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

Morten Nielsen,
CBS, Department of Systems Biology, DTU
and
Instituto de Investigaciones
Biotecnológicas, Universidad de San Martín,
Argentina

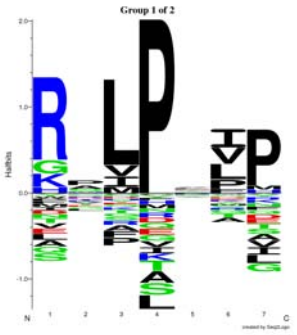
## Why weight matrices?

- 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
  - ….

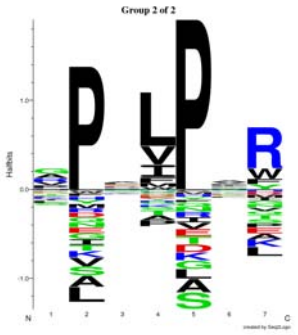
## SH3 domain binding

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Class I



Class II



**class I**
+xϕPxϕP

**class II**
ϕPxϕPx+

## Objectives

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

- 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

## Outline

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

- Pattern recognition
  - Regular expressions and probabilities
- Information content
  - Sequence logos
- Multiple alignment and sequence motifs
- Weight matrix construction
  - Sequence weighting
  - Low (pseudo) counts
- Example from the real world
- Sequence profiles
- Psi-Blast revised ...

## MHC Class I pathway

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

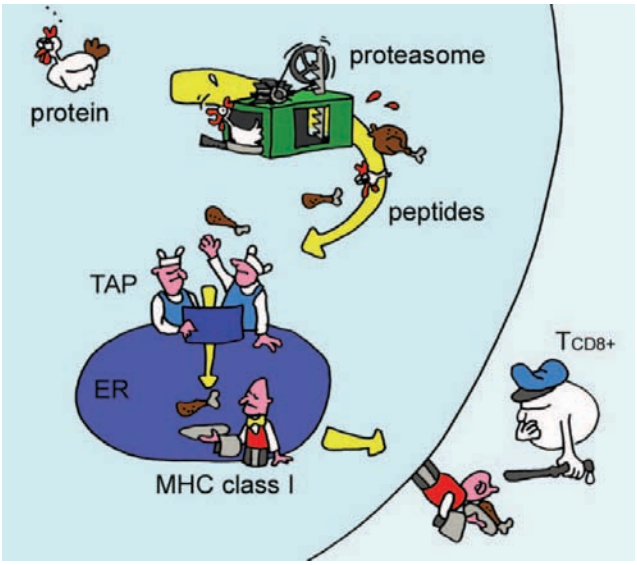
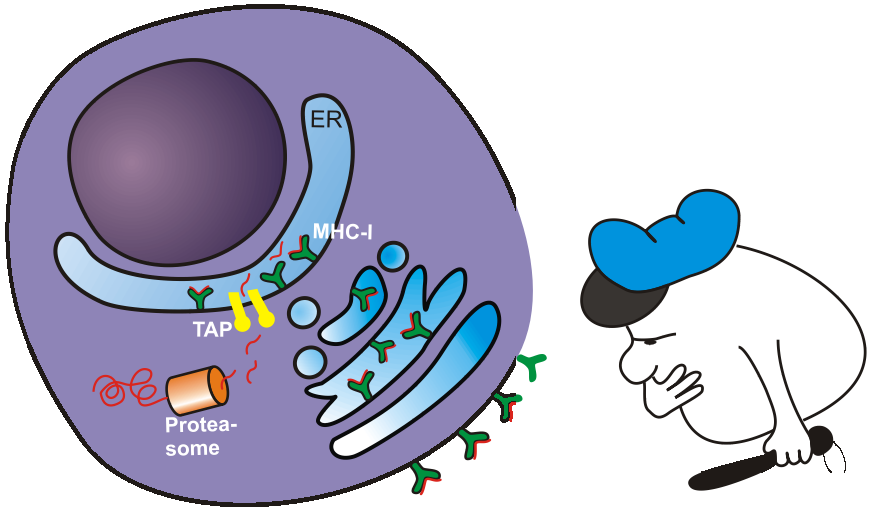



