Sequence motifs, information content, and sequence logos

Morten Nielsen,

Department of Health Technology,

DTU

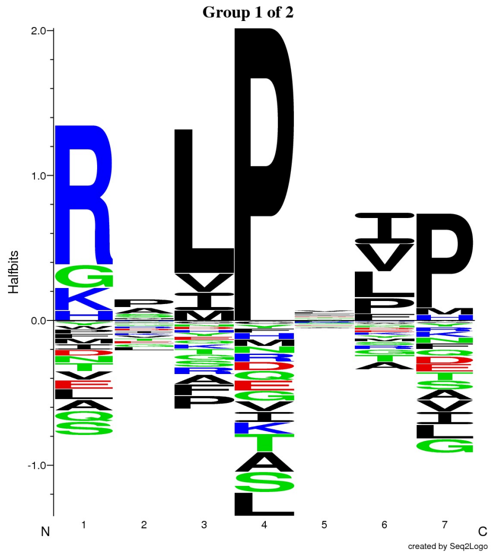
Why weight matrices?

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

- 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)
- Used to prediction protein structure and function

Identifying binding motifs (SH3)

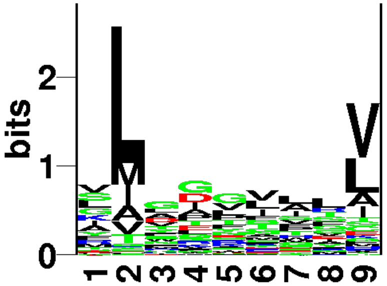
Peptide
LMLSLFEQSLSCQAQ
QGTDATKSIIFEAER
RLEEAQAYLAAGQHD
EISELRTKVQEQQKQ
FAGAKKIFGSLAFLP
VRASSRVSGSFPEDS
CKAFFKRSIQGHNDY
CEGCKAFFKRSIQGH
RLSEADIRGFVAAVV



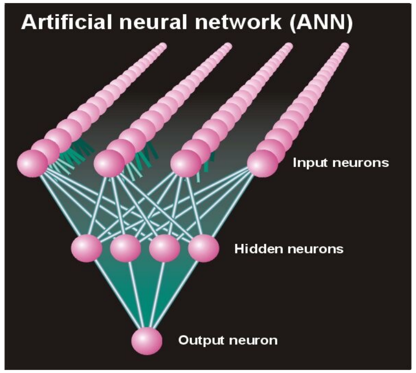
Bioinformatics in a nutshell

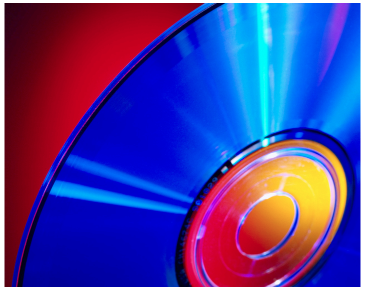
List of peptides that have a given biological feature

```
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
...
```

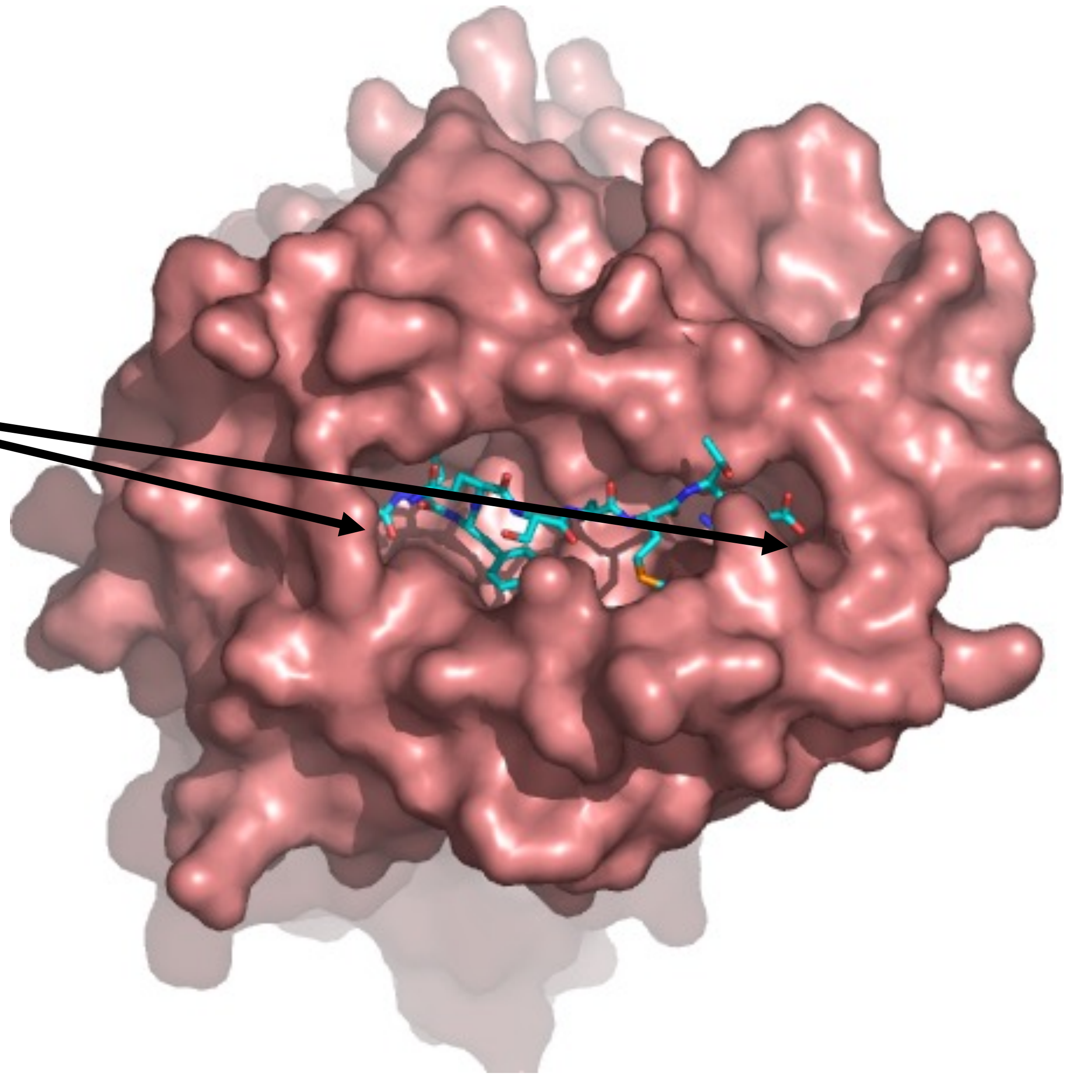
Objectives

- Visualization of binding motifs
  - Construction of sequence logos
- Understand the concepts of weight matrix construction
  - One of the most important methods of bioinformatics
- How to deal with data redundancy
- How to deal with low counts (few observations)
- How to use weight matrices to characterize receptor-ligand interactions
- Case story from the MHC-peptide interactions guiding immune system reactions

