus S, Lund O. *Arch Virol*. 2008;153(10):1833-44. Epub 2008 Sep 12.
 8. Pérez CL, Larsen MV, Gustafsson R, Norström MM, Atlas A, Nixon DF, Nielsen M, Lund O, Karlsson AC. Broadly Immunogenic HLA Class I Supertype-Restricted Elite CTL Epitopes Recognized in a Diverse Population Infected with Different HIV-1 Subtypes. *J Immunol*. 2008 180:5092-100
 9. Wang M, Lambeth K, Harndahl M, Roder G, Stryhn A, Larsen MV, Nielsen M, Lundsgaard C, Tang ST, Dziegial MH, Rosenkvist J, Pedersen AE, Buus S, Claesson MH, Lund O. CTL epitopes for influenza A including the H5N1 bird flu; genome-, pathogen-, and HLA wide screening. *Vaccine*. 2006 25:2823-2831.
 10. Sylvester-Hvid C, Nielsen M, Lambeth K, Roder G, Justesen S, Lundsgaard C, Worming P, Thomadsen H, Lund O, Brunak S, Buus S. SARS CTL vaccine candidates; HLA supertype-, genome-wide scanning and biochemical validation. *Tissue Antigens*. 2004 63:395-400.
-

Sequence Profiles and Weight matrices

- Alignments based on conventional scoring matrices (BLOSUM62) scores all positions in a sequence in an equal manner
 - Some positions are highly conserved, some are highly variable (more than what is described in the BLOSUM matrix)
 - Sequence profile are ideal suited to describe such position specific variations
-

Sequence profiles

Conserved Non-conserved

```
ADDGSLAFVPSEF--SISPGEKIVFKNNAGFPHNIVFDEDSIPSGVDASKIS MSEEEDLLN  
TVNGAI--PGPLIAERLKEGQNVRVTNTLDEDTSIHWHGLLVPFGMDGVPGSFPG---I  
-TSMAPAFGVQEFYRTVKQGDEVTVTIT----NI DQIED-VSHGFVVVNHGVSME---I  
IE--KMKYLTPEVFYTIKAGETVYWVNGEVMPHNVAFKKGIV--GEDAFRGEMMTKD---  
-TSVAPSFSQPSF-LTVKEGDEVTVI VTNLDE----IDDLTHGFTMGNHGVAME---V  
ASAETMVFEPDFLVLEIGPGDRVRFVPTHK-SHNAATIDGMVPEGVEGFKSRINDE---  
TVNGQ--FPGPRLAGVAREGDQVILVKVVNHVAENITIHWHGVQLGTGWADPPAVTQCPI  
TKAVVLTFNTSVEICLVMQGTSIV----AAESHPLHLHGFNFPSNFNLVDGMERNTAGVP
```

Matching any thing
but G => large
negative score

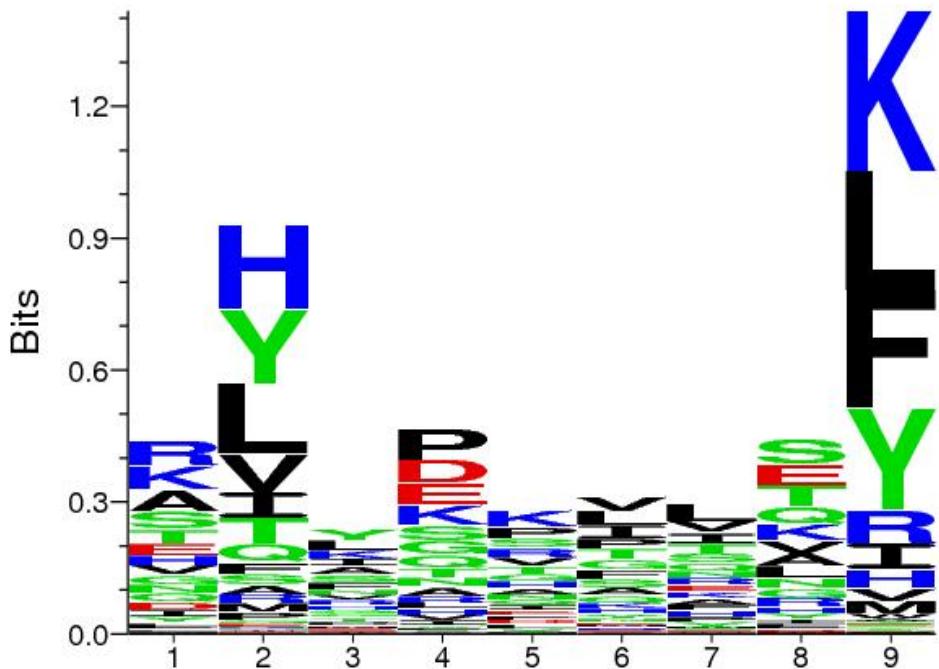
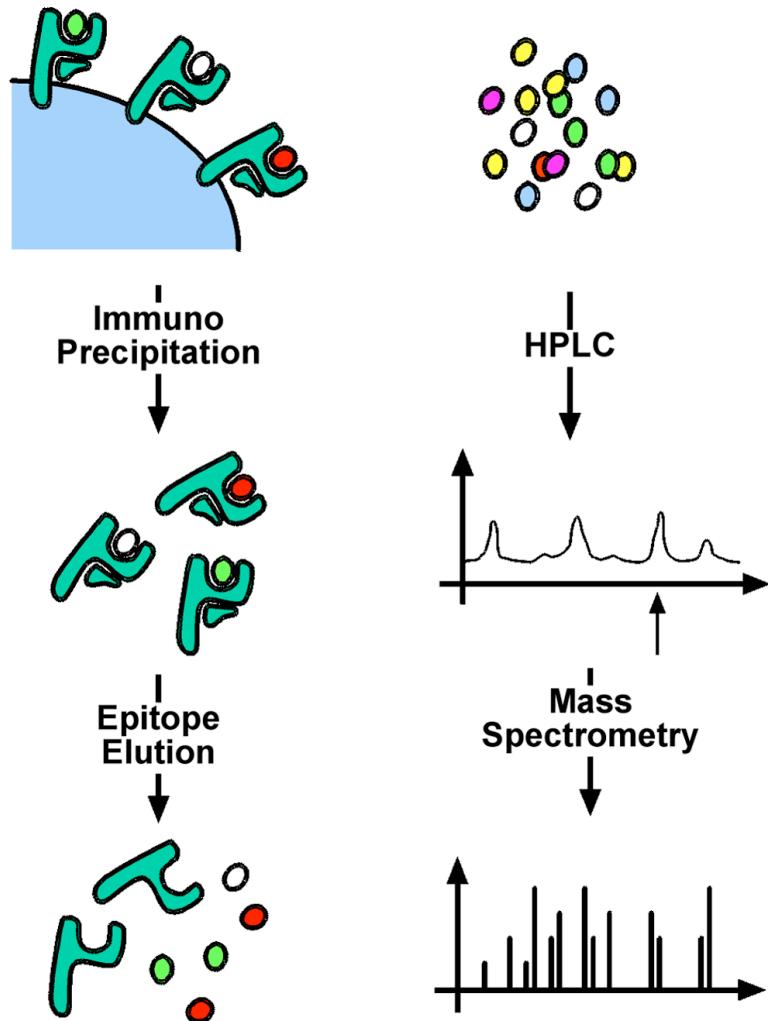
Anything can match