Figure by Eric A.J. Reits

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

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS




## CTL response

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

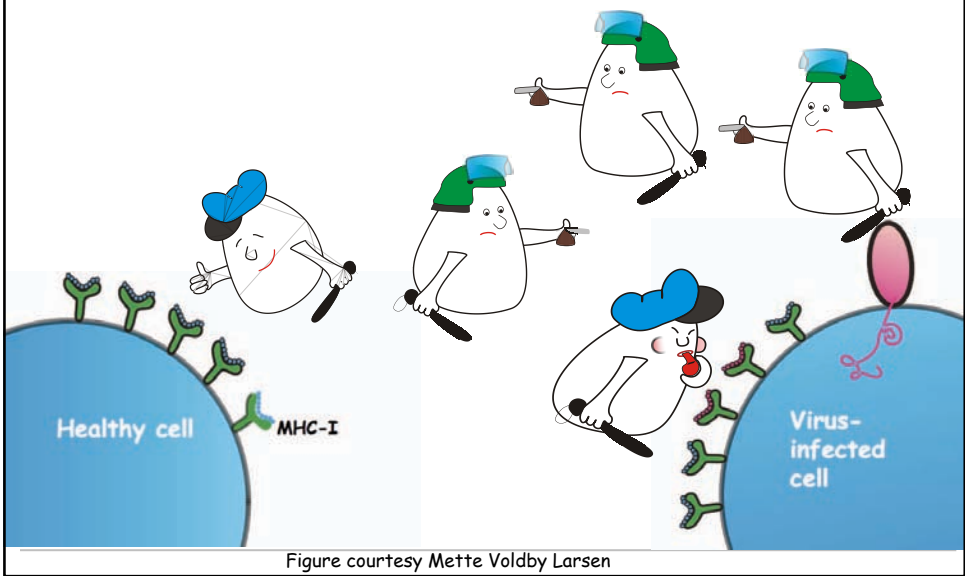


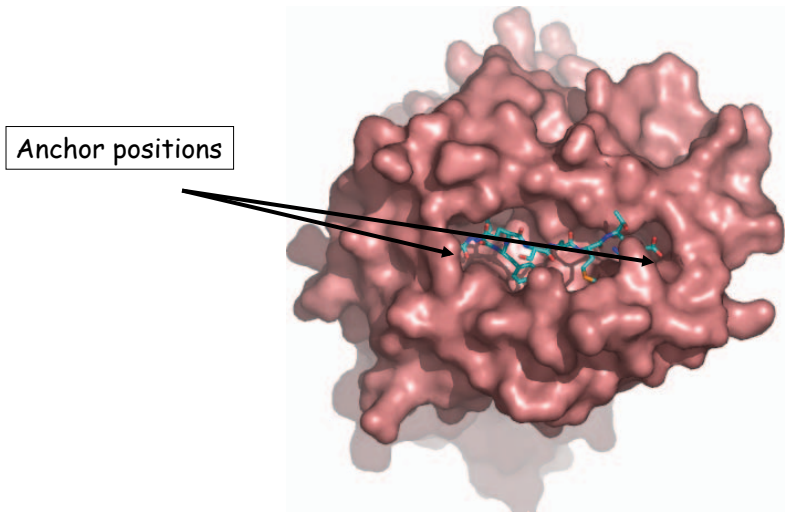

Figure courtesy Mette Voldby Larsen

## Encounter with death

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

## Binding Motif. MHC class I with peptide

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS



## 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
```

## Cost of a motif characterization

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

- 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 …

# Information content

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 | 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

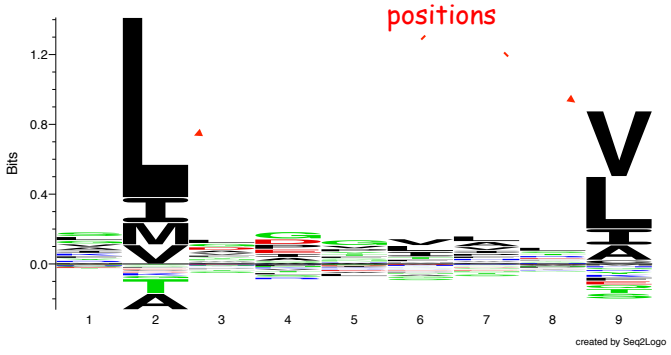
# Sequence logos

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

- 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

## Sequence Information

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

- 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

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

- 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

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

- 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

## Characterizing a binding motif from small data sets

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS 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

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

10 MHC restricted peptides

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

$$[AGTK]_1[LMIV]_2[ANLV]_3 ... [MNRTVL]_9$$

- 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

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

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

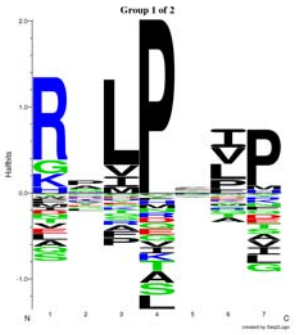
```
RLLDDTPEV 84 nM
GLLGNVSTV 23 nM
ALAKAAAAL 309 nM
```

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

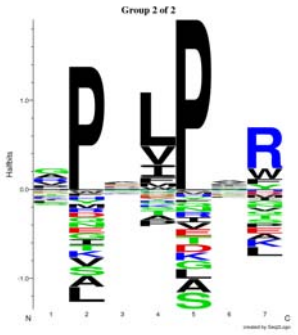
## SH3 domain binding

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Class I




Class II




**class I**
+xϕPxϕP

**class II**
ϕPxϕPx+

## Extended motifs

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

- 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

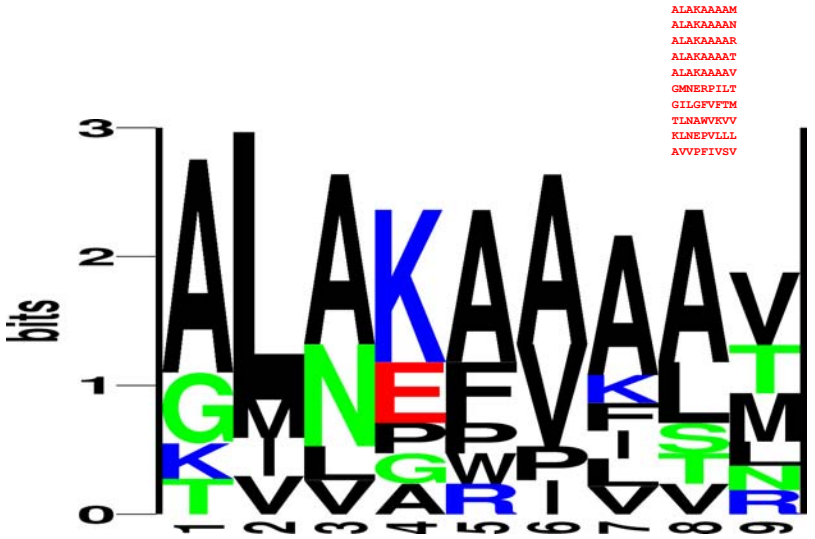
ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV

RLLDDTPEV 84 nM
GLLGNVSTV 23 nM
ALAKAAAAL 309 nM

## Sequence information

Raw sequence counting

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS




## 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

} Similar sequences Weight 1/5

GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV

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

## Sequence weighting

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS


## 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 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 | 0.02 | 0.03 |
| 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)

# The way from log-odds to frequencies

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 | 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 |

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

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

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

## What is a pseudo count?

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 |
| … | | | | | | | | | | | | | | | | | | | | |
| 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

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

- 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

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

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

