# 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?

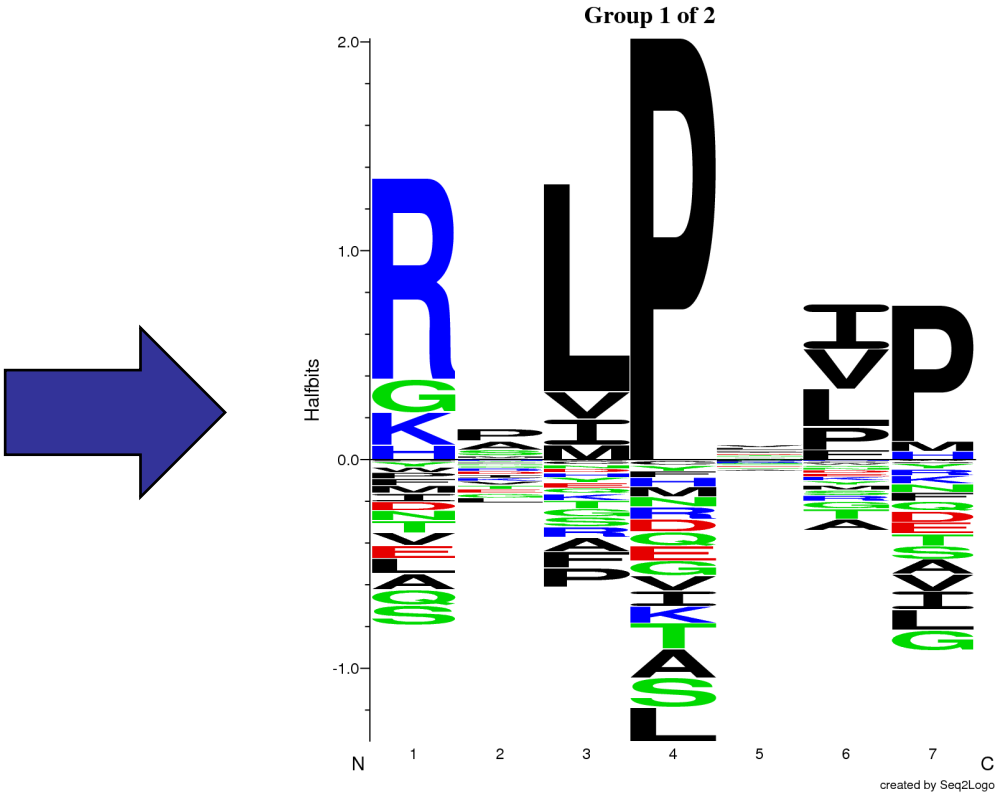


- The vast majority of biological motifs are characterized by a linear motif
  - Post translational modifications
  - Signal peptides
  - T cell epitopes
  - Transcription binding sites
  - SH2/SH3 domain binding
  - MHC binding
  - ….
- Predict impact of sequence variation (SNP)
- …

# Identifying binding motifs (SH3)



| Peptide |
| --- |
| LMLSLFEQSLSCQAQ |
| QGTDATKSIIFEAER |
| RLEEAQAYLAAGQHD |
| EISELRTKVQEQQKQ |
| FAGAKKIFGSLAFLP |
| VRASSRVSGSFPEDS |
| CKAFFKRSIQGHNDY |
| CEGCKAFFKRSIQGH |
| RLSEADIRGFVAAVV |



# Objectives



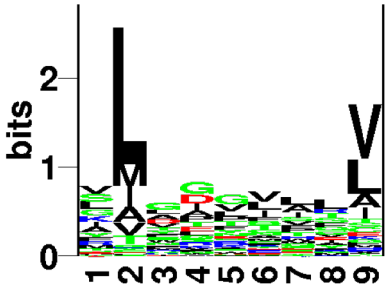
- Understand the concepts of weight matrix construction
  - One of the most important methods of bioinformatics
- Visualization of binding motifs
  - Construction of sequence logos
- How to construct a weight matrix
- How to use weight matrices to characterize receptor-ligand interactions
- Case story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

# Bioinformatics in a nutshell

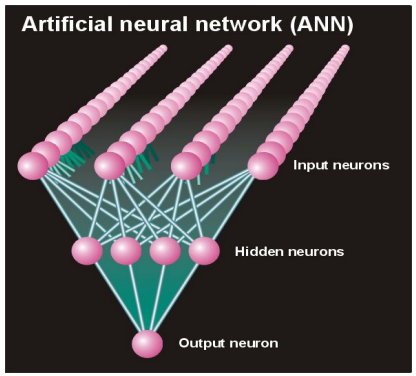


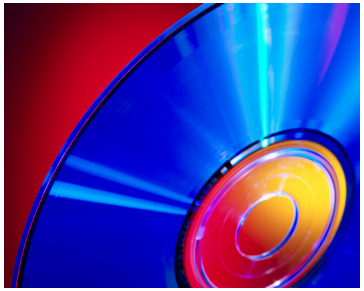
List of peptides that have a given biological feature

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
...

# MHC Class I pathway



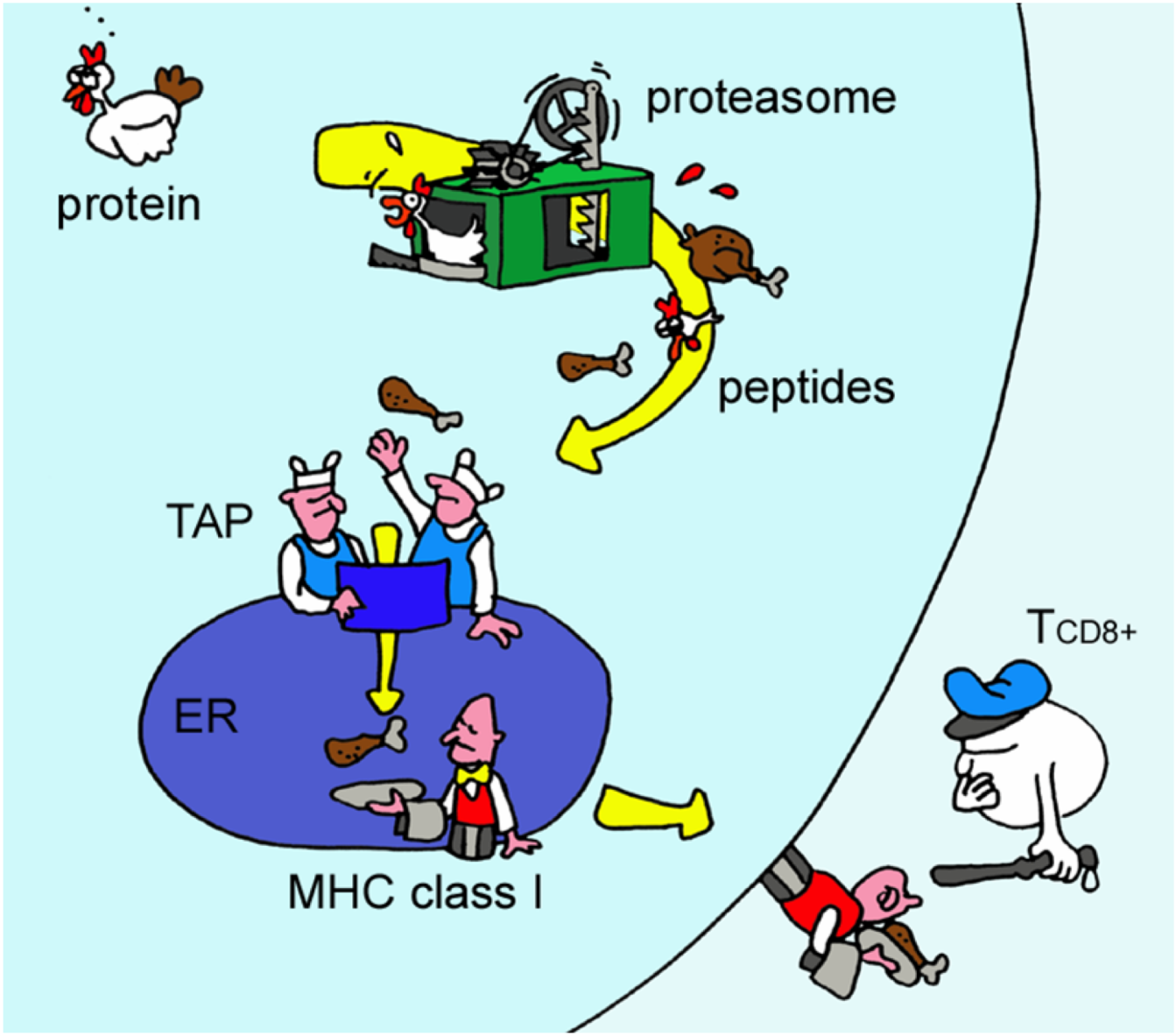
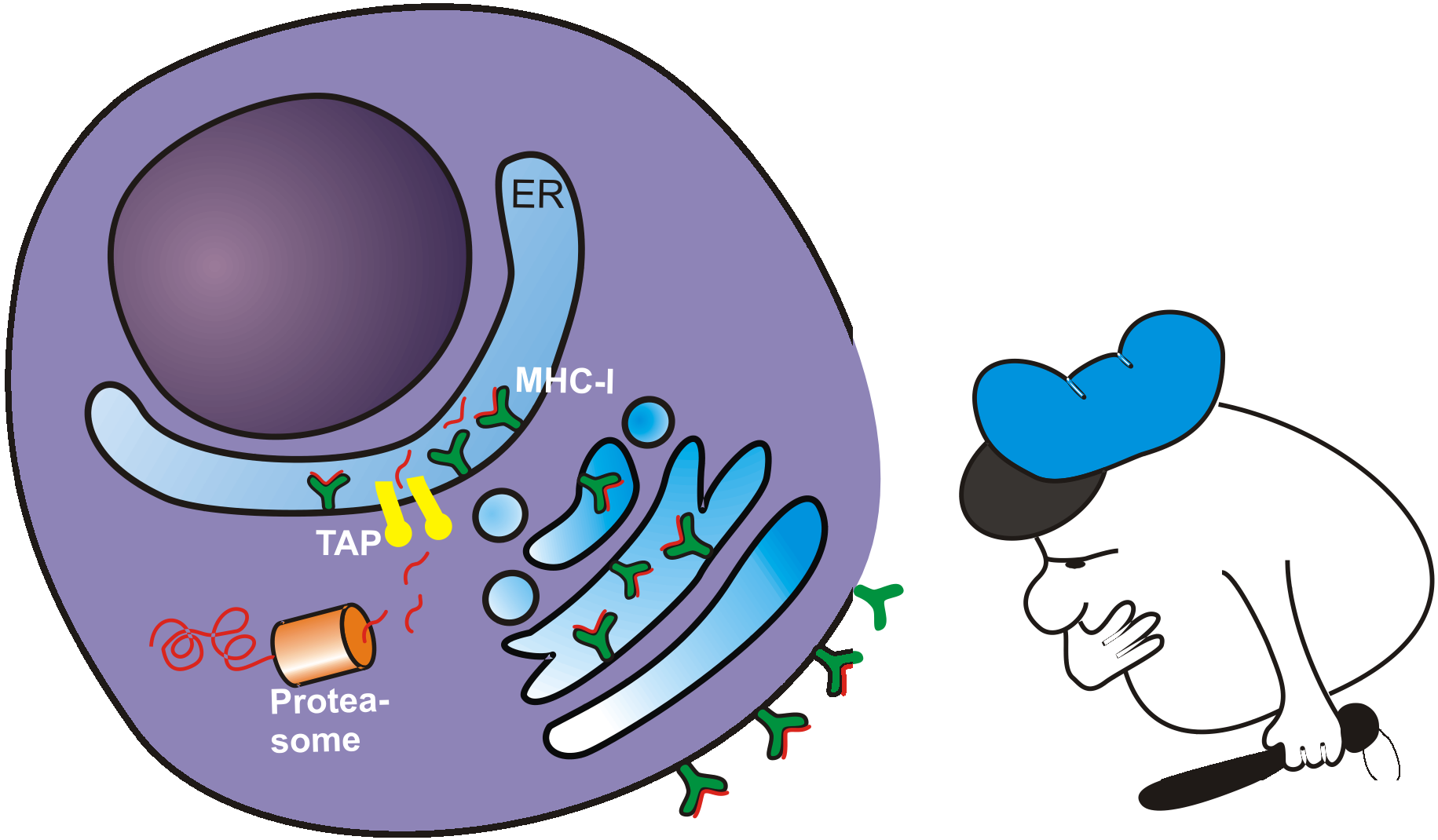