Binding Motif. MHC class I with peptide

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Anchor positions



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	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	S	I
1	0.10	0.06	0.01	0.02	0.01	0.02	0.02	0.09	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**	**0.01**	**0.07**	**0.01**	**0.00**	**0.01**	**0.06**	**0.00**	**0.01**	**0.08**	**2.16**	**2.16**
3	0.08	0.03	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	0.09	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	0.16	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
6	0.04	0.03	0.03	0.01	0.02	0.03	0.03	0.04	0.02	0.14	0.13	0.02	0.03	0.07	0.03	0.05	0.08	0.01	0.03	0.15	3.92	0.40
7	0.14	0.01	0.03	0.03	0.02	0.03	0.04	0.03	0.05	0.07	0.15	0.01	0.03	0.07	0.06	0.07	0.04	0.03	0.02	0.08	3.98	0.34
8	0.05	0.09	0.04	0.01	0.01	0.05	0.07	0.05	0.02	0.04	0.14	0.04	0.02	0.05	0.05	0.08	0.10	0.01	0.04	0.03	4.04	0.28
9	**0.07**	**0.01**	**0.00**	**0.00**	**0.02**	**0.02**	**0.02**	**0.01**	**0.01**	**0.08**	**0.26**	**0.01**	**0.01**	**0.02**	**0.00**	**0.04**	**0.02**	**0.00**	**0.01**	**0.38**	**2.78**	**1.55**

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

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

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?

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

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

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

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, q_a=0.05

or

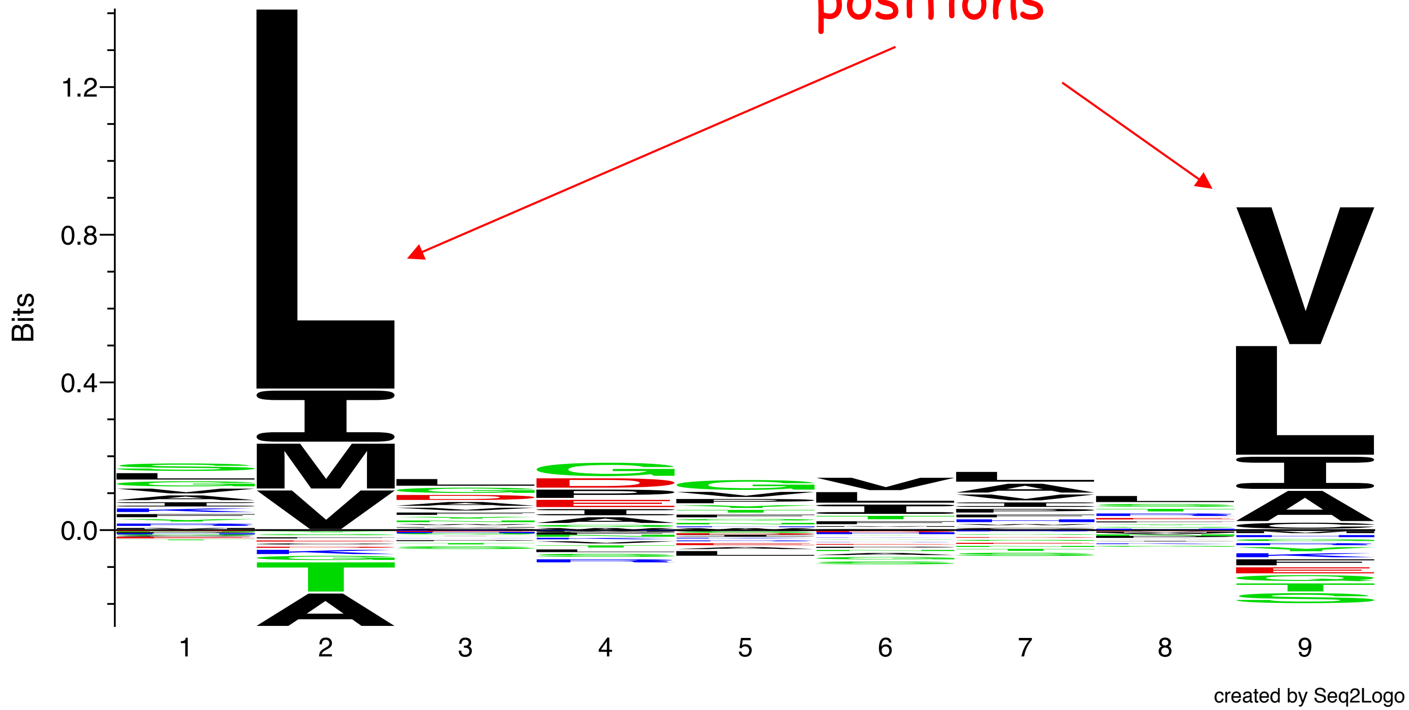
$$I = \sum_a p_a \cdot \log_2\left(\frac{p_a}{q_a}\right)$$ Kullback - Leibler

Sequence logos

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

HLA-A0201

High information positions



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

Logo handout

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

https://teaching.healthtech.dtu.dk/morten_teaching/27625.algo/presentations/PSSM/Ex_Logo.pdf

Introduction to the immune system [PDF]

9.30 - 11.20 (coffee break included)

Weight matrix construction[PDF]. [PPTX]

Logo Handout

Handout. Estimation of pseudo counts

11.20 - 12.00

Logo handout

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

Q1) Below is a multiple alignment of 35 human sequences. The sequences have been aligned around a donor splice. That site is indicated as the boundary between the '**Dark blue**' and '**Dark red**' colours.

```
-----Exon|intron----
01234567890123456789
tatcacaATGGTAGGTAACT
TCAACCAGGAGTAAGTCTTG
GTTGCACCCTGTAAGTCTCA
tatcacaATGGTAGGTAACT
TCAACCAGGAGTAAGTCTTG
CTTGCGAGAGGTGTGACATG
GCTCTACTCGGTAAGGTGAC
GCCTGGAGAGGTAATGACCC
CAAAACCATTGTGAGTAATC
GCCAGAGCAGGTAAAATATC
GAACAGTCAGGTCTGTTGCT
GAAGGCCCAGGTGAGCATAA
TCCTCTACAGGTGGGTACAT
GGCGTCCCGCGTAAGTATGG
CCTCGTGCAGGTAAGATTAA
TGCATGACAGGTGAGTGTTA
GAAATGTACAGTAAGTCTCT
GGTTCTCTGGGTAAGTAGAG
AAATGTACAGGTGAGTACTG
ACCTCGCTTGGTACGTGGGA
AATCAGACAGGTATAGAAAC
AGGACAGAAGGTAATTTTCT
AACTATTTGGGTAGGTAGCA
AAACTTGAAGGTATGTTGTT
CTGGGATAAGGTAAAAGTAT
TTGCACCCAGGTTAGTGGAT
ACTTCAATCGGTATGTTTTC
ACAGAGAAAAGTAAATTCCT
AATGGGAAAGGTAACAACAA
CATGCTACAGGTAGGTGAAT
ggctaggATGGTGAGGGCGC
CGACGCGGGCGTGAGAGGCG
CATTGAGAATGTGAGTTATT
AACAGAGCAGGTACTTGTAT
TGAACCAAAGGTGAAGACAT
```

Calculate the counts and frequencies (P) for positions 6-5. You have each been assigned one column on the upper right corner of the handout.