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV

- Example

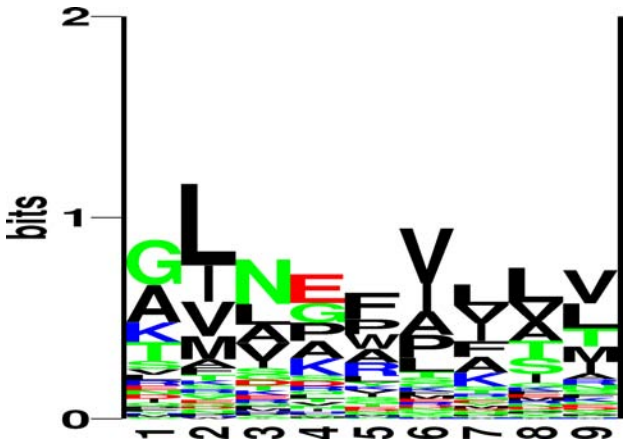
$$p_a = \frac{\alpha \cdot f_a + \beta \cdot g_a}{\alpha + \beta}$$

- If $\alpha$ 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

## Sequence weighting and pseudo counts

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

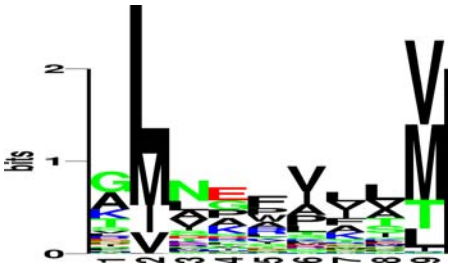
```
ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV
```

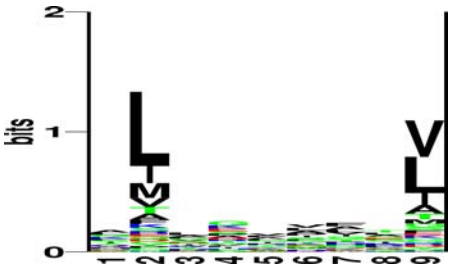



## Position specific weighting

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

- 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

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

- 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

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

- 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 | 0.5 | 0.0 | 1.4 | 1.2 | -2.7 | 1.4 | -1.2 | -2.0 | 1.1 | 0.7 |
| 2 | -1.6 | -0.6 | -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 | -3.1 | -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 | -2.5 | 1.1 | -1.3 | -1.4 | -0.2 | -1.0 | 1.8 | 0.8 | -1.9 | 0.2 | 1.0 | -0.1 | -0.6 | 0.4 | -0.5 | -0.0 | 2.1 |
| 7 | 1.1 | -3.8 | -0.2 | -1.3 | 1.3 | -0.3 | -1.2 | -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 | -2.3 | -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)

## 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 |

## Example from real life

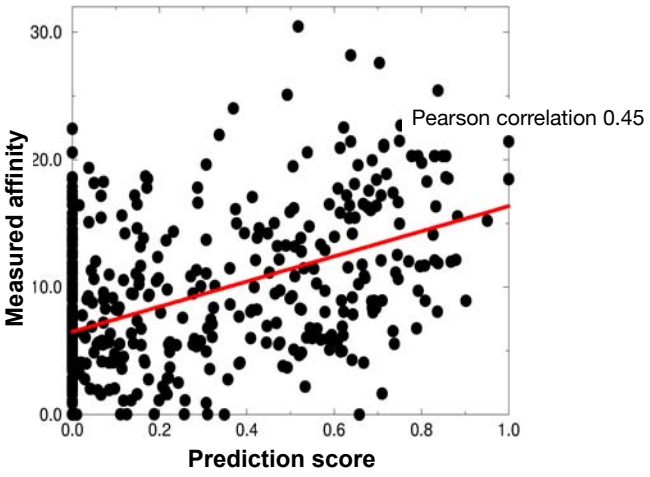


- 10 peptides from MHCpep database
- Bind to the MHC complex
- Relevant for immune system recognition
- Estimate sequence motif and weight matrix
- Evaluate motif "correctness" on 528 peptides

ALAKAAAAM
ALAKAAAAN
ALAKAAAAR
ALAKAAAAT
ALAKAAAAV
GMNERPILT
GILGFVFTM
TLNAWVKVV
KLNEPVLLL
AVVPFIVSV

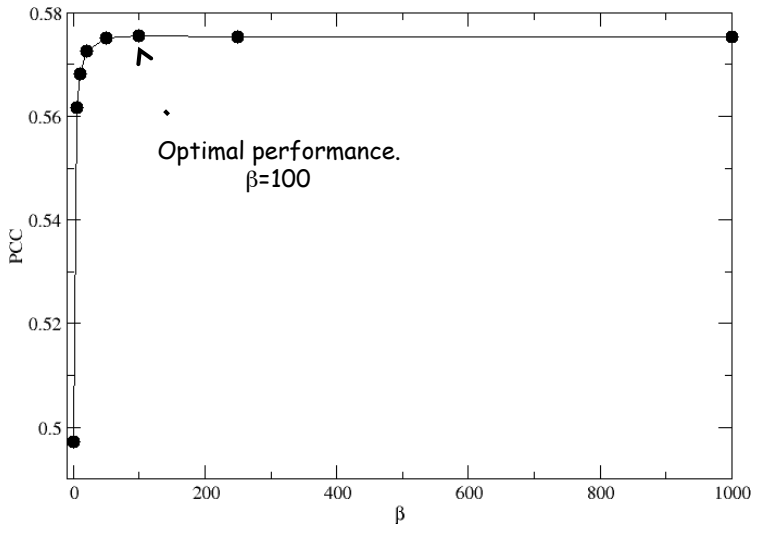
## Prediction accuracy





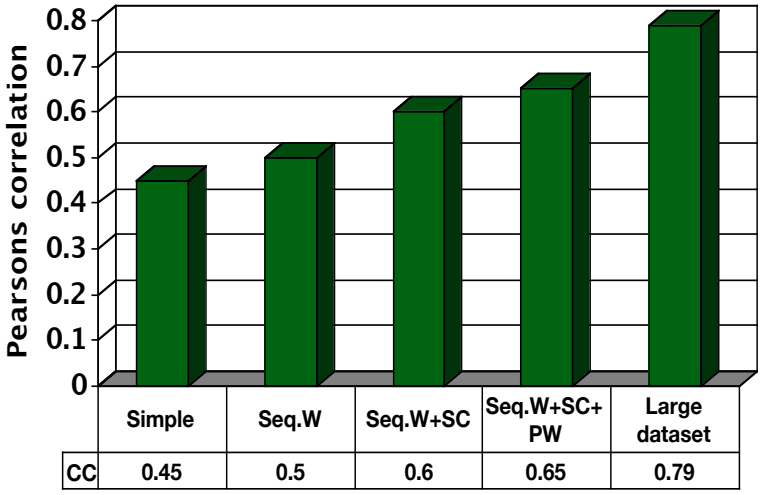

## How to define β?





## 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

## Sequence Profiles and Weight matrices

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

- 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 alignment

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

• Conventional sequence alignment uses a (Blosum) scoring matrix to identify amino acids matches in the two protein sequences

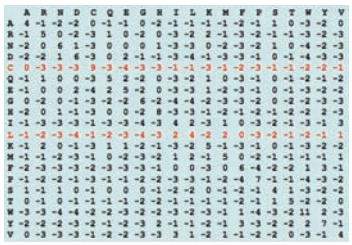
| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 4 | -1 | -2 | -2 | 0 | -1 | -1 | 0 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 0 | -3 | -2 | 0 |
| R | -1 | 5 | 0 | -2 | -3 | 1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

## Alignment scoring matrices

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

- Blosum62 score matrix. Fg=1. Ng=0?

| | L | A | G | D | S | D |
|---|---|---|---|---|---|---|
| F | | | | | | |
| I | | | | | | |
| G | | | | | | |
| D | | | | | | |
| S | | | | | | |
| L | | | | | | |



## Alignment scoring matrices

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

- Blosum62 score matrix. Fg=1. Ng=0?

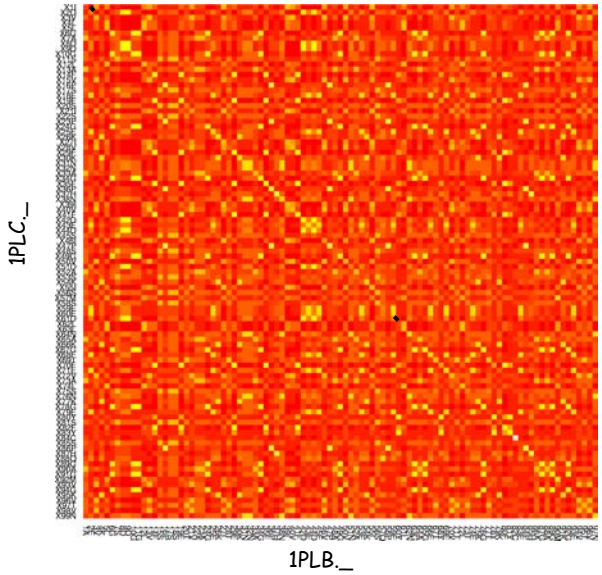
| | L | A | G | D | S | D |
|---|---|---|---|---|---|---|
| F | 0 | -2 | -3 | -3 | -2 | -3 |
| I | 2 | -1 | -4 | -3 | -2 | -3 |
| G | -4 | 0 | 6 | -1 | 0 | -1 |
| D | -4 | -2 | -1 | 6 | 0 | 6 |
| S | -2 | 1 | 0 | 0 | 4 | 0 |
| L | 4 | -1 | -4 | -4 | -2 | -4 |

- Score =2-1+6+6+4=17

**LAGDS**
**I-GDS**

## When Blast works!

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS




## What goes wrong when Blast fails?

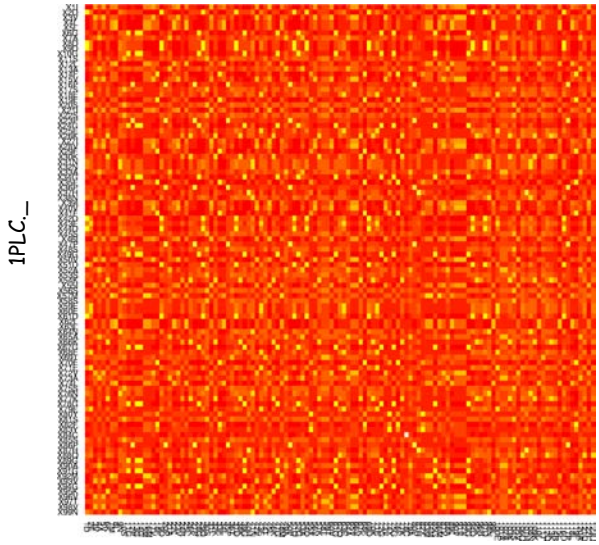
CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

- Conventional sequence alignment uses a (Blosum) scoring matrix to identify amino acids matches in the two protein sequences
- This scoring matrix is identical at all positions in the protein sequence!