Figure by Eric A.J. Reits

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





# CTL response



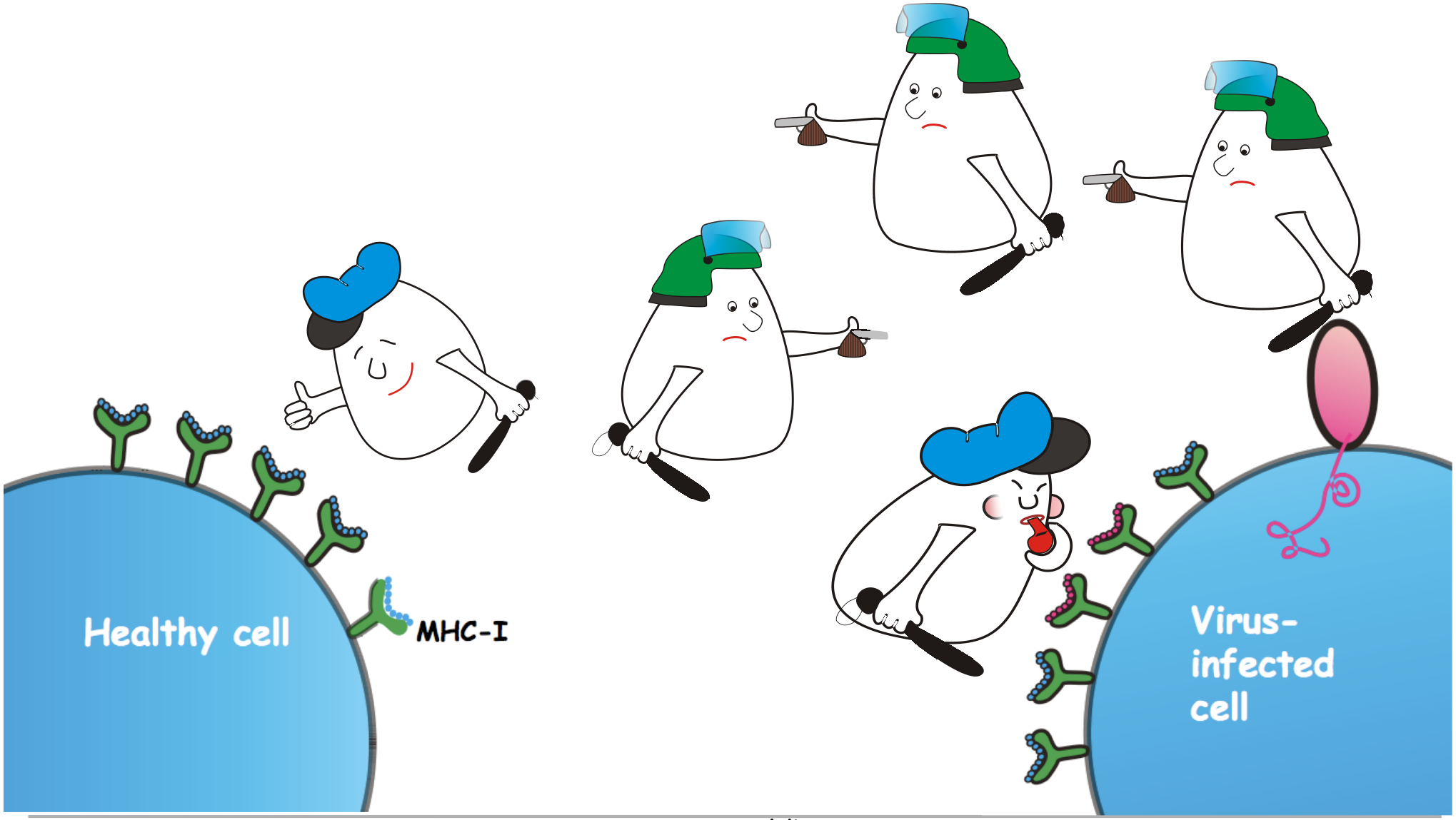
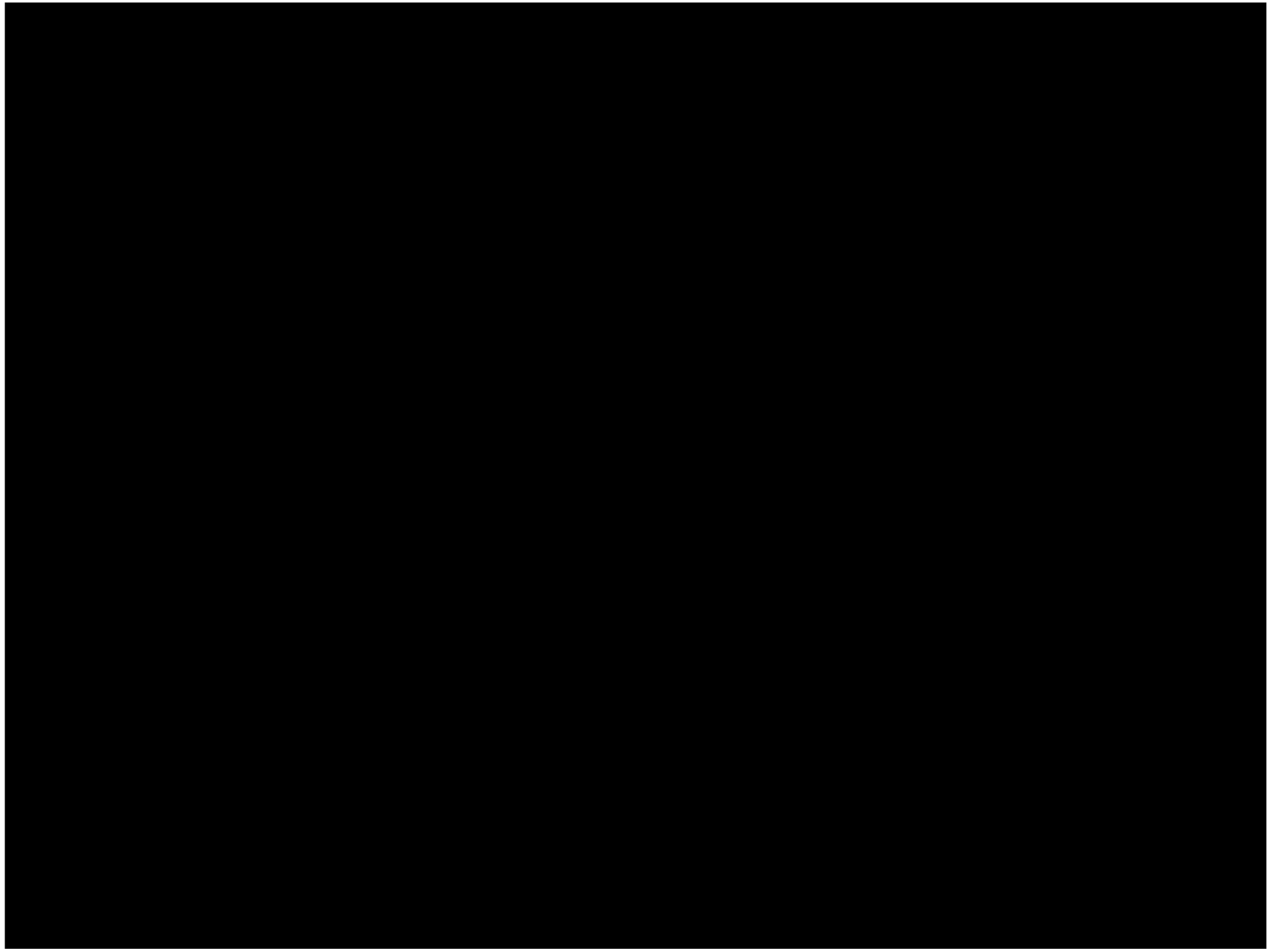



Figure courtesy Mette Voldby Larsen

# Encounter with death

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS



# 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 $P_2$?

**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 $P_2$?
- P1: 4 questions (at most)
- P2: 1 question (L or not)
- P2 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 $P_2$?
- P1: 4 questions (at most)
- P2: 1 question (L or not)
- P2 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*

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

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

# Sequence information

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

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 ILMEHIHKL 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 CINGVCWTV VMNILLQYV
ILTVILGVL KVLEYVIKV FLWGPRALV GLSRYVARL FLLTRILTI HLGNVKYLV GIAGGLALL GLQDCTMLV
TGAPVTYST VIYQYMDDL VLPDVFIRC VLPDVFIRC AVGIGIAVV LVVLGLLAV ALGLGLLPV GIGIGVLAA
GAGIGVAVL IAGIGILAI LIVIGILIL LAGIGLIAA VDGIGILTI GAGIGVLTA AAGIGIIQI QAGIGILLA
KARDPHSGH KACDPHSGH ACDPHSGHF SLYNTVATL RGPGRAFVT NLVPMVATV GLHCYEQLV PLKQHFQIV
AVFDRKSDA LLDFVRFMG VLVKSPNHV GLAPPQHLI LLGRNSFEV PLTFGWCYK VLEWRFDSR TLNAWVKVV
GLCTLVAML FIDSYICQV IISAVVGIL VMAGVGSPY LLWTLVVLL SVRDRLARL LLMDCSGSI CLTSTVQLV
VLHDDLLEA LMWITQCFL SLLMWITQC QLSLLMWIT LLGATCMFV RLTRFLSRV YMDGTMSQV FLTPKKLQC
ISNDVCAQV VKTDGNPPE SVYDFFVWL FLYGALLLA VLFSSDFRI LMWAKIGPV SLLLELEEV SLSRFSWGA
YTAFTIPSI RLMKQDFSV RLPRIFCSC FLWGPRAYA RLLQETELV SLFEGIDFY SLDQSVVEL RLNMFTPYI
NMFTPYIGV LMIIPLINV TLFIGSHVV SLVIVTTFV VLQWASLAV ILAKFLHWL STAPPHVNV LLLLTVLTV
VVLGVVFGI ILHNGAYSL MIMVKCWMI MLGTHTMEV MLGTHTMEV SLADTNSLA LLWAARPRL GVALQTMKQ
GLYDGMEHL KMVELVHFL YLQLVFGIE MLMAQEALA LMAQEALAF VYDGREHTV YLSGANLNL RMFPNAPYL
EAAGIGILT TLDSQVMSL STPPPGTRV KVAELVHFL IMIGVLVGV ALCRWGLLL LLFAGVQCQ VLLCESTAV
YLSTAFARV YLLEMLWRL SLDDYNHLV RTLDKVLEV GLPVEYLQV KLIANNTRV FIYAGSLSA KLVANNTRL
FLDEFMEGV ALQPGTALL VLDGLDVLL SLYSFPEPE ALYVDSLFF SLLQHLIGL ELTLGEFLK MINAYLDKL
AAGIGILTV FLPSDFFPS SVRDRLARL SLREWLLRI LLSAWILTA AAGIGILTV AVPDEIPPL FAYDGKDYI
AAGIGILTV FLPSDFFPS AAGIGILTV FLPSDFFPS AAGIGILTV FLWGPRALV ETVSEQSNV ITLWQRPLV

# Information content



| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V | I |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.09 | 0.06 | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 | 0.09 | 0.01 | 0.08 | 0.11 | 0.07 | 0.04 | 0.07 | 0.01 | 0.12 | 0.04 | 0.01 | 0.06 | 0.09 | 0.20 |
| 2 | 0.06 | 0.00 | 0.00 | 0.01 | 0.01 | 0.00 | 0.01 | 0.01 | 0.00 | 0.09 | 0.62 | 0.01 | 0.08 | 0.01 | 0.00 | 0.01 | 0.05 | 0.00 | 0.01 | 0.07 | 1.59 |
| 3 | 0.08 | 0.03 | 0.05 | 0.10 | 0.02 | 0.02 | 0.01 | 0.10 | 0.02 | 0.03 | 0.12 | 0.01 | 0.04 | 0.06 | 0.04 | 0.07 | 0.04 | 0.04 | 0.05 | 0.07 | 0.17 |
| 4 | 0.08 | 0.05 | 0.02 | 0.11 | 0.01 | 0.04 | 0.09 | 0.15 | 0.01 | 0.08 | 0.04 | 0.04 | 0.01 | 0.02 | 0.10 | 0.05 | 0.04 | 0.02 | 0.00 | 0.04 | 0.30 |
| 5 | 0.05 | 0.04 | 0.04 | 0.02 | 0.01 | 0.04 | 0.05 | 0.15 | 0.04 | 0.03 | 0.09 | 0.04 | 0.01 | 0.06 | 0.08 | 0.02 | 0.06 | 0.03 | 0.06 | 0.09 | 0.21 |
| 6 | 0.04 | 0.03 | 0.04 | 0.01 | 0.03 | 0.03 | 0.03 | 0.05 | 0.02 | 0.13 | 0.14 | 0.03 | 0.03 | 0.06 | 0.04 | 0.06 | 0.06 | 0.01 | 0.03 | 0.16 | 0.19 |
| 7 | 0.13 | 0.01 | 0.04 | 0.03 | 0.02 | 0.03 | 0.04 | 0.04 | 0.06 | 0.08 | 0.14 | 0.01 | 0.03 | 0.06 | 0.07 | 0.06 | 0.04 | 0.04 | 0.03 | 0.09 | 0.21 |
| 8 | 0.04 | 0.09 | 0.03 | 0.01 | 0.01 | 0.05 | 0.07 | 0.06 | 0.03 | 0.04 | 0.15 | 0.05 | 0.02 | 0.06 | 0.04 | 0.09 | 0.09 | 0.01 | 0.05 | 0.03 | 0.18 |
| 9 | 0.08 | 0.01 | 0.00 | 0.00 | 0.02 | 0.02 | 0.02 | 0.01 | 0.01 | 0.09 | 0.28 | 0.01 | 0.01 | 0.02 | 0.00 | 0.03 | 0.03 | 0.00 | 0.01 | 0.35 | 0.98 |

$$I = \log_2(20) + \sum_a p_a \cdot \log_2(p_a)$$ Shannon

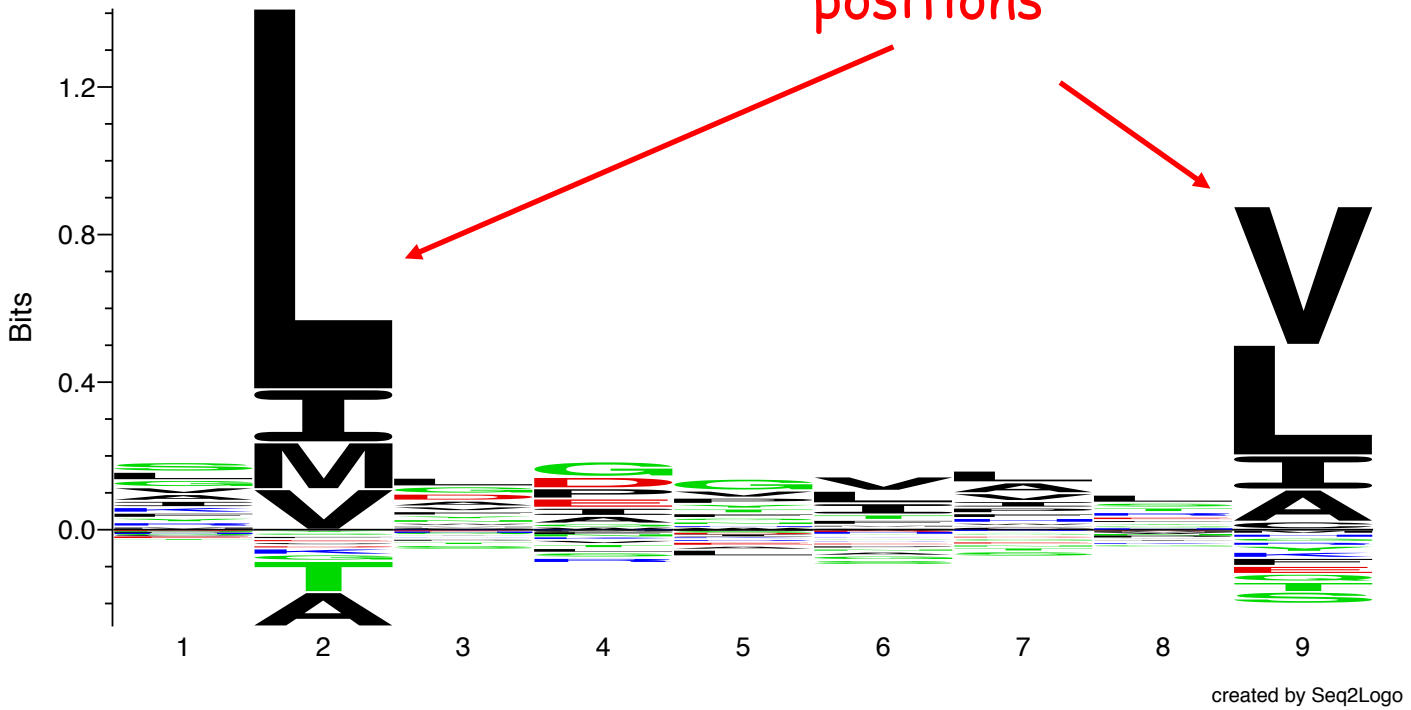
*or*

$$I = \sum_a p_a \cdot \log_2(\frac{p_a}{q_a})$$ Kullback - Leibler (== Shannon is q=0.05)

