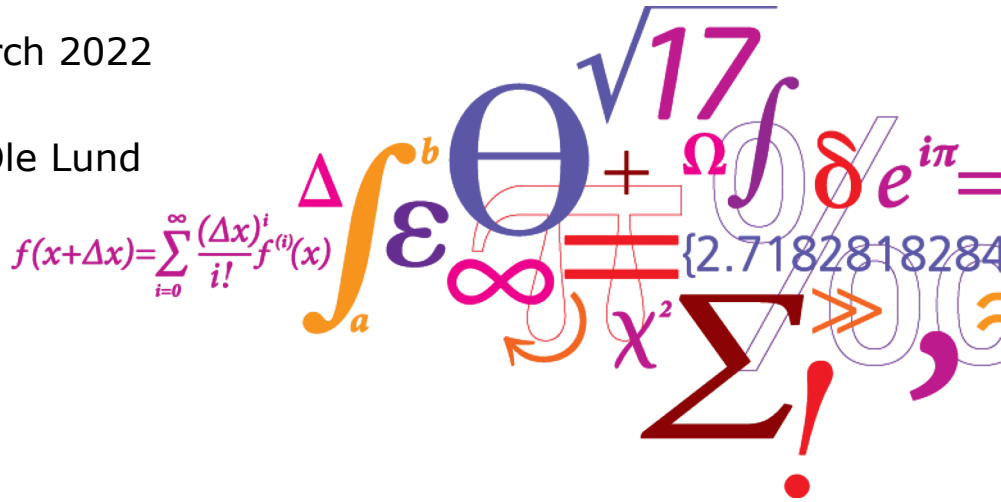
Sequence information and LOGO plots

Or: How to summarize and quantify sequence motifs

Rasmus Wernersson / Henrik Nielsen, March 2022

With examples from Morten Nielsen and Ole Lund



DTU Bioinformatics
Department of Bio and Health Informatics

Outline

- Why bother with LOGOs and matrices?
  - Summarizing information across sequences
  - When consensus sequences fail

- LOGO plots
  - How to construct them
  - How to interpret them

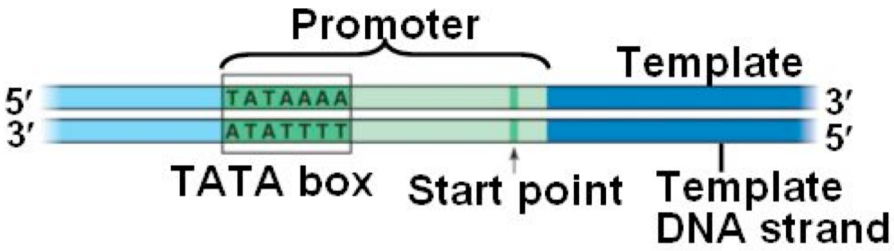
This week

- Weight matrices
  - How to construct them
  - How to apply them

Next week

Consensus sequences

- TATA/Pribnow box
  - "TATAAT"



- Shine-Dalgarno sequence
  - "AGGAGG"

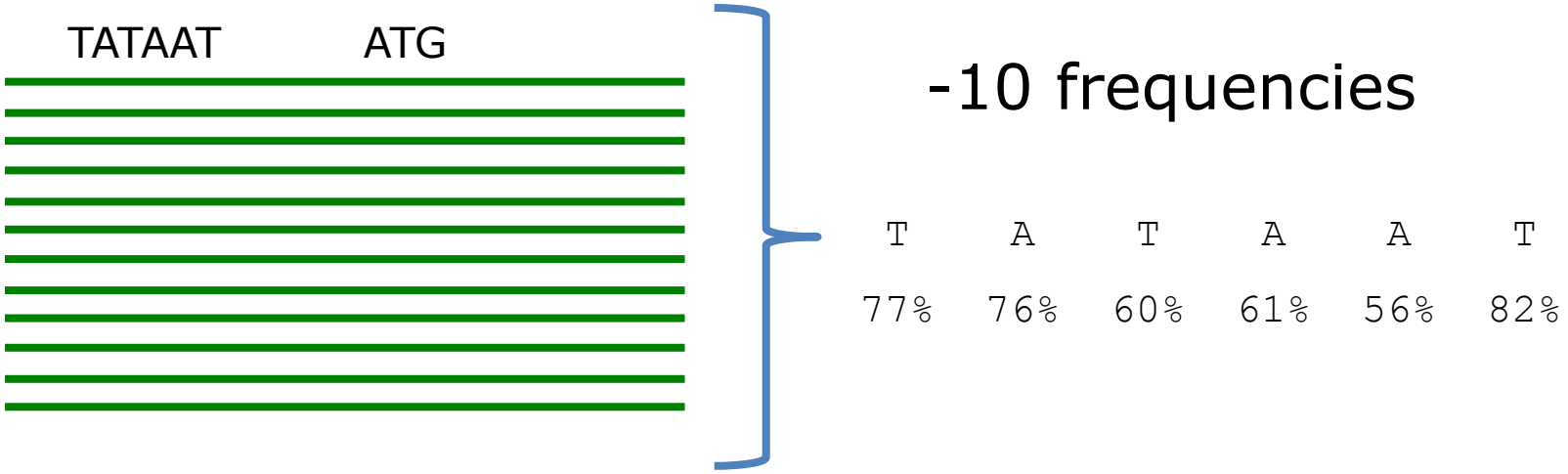
- Where do we get our knowledge from:
  - Observing many sequences
  - Multiple alignments

Why do we care about sequence motifs?

- Points to a molecular mechanism
- We can learn something new directly from comparing a lot of sequences
- Makes it possible to scan new sequences for known elements (e.g. “gene finding”)

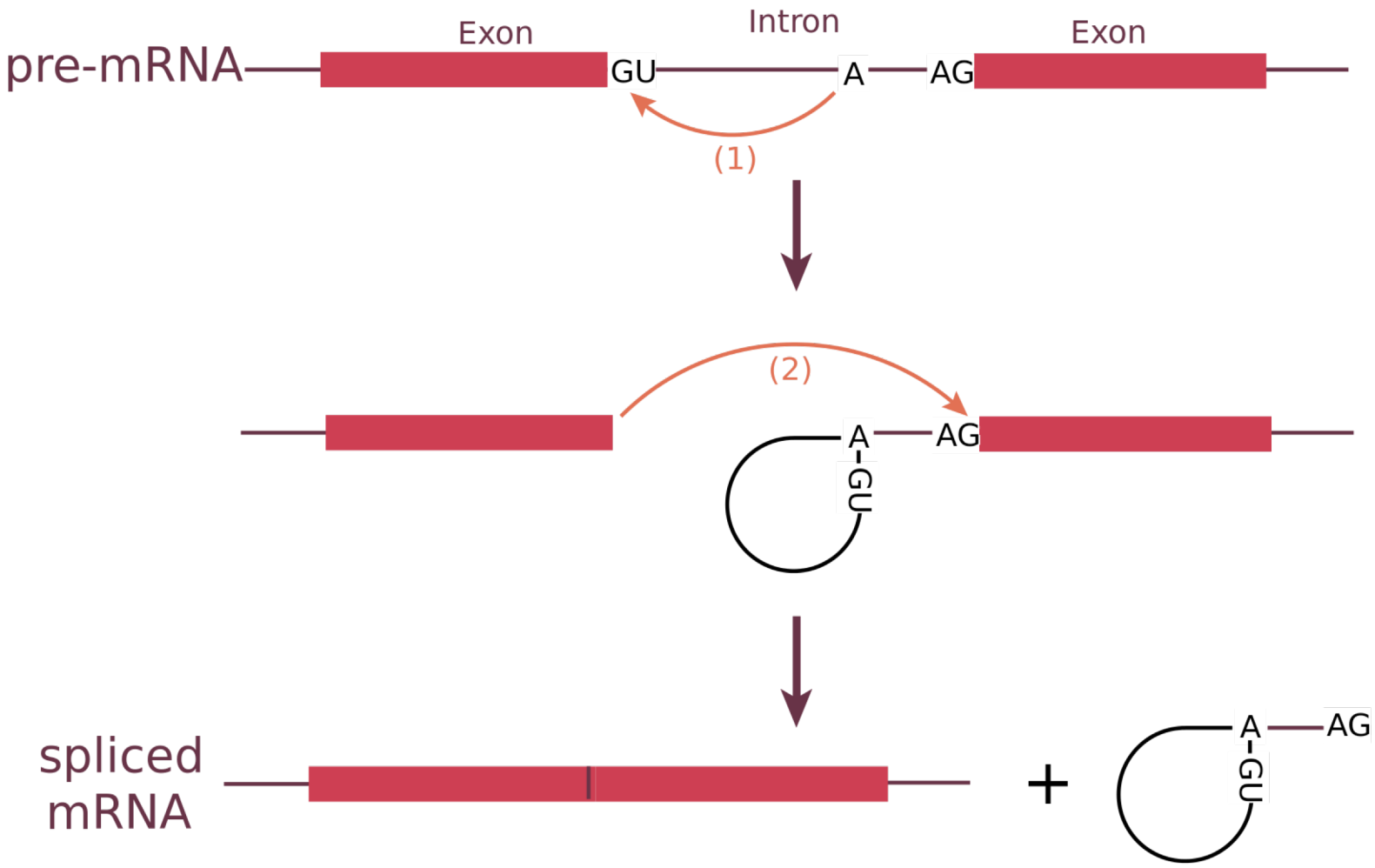
Does one size fit all?