```
  EVVFIGDSLVQLMHQC
A
G    X
D     X
S      X
.
G
G
G    X
D     X
S      X
```

## When Blast fails!

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS




## Alignment match scores

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 | 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 |

TVNGQ--FPGPRLAGVAREGDQVLVKVVNHVAENITIHWHGVQLGTGWADGPAYVTQCPI

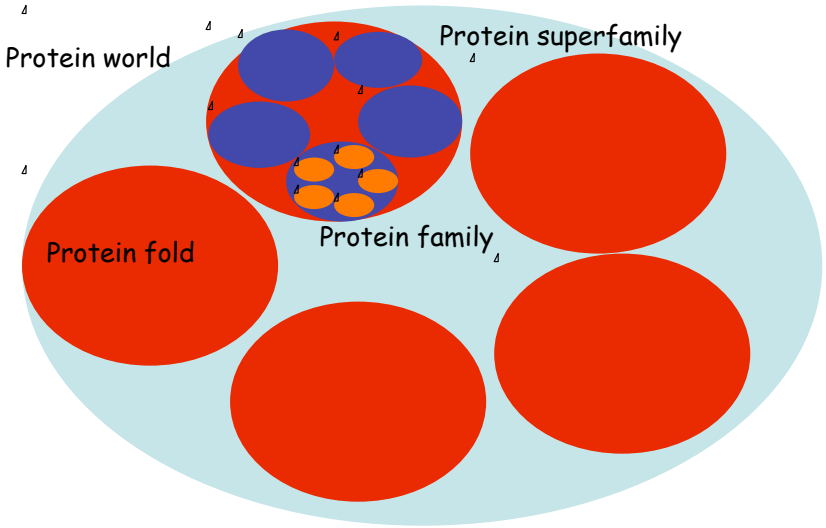
## Sequence profiles

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

- In reality not all positions in a protein are equally likely to mutate
  - Some amino acids (active cites) are highly conserved, and the score for mismatch must be very high
  - Other amino acids can mutate almost for free, and the score for mismatch should be lower than the BLOSUM score
- Sequence profiles can capture these differences

## Protein structure classification

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS



## Sequence profiles

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