# Sequence logos



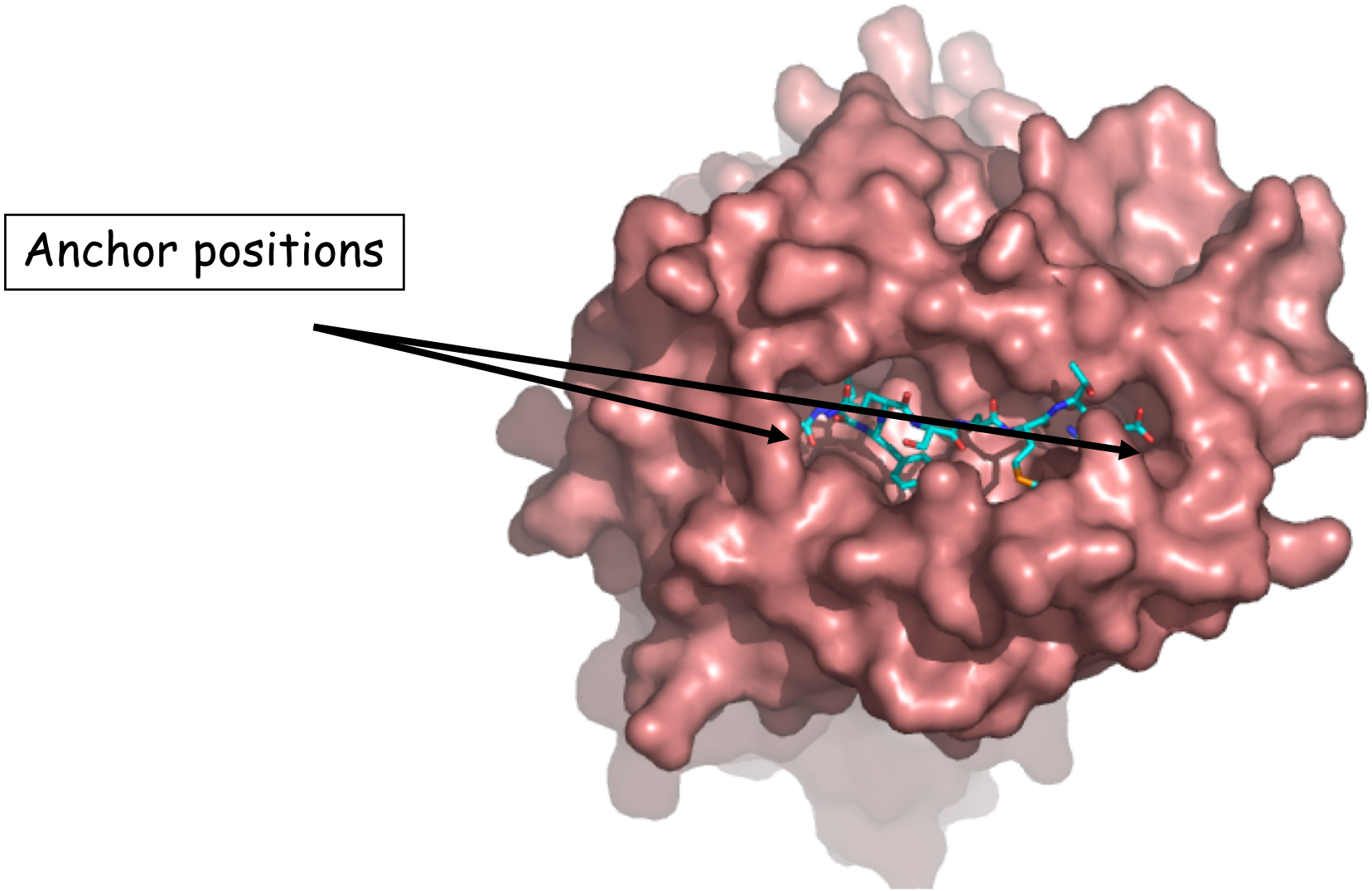
- Height of a column equal to I
- Relative height of a letter is p
- Highly useful tool to visualize sequence motifs




http://www.cbs.dtu.dk/biotools/Seq2Logo

# Binding Motif. MHC class I with peptide





An example!!
(See handout)

| position | 0 |
|---|---|
| Counts A | 0 |
| Counts T | 0 |
| Counts C | 0 |
| Counts G | 35 |
| P(A) | 0.0 |
| P(T) | 0.0 |
| P(c) | 0.0 |
| P(G) | 1.0 |

| position | **0** |
|---|---|
| Entropy | 0 |
| Information content | 2 |

# 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



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

| |
|---|
| ALAKAAAAM |
| ALAKAAAAN |
| ALAKAAAAR |
| ALAKAAAAT |
| ALAKAAAAV |
| GMNERPILT |
| GILGFVFTM |
| TLNAWVKVV |
| KLNEPVLLL |
| AVVPFIVSV |

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

- 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 |

$$[AGTK]_1[LMIV]_2[ANLV]_3 \ldots [AIFKLV]_7 \ldots [MNRTVL]_9$$

- Example

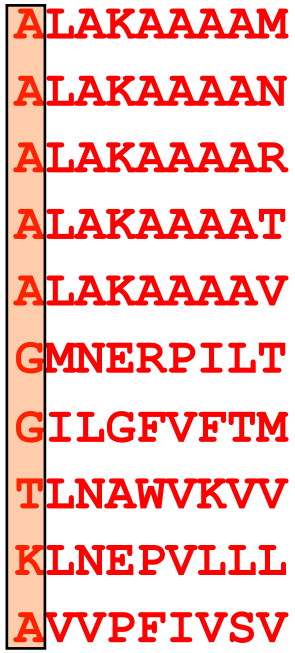
RLLDDTPEV 84 nM
GLLGNVSTV 23 nM
ALAKAAAAL 309 nM

- 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
  $P_A = 6/10$
  $P_G = 2/10$
  $P_T = P_K = 1/10$
  $P_C = P_D = \ldots P_V = 0$
- Problems
  - Few data
  - Data redundancy/duplication




**RLLDDTPEV 84 nM**
**GLLGNVSTV 23 nM**
**ALAKAAAAL 309 nM**

# Sequence information

## Raw sequence counting



# Sequence weighting



- Poor or biased sampling of sequence space
- Example P1
  - $P_A = 2/6$
  - $P_G = 2/6$
  - $P_T = P_K = 1/6$
  - $P_C = P_D = \ldots P_V = 0$

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

# log-odds to frequencies



$S_{ij}$

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 4 | -1 | -2 | -2 | 0 | -1 | -1 | 0 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 0 | -3 | -2 | 0 |
| R | -1 | 5 | 0 | -2 | -3 | 1 | 0 | -2 | 0 | -3 | -2 | 2 | -1 | -3 | -2 | -1 | -1 | -3 | -2 | -3 |
| N | -2 | 0 | 6 | 1 | -3 | 0 | 0 | 0 | 1 | -3 | -3 | 0 | -2 | -3 | -2 | 1 | 0 | -4 | -2 | -3 |
| D | -2 | -2 | 1 | 6 | -3 | 0 | 2 | -1 | -1 | -3 | -4 | -1 | -3 | -3 | -1 | 0 | -1 | -4 | -3 | -3 |
| C | 0 | -3 | -3 | -3 | 9 | -3 | -4 | -3 | -3 | -1 | -1 | -3 | -1 | -2 | -3 | -1 | -1 | -2 | -2 | -1 |
| Q | -1 | 1 | 0 | 0 | -3 | 5 | 2 | -2 | 0 | -3 | -2 | 1 | 0 | -3 | -1 | 0 | -1 | -2 | -1 | -2 |
| E | -1 | 0 | 0 | 2 | -4 | 2 | 5 | -2 | 0 | -3 | -3 | 1 | -2 | -3 | -1 | 0 | -1 | -3 | -2 | -2 |
| G | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | -2 | -4 | -4 | -2 | -3 | -3 | -2 | 0 | -2 | -2 | -3 | -3 |
| H | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | -3 | -3 | -1 | -2 | -1 | -2 | -1 | -2 | -2 | 2 | -3 |
| I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | 2 | -3 | 1 | 0 | -3 | -2 | -1 | -3 | -1 | 3 |
| L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | -2 | 2 | 0 | -3 | -2 | -1 | -2 | -1 | 1 |
| K | -1 | 2 | 0 | -1 | -3 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | -1 | -3 | -1 | 0 | -1 | -3 | -2 | -2 |
| M | -1 | -1 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | 0 | -2 | -1 | -1 | -1 | -1 | 1 |
| F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | -4 | -2 | -2 | 1 | 3 | -1 |
| P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | -1 | -1 | -4 | -3 | -2 |
| S | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | 1 | -3 | -2 | -2 |
| T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | -2 | -2 | 0 |
| W | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | 2 | -3 |
| Y | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | -1 |
| V | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | -2 | 0 | -3 | -1 | 4 |

$$S_{ij} = 2 \cdot \log_2\left(\frac{P_{ij}}{Q_i \cdot Q_j}\right)$$

# The way from log-odds to frequencies

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

$S_{ij}$

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 4 | -1 | -2 | -2 | 0 | -1 | -1 | 0 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 0 | -3 | -2 | 0 |
| R | -1 | 5 | 0 | -2 | -3 | 1 | 0 | -2 | 0 | -3 | -2 | 2 | -1 | -3 | -2 | -1 | -1 | -3 | -2 | -3 |
| N | -2 | 0 | 6 | 1 | -3 | 0 | 0 | 0 | 1 | -3 | -3 | 0 | -2 | -3 | -2 | 1 | 0 | -4 | -2 | -3 |
| D | -2 | -2 | 1 | 6 | -3 | 0 | 2 | -1 | -1 | -3 | -4 | -1 | -3 | -3 | -1 | 0 | -1 | -4 | -3 | -3 |
| C | 0 | -3 | -3 | -3 | 9 | -3 | -4 | -3 | -3 | -1 | -1 | -3 | -1 | -2 | -3 | -1 | -1 | -2 | -2 | -1 |
| Q | -1 | 1 | 0 | 0 | -3 | 5 | 2 | -2 | 0 | -3 | -2 | 1 | 0 | -3 | -1 | 0 | -1 | -2 | -1 | -2 |
| E | -1 | 0 | 0 | 2 | -4 | 2 | 5 | -2 | 0 | -3 | -3 | 1 | -2 | -3 | -1 | 0 | -1 | -3 | -2 | -2 |
| G | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | -2 | -4 | -4 | -2 | -3 | -3 | -2 | 0 | -2 | -2 | -3 | -3 |
| H | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | -3 | -3 | -1 | -2 | -1 | -2 | -1 | -2 | -2 | 2 | -3 |
| I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | 2 | -3 | 1 | 0 | -3 | -2 | -1 | -3 | -1 | 3 |
| L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | -2 | 2 | 0 | -3 | -2 | -1 | -2 | -1 | 1 |
| K | -1 | 2 | 0 | -1 | -3 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | -1 | -3 | -1 | 0 | -1 | -3 | -2 | -2 |
| M | -1 | -1 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | 0 | -2 | -1 | -1 | -1 | -1 | 1 |
| F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | -4 | -2 | -2 | 1 | 3 | -1 |
| P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | -1 | -1 | -4 | -3 | -2 |
| S | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | 1 | -3 | -2 | -2 |
| T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | -2 | -2 | 0 |
| W | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | 2 | -3 |
| Y | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | -1 |
| V | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | -2 | 0 | -3 | -1 | 4 |

P(j|i)

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0.29 | 0.03 | 0.03 | 0.03 | 0.02 | 0.03 | 0.04 | 0.08 | 0.01 | 0.04 | 0.06 | 0.04 | 0.02 | 0.02 | 0.03 | 0.09 | 0.05 | 0.01 | 0.02 | 0.07 |
| R | 0.04 | 0.34 | 0.04 | 0.03 | 0.01 | 0.05 | 0.05 | 0.03 | 0.02 | 0.02 | 0.05 | 0.12 | 0.02 | 0.02 | 0.02 | 0.04 | 0.03 | 0.01 | 0.02 | 0.03 |
| N | 0.04 | 0.04 | 0.32 | 0.08 | 0.01 | 0.03 | 0.05 | 0.07 | 0.03 | 0.02 | 0.03 | 0.05 | 0.01 | 0.02 | 0.02 | 0.07 | 0.05 | 0.00 | 0.02 | 0.03 |
| D | 0.04 | 0.03 | 0.07 | 0.40 | 0.01 | 0.03 | 0.09 | 0.05 | 0.02 | 0.02 | 0.03 | 0.04 | 0.01 | 0.01 | 0.02 | 0.05 | 0.04 | 0.00 | 0.01 | 0.02 |
| C | 0.07 | 0.02 | 0.02 | 0.02 | 0.48 | 0.01 | 0.02 | 0.03 | 0.01 | 0.04 | 0.07 | 0.02 | 0.02 | 0.02 | 0.02 | 0.04 | 0.04 | 0.00 | 0.01 | 0.06 |
| Q | 0.06 | 0.07 | 0.04 | 0.05 | 0.01 | 0.21 | 0.10 | 0.04 | 0.03 | 0.03 | 0.05 | 0.09 | 0.02 | 0.01 | 0.02 | 0.06 | 0.04 | 0.01 | 0.02 | 0.04 |
| E | 0.06 | 0.05 | 0.04 | 0.09 | 0.01 | 0.06 | 0.30 | 0.04 | 0.03 | 0.02 | 0.04 | 0.08 | 0.01 | 0.02 | 0.03 | 0.06 | 0.04 | 0.01 | 0.02 | 0.03 |
| G | 0.08 | 0.02 | 0.04 | 0.03 | 0.01 | 0.02 | 0.03 | 0.51 | 0.01 | 0.02 | 0.03 | 0.03 | 0.01 | 0.02 | 0.02 | 0.05 | 0.03 | 0.01 | 0.01 | 0.02 |
| H | 0.04 | 0.05 | 0.05 | 0.04 | 0.01 | 0.04 | 0.05 | 0.04 | 0.35 | 0.02 | 0.04 | 0.05 | 0.02 | 0.03 | 0.02 | 0.04 | 0.03 | 0.01 | 0.06 | 0.02 |
| I | 0.05 | 0.02 | 0.01 | 0.02 | 0.02 | 0.01 | 0.02 | 0.02 | 0.01 | 0.27 | 0.17 | 0.02 | 0.04 | 0.04 | 0.01 | 0.03 | 0.04 | 0.01 | 0.02 | 0.18 |
| L | 0.04 | 0.02 | 0.01 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.01 | 0.12 | 0.38 | 0.03 | 0.05 | 0.05 | 0.01 | 0.02 | 0.03 | 0.01 | 0.02 | 0.10 |
| K | 0.06 | 0.11 | 0.04 | 0.04 | 0.01 | 0.05 | 0.07 | 0.04 | 0.02 | 0.03 | 0.04 | 0.28 | 0.02 | 0.02 | 0.03 | 0.05 | 0.04 | 0.01 | 0.02 | 0.03 |
| M | 0.05 | 0.03 | 0.02 | 0.02 | 0.02 | 0.03 | 0.03 | 0.03 | 0.02 | 0.10 | 0.20 | 0.04 | 0.16 | 0.05 | 0.02 | 0.04 | 0.04 | 0.01 | 0.02 | 0.09 |
| F | 0.03 | 0.02 | 0.02 | 0.02 | 0.01 | 0.01 | 0.02 | 0.03 | 0.02 | 0.06 | 0.11 | 0.02 | 0.03 | 0.39 | 0.01 | 0.03 | 0.03 | 0.02 | 0.09 | 0.06 |
| P | 0.06 | 0.03 | 0.02 | 0.03 | 0.01 | 0.02 | 0.04 | 0.04 | 0.01 | 0.03 | 0.04 | 0.04 | 0.01 | 0.01 | 0.49 | 0.04 | 0.04 | 0.00 | 0.01 | 0.03 |
| S | 0.11 | 0.04 | 0.05 | 0.05 | 0.02 | 0.03 | 0.05 | 0.07 | 0.02 | 0.03 | 0.04 | 0.05 | 0.02 | 0.02 | 0.03 | 0.22 | 0.08 | 0.01 | 0.02 | 0.04 |
| T | 0.07 | 0.04 | 0.04 | 0.04 | 0.02 | 0.03 | 0.04 | 0.04 | 0.01 | 0.05 | 0.07 | 0.05 | 0.02 | 0.02 | 0.03 | 0.09 | 0.25 | 0.01 | 0.02 | 0.07 |
| W | 0.03 | 0.02 | 0.02 | 0.02 | 0.01 | 0.02 | 0.02 | 0.03 | 0.02 | 0.03 | 0.05 | 0.02 | 0.02 | 0.06 | 0.01 | 0.02 | 0.02 | 0.49 | 0.07 | 0.03 |
| Y | 0.04 | 0.03 | 0.02 | 0.02 | 0.01 | 0.02 | 0.03 | 0.02 | 0.05 | 0.04 | 0.07 | 0.03 | 0.02 | 0.13 | 0.02 | 0.03 | 0.03 | 0.03 | 0.32 | 0.05 |
| V | 0.07 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.01 | 0.16 | 0.13 | 0.03 | 0.03 | 0.04 | 0.02 | 0.03 | 0.05 | 0.01 | 0.02 | 0.27 |