- Consensus sequences are more like a rule of thumb — only a few Pribnow boxes *actually* look like "TATAAT"



- LOGO plots and weight matrices were invented to solve this

CASE: RNA splicing



RNA splicing – what is known?

- The splicing signal is contained WITHIN the intron
- Always* starts with GU ("donor site") and ends with AG ("acceptor site")
  - –GT / AG at the DNA level
- **QUESTION:** can we find any additional signal?

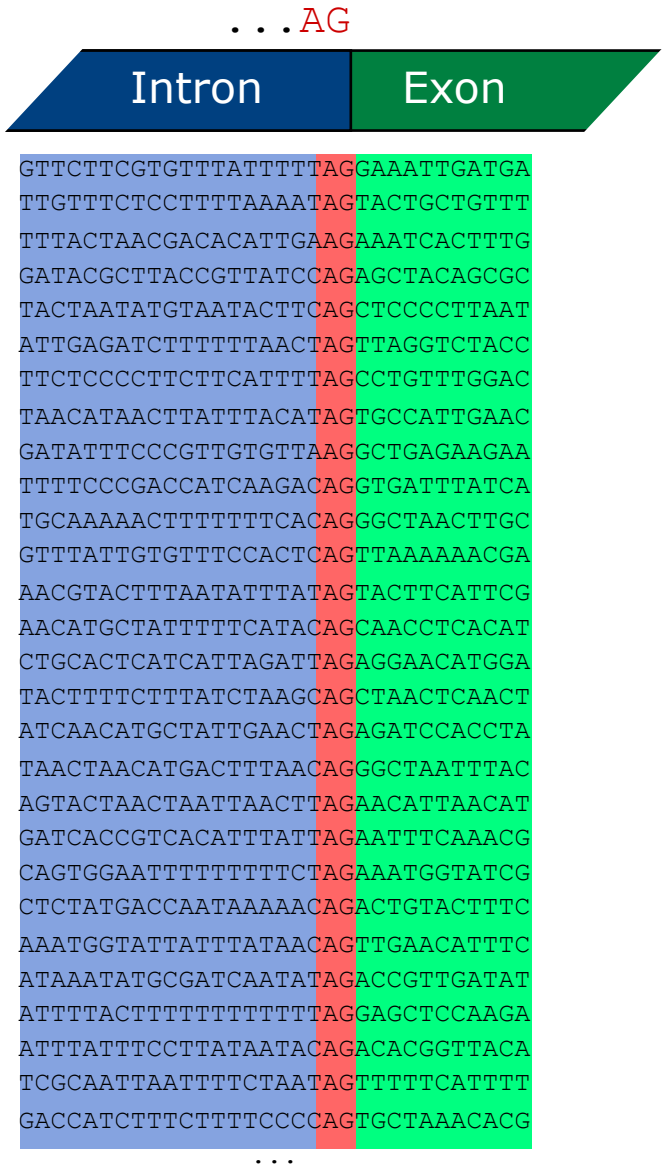
* Terms and conditions apply – batteries not included

Step 1: Define biological question

- Example:
- What is the signal around the **acceptor site** across all **yeast** (*Saccharomyces cerevisiae*) introns?
- This is important: what we find could be different if we compared to other organisms

Step 2: Gather data

- Download data from the yeast genome website
- Write a small program* to extract the intron/exon boundaries
- Stack up the sequences around the acceptor sites to make it easy to compare



* Or team up with a bioinformatician for this step

Step 3: statistics for each position

- Count occurrences

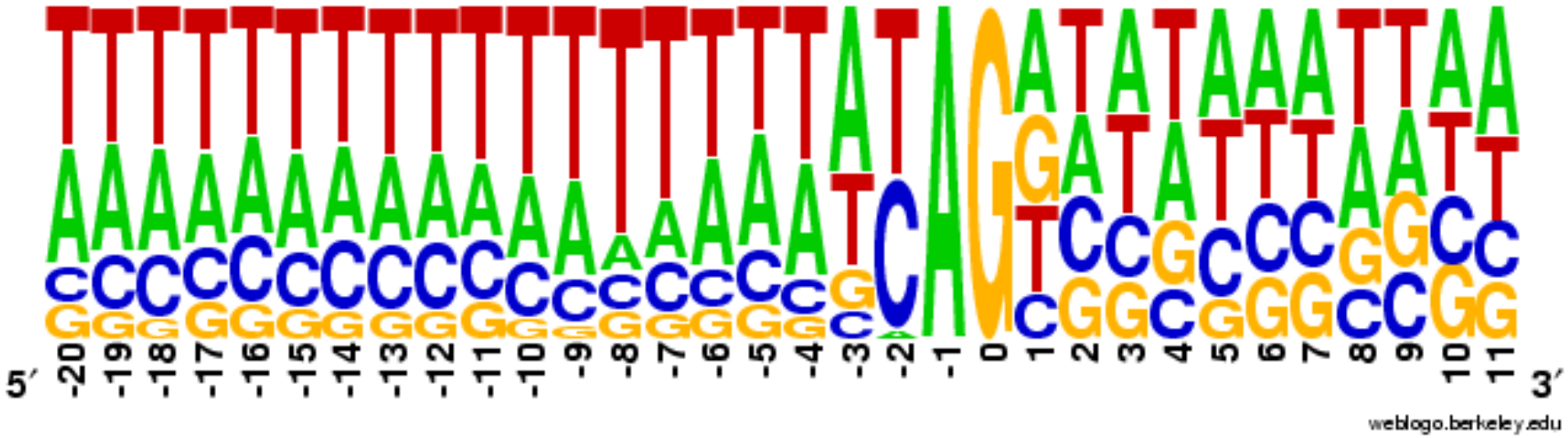
A	94	88	84	75	78	78	71	69	70	60	68	77	32	49	87	93	93	**134**	9	**266**	**0**	86	66	85	81	89	81	88	82
C	31	45	52	44	56	46	62	54	56	51	46	37	30	42	32	44	30	25	**122**	**1**	**0**	38	65	52	43	62	62	57	43
T	**113**	**110**	**113**	**117**	**104**	**117**	**111**	**120**	**118**	**125**	**136**	**140**	**182**	**155**	**122**	**100**	**124**	75	**137**	**0**	**0**	72	85	82	91	83	73	67	96
G	30	25	19	32	30	27	24	25	24	32	18	14	24	22	27	31	21	34	0	**1**	**268**	72	52	49	53	34	52	56	47