Position	6	7	8	9	0	1	2	3	4	5
Counts A										
Counts T										
Counts C										
Counts G										
P(A)										
P(T)										
P(C)										
P(G)										

Note P(A) is the frequency of amino acid A, this number of between 0 and 1, and the sum of P over the four nucleotides is 1.

Logo handout

Q1) Below is a multiple alignment of 35 human sequences. The sequences have been aligned around a donor splice. That site is indicated as the boundary between the '**Dark blue**' and '**Dark red**' colours.

```
-----Exon|intron----
01234567890123456789
tatcacaATGGTAGGTAACT
TCAACCAGGAGTAAGTCTTG
GTTGCACCCTGTAAGTCTCA
tatcacaATGGTAGGTAACT
TCAACCAGGAGTAAGTCTTG
CTTGCGAGAGGTGTGACATG
GCTCTACTCGGTAAGGTGAC
GCCTGGAGAGGTAATGACCC
CAAAACCATTGTGAGTAATC
GCCAGAGCAGGTAAAATATC
GAACAGTCAGGTCTGTTGCT
GAAGGCCCAGGTGAGCATAA
TCCTCTACAGGTGGGTACAT
GGCGTCCCGCGTAAGTATGG
CCTCGTGCAGGTAAGATTAA
TGCATGACAGGTGAGTGTTA
GAAATGTACAGTAAGTCTCT
GGTTCTCTGGGTAAGTAGAG
AAATGTACAGGTGAGTACTG
ACCTCGCTTGGTACGTGGGA
AATCAGACAGGTATAGAAAC
AGGACAGAAGGTAATTTTCT
AACTATTTGGGTAGGTAGCA
AAACTTGAAGGTATGTTGTT
CTGGGATAAGGTAAAAGTAT
TTGCACCCAGGTTAGTGGAT
ACTTCAATCGGTATGTTTTC
ACAGAGAAAAGTAAATTCCT
AATGGGAAAGGTAACAACAA
CATGCTACAGGTAGGTGAAT
ggctaggATGGTGAGGGCGC
CGACGCGGGCGTGAGAGGCG
CATTGAGAATGTGAGTTATT
AACAGAGCAGGTACTTGTAT
TGAACCAAAGGTGAAGACAT
```

Q2) Calculate the Entropy (S) and Information Content (I) using the formula below

Eq.1 $$S(p) = -\sum_a p_a \log_2(p_a) = -\frac{1}{\log(2)} \sum_a p_a \log(p_a)$$

where $\log_2$ is the logarithm with base 2, and log is the logarithm with base 10 (or any base for that sake)

Eq.2 $$I = 2.0 - S(p)$$

position	6	7	8	9	**0**	**1**	**2**	**3**	**4**	**5**
Entropy										
Information content										

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

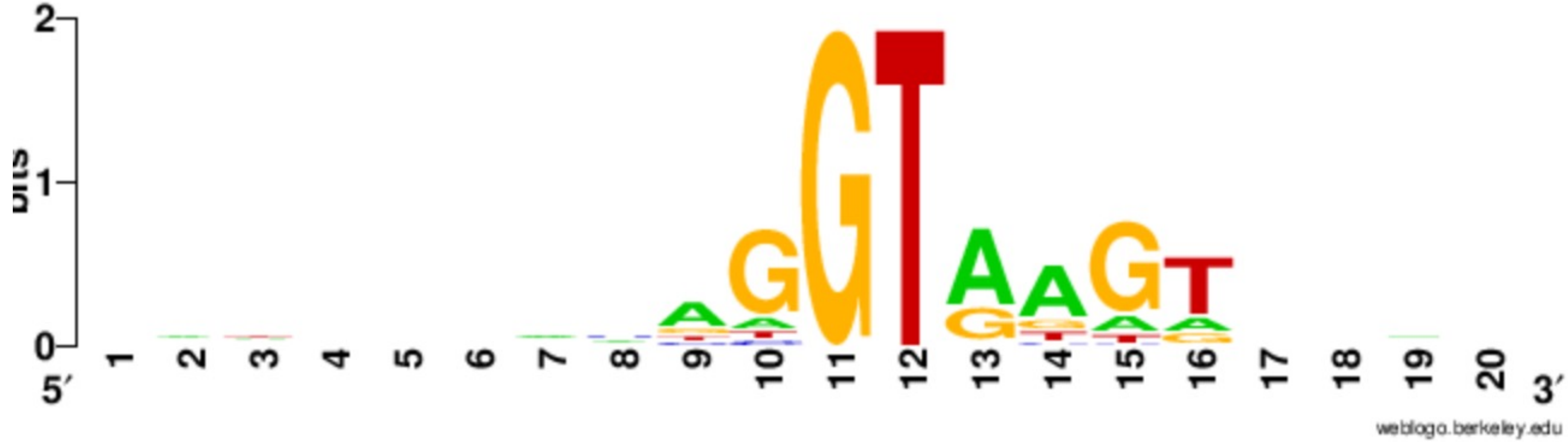
Q3) Where does the constant 2.0 come from in Eq.2?

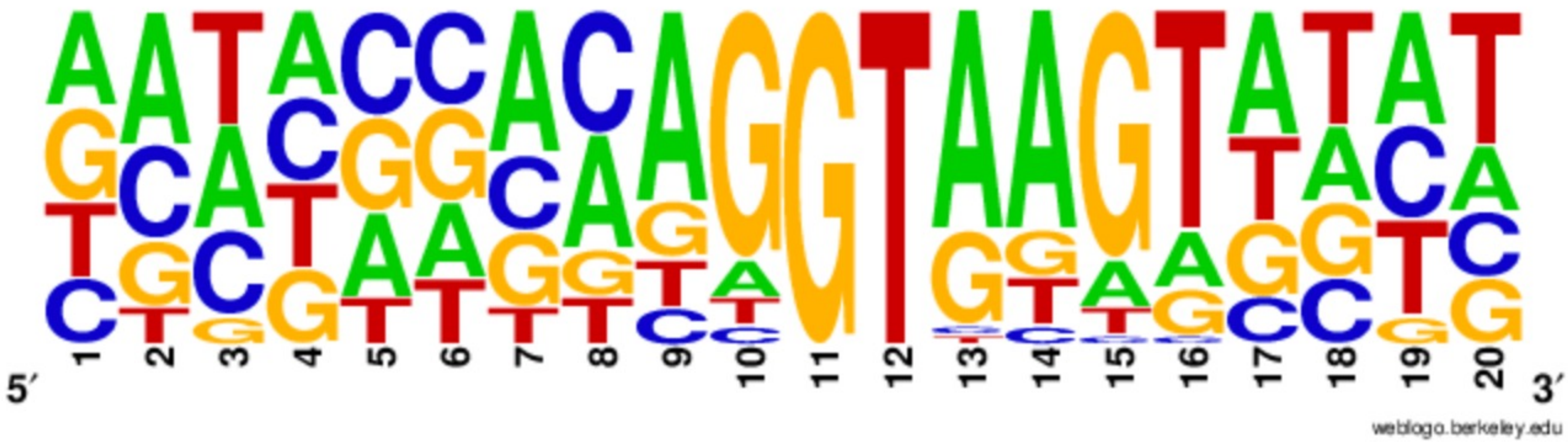
Q4) Draw an approximate Logo Plot by hand on the White board