We will go deeper into this on
Friday 11th

Summary

- Sequence logo is a power tool to visualize (binding) motifs
 - Information content identifies essential residues for function and/or structural stability
- Weight matrices can be derived from very limited number of data using the techniques of
 - Sequence weighting
 - Pseudo counts

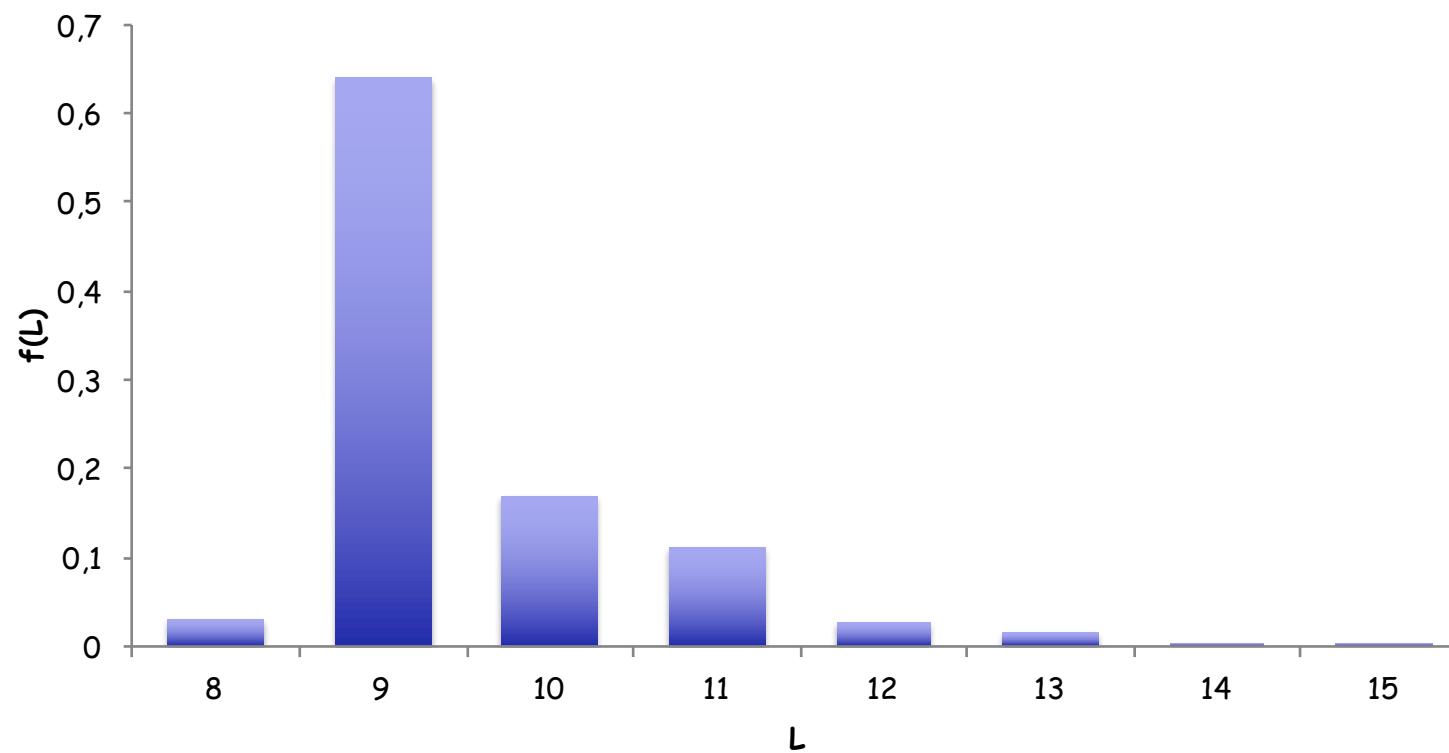
Naturally presented MHC ligands