- Calculate frequencies (calc. for each column)

A	0,35	0,33	0,31	0,28	0,29	0,29	0,26	0,26	0,26	0,22	0,25	0,29	0,12	0,18	0,32	0,35	0,35	**0,50**	0,03	**0,99**	0,00	0,32	0,25	0,32	0,30	0,33	0,30	0,33	0,31
C	0,12	0,17	0,19	0,16	0,21	0,17	0,23	0,20	0,21	0,19	0,17	0,14	0,11	0,16	0,12	0,16	0,11	0,09	**0,46**	0,00	0,00	0,14	0,24	0,19	0,16	0,23	0,23	0,21	0,16
T	**0,42**	**0,41**	**0,42**	**0,44**	**0,39**	**0,44**	**0,41**	**0,45**	**0,44**	**0,47**	**0,51**	**0,52**	**0,68**	**0,58**	**0,46**	**0,37**	**0,46**	0,28	**0,51**	0,00	0,00	0,27	0,32	0,31	0,34	0,31	0,27	0,25	0,36
G	0,11	0,09	0,07	0,12	0,11	0,10	0,09	0,09	0,09	0,12	0,07	0,05	0,09	0,08	0,10	0,12	0,08	0,13	0,00	0,00	**1,00**	0,27	0,19	0,18	0,20	0,13	0,19	0,21	0,18

Step 4: Visualize the data

- Naïve visualization:

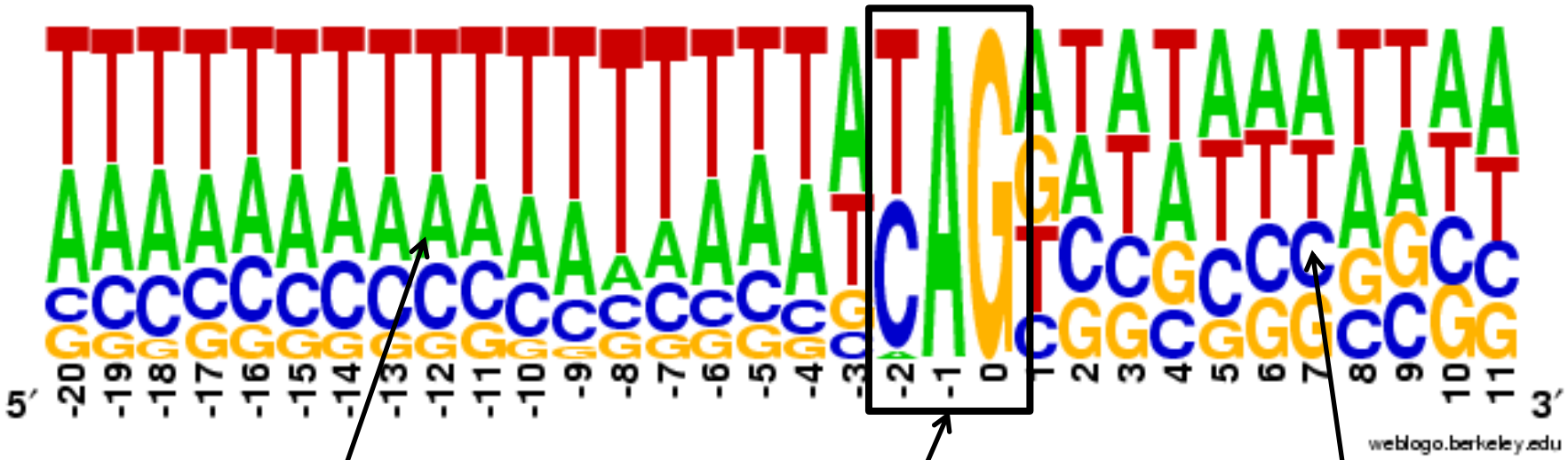


- AKA frequency LOGO
  - Each letter is proportional to the observed frequency
  - Easier overview than just looking at the tables

- BUT …. Are the observations **significant**??

How surprised are we at the observations?

- Frequency logo:



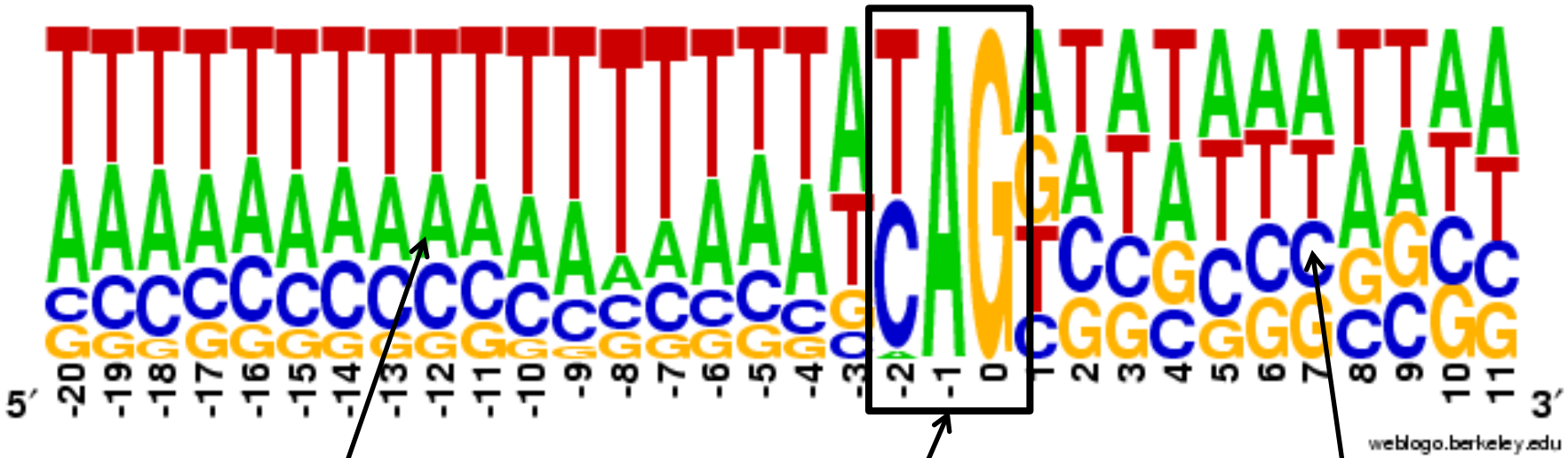
That's a lot of Ts and As

Surely, this cannot be random?

Hmm – looks pretty scrambled over here

How surprised are we at the observations?

- Frequency logo:



That's a lot of Ts

Umm... looks pretty random?

How does this compare to what we should expect by chance?