If you have internet-access

Q5) Submit the multiple alignment to the WebLogo server http://weblogo.berkeley.edu/

Make both the Logo plot and a frequency plot
Explain what you see on the two plots.


bits
5′
3′
weblogo.berkeley.edu


5′
3′
weblogo.berkeley.edu

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. 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 determine binding more than other positions
- Cannot discriminate between good and very good binders

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

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

```
RLLDDTPEV 84 nM
GLLGNVSTV 23 nM
ALAKAAAAL 309 nM
```

Sequence information

Raw sequence counting

ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV

Sequence weighting

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS **CBS**

- 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

RLLDDTPEV 84 nM
GLLGNVSTV 23 nM
ALAKAAAAL 309 nM

Sequence weighting

- How to define clusters?
  - Hobohm algorithm
    - We will work on Hobohm later in the course
    - Slow when data sets are large
  - Heuristics
    - Less accurate
    - Fast

Sequence weighting - Clustering, Hobohm 1

Peptide	Weight	
ALAKAAAAM	0.20	Similar sequences; Weight 1/5
ALAKAAAAN	0.20	
ALAKAAAAR	0.20	
ALAKAAAAT	0.20	
ALAKAAAAV	0.20	
GMNERPILT	1.00	
GILGFVFTM	1.00	
TLNAWVKVV	1.00	
KLNEPVLLL	1.00	
AVVPFIVSV	1.00	

Sequence weighting

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

- Heuristics - weight on peptide k at position p

$$w_{kp} = \frac{1}{r \cdot s}$$

  – where r is the number of different amino acids in the column p, and s is the number occurrence of amino acid a in that column

- Weight of sequence k is the sum of the weights over all positions

$$w_k = \sum_p w_{kp} = \sum_p \frac{1}{r_p \cdot s_p}$$

Sequence weighting

$$w_{kp} = \frac{1}{r \cdot s}$$

r is the number of different amino acids in the column p, and s is the number occurrence of amino acid a in that column

In random sequences r=20, and s=0.05*N

$$w_{kp} = \frac{1}{20 \cdot 0.05 \cdot N} = \frac{1}{N}$$

where N is the number of sequences

Example

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

$$w_{kp} = \frac{1}{r \cdot s}$$

r is the number of different amino acids in the column p, and s is the number occurrence of amino acids a in that column

Peptide
ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV

Example (weight on each sequence)

CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS

$$w_{kp} = \frac{1}{r \cdot s}$$

r is the number of different amino acids in the column p, and s is the number occurrence of amino acids a in that column

W_{11}= 1/(4*6) = 0.042
W_{12}= 1/(4*7) = 0.036
W_{13}= 1/(4*5) = 0.050
W_{14}= 1/(5*5) = 0.040
W_{15}= 1/(5*5) = 0.040
W_{16}= 1/(4*5) = 0.050
W_{17}= 1/(6*5) = 0.033
W_{18}= 1/(5*5) = 0.040
W_{19}= 1/(6*2) = 0.083
Sum = 0.414

Peptide
ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV

Example (weight on each column)

$$w_{kp} = \frac{1}{r \cdot s}$$

r is the number of different amino acids in the column p, and s is the number occurrence of amino acids a in that column

W_{11}= 1/(4*6) = 0.042
W_{21}= 1/(4*6) = 0.042
W_{31}= 1/(4*6) = 0.042
W_{41}= 1/(4*6) = 0.042
W_{51}= 1/(4*6) = 0.042
W_{61}= 1/(4*2) = 0.125
W_{71}= 1/(4*2) = 0.125
W_{81}= 1/(4*1) = 0.250
W_{91}= 1/(4*1) = 0.250
W_{101}= 1/(4*6) = 0.042
Sum = 1.000

Peptide	Weight
ALAKAAAAM	0.41
ALAKAAAAN	0.50
ALAKAAAAR	0.50
ALAKAAAAT	0.41
ALAKAAAAV	0.39
GMNERPILT	1.36
GILGFVFTM	1.46
TLNAWVKVV	1.27
KLNEPVLLL	1.19
AVVPFIVSV	1.51

Example (weight on each column)

$$w_{kp} = \frac{1}{r \cdot s}$$

r is the number of different amino acids in the column p, and s is the number occurrence of amino acids a in that column

W_{11}= 1/(4*6) = 0.042
W_{21}= 1/(4*6) = 0.042
W_{31}= 1/(4*6) = 0.042
W_{41}= 1/(4*6) = 0.042
W_{51}= 1/(4*6) = 0.042
W_{61}= 1/(4*2) = 0.125
W_{71}= 1/(4*2) = 0.125
W_{81}= 1/(4*1) = 0.250
W_{91}= 1/(4*1) = 0.250
W_{101}= 1/(4*6) = 0.042
Sum = 1.000

Peptide	Weight
ALAKAAAAM	0.41
ALAKAAAAN	0.50
ALAKAAAAR	0.50
ALAKAAAAT	0.41
ALAKAAAAV	0.39
GMNERPILT	1.36
GILGFVFTM	1.46
TLNAWVKVV	1.27
KLNEPVLLL	1.19
AVVPFIVSV	1.51
Sum =	9.00

Sum of weights for all sequences is hence L (=9)

Sequence weighting

Pseudo counts

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

- 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

The Blosum (substitution frequency) matrix

	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

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

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

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

- Example

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

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

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

$$g_D = 0.2 \cdot + 0.1 \cdot 0.02 + \ldots + 0.3 \cdot 0.16 + 0.1 \cdot 0.12 = 0.020$$

Weight on pseudo count

CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS

- 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 -1, β is the **weight on prior/weght on pseudo count**
  - In clustering α= #clusters -1
  - In heuristics α= <# different amino acids in each column> -1

ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV

Example

In heuristics

- α= <# different amino acids in each column> -1

α =(4+4+4+5+5+4+6+5+6)/9
= 4.8

Note: α <= 20!

