3. Predicting binding sites

- basics of TFBS identification
- defining a background model
- tools
- phylogenetic footprinting
- including "in-vivo features"

IPMB
Institut für Pharmazie und Molekulare Biotechnologie

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Predicting binding sites in sequences

a|	0.55	0.02	0.95	0.02	0.95	0.95	0.02	0.95	0.02	0.88	0.75
c|	0.08	0.02	0.02	0.48	0.02	0.02	0.02	0.02	0.22	0.08	0.02
g|	0.15	0.88	0.02	0.02	0.02	0.02	0.95	0.02	0.02	0.02	0.22
t|	0.22	0.08	0.02	0.48	0.02	0.02	0.02	0.02	0.75	0.02	0.02
	T	**G**	**A**	**C**	**A**	**C**	**G**	**A**	**C**	**C**	**G**

p(S|M) = 0.22 * 0.88 * 0.95 * 0.48 * 0.95 * 0.02 * 0.95 * 0.95 * 0.22 * 0.08 * 0.22

= **5.8e-6**

S = sequence
M = TFBS model

$$P(S \mid M) = \prod_{j=1}^{w} f'_{i(j)j}$$

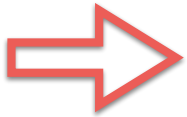
i(j) = nucleotide *i* at position *j* of sequence

Predicting binding sites in sequences

a|	0.55	0.02	0.95	0.02	0.95	0.95	0.02	0.95	0.02	0.88	0.75
c|	0.08	0.02	0.02	0.48	0.02	0.02	0.02	0.02	0.22	0.08	0.02
g|	0.15	0.88	0.02	0.02	0.02	0.02	0.95	0.02	0.02	0.02	0.22
t|	0.22	0.08	0.02	0.48	0.02	0.02	0.02	0.02	0.75	0.02	0.02
	T	**G**	**A**	**C**	**A**	**C**	**G**	**A**	**C**	**C**	**G**

$$p(S|M) = 5.8e\text{-}6$$

Is this sequence a binding site for this transcription factor ?



Is it more likely to be a binding site than a background sequence ? p(S|M) < > p(S|B)

Predicting binding sites in sequences

- Likelihood of a sequence being a binding site

$$P(S \mid M) = \prod_{j=1}^{w} f'_{i(j)j}$$

S = sequence
M = TFBS model

- Likelihood of a sequence being a background sequence (*j(i)* = nucleotide at position i)

$$P(S \mid B) = \prod_{j=1}^{w} p_{i(j)}$$

S = sequence
B = background model

- Log-likelihood ratio

$$LLR = \ln \frac{P(S \mid M)}{P(S \mid B)} = \ln \frac{\prod_{j=1}^{w} f'_{i(j)j}}{\prod_{j=1}^{w} p_{i(j)}}$$

Sum of weights of the position weight matrix

$$LLR = \ln \frac{P(S \mid M)}{P(S \mid B)} = \sum_{j=1}^{w} \ln \frac{f'_{i(j)j}}{p_{i(j)}}$$

Predicting binding sites in sequences

a|	0.55	0.02	0.95	0.02	0.95	0.95	0.02	0.95	0.02	0.88	0.75
c|	0.08	0.02	0.02	0.48	0.02	0.02	0.02	0.02	0.22	0.08	0.02
g|	0.15	0.88	0.02	0.02	0.02	0.02	0.95	0.02	0.02	0.02	0.22
t|	0.22	0.08	0.02	0.48	0.02	0.02	0.02	0.02	0.75	0.02	0.02
	T	**G**	**A**	**C**	**A**	**C**	**G**	**A**	**C**	**C**	**G**

$p(S|M) = \mathbf{5.8e\text{-}6}$ $p(S|B) = p_A^3\ p_C^4\ p_G^3\ p_T = \mathbf{1.9e\text{-}7}$

LLR = 3.41

Predicting binding sites in sequences

	A	**T**	**G**	**A**	**C**	**A**	**C**	**G**	**A**	**C**	**C**	**G**	**T**	. . .
a|	**0.55**	0.02	0.95	**0.02**	0.95	**0.95**	0.02	0.95	**0.02**	0.88	0.75			
c|	0.08	0.02	0.02	0.48	**0.02**	0.02	**0.02**	0.02	0.22	**0.08**	**0.02**			
g|	0.15	0.88	**0.02**	0.02	0.02	0.02	0.95	**0.02**	0.02	0.02	0.22			
t|	0.22	**0.08**	0.02	0.48	0.02	0.02	0.02	0.02	0.75	0.02	0.02			

$p(S|M) = 4.2e{-}15$; $P(S|B) = p_A^4 * p_C^4 * p_G^2 * p_T = 2.1e{-}7$ → **LLR = -17.7**