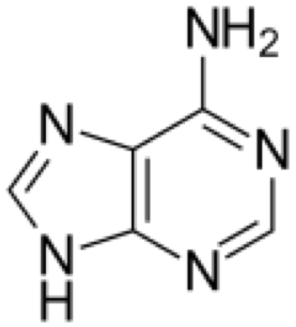
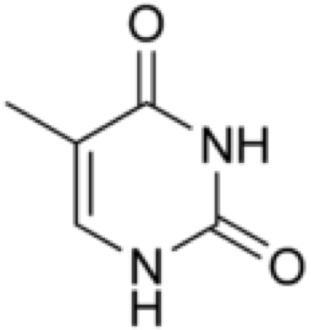
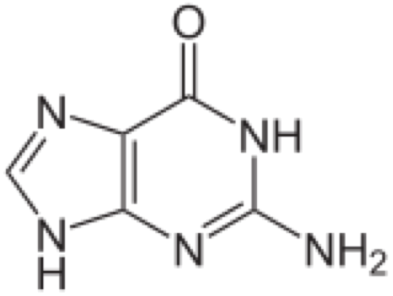
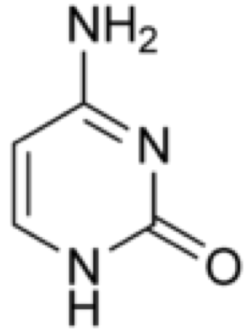
Information theory to the rescue

- Assumption (for now) – each letter (A, T, G, C) has the same background frequency
  - If you pick a **random** position each letter will be picked with 25% probability

- But – if there actually is a signal your **observed** probabilities will deviate from the **expected**

- This can be quantified by calculating the **information content** in each position in the data set (multiple alignment)

The bit as a yes/no answer

	Purine	Pyrimidine
"Weak" (2 H bonds)	 Adenine (**A**)	 Thymine (**T**)
"Strong" (3 H bonds)	 Guanine (**G**)	 Cytosine (**C**)

Question #1:

Is it a purine?

(yes/no => 1/0)

Question #2:

Is it a weak bond?

(yes/no => 1/0)

Q1 Q2

0,0 = C (no, no)

0,1 = T (no, yes)

1,0 = G (yes, no)

1,1 = A (yes, yes)

The bit as a yes/no answer

- To specify one out of eight possibilities, you need to answer three yes/no questions
- *In other words:* Having eight (equally probable) possibilities yields an uncertainty of three bits

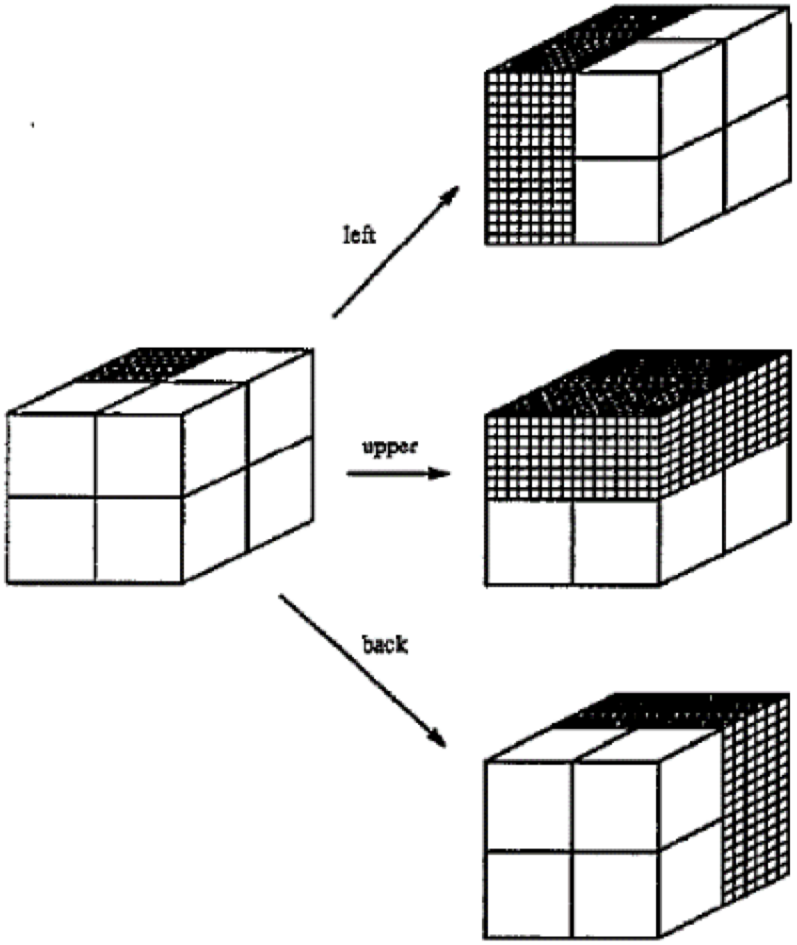


Figure 1. Three independent choices specify 1 box in 8.

N equally probable possibilities

N	H (bits)
2	1
4	2
8	3
16	4
32	5

$$N = 2^H$$

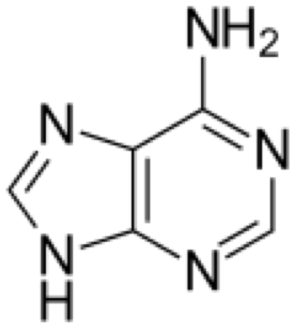
$$H = \log_2 N$$

But what happens if we already have *some* information?

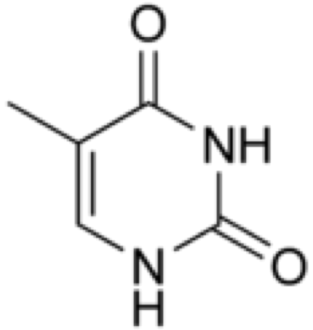
Purine Pyrimidine

"Weak"

(2 H bonds)



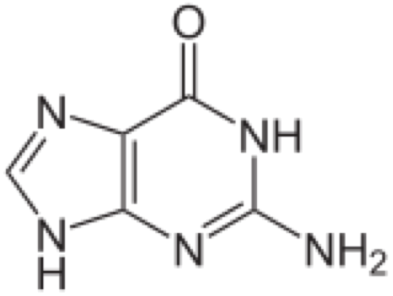
Adenine (**A**)



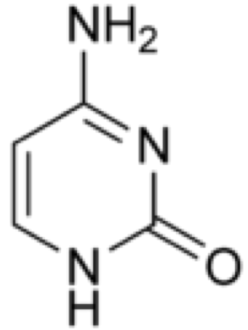
Thymine (**T**)

"Strong"

(3 H bonds)



Guanine (**G**)



Cytosine (**C**)

Question #1:

Is it a purine?

(yes/no => 1/0)

Question #2:

Is it a weak bond?

(yes/no => 1/0)

Q1 Q2

0,0 = C (no, no)

0,1 = T (no, yes)

1,0 = G (yes, no)

1,1 = A (yes, yes)

But what happens if we already have *some* information?

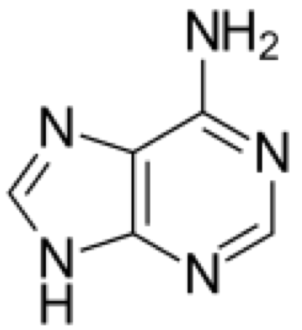
"Weak"

(2 H bonds)

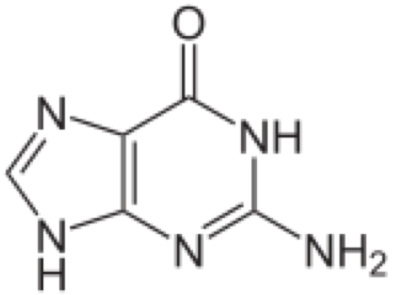
"Strong"

(3 H bonds)

Purine

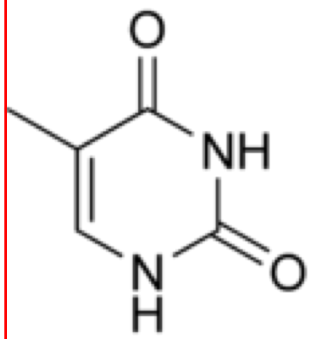


Adenine (**A**)

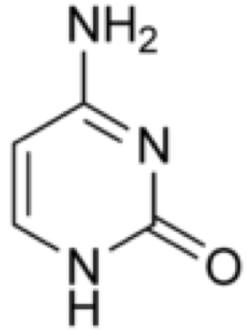


Guanine (**G**)

Pyrimidine



Thymine (**T**)



Cytosine (**C**)

Question #1:

Is it a purine?