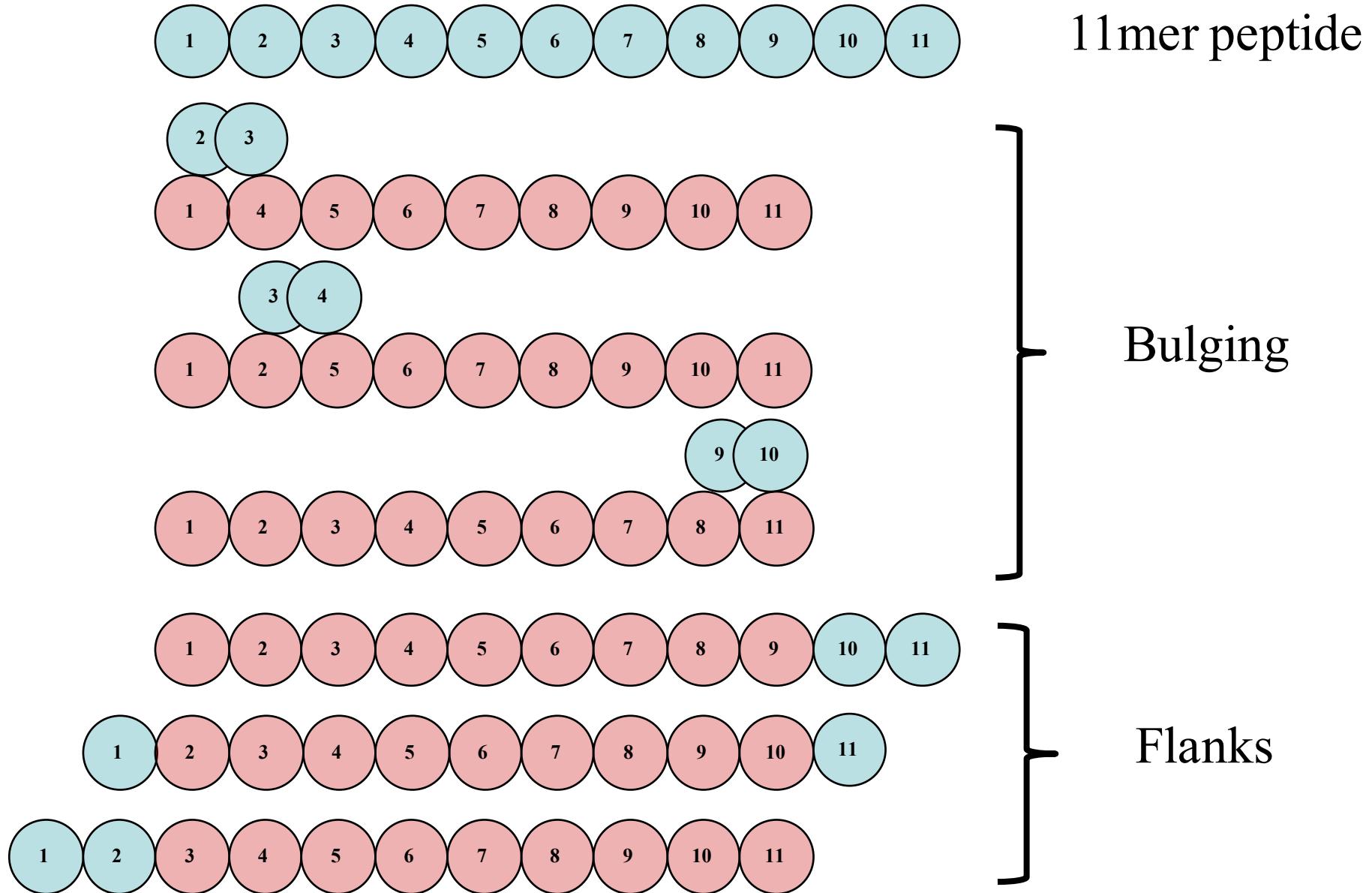
Length distribution

Cell line with HLA's:

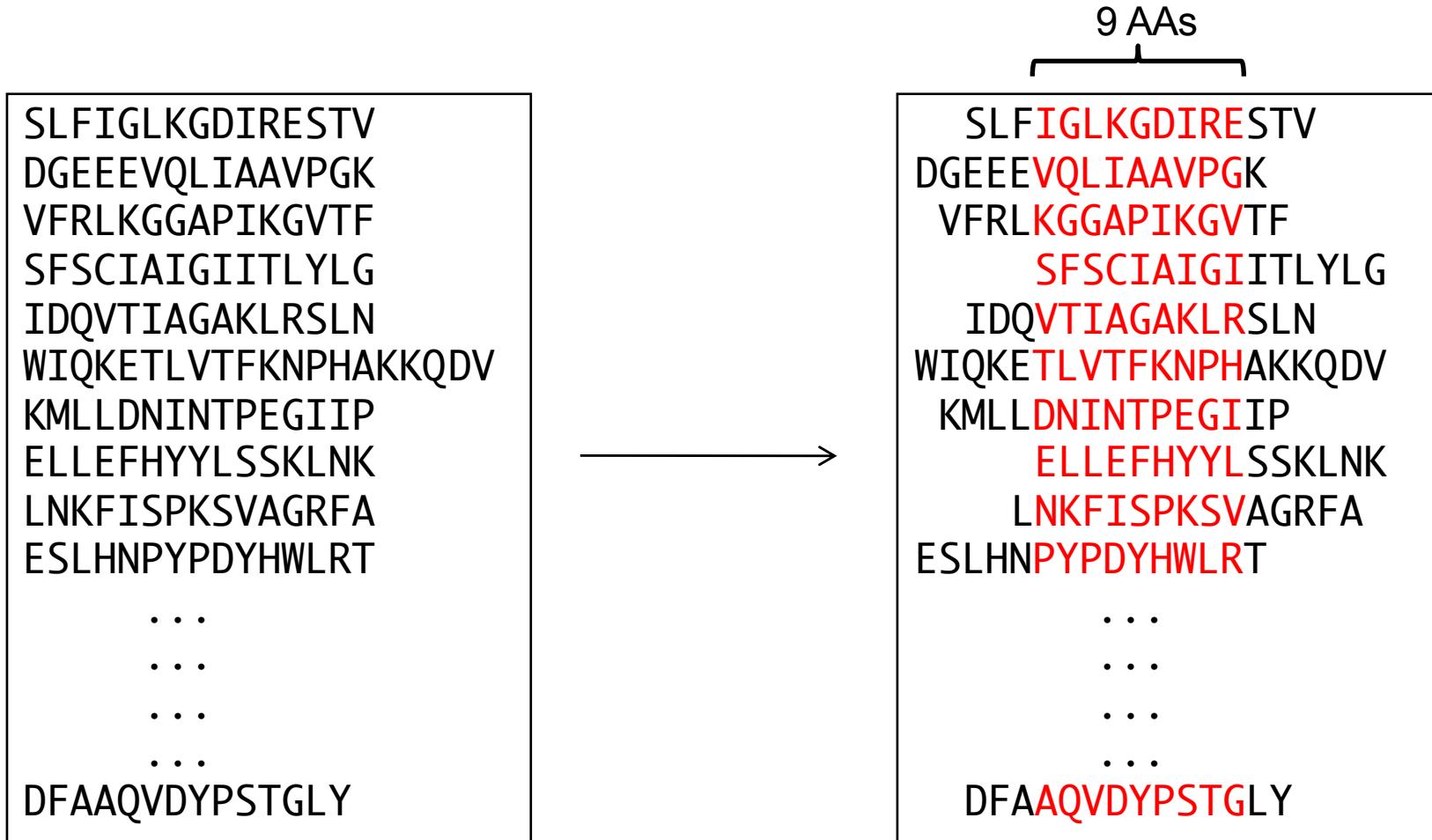
HLA-A03:01, HLA-A23:01, HLA-B08:01, HLA-B15:18, HLA-C07:01, HLA-C07:04



Dealing ligands of variable length



Conventional Gibbs sampling MHC class II binding



$$E = \sum_{p,a} C_{p,a} \log \frac{p_{p,a}}{q_a}$$

Gibbs sampling - sequence alignment

Why sampling?