Peptide

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

Weight on pseudo count

ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV

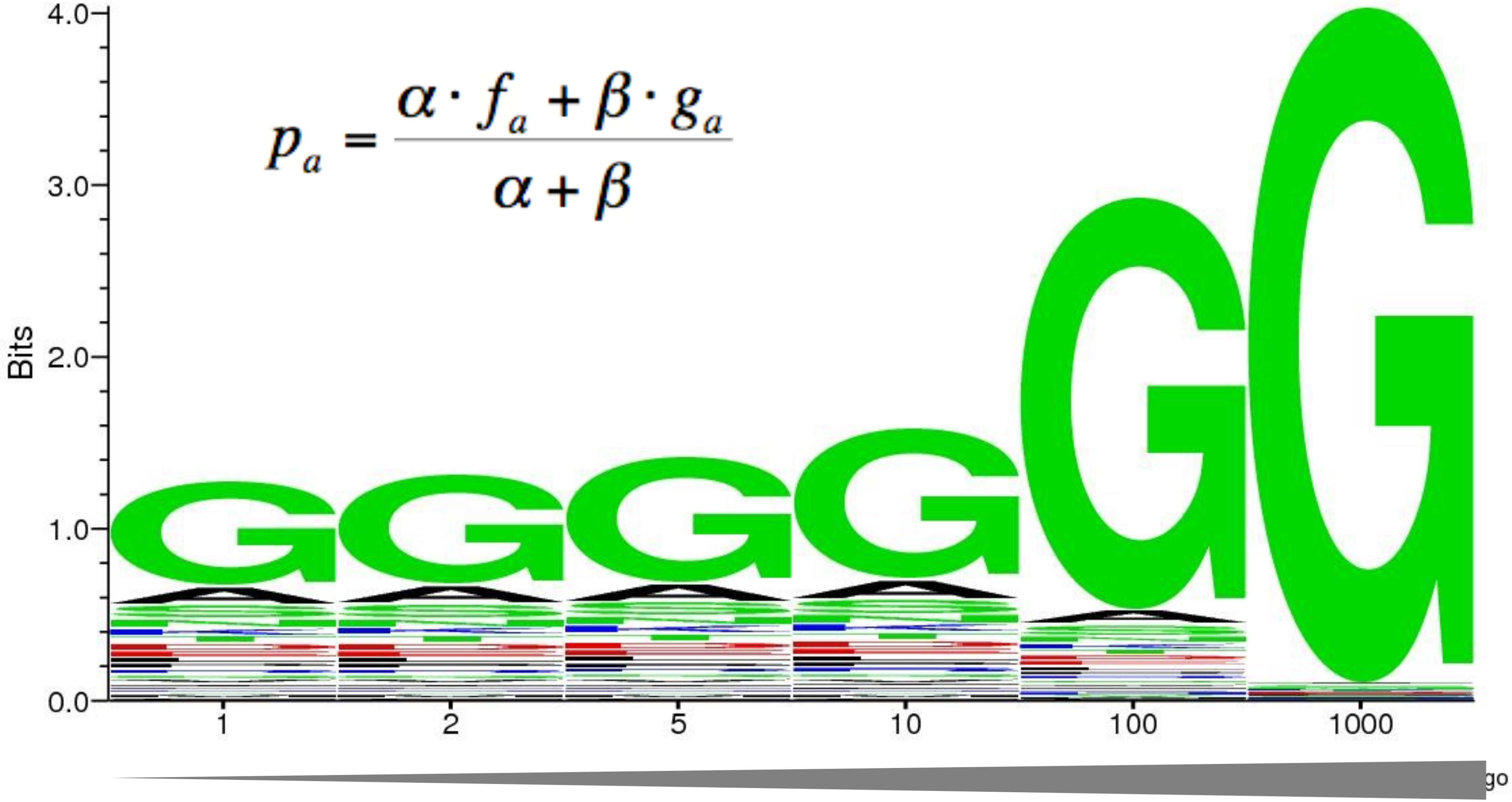
- Example

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

- If α large, p ≈ f and only the observed data defines the motif
- If α small, p ≈ g and the pseudo counts (or prior) defines the motif
- β is [50-200] normally