$$S_{ij} = 2 \cdot \log_2\left(\frac{P_{ij}}{Q_i \cdot Q_j}\right) = 2 \cdot \log_2\left(\frac{P(j \mid i)}{Q_j}\right)$$

$$S_{AA} = 2 \cdot \log_2\left(\frac{P(A \mid A)}{Q_A}\right) = 2 \cdot \log_2\left(\frac{0.29}{0.074}\right) = 3.9$$

$$S_{AR} = 2 \cdot \log_2\left(\frac{P(R \mid A)}{Q_R}\right) = 2 \cdot \log_2\left(\frac{0.03}{0.052}\right) = -1.6$$

# The Blosum matrix

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0.29 | 0.03 | 0.03 | 0.03 | 0.02 | 0.03 | 0.04 | 0.08 | 0.01 | 0.04 | 0.06 | 0.04 | 0.02 | 0.02 | 0.03 | 0.09 | 0.05 | 0.01 | 0.02 | 0.07 |
| R | 0.04 | 0.34 | 0.04 | 0.03 | 0.01 | 0.05 | 0.05 | 0.03 | 0.02 | 0.02 | 0.05 | 0.12 | 0.02 | 0.02 | 0.02 | 0.04 | 0.03 | 0.01 | 0.02 | 0.03 |
| N | 0.04 | 0.04 | 0.32 | 0.08 | 0.01 | 0.03 | 0.05 | 0.07 | 0.03 | 0.02 | 0.03 | 0.05 | 0.01 | 0.02 | 0.02 | 0.07 | 0.05 | 0.00 | 0.02 | 0.03 |
| D | 0.04 | 0.03 | 0.07 | 0.40 | 0.01 | 0.03 | 0.09 | 0.05 | 0.02 | 0.02 | 0.03 | 0.04 | 0.01 | 0.01 | 0.02 | 0.05 | 0.04 | 0.00 | 0.01 | 0.02 |
| C | 0.07 | 0.02 | 0.02 | 0.02 | 0.48 | 0.01 | 0.02 | 0.03 | 0.01 | 0.04 | 0.07 | 0.02 | 0.02 | 0.02 | 0.02 | 0.04 | 0.04 | 0.00 | 0.01 | 0.06 |
| Q | 0.06 | 0.07 | 0.04 | 0.05 | 0.01 | 0.21 | 0.10 | 0.04 | 0.03 | 0.03 | 0.05 | 0.09 | 0.02 | 0.01 | 0.02 | 0.06 | 0.04 | 0.01 | 0.02 | 0.04 |
| E | 0.06 | 0.05 | 0.04 | 0.09 | 0.01 | 0.06 | 0.30 | 0.04 | 0.03 | 0.02 | 0.04 | 0.08 | 0.01 | 0.02 | 0.03 | 0.06 | 0.04 | 0.01 | 0.02 | 0.03 |
| G | 0.08 | 0.02 | 0.04 | 0.03 | 0.01 | 0.02 | 0.03 | 0.51 | 0.01 | 0.02 | 0.03 | 0.03 | 0.01 | 0.02 | 0.02 | 0.05 | 0.03 | 0.01 | 0.01 | 0.02 |
| H | 0.04 | 0.05 | 0.05 | 0.04 | 0.01 | 0.04 | 0.05 | 0.04 | 0.35 | 0.02 | 0.04 | 0.05 | 0.02 | 0.03 | 0.02 | 0.04 | 0.03 | 0.01 | 0.06 | 0.02 |
| I | 0.05 | 0.02 | 0.01 | 0.02 | 0.02 | 0.01 | 0.02 | 0.02 | 0.01 | 0.27 | 0.17 | 0.02 | 0.04 | 0.04 | 0.01 | 0.03 | 0.04 | 0.01 | 0.02 | 0.18 |
| L | 0.04 | 0.02 | 0.01 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.01 | 0.12 | 0.38 | 0.03 | 0.05 | 0.05 | 0.01 | 0.02 | 0.03 | 0.01 | 0.02 | 0.10 |
| K | 0.06 | 0.11 | 0.04 | 0.04 | 0.01 | 0.05 | 0.07 | 0.04 | 0.02 | 0.03 | 0.04 | 0.28 | 0.02 | 0.02 | 0.03 | 0.05 | 0.04 | 0.01 | 0.02 | 0.03 |
| M | 0.05 | 0.03 | 0.02 | 0.02 | 0.02 | 0.03 | 0.03 | 0.03 | 0.02 | 0.10 | 0.20 | 0.04 | 0.16 | 0.05 | 0.02 | 0.04 | 0.04 | 0.01 | 0.02 | 0.09 |
| F | 0.03 | 0.02 | 0.02 | 0.02 | 0.01 | 0.01 | 0.02 | 0.03 | 0.02 | 0.06 | 0.11 | 0.02 | 0.03 | 0.39 | 0.01 | 0.03 | 0.03 | 0.02 | 0.09 | 0.06 |
| P | 0.06 | 0.03 | 0.02 | 0.03 | 0.01 | 0.02 | 0.04 | 0.04 | 0.01 | 0.03 | 0.04 | 0.04 | 0.01 | 0.01 | 0.49 | 0.04 | 0.04 | 0.00 | 0.01 | 0.03 |
| S | 0.11 | 0.04 | 0.05 | 0.05 | 0.02 | 0.03 | 0.05 | 0.07 | 0.02 | 0.03 | 0.04 | 0.05 | 0.02 | 0.02 | 0.03 | 0.22 | 0.08 | 0.01 | 0.02 | 0.04 |
| T | 0.07 | 0.04 | 0.04 | 0.04 | 0.02 | 0.03 | 0.04 | 0.04 | 0.01 | 0.05 | 0.07 | 0.05 | 0.02 | 0.02 | 0.03 | 0.09 | 0.25 | 0.01 | 0.02 | 0.07 |
| W | 0.03 | 0.02 | 0.02 | 0.02 | 0.01 | 0.02 | 0.02 | 0.03 | 0.02 | 0.03 | 0.05 | 0.02 | 0.02 | 0.06 | 0.01 | 0.02 | 0.02 | 0.49 | 0.07 | 0.03 |
| Y | 0.04 | 0.03 | 0.02 | 0.02 | 0.01 | 0.02 | 0.03 | 0.02 | 0.05 | 0.04 | 0.07 | 0.03 | 0.02 | 0.13 | 0.02 | 0.03 | 0.03 | 0.03 | 0.32 | 0.05 |
| V | 0.07 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.01 | 0.16 | 0.13 | 0.03 | 0.03 | 0.04 | 0.02 | 0.03 | 0.05 | 0.01 | 0.02 | 0.27 |

Some amino acids are highly conserved (i.e. C),
some have a high change of mutation (i.e. I)

# What is a pseudo count?



| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0.29 | 0.03 | 0.03 | 0.03 | 0.02 | 0.03 | 0.04 | 0.08 | 0.01 | 0.04 | 0.06 | 0.04 | 0.02 | 0.02 | 0.03 | 0.09 | 0.05 | 0.01 | 0.02 | 0.07 |
| R | 0.04 | 0.34 | 0.04 | 0.03 | 0.01 | 0.05 | 0.05 | 0.03 | 0.02 | 0.02 | 0.05 | 0.12 | 0.02 | 0.02 | 0.02 | 0.04 | 0.03 | 0.01 | 0.02 | 0.03 |
| N | 0.04 | 0.04 | 0.32 | 0.08 | 0.01 | 0.03 | 0.05 | 0.07 | 0.03 | 0.02 | 0.03 | 0.05 | 0.01 | 0.02 | 0.02 | 0.07 | 0.05 | 0.00 | 0.02 | 0.03 |
| D | 0.04 | 0.03 | 0.07 | 0.40 | 0.01 | 0.03 | 0.09 | 0.05 | 0.02 | 0.02 | 0.03 | 0.04 | 0.01 | 0.01 | 0.02 | 0.05 | 0.04 | 0.00 | 0.01 | 0.02 |
| C | 0.07 | 0.02 | 0.02 | 0.02 | 0.48 | 0.01 | 0.02 | 0.03 | 0.01 | 0.04 | 0.07 | 0.02 | 0.02 | 0.02 | 0.02 | 0.04 | 0.04 | 0.00 | 0.01 | 0.06 |
| … | | | | | | | | | | | | | | | | | | | | |
| Y | 0.04 | 0.03 | 0.02 | 0.02 | 0.01 | 0.02 | 0.03 | 0.02 | 0.05 | 0.04 | 0.07 | 0.03 | 0.02 | 0.13 | 0.02 | 0.03 | 0.03 | 0.03 | 0.32 | 0.05 |
| V | 0.07 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.01 | 0.16 | 0.13 | 0.03 | 0.03 | 0.04 | 0.02 | 0.03 | 0.05 | 0.01 | 0.02 | 0.27 |

- Say V is observed at P2
- Knowing that V at P2 binds, what is the probability that a peptide could have I at P2?
- P(I|V) = 0.16

# 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

ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV

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

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

# Weight on pseudo count



ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV

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

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

- $\alpha$ is the effective number of sequences (N-1), $\beta$ is the **weight on prior** or **weight on pseudo count**

# Weight on pseudo count



- Example

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

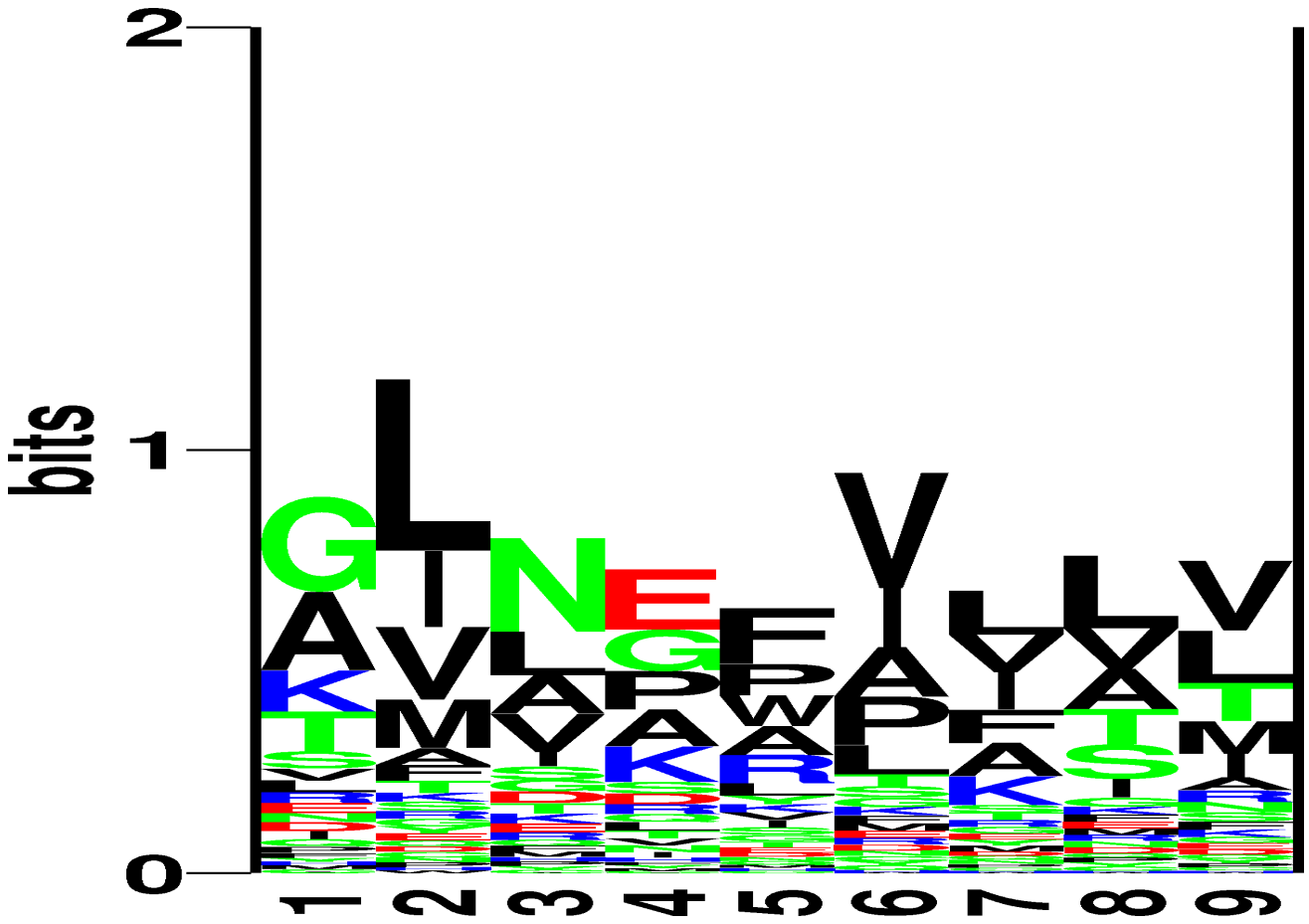
- If $\alpha$ large, p ≈ f and only the observed data defines the motif
- If $\alpha$ small, p ≈ g and the pseudo counts (or prior) defines the motif
- $\beta$ is [50-200] normally (see how $\beta$ is estimated later)

ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV

# Sequence weighting and pseudo counts



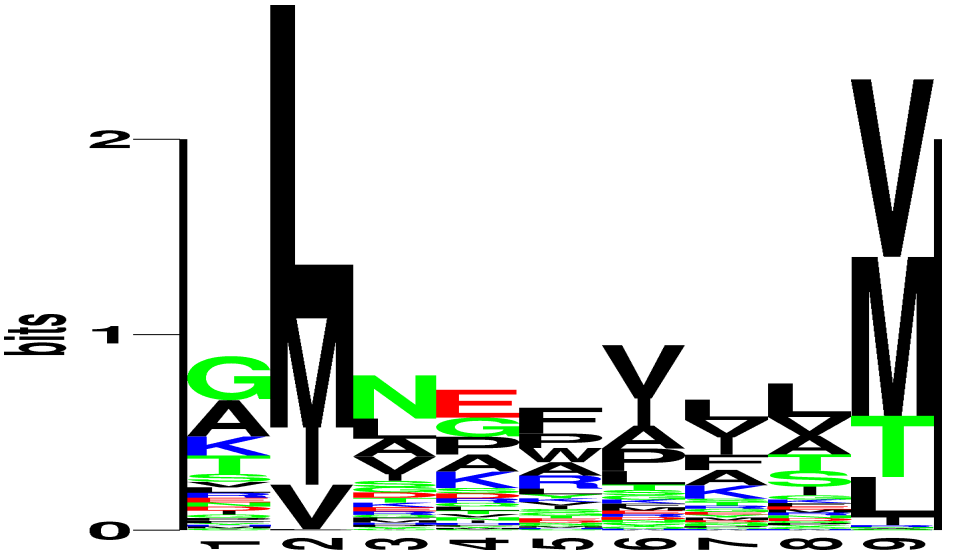
ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV



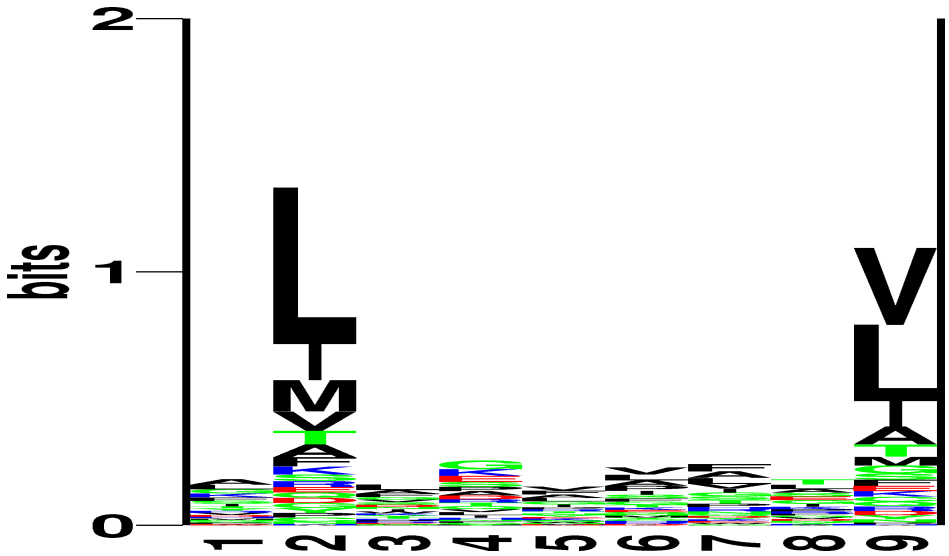
# Position specific weighting



- 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



# Weight matrices



- Estimate amino acid frequencies from alignment including sequence weighting and pseudo count

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.08 | 0.06 | 0.02 | 0.03 | 0.02 | 0.02 | 0.03 | 0.08 | 0.02 | 0.08 | 0.11 | 0.06 | 0.04 | 0.06 | 0.02 | 0.09 | 0.04 | 0.01 | 0.04 | 0.08 |
| 2 | 0.04 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 | 0.02 | 0.01 | 0.11 | 0.44 | 0.02 | 0.06 | 0.03 | 0.01 | 0.02 | 0.05 | 0.00 | 0.01 | 0.10 |
| 3 | 0.08 | 0.04 | 0.05 | 0.07 | 0.02 | 0.03 | 0.03 | 0.08 | 0.02 | 0.05 | 0.11 | 0.03 | 0.03 | 0.06 | 0.04 | 0.06 | 0.05 | 0.03 | 0.05 | 0.07 |
| 4 | 0.08 | 0.05 | 0.03 | 0.10 | 0.01 | 0.05 | 0.08 | 0.13 | 0.01 | 0.05 | 0.06 | 0.05 | 0.01 | 0.03 | 0.08 | 0.06 | 0.04 | 0.02 | 0.01 | 0.05 |
| 5 | 0.06 | 0.04 | 0.05 | 0.03 | 0.01 | 0.04 | 0.05 | 0.11 | 0.03 | 0.04 | 0.09 | 0.04 | 0.02 | 0.06 | 0.06 | 0.04 | 0.05 | 0.02 | 0.05 | 0.08 |
| 6 | 0.06 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.04 | 0.06 | 0.02 | 0.10 | 0.14 | 0.04 | 0.03 | 0.05 | 0.04 | 0.06 | 0.06 | 0.01 | 0.03 | 0.13 |
| 7 | 0.10 | 0.02 | 0.04 | 0.04 | 0.02 | 0.03 | 0.04 | 0.05 | 0.04 | 0.08 | 0.12 | 0.02 | 0.03 | 0.06 | 0.07 | 0.06 | 0.05 | 0.03 | 0.03 | 0.08 |
| 8 | 0.05 | 0.07 | 0.04 | 0.03 | 0.01 | 0.04 | 0.06 | 0.06 | 0.03 | 0.06 | 0.13 | 0.06 | 0.02 | 0.05 | 0.04 | 0.08 | 0.07 | 0.01 | 0.04 | 0.05 |
| 9 | 0.08 | 0.02 | 0.01 | 0.01 | 0.02 | 0.02 | 0.03 | 0.02 | 0.01 | 0.10 | 0.23 | 0.03 | 0.02 | 0.04 | 0.01 | 0.04 | 0.04 | 0.00 | 0.02 | 0.25 |

- What do the numbers mean?
  - P2(V)>P2(M). Does this mean that V enables binding more than M.
  - In nature not all amino acids are found equally often
    - In nature V is found more often than M, so we must somehow rescale with the background
    - $q_M = 0.025$, $q_V = 0.073$
    - Finding 7% V is hence not significant, but 7% M highly significant

# Weight matrices



- A weight matrix is given as

  $W_{ij} = \log(p_{ij}/q_j)$

  - where i is a position in the motif, and j an amino acid. $q_j$ is the background frequency for amino acid j.

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.6 | 0.4 | -3.5 | -2.4 | -0.4 | -1.9 | -2.7 | 0.3 | -1.1 | 1.0 | 0.3 | 0.0 | 1.4 | 1.2 | -2.7 | 1.4 | -1.2 | -2.0 | 1.1 | 0.7 |
| 2 | -1.6 | -6.6 | -6.5 | -5.4 | -2.5 | -4.0 | -4.7 | -3.7 | -6.3 | 1.0 | 5.1 | -3.7 | 3.1 | -4.2 | -4.3 | -4.2 | -0.2 | -5.9 | -3.8 | 0.4 |
| 3 | 0.2 | -1.3 | 0.1 | 1.5 | 0.0 | -1.8 | -3.3 | 0.4 | 0.5 | -1.0 | 0.3 | -2.5 | 1.2 | 1.0 | -0.1 | -0.3 | -0.5 | 3.4 | 1.6 | 0.0 |
| 4 | -0.1 | -0.1 | -2.0 | 2.0 | -1.6 | 0.5 | 0.8 | 2.0 | -3.3 | 0.1 | -1.7 | -1.0 | -2.2 | -1.6 | 1.7 | -0.6 | -0.2 | 1.3 | -6.8 | -0.7 |
| 5 | -1.6 | -0.1 | 0.1 | -2.2 | -1.2 | 0.4 | -0.5 | 1.9 | 1.2 | -2.2 | -0.5 | -1.3 | -2.2 | 1.7 | 1.2 | -2.5 | -0.1 | 1.7 | 1.5 | 1.0 |
| 6 | -0.7 | -1.4 | -1.0 | -2.3 | 1.1 | -1.3 | -1.4 | -0.2 | -1.0 | 1.8 | 0.8 | -1.9 | 0.2 | 1.0 | -0.4 | -0.6 | 0.4 | -0.5 | -0.0 | 2.1 |
| 7 | 1.1 | -3.8 | -0.2 | -1.3 | 1.3 | -0.3 | -1.3 | -1.4 | 2.1 | 0.6 | 0.7 | -5.0 | 1.1 | 0.9 | 1.3 | -0.5 | -0.9 | 2.9 | -0.4 | 0.5 |
| 8 | -2.2 | 1.0 | -0.8 | -2.9 | -1.4 | 0.4 | 0.1 | -0.4 | 0.2 | -0.0 | 1.1 | -0.5 | -0.5 | 0.7 | -0.3 | 0.8 | 0.8 | -0.7 | 1.3 | -1.1 |
| 9 | -0.2 | -3.5 | -6.1 | -4.5 | 0.7 | -0.8 | -2.5 | -4.0 | -2.6 | 0.9 | 2.8 | -3.0 | -1.8 | -1.4 | -6.2 | -1.9 | -1.6 | -4.9 | -1.6 | 4.5 |

- W is a L x 20 matrix, L is motif length

# Scoring a sequence to a weight matrix



- Score sequences to weight matrix by looking up and adding L values from the matrix

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.6 | 0.4 | -3.5 | -2.4 | -0.4 | -1.9 | -2.7 | 0.3 | -1.1 | 1.0 | 0.3 | 0.0 | 1.4 | 1.2 | -2.7 | 1.4 | -1.2 | -2.0 | 1.1 | 0.7 |
| 2 | -1.6 | -6.8 | -6.5 | -5.4 | -2.5 | -4.0 | -4.7 | -3.7 | -6.3 | 1.0 | 5.1 | -3.7 | 3.1 | -4.2 | -4.3 | -4.2 | -0.2 | -5.9 | -3.8 | 0.4 |
| 3 | 0.2 | -1.3 | 0.1 | 1.5 | 0.0 | -1.8 | -3.3 | 0.4 | 0.5 | -1.0 | 0.3 | -2.5 | 1.2 | 1.0 | -0.1 | -0.3 | -0.5 | 3.4 | 1.6 | 0.0 |
| 4 | -0.1 | -0.1 | -2.0 | 2.0 | -1.6 | 0.5 | 0.8 | 2.0 | -3.3 | 0.1 | -1.7 | -1.0 | -2.2 | -1.6 | 1.7 | -0.6 | -0.2 | 1.3 | -6.8 | -0.7 |
| 5 | -1.6 | -0.1 | 0.1 | -2.2 | -1.2 | 0.4 | -0.5 | 1.9 | 1.2 | -2.2 | -0.5 | -1.3 | -2.2 | 1.7 | 1.2 | -2.5 | -0.1 | 1.7 | 1.5 | 1.0 |
| 6 | -0.7 | -1.4 | -1.0 | -2.3 | 1.1 | -1.3 | -1.4 | -0.2 | -1.0 | 1.8 | 0.8 | -1.9 | 0.2 | 1.0 | -0.4 | -0.6 | 0.4 | -0.5 | -0.0 | 2.1 |
| 7 | 1.1 | -3.8 | -0.2 | -1.3 | 1.3 | -0.3 | -1.3 | -1.4 | 2.1 | 0.6 | 0.7 | -5.0 | 1.1 | 0.9 | 1.3 | -0.5 | -0.9 | 2.9 | -0.4 | 0.5 |
| 8 | -2.2 | 1.0 | -0.8 | -2.9 | -1.4 | 0.4 | 0.1 | -0.4 | 0.2 | -0.0 | 1.1 | -0.5 | -0.5 | 0.7 | -0.3 | 0.8 | 0.8 | -0.7 | 1.3 | -1.1 |
| 9 | -0.2 | -3.5 | -6.1 | -4.5 | 0.7 | -0.8 | -2.5 | -4.0 | -2.6 | 0.9 | 2.8 | -3.0 | -1.8 | -1.4 | -6.2 | -1.9 | -1.6 | -4.9 | -1.6 | 4.5 |

| | | |
|---|---|---|
| RLLDDTPEV | 11.9 | 84nM |
| GLLGNVSTV | 14.7 | 23nM |
| ALAKAAAAL | 4.3 | 309nM |

Which peptide is most likely to bind?
Which peptide second?