50 sequences 12 amino acids long

try all possible combinations with a 9-mer overlap



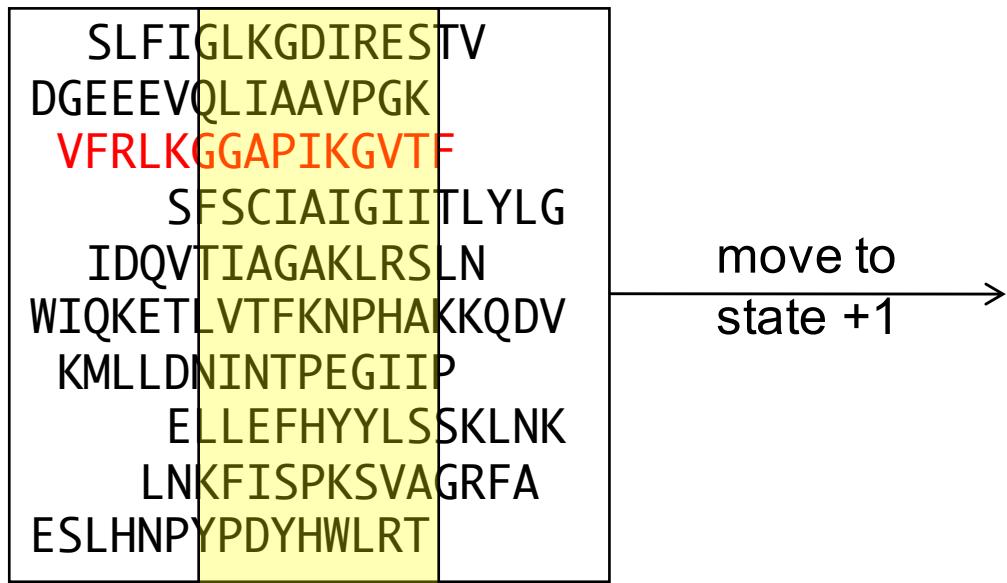
$4^{50} \sim 10^{30}$ possible combinations

...computationally unfeasible

```
SLFIGLKGDIRESTV
DGEDEVQLIAAVPGK
VFRLKGGAPIKGVTF
SFSCIAIGIITLYLG
IDQVTIAGAKLRSLN
WIQKETLVTFKNPHAKKQDV
KMlldnintpegiip
ELLEFHYYLSSKLNK
LNKFISPksVAGRFA
ESLHNPyPDYHWLRT
...
...
...
...
DFAAQVDYPSTGLY
```

Gibbs sampling - sequence alignment

State transition

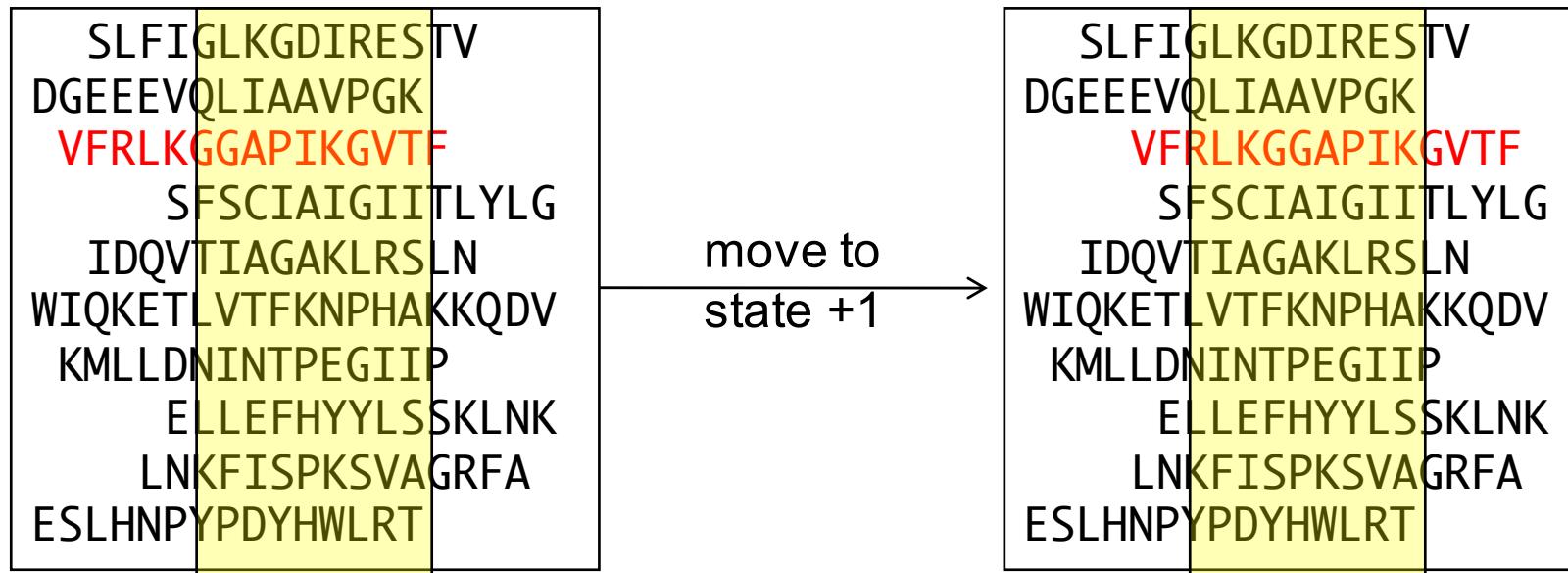


$$E = \sum_{p,a} C_{p,a} \log \frac{p_{p,a}}{q_a}$$

$$dE = E_i - E_{i-1}$$

Gibbs sampling - sequence alignment

State transition



$$E = \sum_{p,a} C_{p,a} \log \frac{p_{p,a}}{q_a}$$

$$dE = E_i - E_{i-1}$$

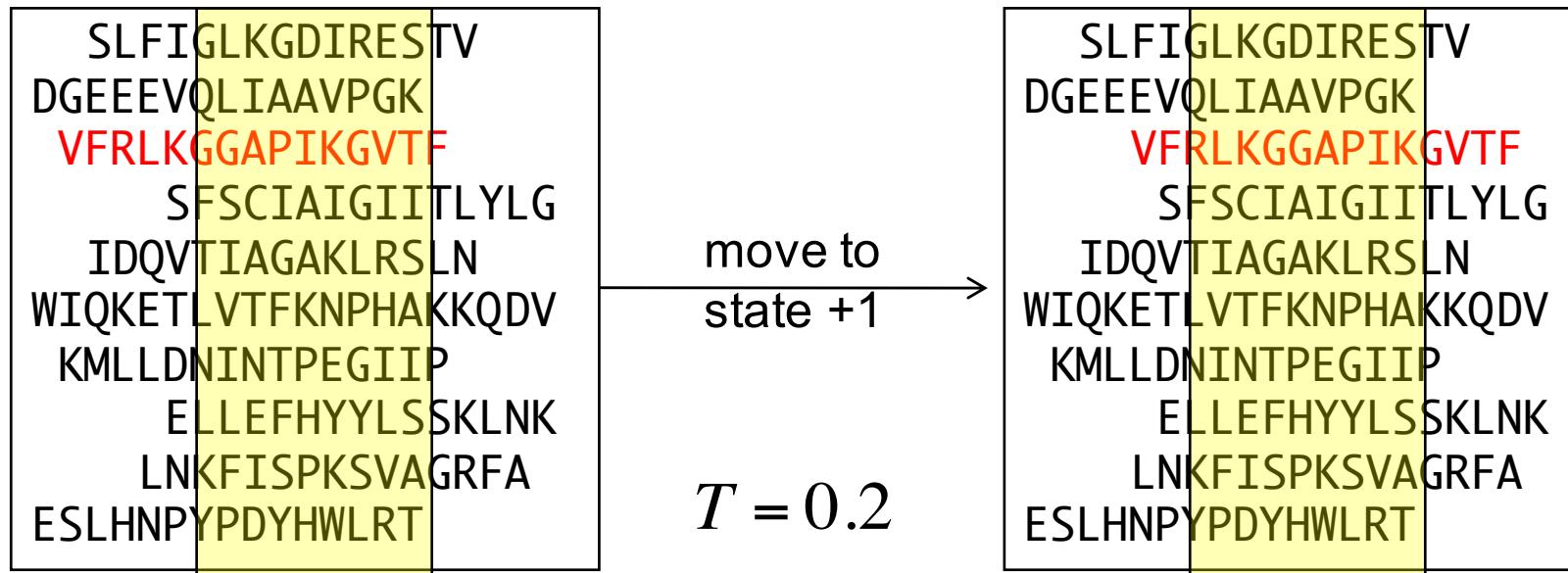
Accept or reject the move?