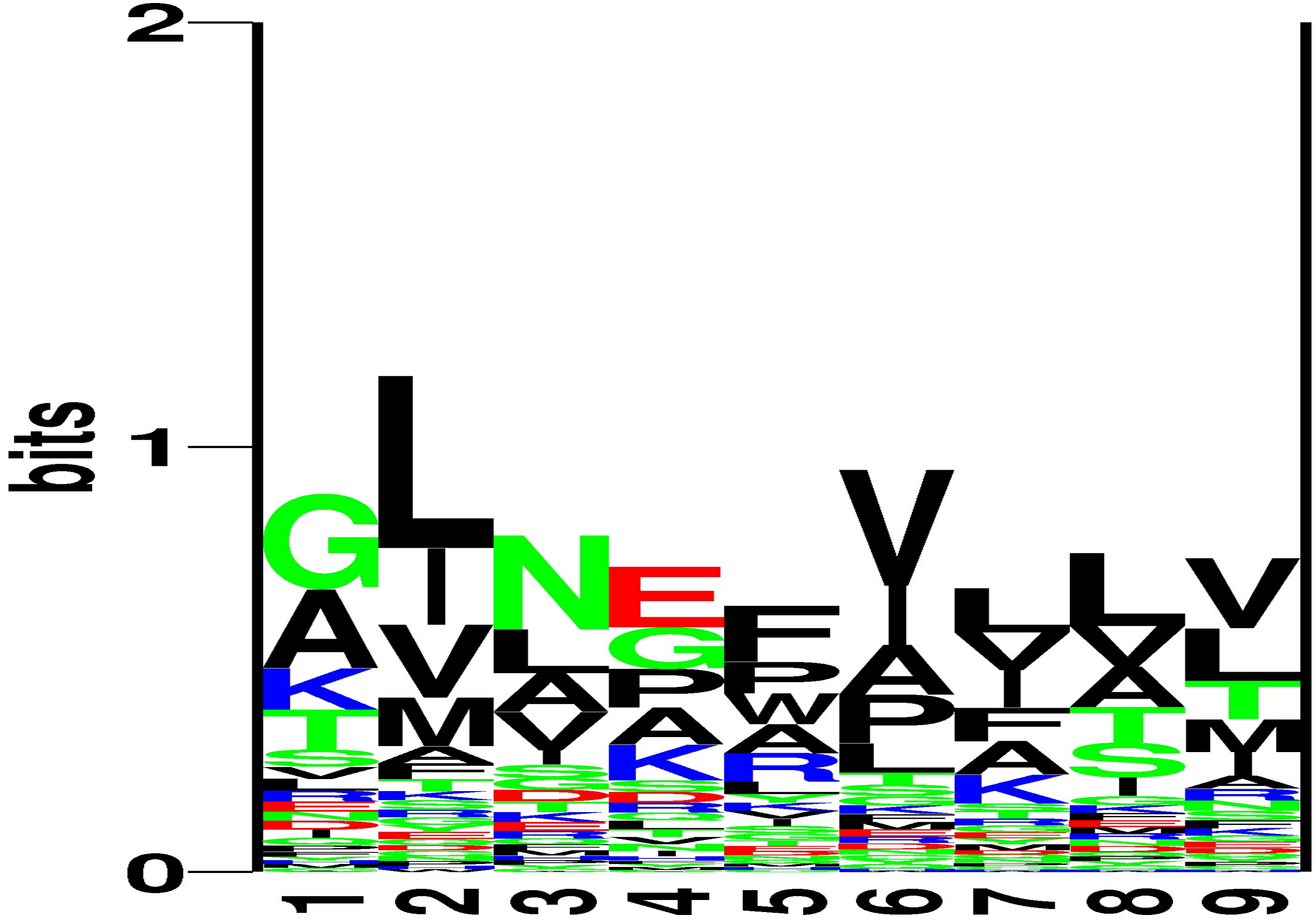
Gaining confidence

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



α

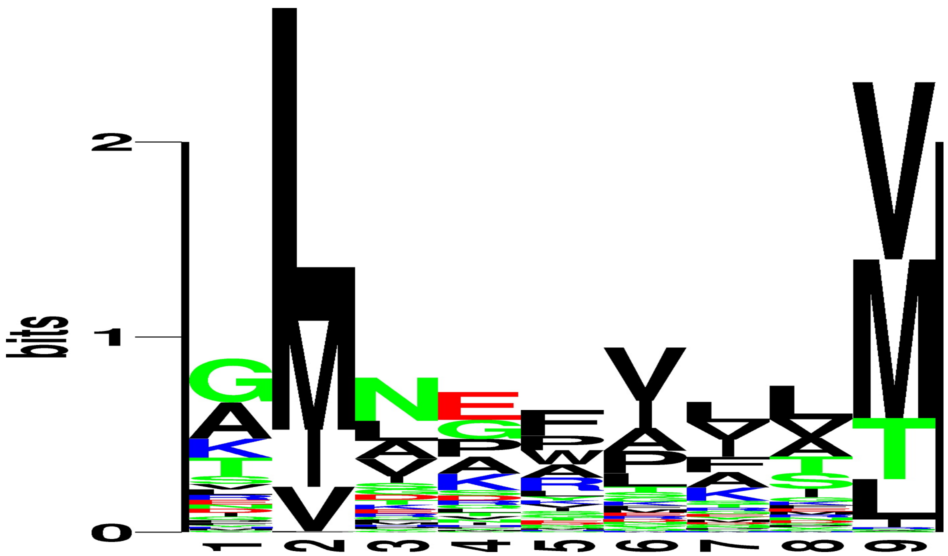
Sequence weighting and pseudo counts



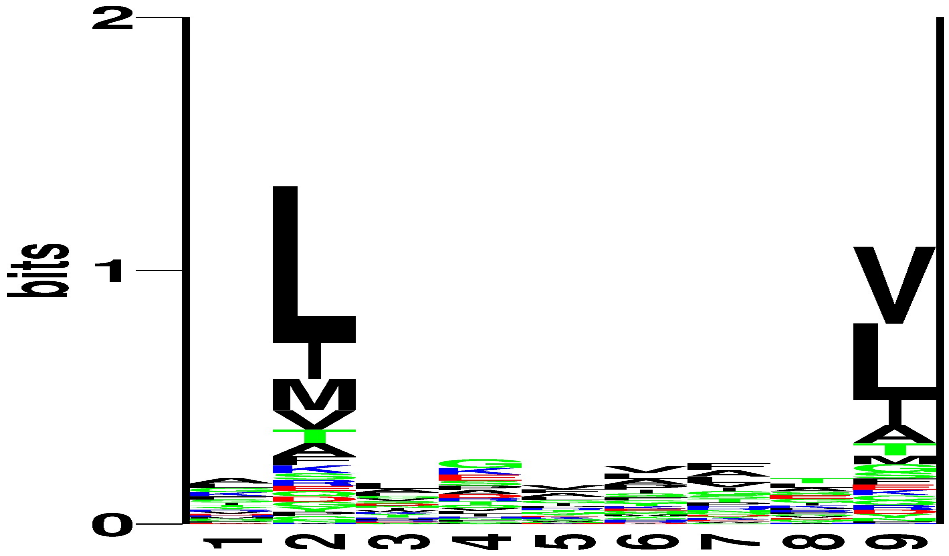
ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV

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

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS **CBS**

- 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	-6.6	-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

RLLDDTPEV	**11.9**	**84nM**
$\boxed{\text{GLLGNVSTV}}$	**14.7**	**23nM**
ALAKAAAAL	**4.3**	**309nM**

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

An example!!
(See handout)

Introduction to the immune system [PDF]

9.30 - 11.20 (coffee break included)

Weight matrix construction[PDF]. [PPTX]

Logo Handout

Handout. Estimation of pseudo counts

Logo handout - 2

The equation used to estimate frequencies in a weight matrix is

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

where α is the number of sequence in the multiple alignment (minus 1), β is the weight on prior (or weight on pseudo counts), f_a is the observed frequency for amino acid a and g_a is the pseudo frequency for amino acid a.

The pseudo frequency is estimated using the relation

$$g_a = \sum_b f_b \cdot q(a \mid b)$$

where f_b is the observed frequency for amino acid b, and $q(a|b)$ is the Blosum substitution frequency for the amino acid a, conditional on the observation of amino acid b.

Once you have estimated the frequency p_a, the weight matrix values are calculated using the relation

$$W_{ia} = 2 * \frac{\log(\frac{p_{ia}}{q_a})}{\log(2)}$$

where p_{ia} is the frequencies of amino acid a at position i in the motif, and q_a is the background frequency of amino acid a (see last page).

The Blosum62 substitution matrix and a table of the 20 background frequencies are given on the last page.

Logo handout - 2

CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS

Say, you have the following 6 sequences

EDRYK
EHYLK
QGHLP
EHLYR
EHQEA
EHYLR

Estimate the observed frequencies (f_a), the pseudo frequencies (g_a), and the combined frequencies p_a at P1 for the 20 amino acids (fill out the table below). Use β=5 and no sequence weighting.

Logo handout - 2

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

Table. The Blosum frequency substitution matrix. Each row gives the probabilities for substituting an amino acid to each of the 20 conventional amino acids. That is, the first row gives the probabilities P(aa|A) etc..

```
# Background frequencies
A 0.07400
R 0.05200
N 0.04500
D 0.05400
C 0.02500
Q 0.03400
E 0.05400
G 0.07400
H 0.02600
I 0.06800
L 0.09900
K 0.05800
M 0.02500
F 0.04700
P 0.03900
S 0.05700
T 0.05100
W 0.01300
Y 0.03200
V 0.07300
```

Logo handout - 2

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS **CBS**

	f_a	g_a	p_a	w_a
A	0	0.06	0.03	-2.61
R	0			
N	0			
D	0			
C	0			
Q	0.167			
E	0.833			
G	0			
H	0			
I	0			
L	0			
K	0			
M	0			
F	0			
P	0			

$g(A) = f(E)*q(A|E)+f(Q)*q(A|Q) = 5/6*0.06+1/6*0.06 = 0.06$

$p(A) = (5*0.0+5*0.06)/10 = 0.03$

$w(A) = 2*\log(0.03/0.074)/\log(2) = -2.61$

Logo handout – 2 - Answers

	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)

	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

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

KL == Shannon if q=0.05 for all AA

Special case

- What happens when $\alpha = 0$?
  - we only have one sequence, ILVKAIPHL

$$p_{1,A} = \frac{\alpha \cdot f_{1,A} + \beta \cdot g_{1,A}}{\alpha + \beta} = g_{1,A}$$

$$g_{1,A} = \sum_a f_a \cdot q(A \mid a) = q(A \mid I) = \frac{q_{IA}}{q_I}$$

$$W_{1,A} = \log(\frac{p_{1,A}}{q_{1,A}}) = \log(\frac{g_{1,A}}{q_{1,A}}) = \log(\frac{q_{IA}}{q_I \cdot q_A}) = Bl(A,I)$$