An example!!
(See handout)

Say you want to calculate the values for A. Then

$$f(A) = 0.0$$

$$g(A) = f(E)*q(A|E)+f(Q)*q(A|Q) = 5/6*0.06+1/6*0.06 = 0.06$$
$$p(A) = (5*0.0+5*0.06)/10 = 0.03$$
$$w(A) = 2*\log(0.03/0.074)/\log(2) = -2.61$$

since the other 18 amino acids have frequency values of 0.

| | $f_a$ | $g_a$ | $p_a$ | $w_a$ |
|---|---|---|---|---|
| A | 0 | 0.06 | 0.03 | -2.61 |

P(J|I)

J

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0.29 | 0.03 | 0.03 | 0.03 | 0.02 | 0.03 | 0.04 | 0.08 | 0.01 | 0.04 | 0.06 | 0.04 | 0.02 | 0.02 | 0.03 | 0.09 | 0.05 | 0.01 | 0.02 | 0.07 |
| R | 0.04 | 0.34 | 0.04 | 0.03 | 0.01 | 0.05 | 0.05 | 0.03 | 0.02 | 0.02 | 0.05 | 0.12 | 0.02 | 0.02 | 0.02 | 0.04 | 0.03 | 0.01 | 0.02 | 0.03 |
| N | 0.04 | 0.04 | 0.32 | 0.08 | 0.01 | 0.03 | 0.05 | 0.07 | 0.03 | 0.02 | 0.03 | 0.05 | 0.01 | 0.02 | 0.02 | 0.07 | 0.05 | 0.00 | 0.02 | 0.03 |
| D | 0.04 | 0.03 | 0.07 | 0.40 | 0.01 | 0.03 | 0.09 | 0.05 | 0.02 | 0.02 | 0.03 | 0.04 | 0.01 | 0.01 | 0.02 | 0.05 | 0.04 | 0.00 | 0.01 | 0.02 |
| C | 0.07 | 0.02 | 0.02 | 0.02 | 0.48 | 0.01 | 0.02 | 0.03 | 0.01 | 0.04 | 0.07 | 0.02 | 0.02 | 0.02 | 0.02 | 0.04 | 0.04 | 0.00 | 0.01 | 0.06 |
| Q | 0.06 | 0.07 | 0.04 | 0.05 | 0.01 | 0.21 | 0.10 | 0.04 | 0.03 | 0.03 | 0.05 | 0.09 | 0.02 | 0.01 | 0.02 | 0.06 | 0.04 | 0.01 | 0.02 | 0.04 |
| E | 0.06 | 0.05 | 0.04 | 0.09 | 0.01 | 0.06 | 0.30 | 0.04 | 0.03 | 0.02 | 0.04 | 0.08 | 0.01 | 0.02 | 0.03 | 0.06 | 0.04 | 0.01 | 0.02 | 0.03 |
| G | 0.08 | 0.02 | 0.04 | 0.03 | 0.01 | 0.02 | 0.03 | 0.51 | 0.01 | 0.02 | 0.03 | 0.03 | 0.01 | 0.02 | 0.02 | 0.05 | 0.03 | 0.01 | 0.01 | 0.02 |
| H | 0.04 | 0.05 | 0.05 | 0.04 | 0.01 | 0.04 | 0.05 | 0.04 | 0.35 | 0.02 | 0.04 | 0.05 | 0.02 | 0.03 | 0.02 | 0.04 | 0.03 | 0.01 | 0.06 | 0.02 |
| I | 0.05 | 0.02 | 0.01 | 0.02 | 0.02 | 0.01 | 0.02 | 0.02 | 0.01 | 0.27 | 0.17 | 0.02 | 0.04 | 0.04 | 0.01 | 0.03 | 0.04 | 0.01 | 0.02 | 0.18 |
| L | 0.04 | 0.02 | 0.01 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.01 | 0.12 | 0.38 | 0.03 | 0.05 | 0.05 | 0.01 | 0.02 | 0.03 | 0.01 | 0.02 | 0.10 |
| K | 0.06 | 0.11 | 0.04 | 0.04 | 0.01 | 0.05 | 0.07 | 0.04 | 0.02 | 0.03 | 0.04 | 0.28 | 0.02 | 0.02 | 0.03 | 0.05 | 0.04 | 0.01 | 0.02 | 0.03 |
| M | 0.05 | 0.03 | 0.02 | 0.02 | 0.02 | 0.03 | 0.03 | 0.03 | 0.02 | 0.10 | 0.20 | 0.04 | 0.16 | 0.05 | 0.02 | 0.04 | 0.04 | 0.01 | 0.02 | 0.09 |
| F | 0.03 | 0.02 | 0.02 | 0.02 | 0.01 | 0.01 | 0.02 | 0.03 | 0.02 | 0.06 | 0.11 | 0.02 | 0.03 | 0.39 | 0.01 | 0.03 | 0.03 | 0.02 | 0.09 | 0.06 |
| P | 0.06 | 0.03 | 0.02 | 0.03 | 0.01 | 0.02 | 0.04 | 0.04 | 0.01 | 0.03 | 0.04 | 0.04 | 0.01 | 0.01 | 0.49 | 0.04 | 0.04 | 0.00 | 0.01 | 0.03 |
| S | 0.11 | 0.04 | 0.05 | 0.05 | 0.02 | 0.03 | 0.05 | 0.07 | 0.02 | 0.03 | 0.04 | 0.05 | 0.02 | 0.02 | 0.03 | 0.22 | 0.08 | 0.01 | 0.02 | 0.04 |
| T | 0.07 | 0.04 | 0.04 | 0.04 | 0.02 | 0.03 | 0.04 | 0.04 | 0.01 | 0.05 | 0.07 | 0.05 | 0.02 | 0.02 | 0.03 | 0.09 | 0.25 | 0.01 | 0.02 | 0.07 |
| W | 0.03 | 0.02 | 0.02 | 0.02 | 0.01 | 0.02 | 0.02 | 0.03 | 0.02 | 0.03 | 0.05 | 0.02 | 0.02 | 0.06 | 0.01 | 0.02 | 0.02 | 0.49 | 0.07 | 0.03 |
| Y | 0.04 | 0.03 | 0.02 | 0.02 | 0.01 | 0.02 | 0.03 | 0.02 | 0.05 | 0.04 | 0.07 | 0.03 | 0.02 | 0.13 | 0.02 | 0.03 | 0.03 | 0.03 | 0.32 | 0.05 |
| V | 0.07 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.01 | 0.16 | 0.13 | 0.03 | 0.03 | 0.04 | 0.02 | 0.03 | 0.05 | 0.01 | 0.02 | 0.27 |

# Background frequencies ($q_a$)

```
A 0.07400
R 0.05200
N 0.04500
D 0.05400
```

# 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 |

# Information content (Kullback-Leibler)



KL == Shannon if q=0.05 for all AA

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V | I |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.09 | 0.06 | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 | 0.09 | 0.01 | 0.08 | 0.11 | 0.07 | 0.04 | 0.07 | 0.01 | 0.12 | 0.04 | 0.01 | 0.06 | 0.09 | 0.20 |
| 2 | 0.06 | 0.00 | 0.00 | 0.01 | 0.01 | 0.00 | 0.01 | 0.01 | 0.00 | 0.09 | 0.62 | 0.01 | 0.08 | 0.01 | 0.00 | 0.01 | 0.05 | 0.00 | 0.01 | 0.07 | 1.59 |
| 3 | 0.08 | 0.03 | 0.05 | 0.10 | 0.02 | 0.02 | 0.01 | 0.10 | 0.02 | 0.03 | 0.12 | 0.01 | 0.04 | 0.06 | 0.04 | 0.07 | 0.04 | 0.04 | 0.05 | 0.07 | 0.17 |
| 4 | 0.08 | 0.05 | 0.02 | 0.11 | 0.01 | 0.04 | 0.09 | 0.15 | 0.01 | 0.08 | 0.04 | 0.04 | 0.01 | 0.02 | 0.10 | 0.05 | 0.04 | 0.02 | 0.00 | 0.04 | 0.30 |
| 5 | 0.05 | 0.04 | 0.04 | 0.02 | 0.01 | 0.04 | 0.05 | 0.15 | 0.04 | 0.03 | 0.09 | 0.04 | 0.01 | 0.06 | 0.08 | 0.02 | 0.06 | 0.03 | 0.06 | 0.09 | 0.21 |
| 6 | 0.04 | 0.03 | 0.04 | 0.01 | 0.03 | 0.03 | 0.03 | 0.05 | 0.02 | 0.13 | 0.14 | 0.03 | 0.03 | 0.06 | 0.04 | 0.06 | 0.06 | 0.01 | 0.03 | 0.16 | 0.19 |
| 7 | 0.13 | 0.01 | 0.04 | 0.03 | 0.02 | 0.03 | 0.04 | 0.04 | 0.06 | 0.08 | 0.14 | 0.01 | 0.03 | 0.06 | 0.07 | 0.06 | 0.04 | 0.04 | 0.03 | 0.09 | 0.21 |
| 8 | 0.04 | 0.09 | 0.03 | 0.01 | 0.01 | 0.05 | 0.07 | 0.06 | 0.03 | 0.04 | 0.15 | 0.05 | 0.02 | 0.06 | 0.04 | 0.09 | 0.09 | 0.01 | 0.05 | 0.03 | 0.18 |
| 9 | 0.08 | 0.01 | 0.00 | 0.00 | 0.02 | 0.02 | 0.02 | 0.01 | 0.01 | 0.09 | 0.28 | 0.01 | 0.01 | 0.02 | 0.00 | 0.03 | 0.03 | 0.00 | 0.01 | 0.35 | 0.98 |

$$I = \log_2(20) + \sum_a p_a \cdot \log_2(p_a)$$

Shannon

*or*

$$I = \sum_a p_a \cdot \log_2\left(\frac{p_a}{q_a}\right) = \sum_a p_a \cdot (\log_2(p_a) - \log_2(q_a))$$

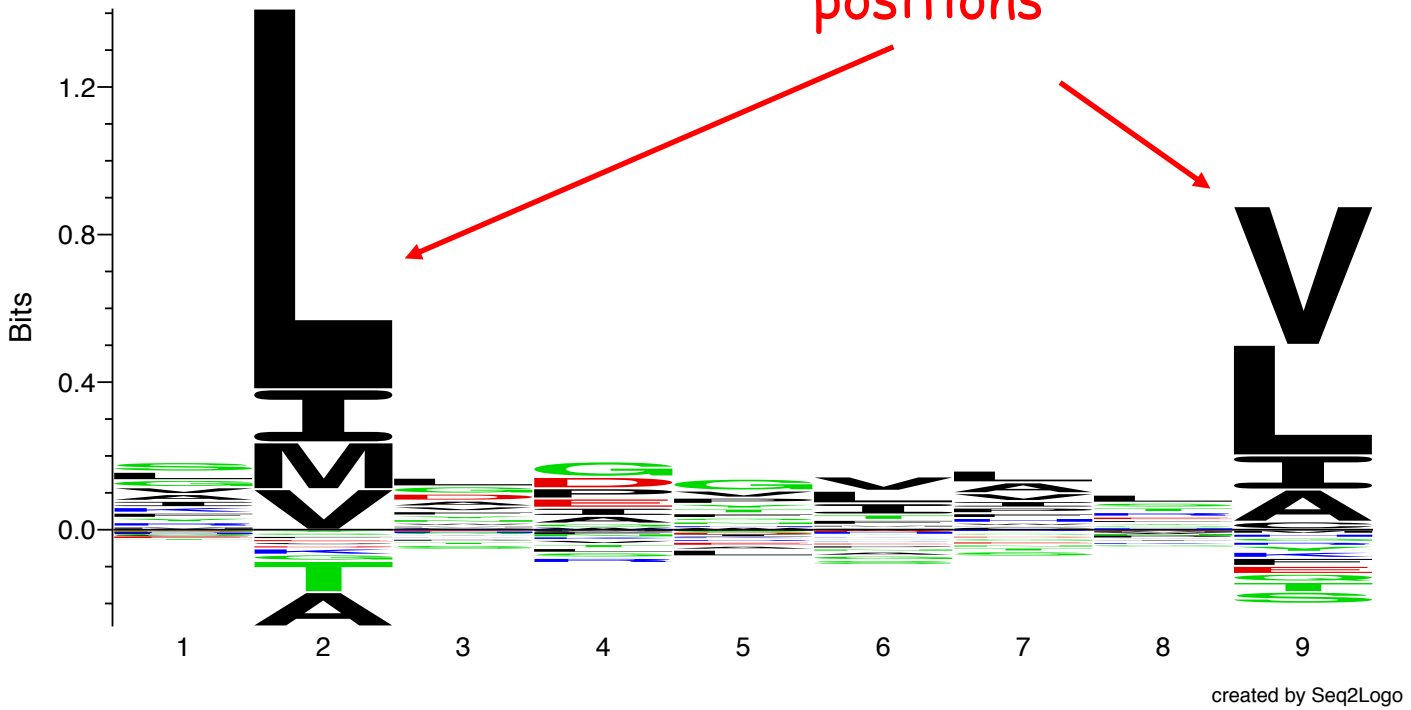
Kullback - Leibler

$$= -\log_2(0.05) + \sum_a p_a \cdot (\log_2(p_a) = \log_2(20) + \sum_a p_a \cdot (\log_2(p_a)$$

# Sequence logos



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




http://www.cbs.dtu.dk/biotools/Seq2Logo

# Example from real life

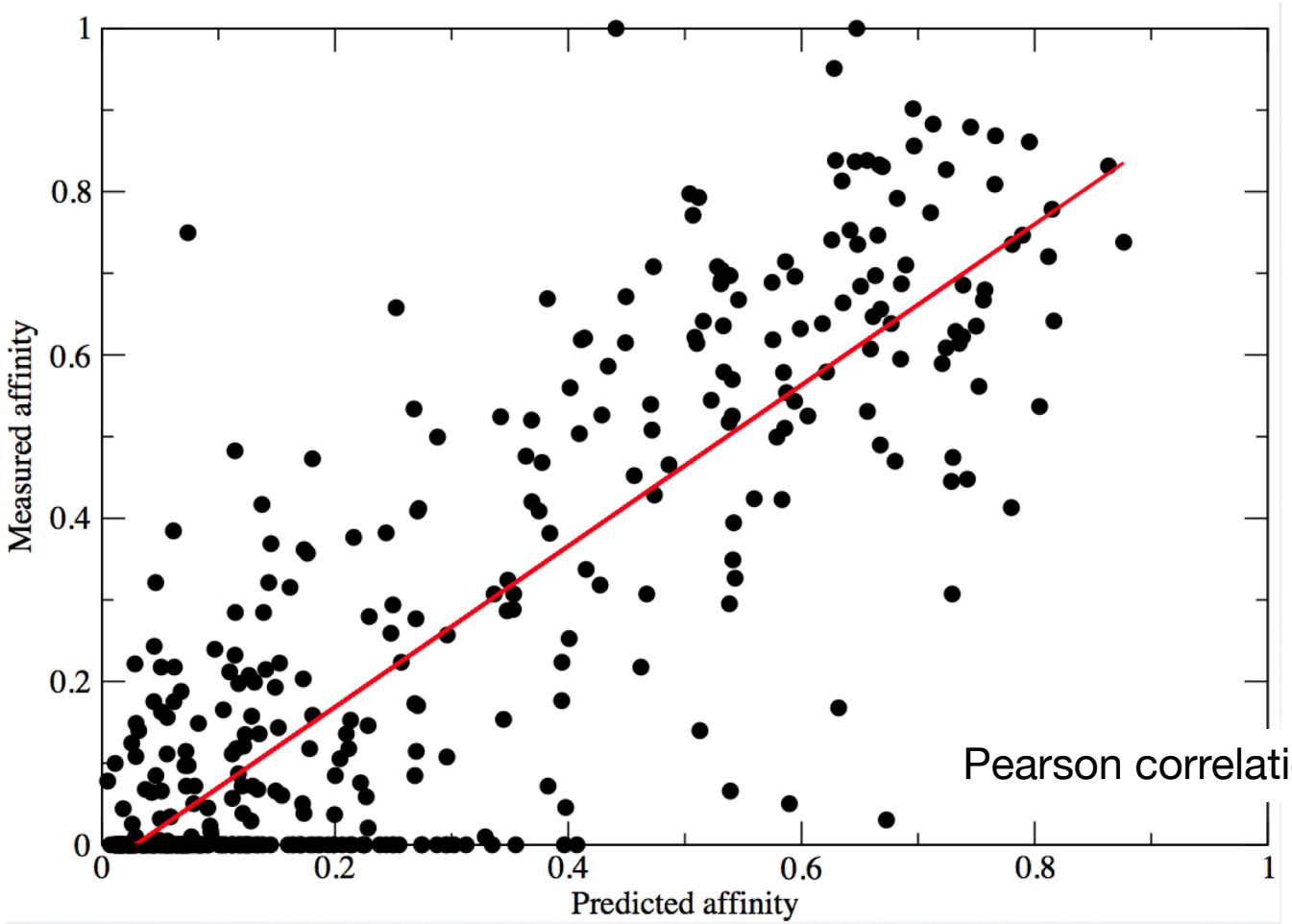


- 10 peptides from MHCpep database
- Bind to the MHC complex
- Relevant for immune system recognition
- Estimate sequence motif and weight matrix
- Evaluate motif "correctness" on 528 peptides