$$P = \min\left[1, \exp\left(\frac{dE}{T}\right)\right]$$

Note that the probability of going to the new state only depends on the previous state

Gibbs sampling - sequence alignment

Numerical example - 1



$$E_{i-1} = 2.44$$

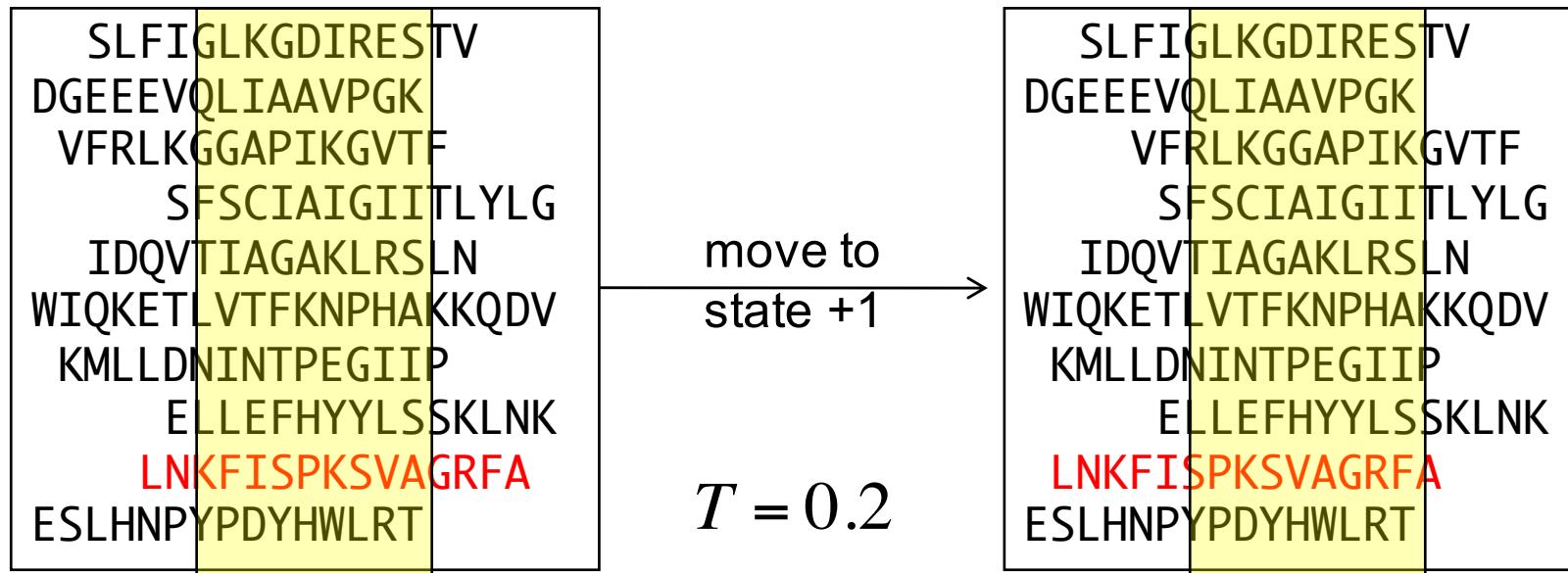
$$E_i = 2.52$$

$$P = \min\left[1, \exp\left(\frac{0.08}{0.2}\right)\right] = \min[1, 1.49] = 1$$

Accept move with
Prob = 100%

Gibbs sampling - sequence alignment

Numerical example - 2



$$E_{i-1} = 2.44$$

$$E_i = 2.35$$

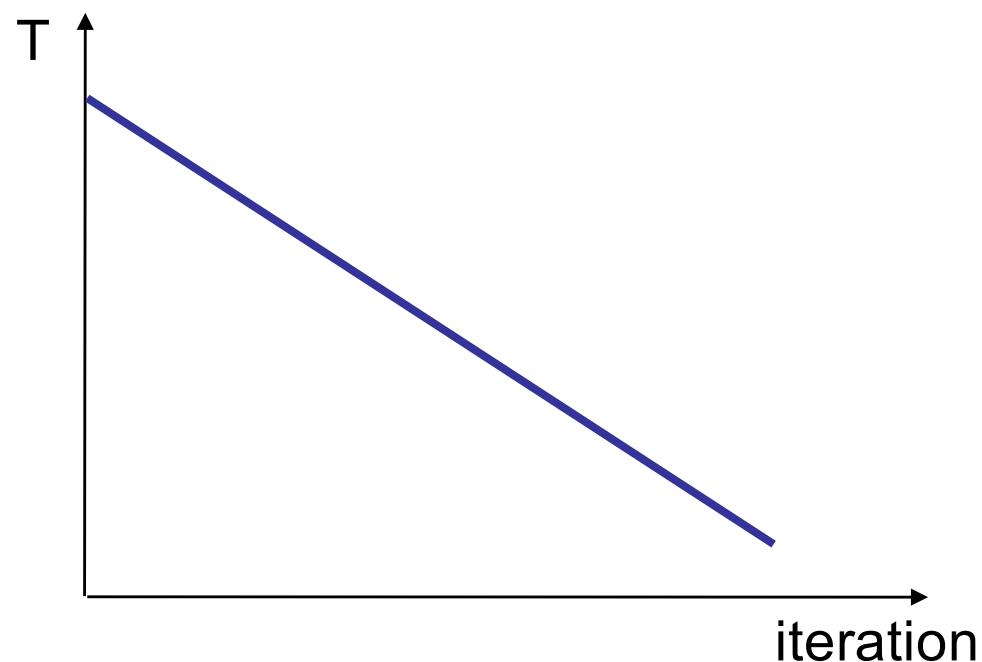
$$P = \min\left[1, \exp\left(\frac{-0.09}{0.2}\right)\right] = \min[1, 0.638] = 0.638$$