ILVKAIPHL

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	I	-1.3	-3.1	-3.2	-3.2	-1.3	-2.7	-3.2	-3.7	-3.1	4.0	1.5	-2.6	1.1	-0.2	-2.8	-2.4	-0.7	-2.3	-1.3	2.6
2	L	-1.5	-2.2	-3.3	-3.7	-1.3	-2.1	-2.8	-3.6	-2.7	1.5	3.8	-2.4	2.0	0.4	-2.9	-2.5	-1.2	-1.7	-1.0	0.8
3	V	-0.2	-2.5	-2.9	-3.2	-0.8	-2.1	-2.4	-3.2	-3.3	2.5	0.8	-2.3	0.7	-0.8	-2.5	-1.6	-0.1	-2.5	-1.3	3.8
4	K	-0.8	2.1	-0.2	-0.8	-3.1	1.3	0.8	-1.6	-0.7	-2.6	-2.4	4.5	-1.4	-3.2	-1.0	-0.2	-0.7	-2.6	-1.8	-2.3
5	A	3.9	-1.5	-1.6	-1.7	-0.4	-0.8	-0.8	0.2	-1.6	-1.3	-1.5	-0.8	-1.0	-2.2	-0.8	1.2	-0.1	-2.5	-1.7	-0.2
6	I	-1.3	-3.1	-3.2	-3.2	-1.3	-2.7	-3.2	-3.7	-3.1	4.0	1.5	-2.6	1.1	-0.2	-2.8	-2.4	-0.7	-2.3	-1.3	2.6
7	P	-0.8	-2.0	-1.9	-1.6	-2.6	-1.4	-1.2	-2.1	-2.0	-2.8	-2.9	-1.0	-2.6	-3.7	7.3	-0.8	-1.0	-4.6	-2.6	-2.5
8	H	-1.6	-0.4	0.5	-1.0	-3.4	0.3	-0.0	-1.9	7.5	-3.1	-2.7	-0.7	-1.4	-1.2	-2.1	-0.9	-1.9	-1.5	1.7	-3.3
9	L	-1.5	-2.2	-3.3	-3.7	-1.3	-2.1	-2.8	-3.6	-2.7	1.5	3.8	-2.4	2.0	0.4	-2.9	-2.5	-1.2	-1.7	-1.0	0.8