**ALAKAAAAM**
**ALAKAAAAN**
**ALAKAAAAR**
**ALAKAAAAT**
**ALAKAAAAV**
**GMNERPILT**
**GILGFVFTM**
**TLNAWVKVV**
**KLNEPVLLL**
**AVVPFIVSV**

# Prediction accuracy

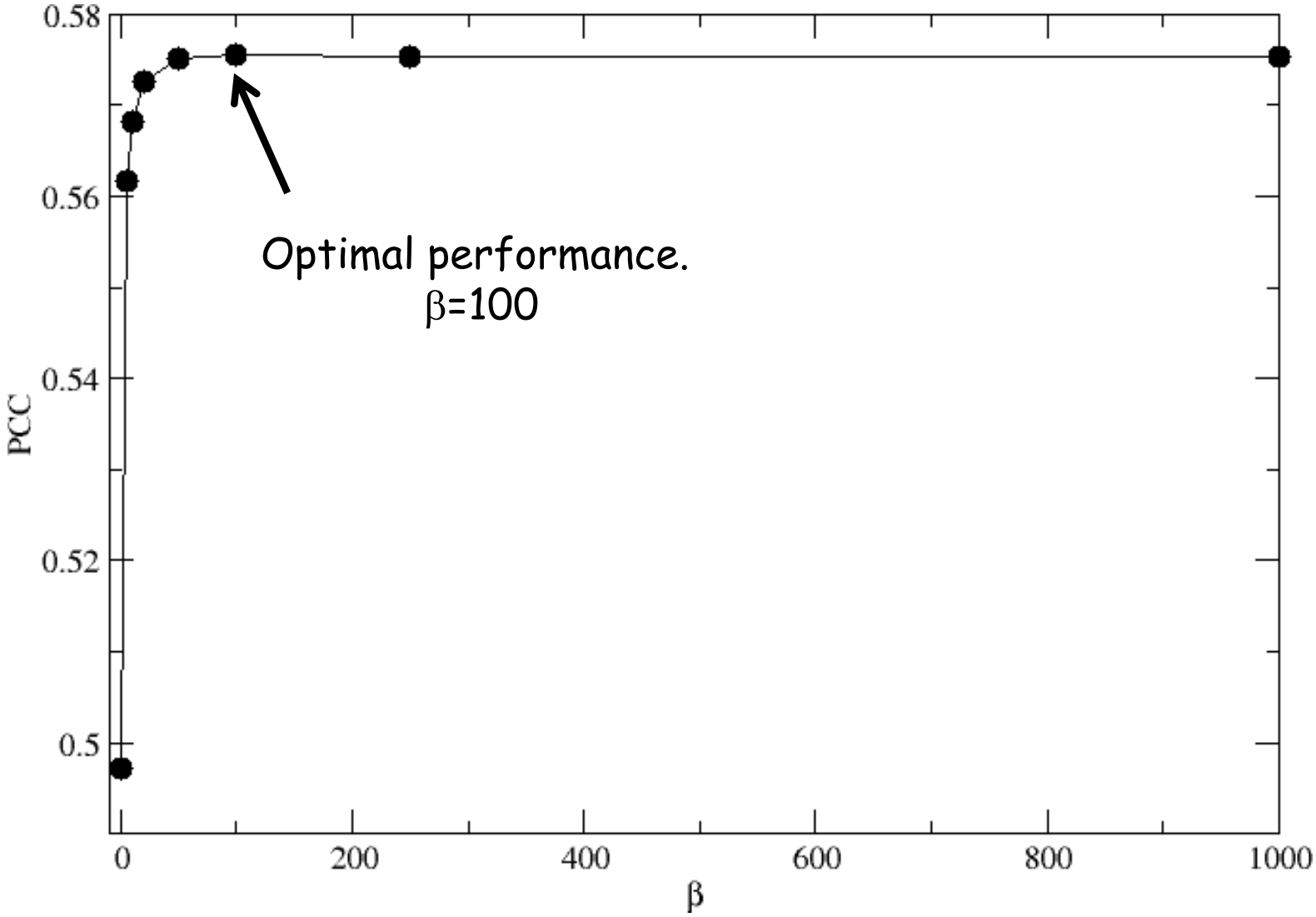





Pearson correlation 0.62

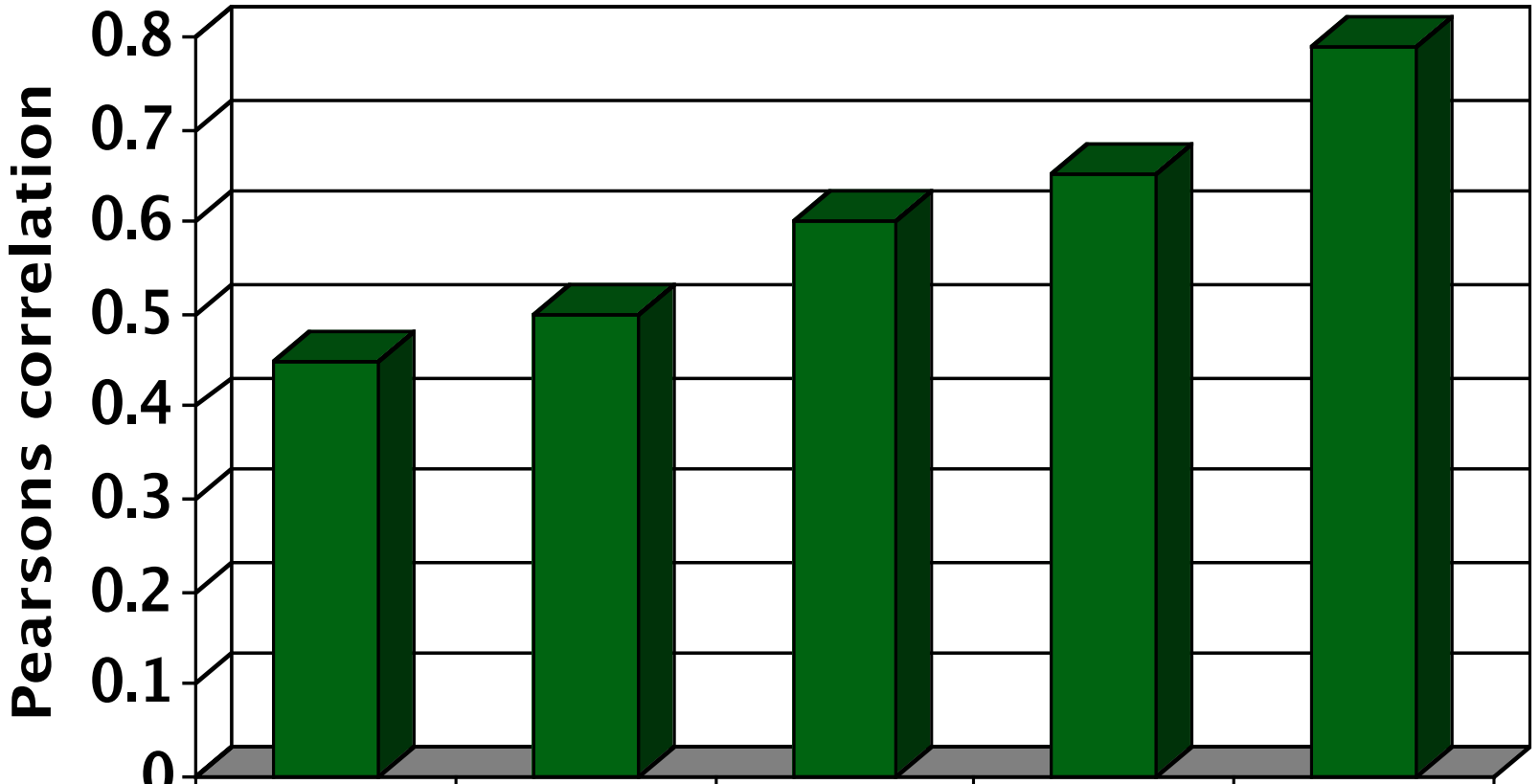
# How to define β?





# Predictive performance






| | Simple | Seq.W | Seq.W+SC | Seq.W+SC+ PW | Large dataset |
|---|---|---|---|---|---|
| CC | 0.45 | 0.5 | 0.6 | 0.65 | 0.79 |

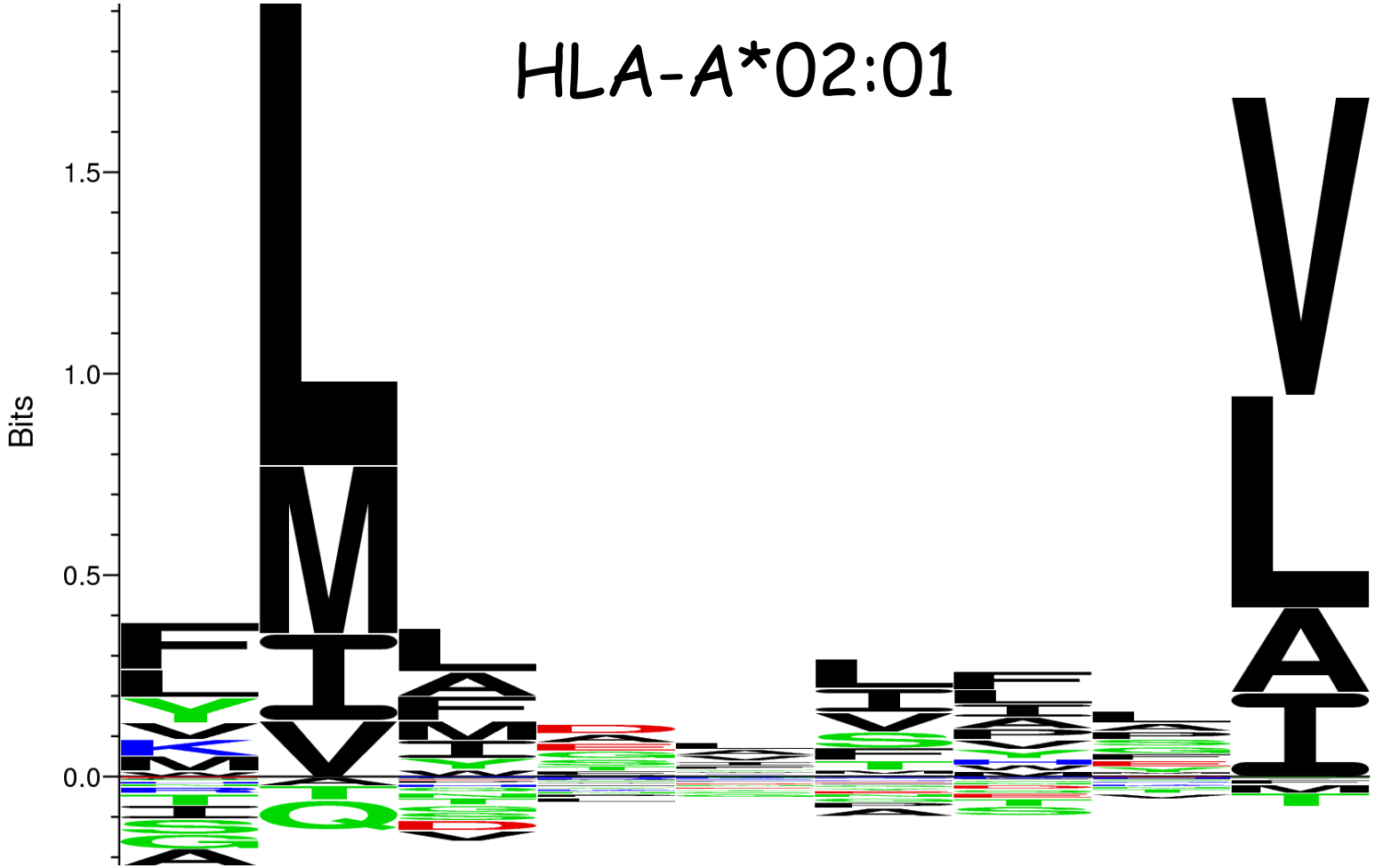
# MHC binding motifs

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

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 ILMEHIHKL 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 CINGVCWTV VMNILLQYV
ILTVILGVL KVLEYVIKV FLWGPRALV GLSRYVARL FLLTRILTI HLGNVKYLV GIAGGLALL GLQDCTMLV
TGAPVTYST VIYQYMDDL VLPDVFIRC VLPDVFIRC AVGIGIAVV LVVLGLLAV ALGLGLLPV GIGIGVLAA
GAGIGVAVL IAGIGILAI LIVIGILIL LAGIGLIAA VDGIGILTI GAGIGVLTA AAGIGIIQI QAGIGILLA
KARDPHSGH KACDPHSGH ACDPHSGHF SLYNTVATL RGPGRAFVT NLVPMVATV GLHCYEQLV PLKQHFQIV
AVFDRKSDA LLDFVRFMG VLVKSPNHV GLAPPQHLI LLGRNSFEV PLTFGWCYK VLEWRFDSR TLNAWVKVV
GLCTLVAML FIDSYICQV IISAVVGIL VMAGVGSPY LLWTLVVLL SVRDRLARL LLMDCSGSI CLTSTVQLV
VLHDDLLEA LMWITQCFL SLLMWITQC QLSLLMWIT LLGATCMFV RLTRFLSRV YMDGTMSQV FLTPKKLQC
ISNDVCAQV VKTDGNPPE SVYDFFVWL FLYGALLLA VLFSSDFRI LMWAKIGPV SLLLELEEV SLSRFSWGA
YTAFTIPSI RLMKQDFSV RLPRIFCSC FLWGPRAYA RLLQETELV SLFEGIDFY SLDQSVVEL RLNMFTPYI
NMFTPYIGV LMIIPLINV TLFIGSHVV SLVIVTTFV VLQWASLAV ILAKFLHWL STAPPHVNV LLLLTVLTV
VVLGVVFGI ILHNGAYSL MIMVKCWMI MLGTHTMEV MLGTHTMEV SLADTNSLA LLWAARPRL GVALQTMKQ
GLYDGMEHL KMVELVHFL YLQLVFGIE MLMAQEALA LMAQEALAF VYDGREHTV YLSGANLNL RMFPNAPYL
EAAGIGILT TLDSQVMSL STPPPGTRV KVAELVHFL IMIGVLVGV ALCRWGLLL LLFAGVQCQ VLLCESTAV
YLSTAFARV YLLEMLWRL SLDDYNHLV RTLDKVLEV GLPVEYLQV KLIANNTRV FIYAGSLSA KLVANNTRL
FLDEFMEGV ALQPGTALL VLDGLDVLL SLYSFPEPE ALYVDSLFF SLLQHLIGL ELTLGEFLK MINAYLDKL
AAGIGILTV FLPSDFFPS SVRDRLARL SLREWLLRI LLSAWILTA AAGIGILTV AVPDEIPPL FAYDGKDYI
AAGIGILTV FLPSDFFPS AAGIGILTV FLPSDFFPS AAGIGILTV FLWGPRALV ETVSEQSNV ITLWQRPLV