Accept move with
Prob = 63.8%

Gibbs sampling - sequence alignment

What is the MC temperature?

it's a scalar decreased during the simulation

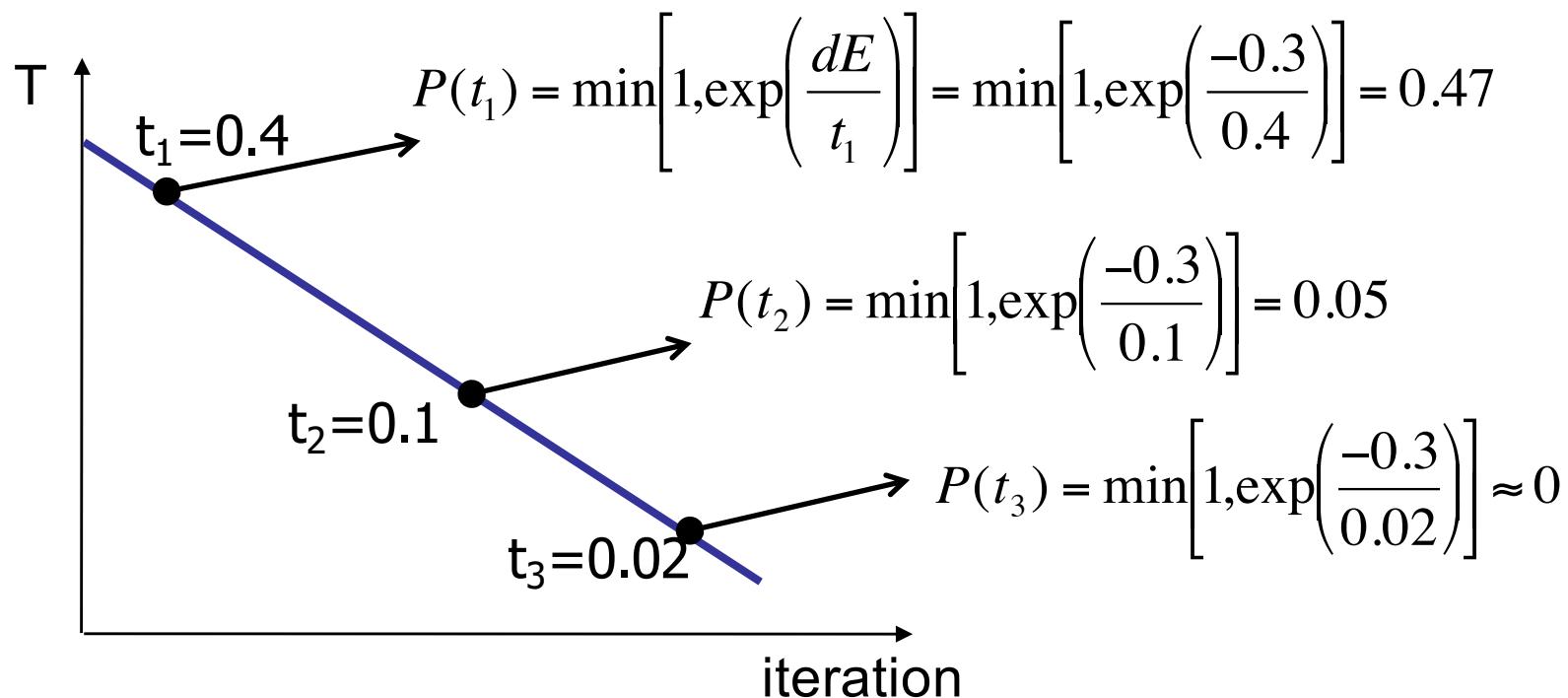


Gibbs sampling - sequence alignment

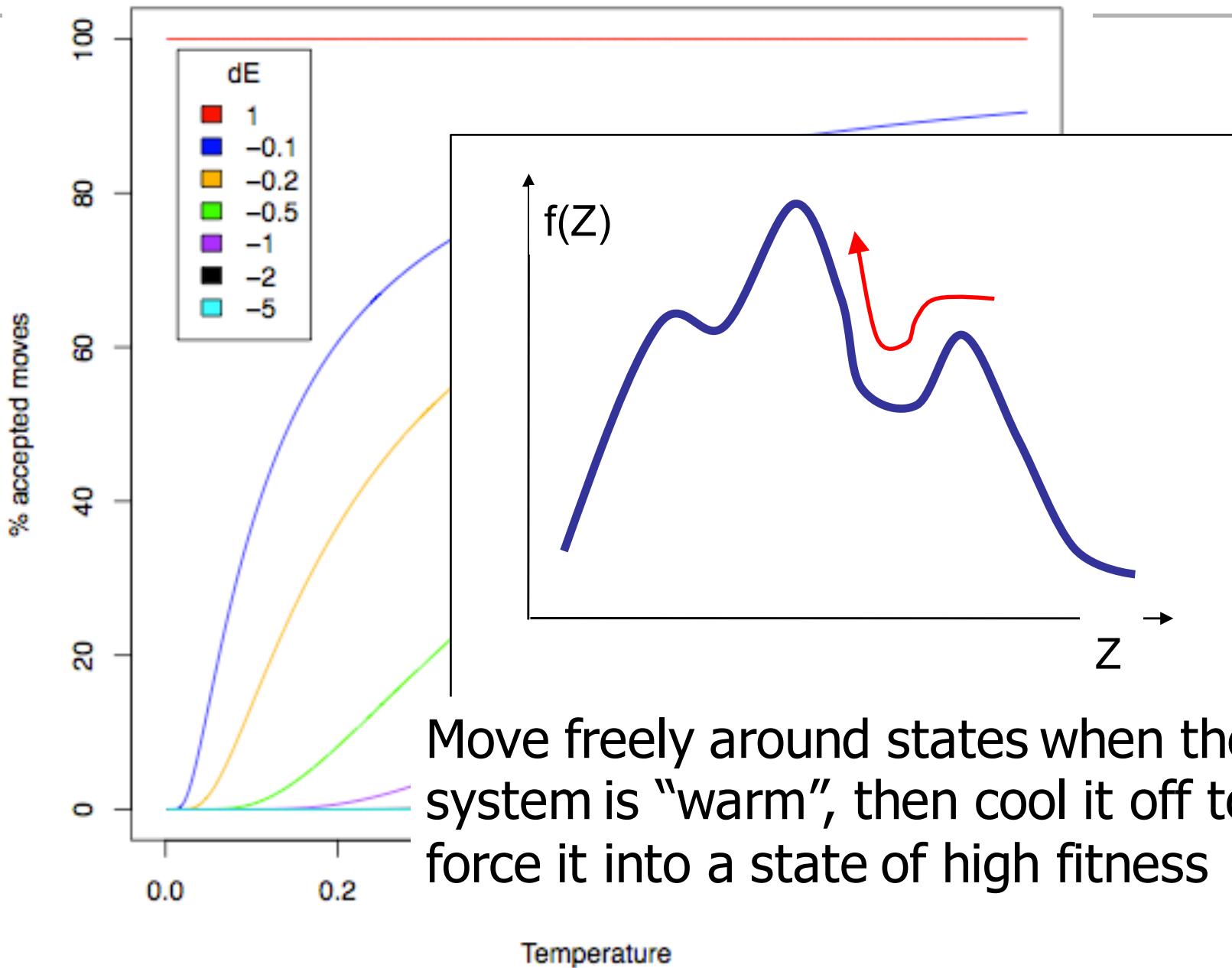
What is the MC temperature?