```
TVNGQ--FPGPRLAGVAREGDQVLVKVVNHVAENITIHWHGVQLGTGWADPPAYVTQCPI
TKAVVLTFNTSVEICLVMQGTSIV----AAESHPLHLHGFNFPSNFNLVDGMERNTAGVP
```

## Sequence profiles

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Conserved

Non-conserved

```
ADDGSLAFVPSEF--SISPGEKIVFKNNAGFPHNIVFDEDSIPSGVDASKISMSEEDLLN
TVNGAI--PGPLIAERLKEGQNVRVTNTLDEDTSIHWHGLLVPFGMDGVPGVSFPG---I
-TSMAPAFGVQEFYRTVKQGDEVTVTIT-----NIDQIED-VSHGFVVVNHGVSME---I
IE--KMKYLTPEVFYTIKAGETVYWVNGEVMPHNVAFKKGIV--GEDAFRGEMMTKD---
-TSVAPSFSQPSF-LTVKEGDEVTVIVTNLDE------IDDLTHGFTMGNHGVAME---V
ASAETMVFEPDFLVLEIGPGDRVRFVPTHK-SHNAATIDGMVPEGVEGFKSRINDE----
TVNGQ--FPGPRLAGVAREGDQVLVKVVNHVAENITIHWHGVQLGTGWADPPAYVTQCPI
TKAVVLTFNTSVEICLVMQGTSIV----AAESHPLHLHGFNFPSNFNLVDGMERNTAGVP
```

Matching any thing but *G* => large negative score

Any thing can match

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

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

## How to make sequence profiles

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

1. Align (BLAST) sequence against large sequence database (Swiss-Prot)
2. Select significant alignments and make sequence profile
3. Use profile to align against sequence database to find new significant hits
4. Repeat 2 and 3 (normally 3 times!)

## Sequence logos. Visualization of sequence profiles

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

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

$P_A$ = 6/10 = 0.6
$P_G$ = 2/10 = 0.2
$P_T = P_K$ = 1/10 = 0.1
$P_C = P_D = \ldots P_V$ = 0.0

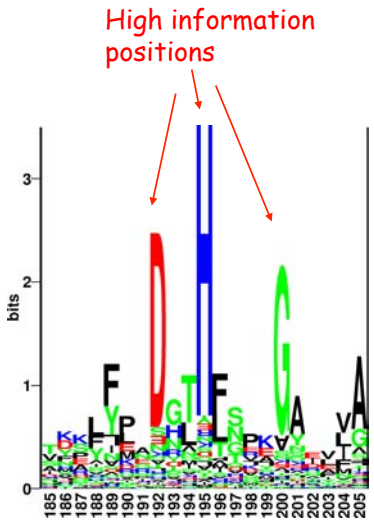
$q_A$ = 0.07
$q_G$ = 0.07
$q_T$ = 0.05
$q_K$ = 0.06

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

## Sequence logos

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

- Height of a column equal to I
- Relative height of a letter is p (letters are upside down if q>p)




## Sequence profiles (1J2J.B)

**IFEDEEKSKMLARLLKSSHPEDLRAANKLIKELVQEDQKRLEK**

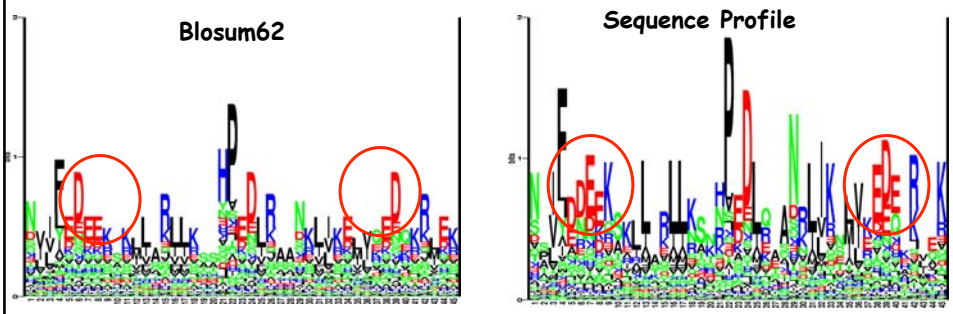