(**yes**/no => 1/0)

Question #2:

Is it a weak bond?

(yes/no => 1/0)

Q1 Q2

~~0,0 = C (no, no)~~

~~0,1 = T (no, yes)~~

1,0 = G (yes, no)

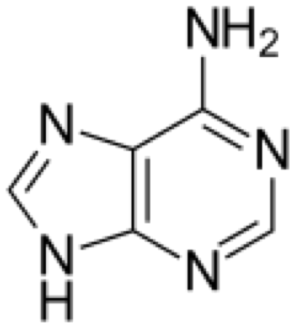
1,1 = A (yes, yes)

But what happens if we already have *some* information?

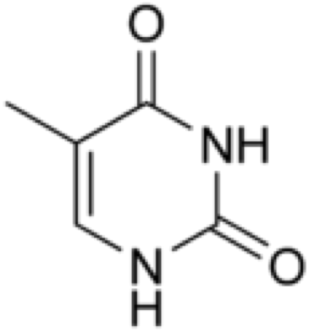
Purine Pyrimidine

"Weak"

(2 H bonds)



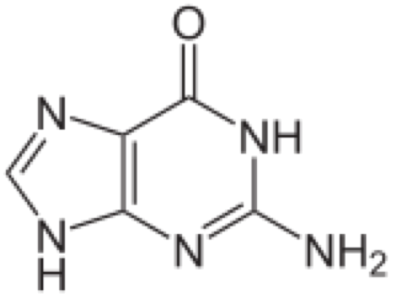
Adenine (**A**)



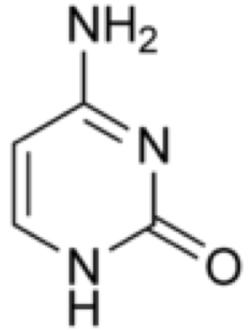
Thymine (**T**)

"Strong"

(3 H bonds)



Guanine (**G**)



Cytosine (**C**)

Question #1:

Is it a purine?

(yes/no => 1/0)

Question #2:

Is it a weak bond?

(yes/**no** => 1/0)

Q1 Q2

0,0 = C (no, no)

~~0,1 = T (no, yes)~~

1,0 = G (yes, no)

~~1,1 = A (yes, yes)~~

Generalized:
What if probabilities are not equal?

- If one possibility is more probable than the others, uncertainty will be lower:

$$H = -\sum_{n=1}^{N} p_n \log_2 p_n$$

N : number of symbols

(A,T,G,C) = 4

1. For each symbol calculate: Frequency * $\log_2$(frequency)
2. Sum it all up

Information content

$$R_{seq} = H_{\max} - H_{obs}$$

Maximum entropy

Observed entropy

$$\log_2 N - \left(- \sum_{n=1}^{N} p_n \log_2 p_n \right)$$

N : number of symbols

(A,T,G,C) = 4

1. For each symbol calculate: Frequency * $\log_2$(frequency)
2. Sum it all up

Information content

Theoretical questions:

1. What is the maximum R_{seq}? (we are most surprised)
2. What is the minimum R_{seq} (we are NOT surprised)

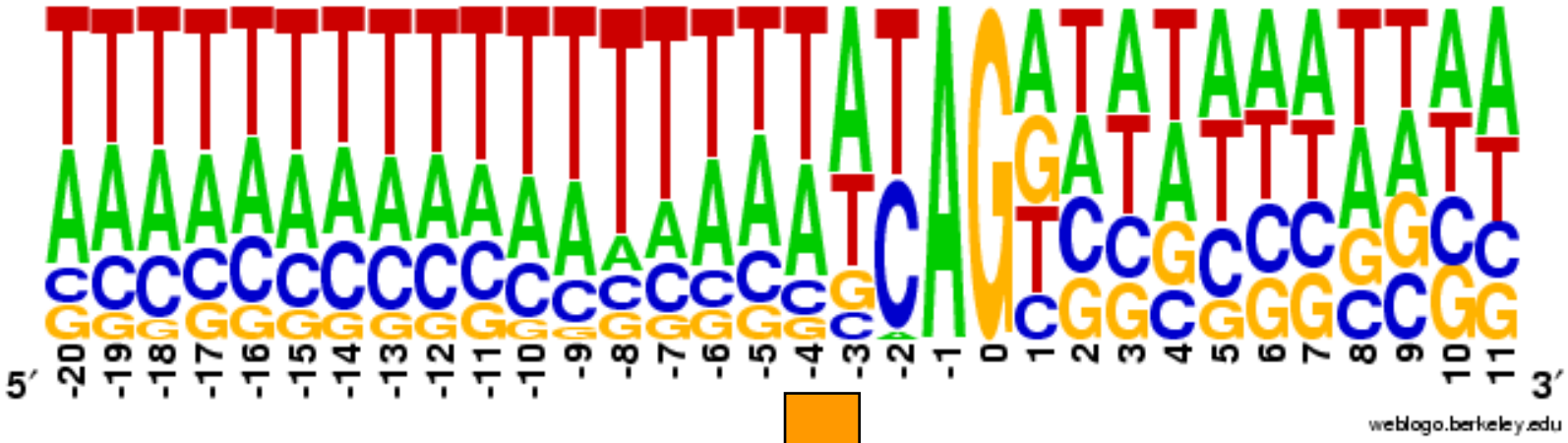
$$R_{seq} = H_{\max} - H_{obs} = \log_2 N - \left(- \sum_{n=1}^{N} p_n \log_2 p_n \right)$$