it's a scalar decreased during the simulation

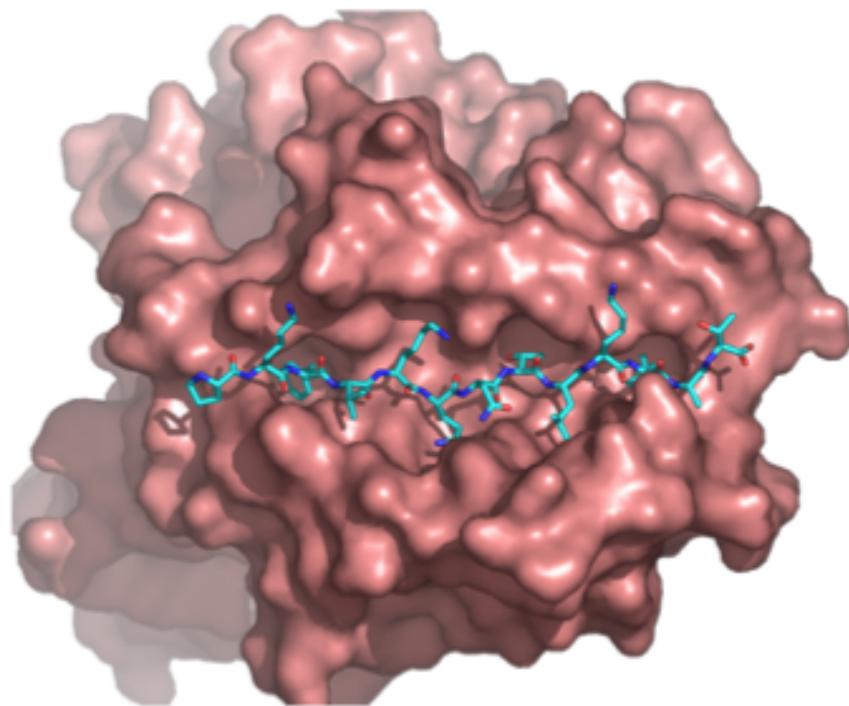
E.g. same **dE=-0.3** but at different temperatures