**Blosum62**

| | 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 |

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

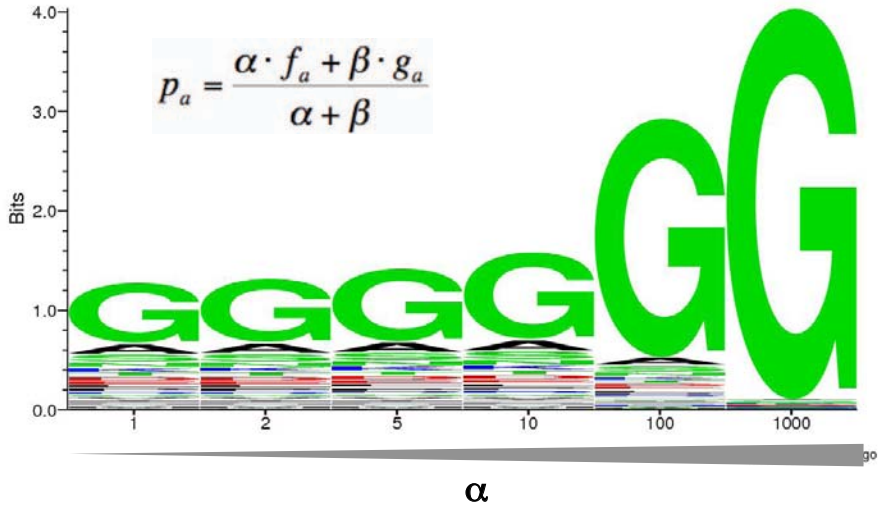
## Sequence profiles (1J2J.B)

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS




## Sequence profiles or Gaining confidence

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS



## Example.

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

>1K7C.A
TTVYLAGDSTMAKNGGGSGTNGWGEYLASYLSATVVNDAVAGRSARSYTREGRFENIADV
VTAGDYVIVEFGHNDGGSLSTDNGRTDCSGTGAEVCYSVYDGVNETILTFPAYLENAAKL
FTAKGAKVILSSQTPNNPWETGTFVNSPTRFVEYAELAAEVAGVEYVDHWSYVDSIYETL
GNATVNSYFPIDHTHTSPAGAEVVAEAFLKAVVCTGTSLKSVLTTTSFEGTCL

- What is the function
- Where is the active site?

## What would you do?

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

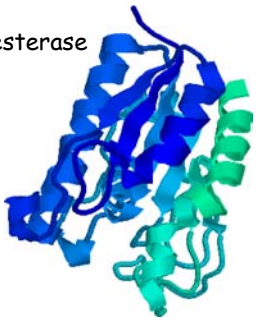
- Function
  - Run Blast against PDB
    - No significant hits
  - Run Blast against NR (Sequence database)
    - Function is Acetylesterase?
- Where is the active site?

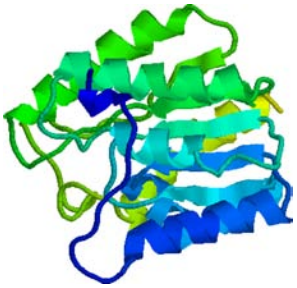
## Example. Where is the active site?

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

1USW Hydrolase

1G66 Acetylxylan esterase

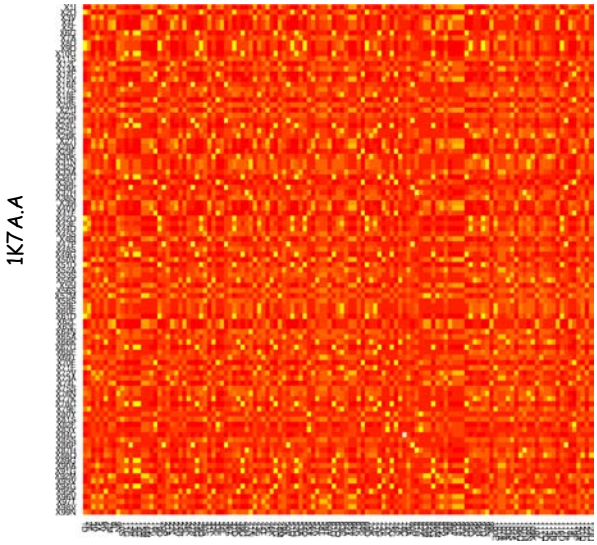




1WAB Acetylhydrolase

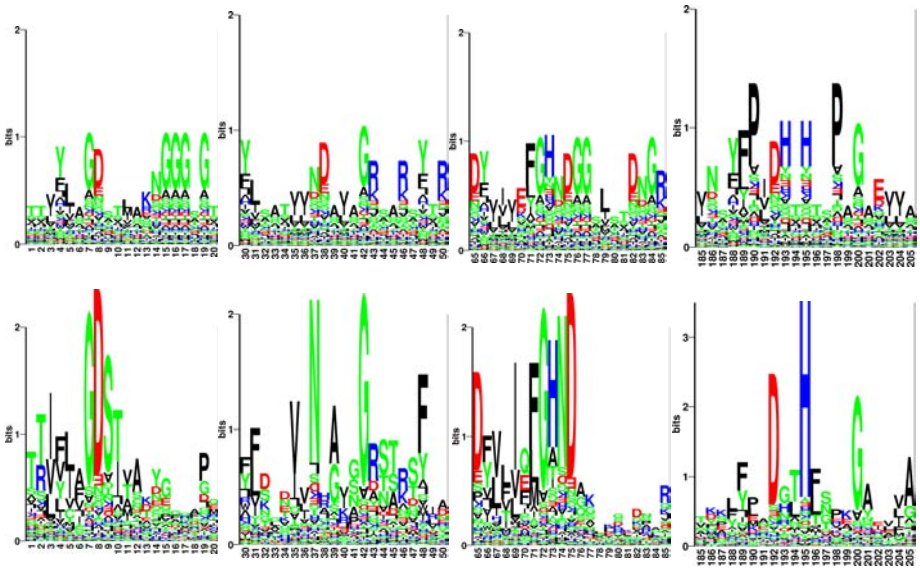
## When Blast fails!

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS



## Example. (SGNH active site)

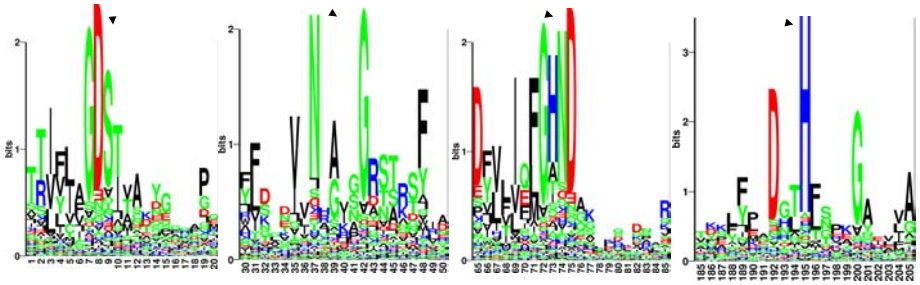
CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS



## Example. Where is the active site?

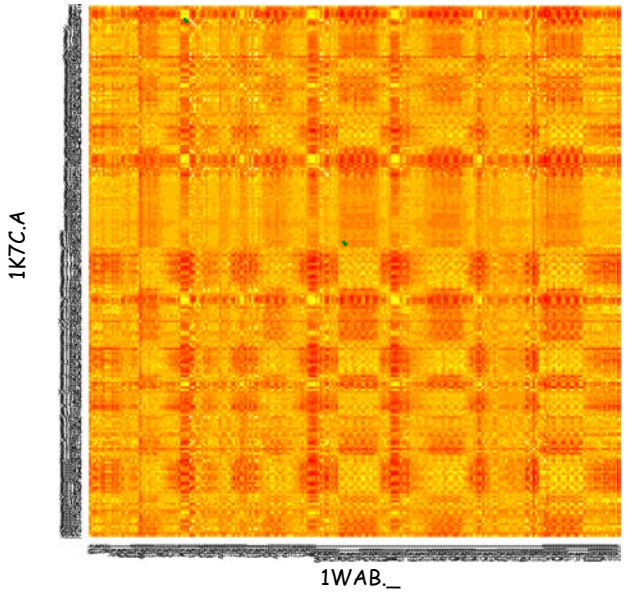
CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

- Sequence profiles might show you where to look!
- The active site could be around
  - S9, G42, N74, and H195



## Profile-profile scoring matrix

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS




## Example. Where is the active site?

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

### Align using sequence profiles

ALN 1K7C.A 1WAB._ RMSD = 5.29522. 14% ID