N : number of symbols

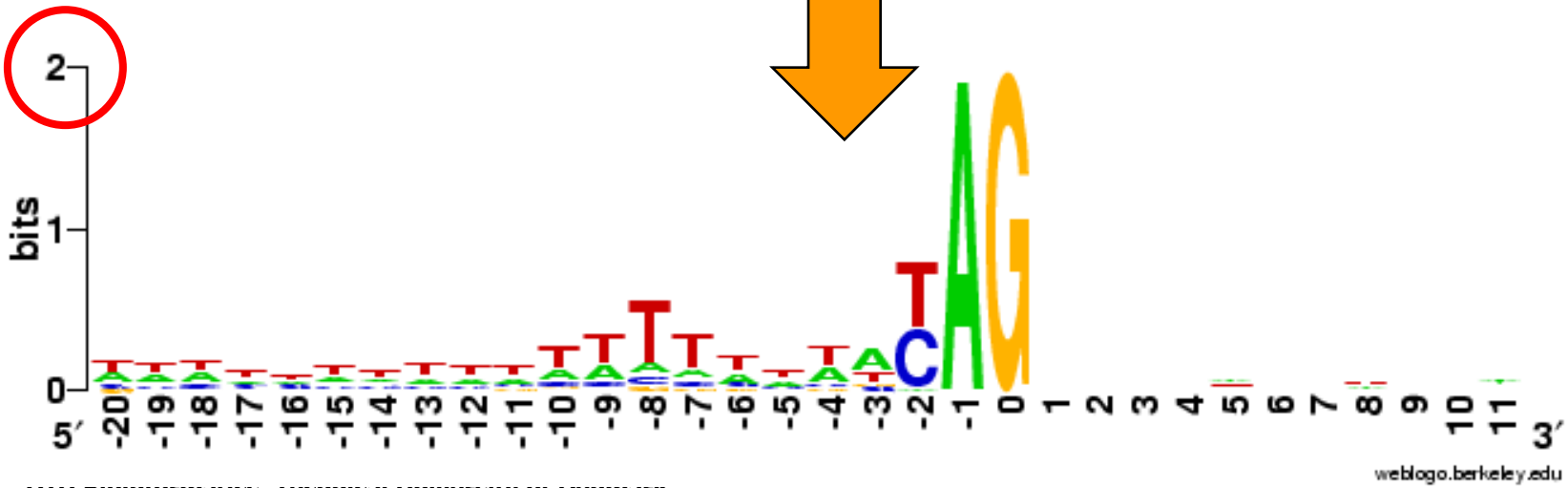
(A,T,G,C) = 4

1. For each symbol calculate: Frequency * $\log_2$(frequency)
2. Sum it all up

Step 5: Scale the visualization



Scale height by **information content**

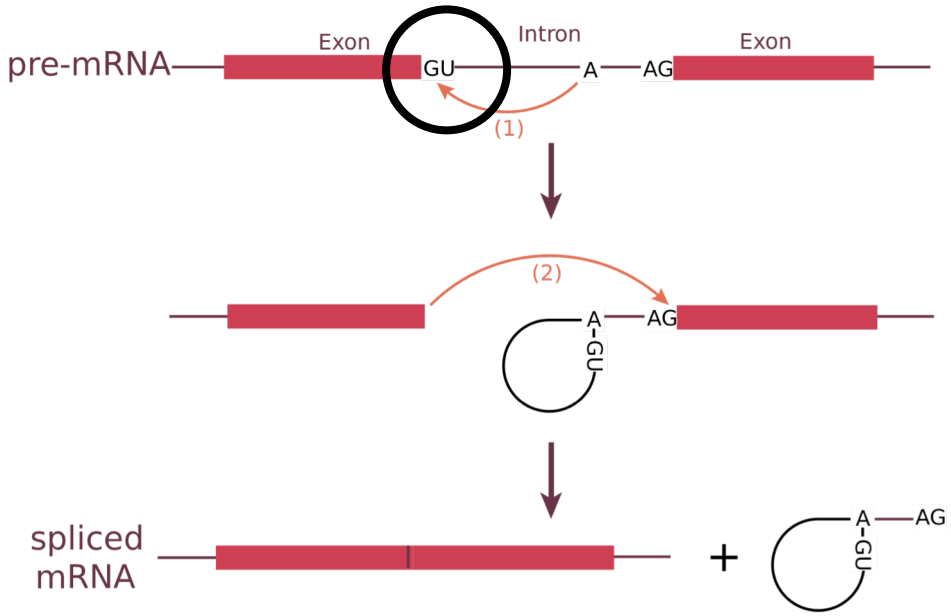


Making sequence logos – handout

Making Sequence logos

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
```



Hand-out, instructions 2022

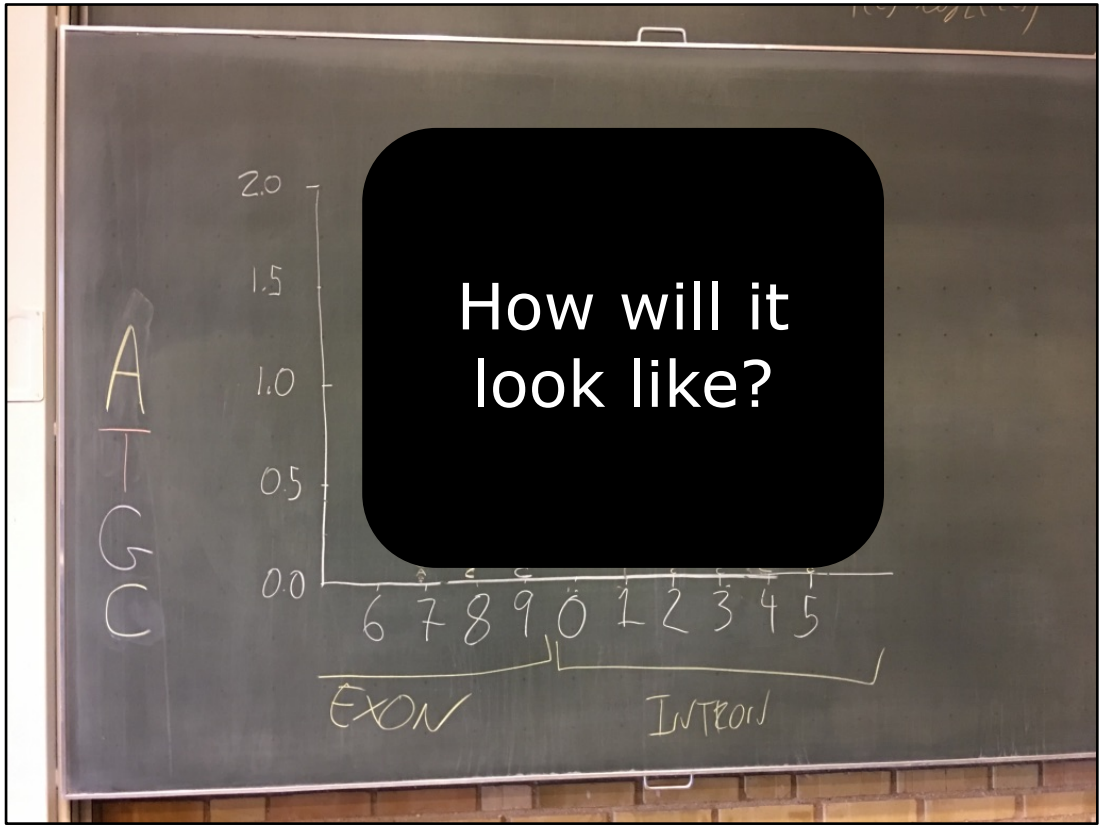
Work in groups of 2-3

Choose 3 positions

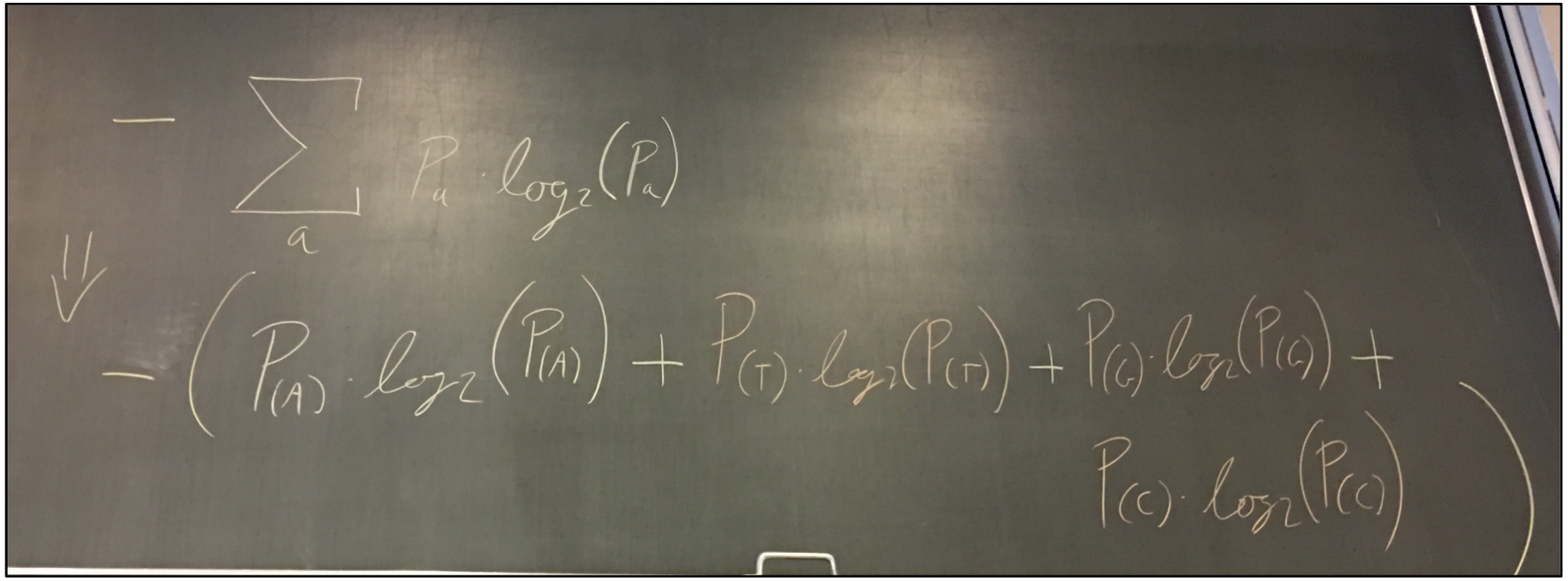
(e.g. 8, 0 and 2)

Do the calculations (by hand)

We'll draw the plot on the blackboard together after ~15 mins



Hand-out, instructions 2022



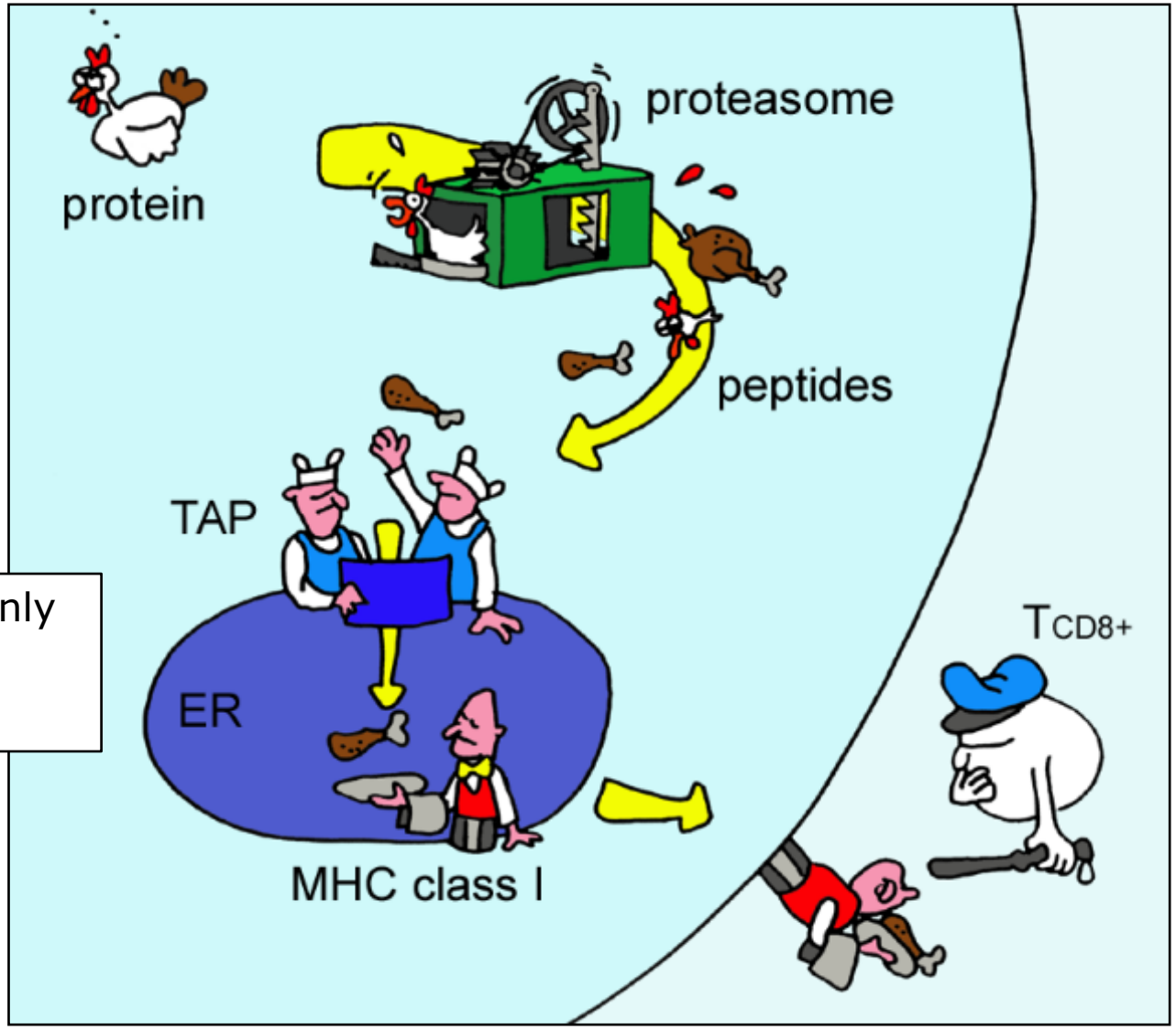
Notice: Since we work with DNA, the equation "expands" to these four elements

P(A) = observed frequency of "A" ***at that position***.

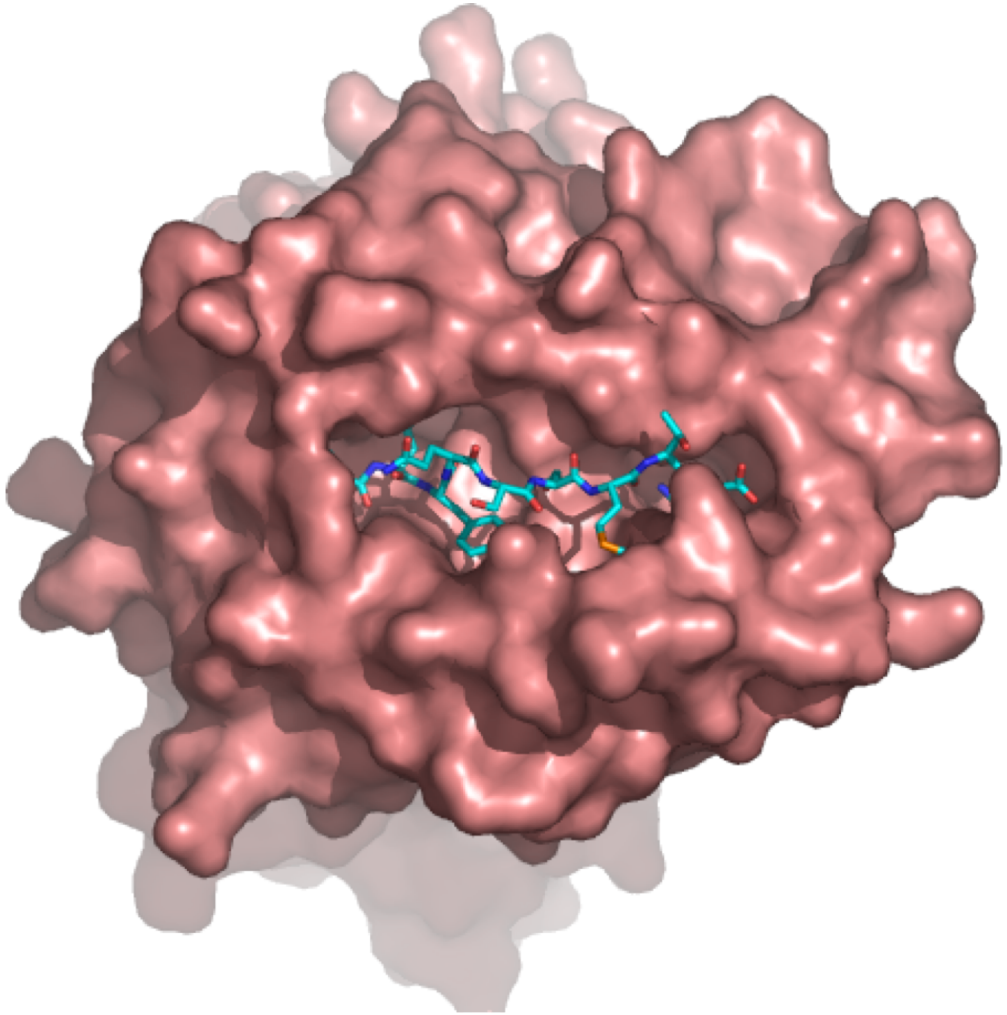
P(T) = observed frequency of "T" … (and so on).

CASE: MHC class 1 epitopes

The "snobby waiter" – only accepts the very best peptides



Only a few peptides will be "seen"



In this case 9 amino acid peptides

Step 1: Define biological question

- **Prior knowledge:** it is known that the sequence of the 9 amino acids is critical for the binding to MHC class 1

- **Question:** Can we describe the sequence pattern (motif) needed for MHC class 1 binding?

- (This can help us in vaccine design !!!)

Step 2: Build data set

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

"Known binders" - from experimental studies

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 for proteins

- **Basics:** same as for DNA but with a larger alphabet:

- Calculate p_a at each position
- Entropy

$$H = -\sum_{a=1}^{N} p_a \log_2 p_a$$

- Information content

$$R_{seq} = \log_2 20 + \sum_{a=1}^{20} p_a \log_2 p_a$$

- Conserved positions
  - $p_V=1$, $p_{!V}=0 \Rightarrow H=0$, $R=\log_2(20)\approx 4.3$

- Mutable positions
  - $p_a=1/20 \Rightarrow H=\log_2(20)$, $R=0$