	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

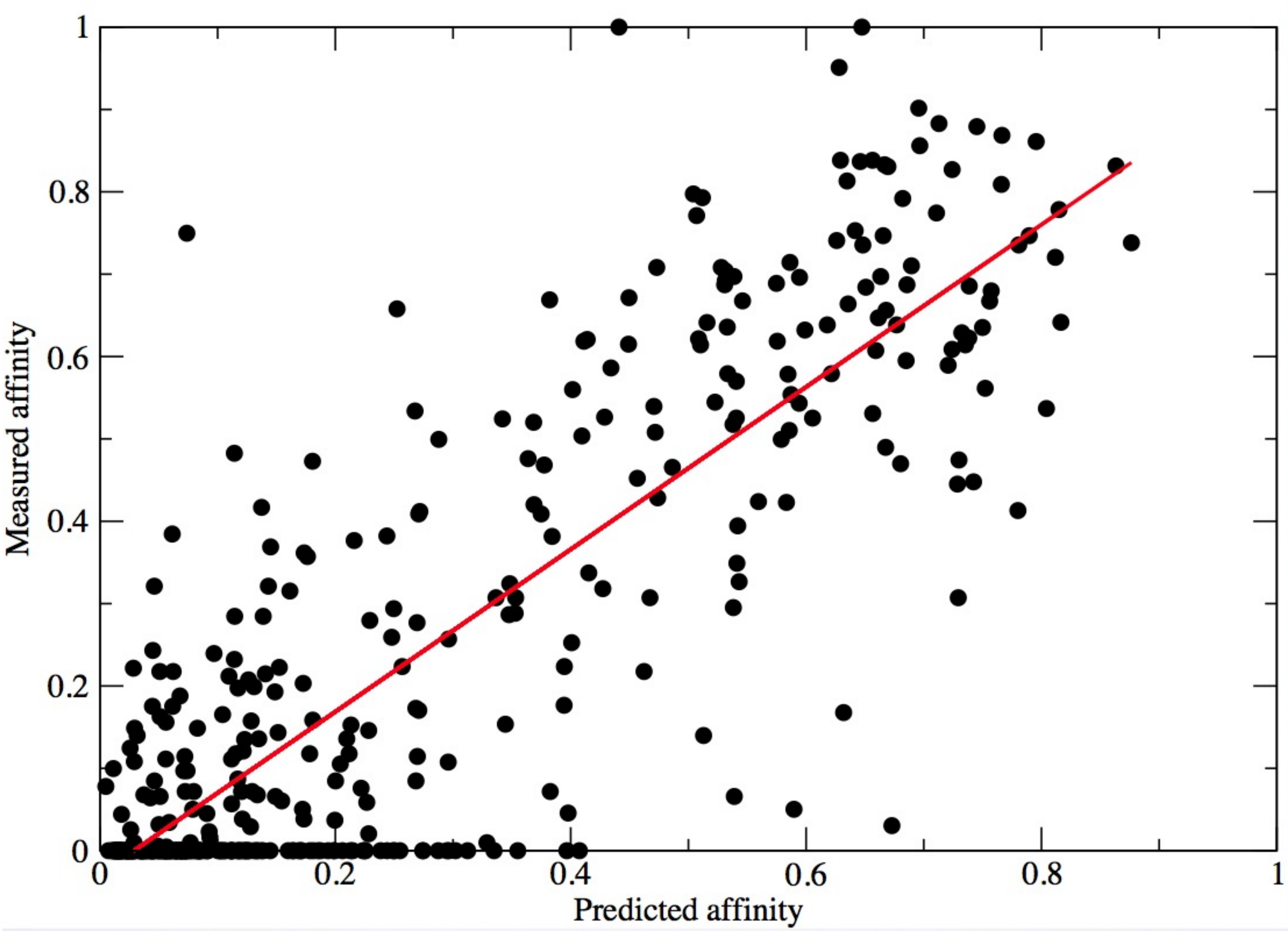
Example from real life

- 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

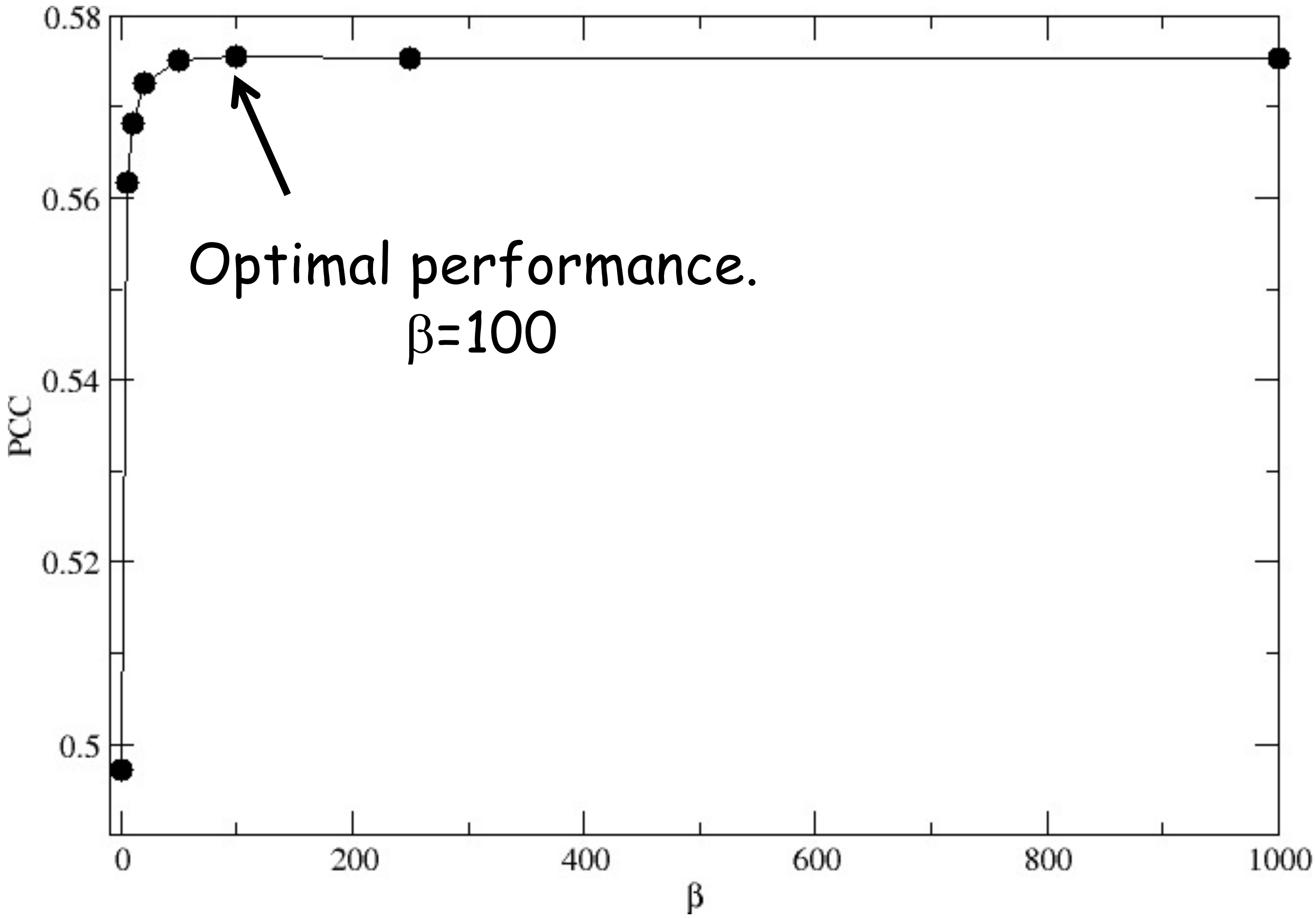
ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV

Prediction accuracy

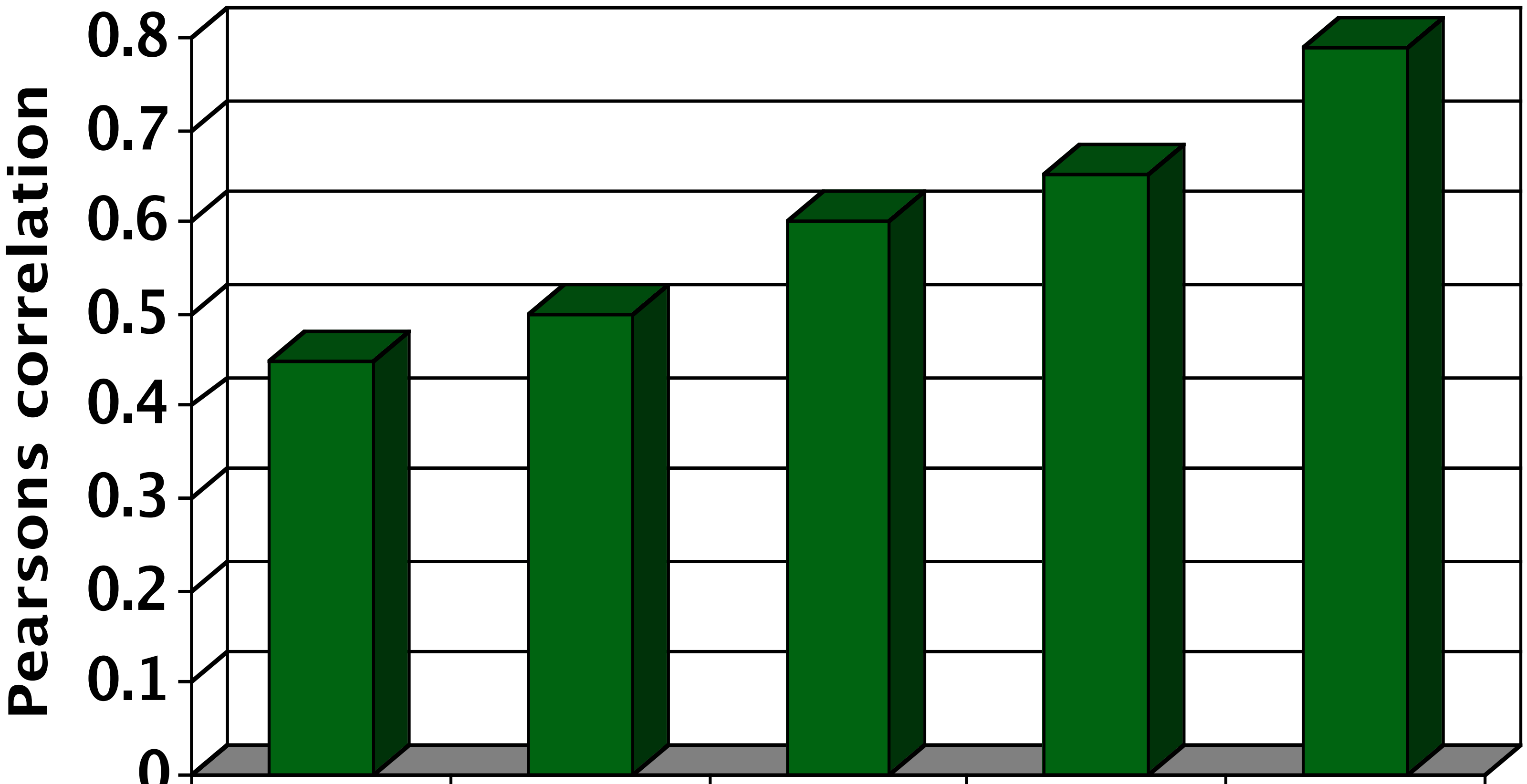
Pearson correlation 0.62



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

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