```
1K7C.A TVYLAGDSTMAKNGGGSGTNGWGEYLASYLSATVVNDAVAGRSARSYTREGRFENIADVVTAGDYVIVEFGHNDGGSLSTDN
          S                                 G                                 N
1WAB._ EVVFIGDSLVQLMHQCE---IWRELFS---PLHALNFGIGGDSTQHVLW--RLENGELEHIRPKIVVVWVGTNNHG------

1K7C.A GRTDCSGTGAEVCYSVYDGVNETILTFPAYLENAAKLFTAK--GAKVILSSQTPNNPWETGTFVNSPTRFVEYAEL-AAEVA
1WAB._ ----------------------HTAEQVTGGIKAIVQLVNERQPQARVVVLGLLPRGQ-HPNPLREKNRRVNELVRAALAGHP

1K7C.A GVEYVDHWSYVDSIYETLGNATVNSYFPIDHTHTSPAGAEVVAEAFLKAVVCTGTSL
                                        H
1WAB._ RAHFLDADPG---FVHSDG--TISHHDMYDYLHLSRLGYTPVCRALHSLLLRL---L
```

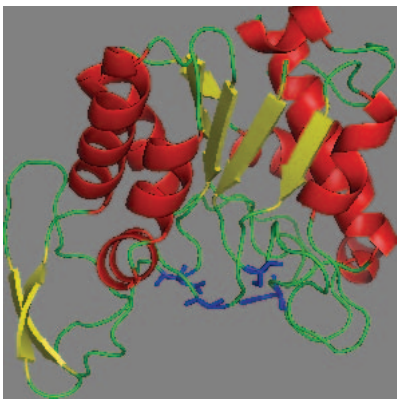
CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

Handout exercise

Using Psi-Blast Profiles

## Where is the active site?

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS



Rhamnogalacturonan
acetylesterase (1k7c)

# Blast2logo

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

**Blast2logo 1.0 Server**

Instructions | Output format

**SUBMISSION**

*Paste a single sequence in FASTA format into the field below:*