```
LLDVPTAAV
LLDVPTAAV
VLFRGGPRG
MVDGTLLLL
YMNGTMSQV
MLLSVPLLL
SLLGLLVEV
ALLPPINIL
TLIKIQHTL
HLIDYLVTS
ILAPPVVKL
ALFPQLVIL
GILGFVFTL
STNRQSGRQ
GLDVLTAKV
RILGAVAKV
QVCERIPTI
```

Issue: Background frequencies

- Amino acid frequencies are far from equal
- We need to take this into account in the information content calculation

Amino acid			%
Alanine	Ala	A	7.85
Arginine	Arg	R	5.33
Asparagine	Asn	N	4.55
Aspartic acid	Asp	D	5.37
Cysteine	Cys	C	1.88
Glutamine	Gln	Q	3.77
Glutamic acid	Glu	E	5.83
Glycine	Gly	G	7.35
Histidine	His	H	2.35
Isoleucine	Ile	I	5.80
Leucine	Leu	L	9.43
Lysine	Lys	K	5.88
Methionine	Met	M	2.28
Phenylalanine	Phe	F	4.07
Proline	Pro	P	4.56
Serine	Ser	S	6.04
Threonine	Thr	T	6.17
Tryptophan	Trp	W	1.31
Tyrosine	Tyr	Y	3.27
Valine	Val	V	6.92
Unknown		X	

Relative information content

- Not all amino acids are found equally frequent in nature. L is found 10% and W only 1.3% of the time.
- The relative information content (also called the Kullback-Leibler divergence) takes this into account

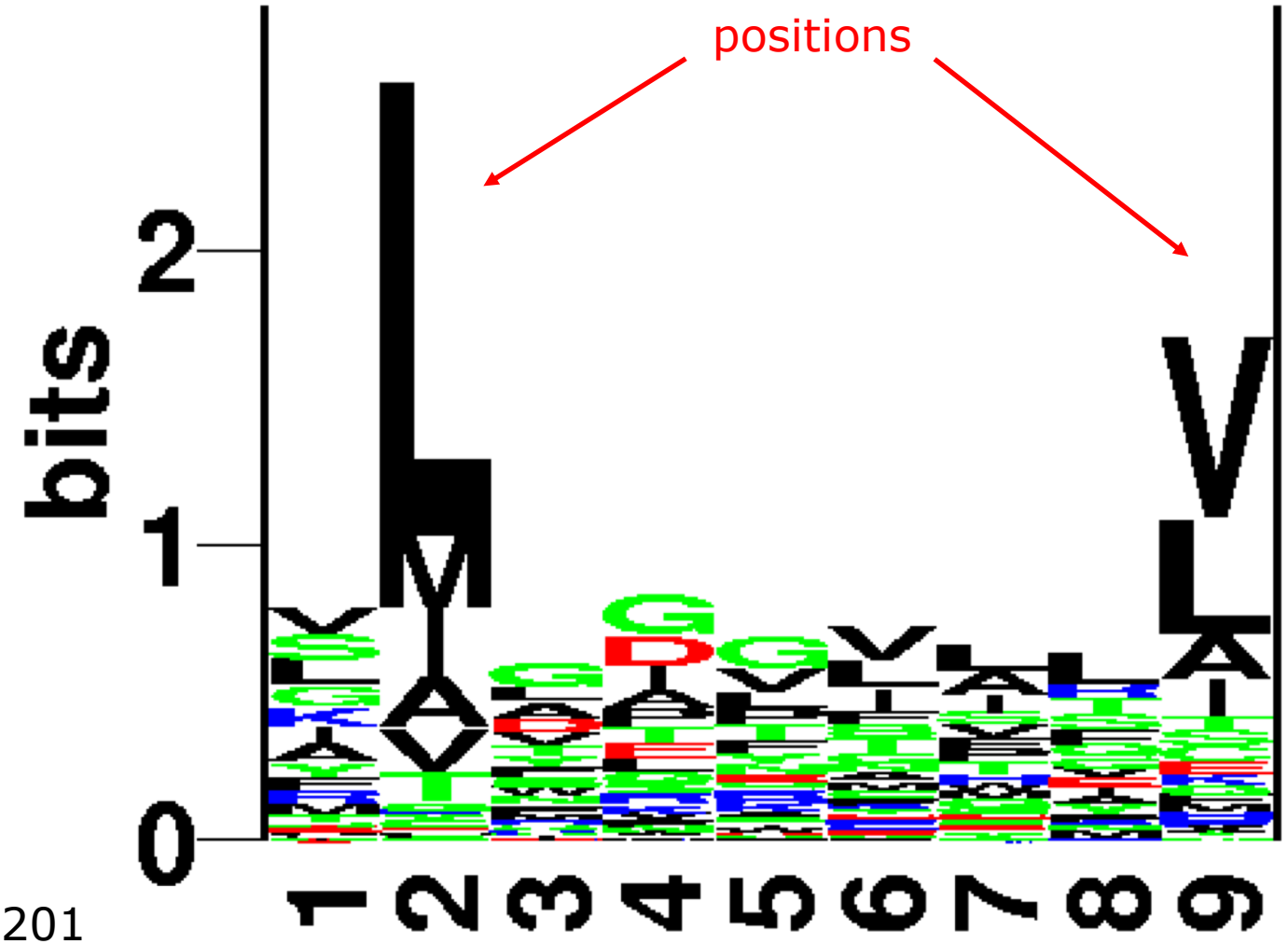
$$I_{KL} = \sum_{a=1}^{20} p_a \log_2 \left(\frac{p_a}{q_a} \right)$$

$$I_{KL} = \sum_{a=1}^{20} p_a \log_2 p_a - \sum_{a=1}^{20} p_a \log_2 q_a$$

$$I_{KL} = \log_2 20 + \sum_{a=1}^{20} p_a \log_2 p_a$$

If $q_a = 0.05$ for all amino acids

Step 3: epitope LOGO



HLA-A0201

Take home messages

- ”Consensus sequences” are very incomplete descriptions of motifs
- Sequence logos are better descriptions
- The information content of a position is a measure of conservation
- The information content of a position is calculated as a difference in uncertainty