# HLA binding motifs

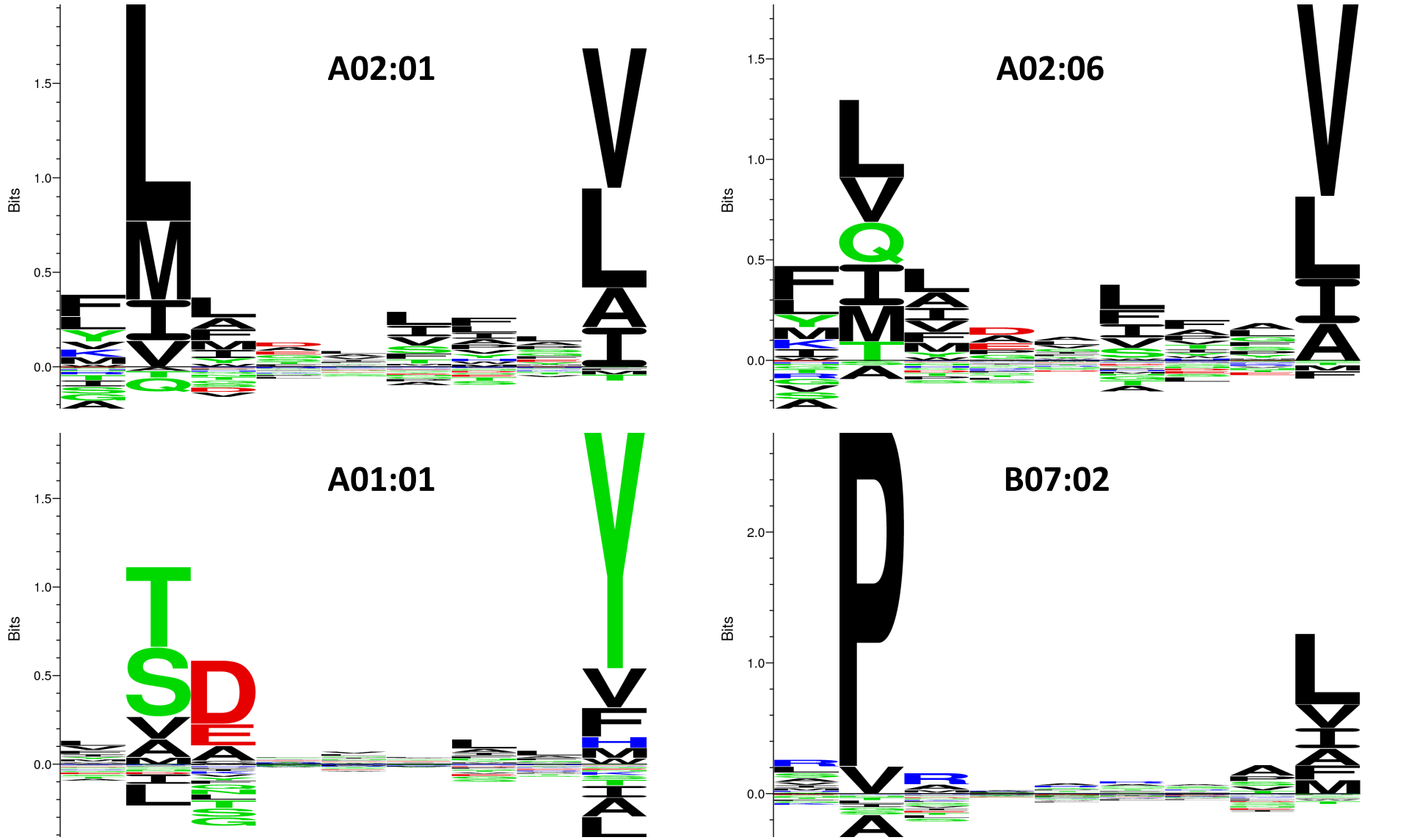





http://www.cbs.dtu.dk/biotools/Seq2Logo

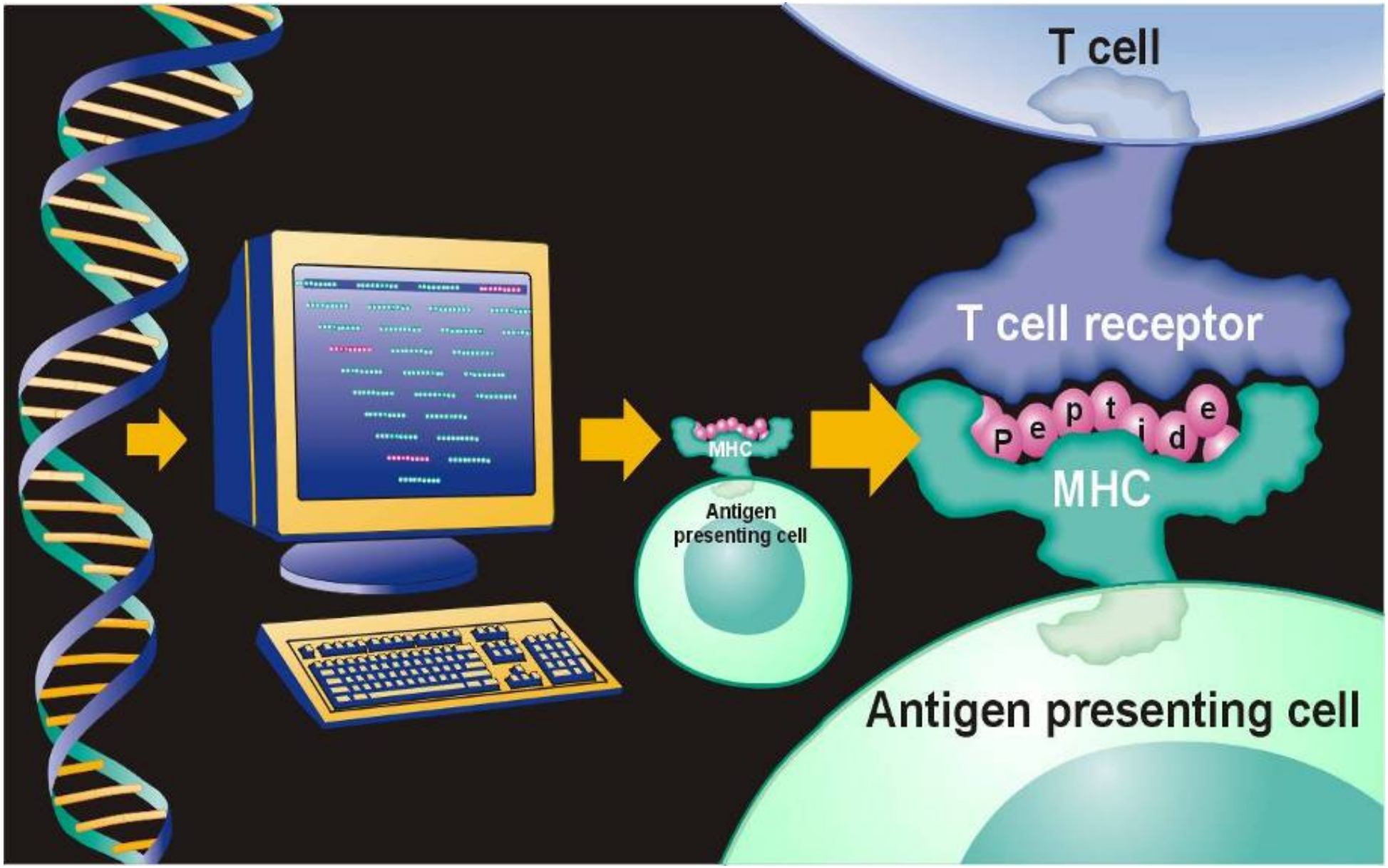
# HLA specificities





# Antigen Discovery






Lauemøller et al., 2000

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

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

## Genome

```
>Segment 1
agcaaaagcaggtcaattatattcaatatggaaagaataaaagaactaagagatctaatg
tcgcagtcccgcactcgcgagatactaacaaaaaccactgtggatcatatggccataatc
aagaaatacacatcaggaagacaagagaagaaccctgctctcagaatgaaatggatgatg
gcaatgaaatatccaatcacagcagacaagagaataatggagatgattcctgaaaggaat
```

**and 13350 other nucleotides on 8 segments**

## Proteins

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

**and 9 other proteins**

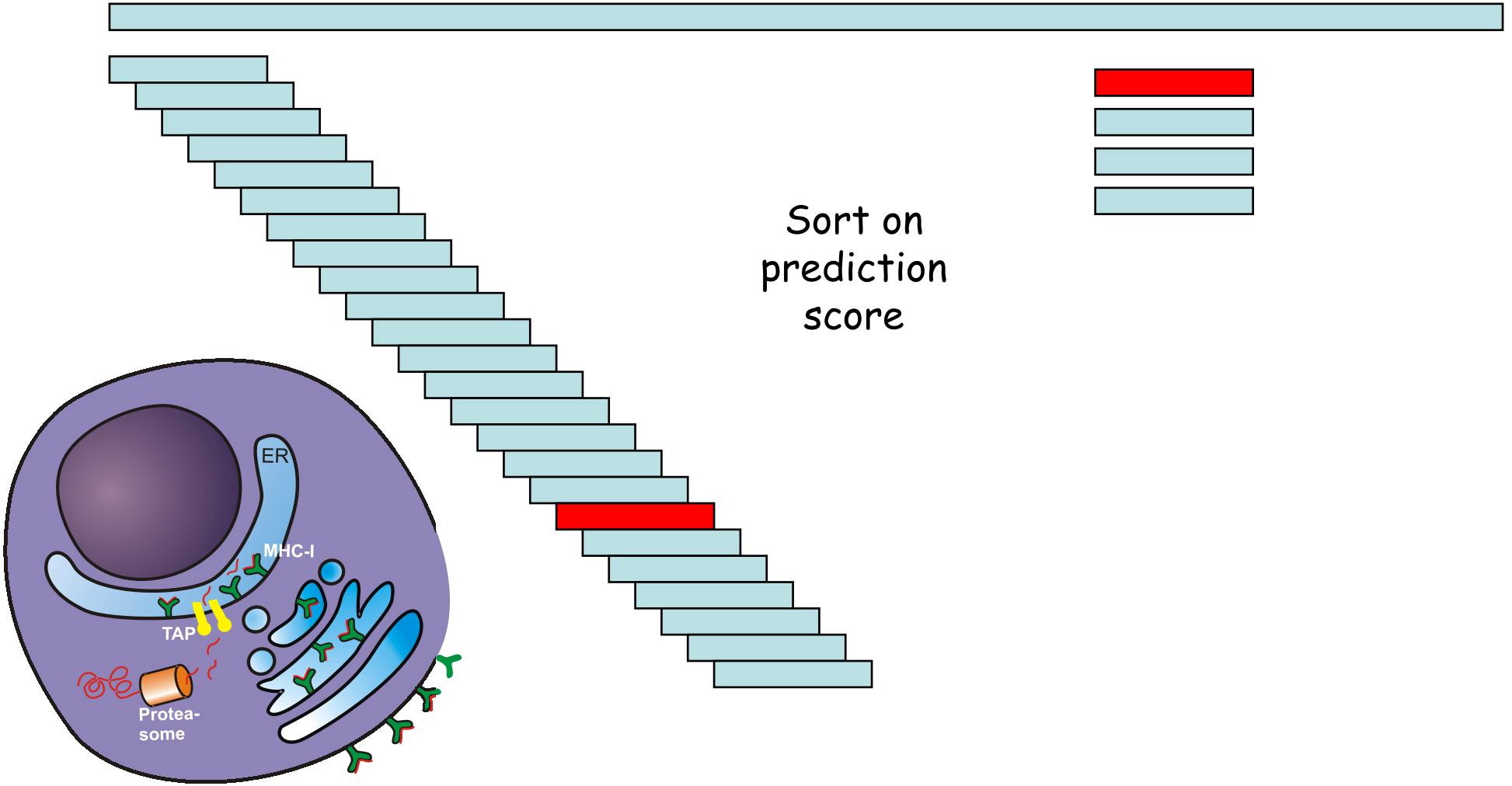
## 9mer peptides

```
MERIKELRD
ERIKELRDL
RIKELRDLM
IKELRDLMS
KELRDLMSQ
ELRDLMSQS
LRDLMSQSR
RDLMSQSRT
DLMSQSRTR
LMSQSRTRE
```

**and 4376 other 9mers**

# Predicting T cell epitopes?





# Validation of binding predictions

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS


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, Dziegiel 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, Lamberth 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, Dziegiel 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, Dziegiel 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, Lamberth K, Harndahl M, Dziegiel 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, Lamberth K, Harndahl M, Roder G, Stryhn A, Larsen MV, Nielsen M, Lundegaard C, Tang ST, Dziegiel 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, Lamberth K, Roder G, Justesen S, Lundegaard C, Worning 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.

# 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