```
>Ex
VALAELYIPEVARRLGQGWHEDECTFAEVTIGTARLQAILRDIATSWSADEGGMRDGPAVLVLLPPG
EQHTLGAMVAVAKLRRLGVSVCLRMSTGPAELRELFGKRRFDAIMISLAHAEMLEVGRKLVKTLKD
MTGGRIPVAMGGALFLDGTEAASIPEADIVTNDIEAALQ
```

*Submit a file in FASTA format directly from your local disk:*

Choose File no file selected

*Upload a file in BLAST PROFILE format:*

Choose File no file selected

***Blast Database*** SP

***Number of Blast iterations*** 1

***Blast E-value cutoff*** 0.00001

***Stack Linesize*** 90

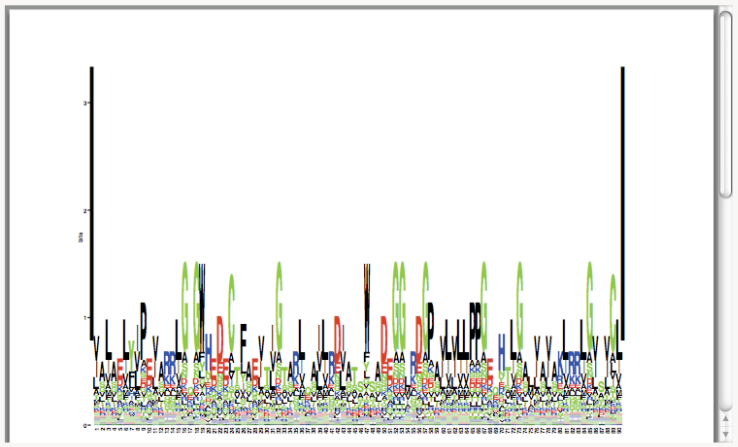
***Plot Kullback-Leibler logo*** ☐

***File format for logo file*** PDF

Submit Clear fields

# Blast2logo

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS



Download logo file Logo

Link to Blastprofile output file Blast.prof

# Blast2logo

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Last position-specific scoring matrix computed

| | | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | V | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | -2 | 0 | -3 | -1 | 4 |
| 2 | A | 4 | -1 | -2 | -2 | 0 | -1 | -1 | 0 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 0 | -3 | -2 | 0 |
| 3 | L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | -2 | 2 | 0 | -3 | -2 | -1 | -2 | -1 | 1 |
| 4 | A | 4 | -1 | -2 | -2 | 0 | -1 | -1 | 0 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 0 | -3 | -2 | 0 |
| 5 | E | -1 | 0 | 0 | 2 | -4 | 2 | 5 | -2 | 0 | -3 | -3 | 1 | -2 | -3 | -1 | 0 | -1 | -3 | -2 | -2 |
| 6 | L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | -2 | 2 | 0 | -3 | -2 | -1 | -2 | -1 | 1 |
| 7 | Y | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | -1 |
| 8 | I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | 2 | -3 | 1 | 0 | -3 | -2 | -1 | -3 | -1 | 3 |
| 9 | P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | -1 | -1 | -4 | -3 | -2 |
| 10 | E | -1 | 0 | 0 | 2 | -4 | 2 | 5 | -2 | 0 | -3 | -3 | 1 | -2 | -3 | -1 | 0 | -1 | -3 | -2 | -2 |

.
.

| | 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 |

# Blast2logo

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

**Blast2logo 1.0 Server**

Instructions | Output format

**SUBMISSION**

*Paste a single sequence in FASTA format into the field below:*

>Ex
VALAELYIPEVARRLCQGWHEDECTFAEVTIGTARLQAILRDIATSWSADEGCMRDCPAVLVLLPPC
EQHTLGAMVAVAKLRRLCVSVCLRMSTGPAELRLLFCKRRFDAIMISLAHAEMLEVGRKLVKTLKD
MTCGRIPVAMCGALFLDCTEAASIPEADIVTNDIEAALQ

*Submit a file in FASTA format directly from your local disk:*

Choose File | no file selected

*Upload a file in BLAST PROFILE format:*

Choose File | no file selected

*Blast Database* NR70

*Number of Blast iterations* 3

*Blast E-value cutoff* 0.00001

*Stack Linesize* 90

*Plot Kullback-Leibler logo*

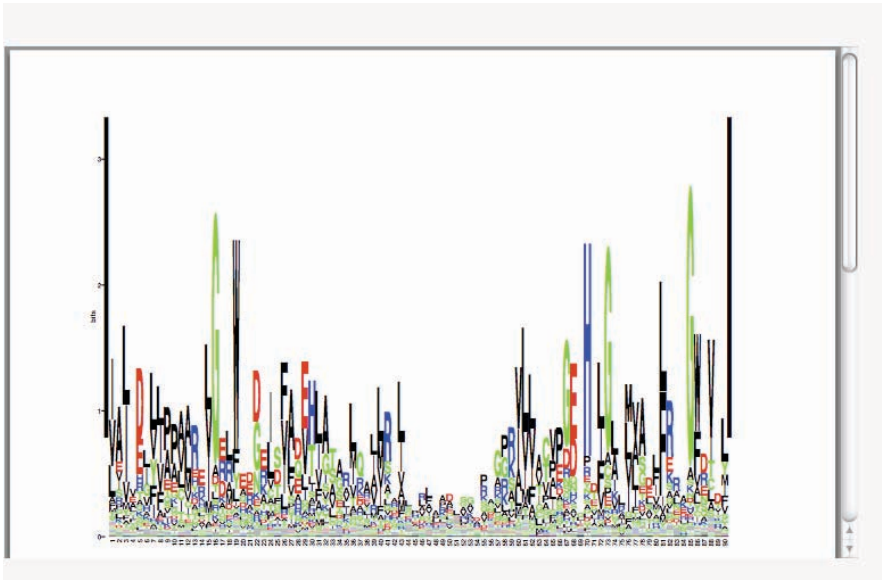
*File format for logo file* PDF

Submit | Clear fields

*Restrictions:*
*At most 1 sequences per submission; each sequence not more than 20,000 amino acids.*

# Blast2logo

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS



# Blast2logo

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Last position-specific scoring matrix computed,

| | | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | V | -2 | -4 | -4 | -5 | -2 | -4 | -4 | -5 | -4 | 5 | 2 | -4 | 0 | -1 | -4 | -3 | -2 | -4 | -2 | 4 |
| 2 | A | 5 | 0 | -3 | -3 | -3 | -2 | 1 | -2 | -3 | 0 | -3 | -2 | -2 | -4 | 0 | 0 | -2 | -4 | -3 | 0 |
| 3 | L | -4 | -5 | -6 | -6 | -4 | -5 | -5 | -6 | -5 | 5 | 4 | -5 | 1 | -2 | -5 | -5 | -3 | -4 | 0 | 1 |
| 4 | A | 1 | -4 | -1 | -1 | 3 | -1 | 2 | -4 | -3 | 0 | -1 | -2 | -3 | 1 | -4 | 0 | 0 | -4 | 2 | 2 |
| 5 | E | -2 | 0 | -2 | 6 | -6 | 0 | 4 | -4 | 2 | -5 | -5 | -2 | -5 | -6 | -4 | -2 | 0 | -6 | -4 | -5 |
| 6 | L | -1 | -2 | -4 | -4 | -4 | -2 | -1 | 2 | 3 | 3 | 2 | -1 | 0 | -2 | -5 | -1 | -1 | -5 | -3 | 1 |
| 7 | Y | -4 | -5 | -5 | -6 | -4 | -5 | -5 | -4 | 0 | 1 | 4 | -5 | -1 | 3 | -5 | -5 | -4 | -3 | 5 | 3 |
| 8 | I | -1 | -2 | -5 | -5 | -4 | -5 | -2 | -6 | -5 | 4 | 3 | -5 | -1 | 3 | -5 | -4 | -2 | -4 | -1 | 3 |
| 9 | P | 3 | -4 | -4 | -3 | -4 | 1 | 1 | -4 | -2 | -2 | -3 | -2 | -4 | -5 | 6 | -1 | 0 | -5 | -5 | -2 |
| 10 | E | 2 | -2 | -3 | -2 | -3 | 0 | 1 | -1 | -3 | -4 | -3 | -1 | -1 | -4 | 6 | -2 | -2 | -4 | -4 | -3 |

## Sequence profiles take home message

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

- Blast will often fail to recognize sequence relationships for low homology sequence pairs
- Sequence profiles contain information on conserved/ variable residues in a protein sequence
- Sequence profiles are calculated from (multiple) sequence alignments
- Iterative Blast enables homology recognition also for low sequence similarity
- Sequence profiles give information on residues essential for protein function and protein structure

## Summary

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

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

## The Beauty of Sequence profiles



1) $\alpha$ = N-1 = 0
2) $p_a = g_a$
3) $f_G = 1$, $f_{!G} = 0$
4) $p_R = f_G * q(R|G) = \underline{0.02}$
5) $q_R = 0.052$
6) Log-odd = $2*\log(p_a/q_a)/\log(2) = -2.7$
7) Blosum62(G,R) = -2

TKAVVLTFNTSVEICLVMQGTSIV----AAESHPLHLHGFNFPSNFNLVDPMERNTAGVP

$$p_a = \frac{\alpha \cdot f_a + \beta \cdot g_a}{\alpha + \beta} \qquad g_b = \sum_a f_a \cdot q_{b|a}$$

## The Blosum matrix



| | 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 |