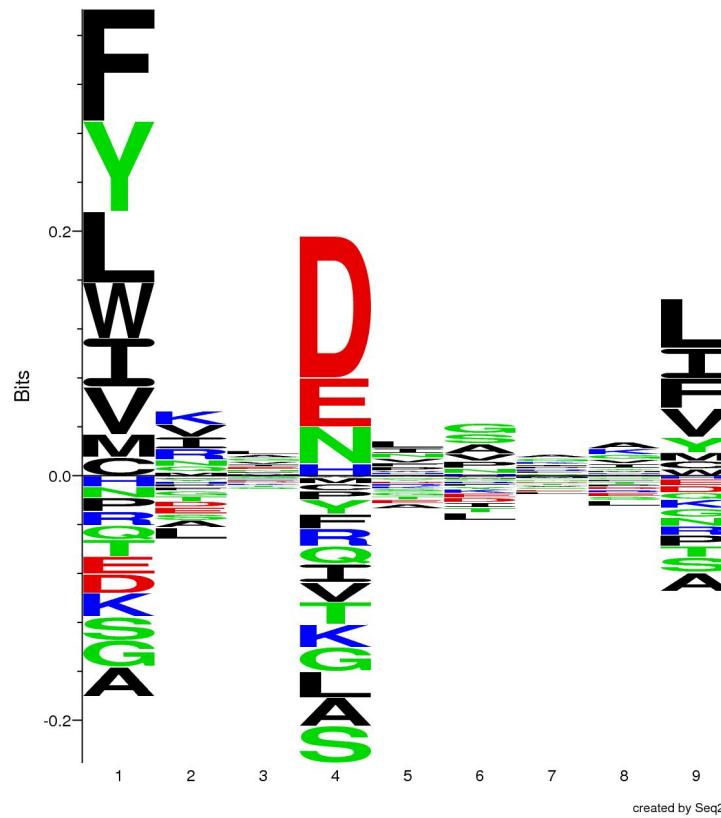
Monte Carlo probability



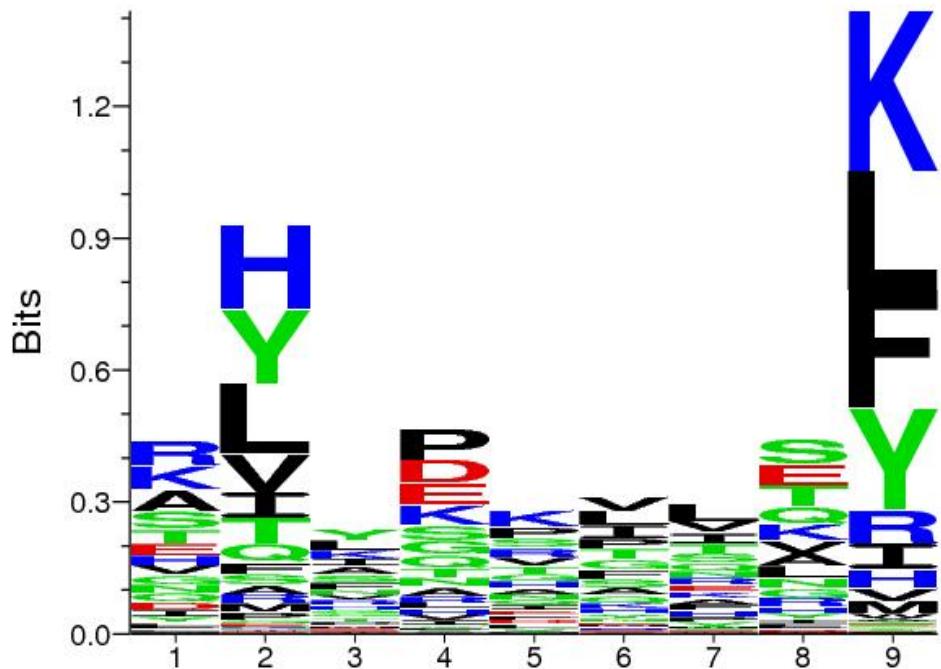
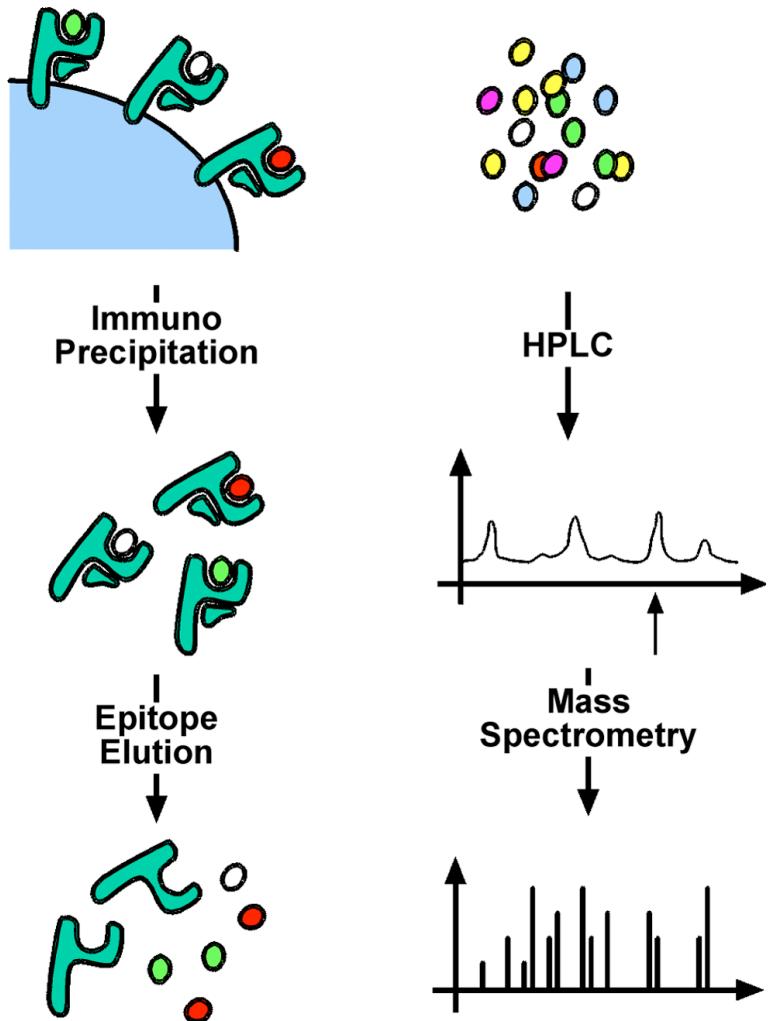
Does it work?



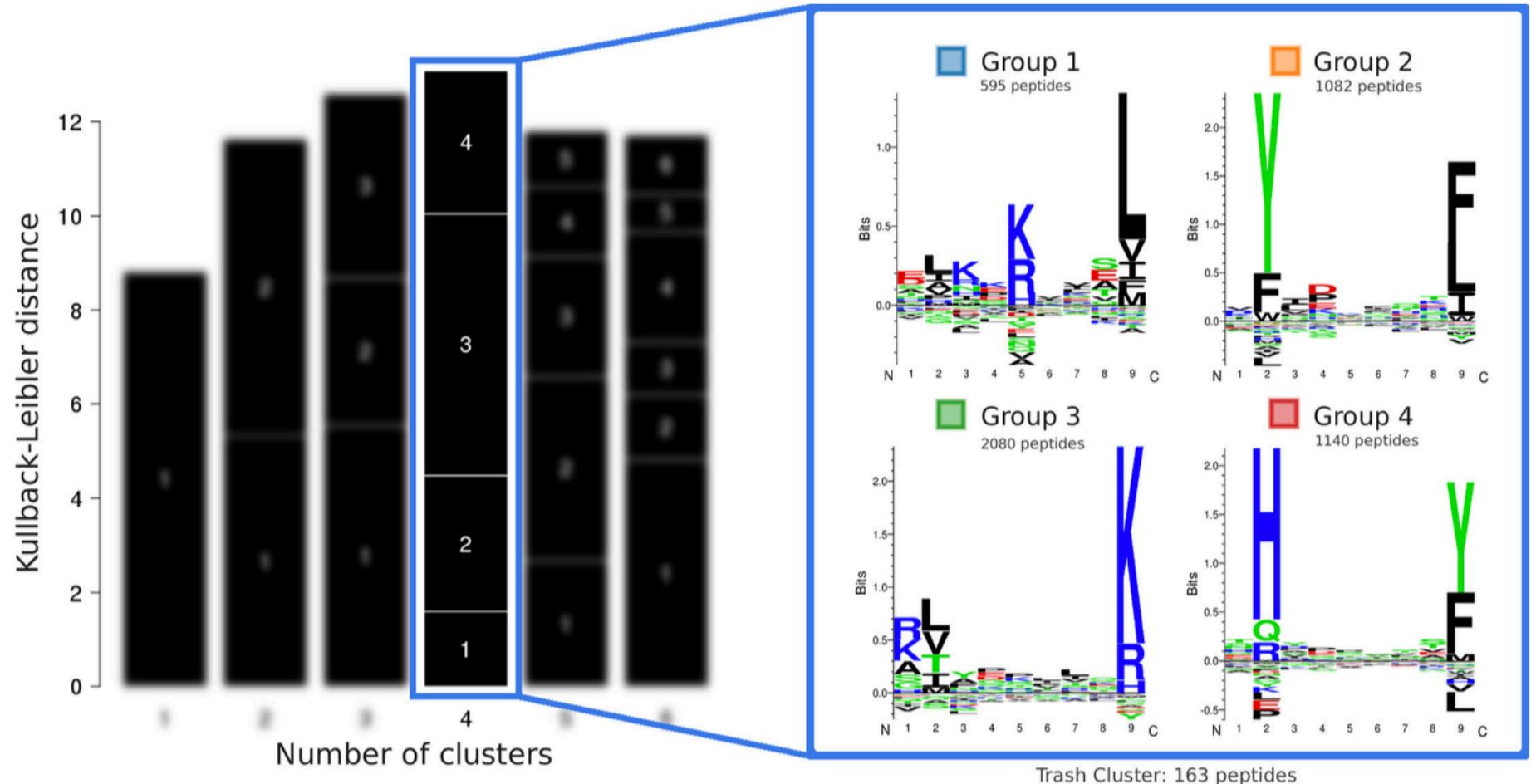
HLA-DRB3*01:01



Naturally presented MHC ligands



Dealing ligands of variable length



Summary

- Simple heuristics combined with conventional sequence motif identical can be used to characterized binding motifs in convoluted data sets (*GibbsCluster*)