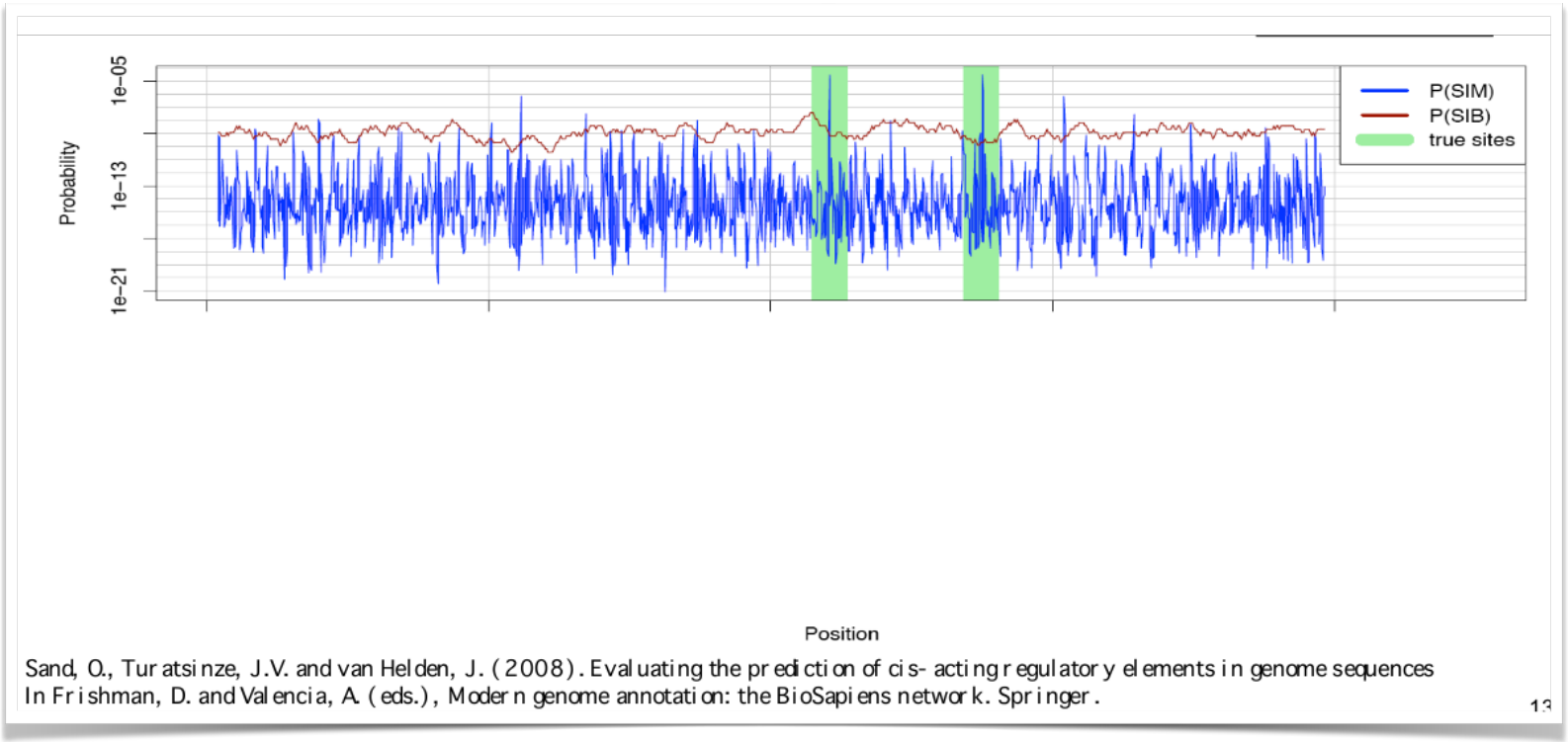
	A	**T**	**G**	**A**	**C**	**A**	**C**	**G**	**A**	**C**	**C**	**G**	**T**	. . .
a|	0.55	0.02	**0.95**	0.02	**0.95**	0.95	0.02	**0.95**	0.02	0.88	0.75			
c|	0.08	0.02	0.02	**0.48**	0.02	**0.02**	0.02	0.02	**0.22**	**0.08**	0.02			
g|	0.15	**0.88**	0.02	0.02	0.02	0.02	**0.95**	0.02	0.02	0.02	**0.22**			
t|	**0.22**	0.08	0.02	0.48	0.02	0.02	0.02	0.02	0.75	0.02	0.02			

$p(S|M) = 5.8e{-}6$; ; $P(S|B) = p_A^3 * p_C^4 * p_G^3 * p_T = 1.9e{-}7$ → **LLR = 3.41**

	A	**T**	**G**	**A**	**C**	**A**	**C**	**G**	**A**	**C**	**C**	**G**	**T**	. . .
a|	0.55	**0.02**	0.95	**0.02**	0.95	0.95	**0.02**	0.95	0.02	0.88	0.75			
c|	0.08	0.02	**0.02**	0.48	**0.02**	0.02	0.02	**0.02**	**0.22**	0.08	0.02			
g|	**0.15**	0.88	0.02	0.02	0.02	**0.02**	0.95	0.02	0.02	**0.02**	0.22			
t|	0.22	0.08	0.02	0.48	0.02	0.02	0.02	0.02	0.75	0.02	**0.02**			

$p(S|M) = 1.6e{-}17$; ; $P(S|B) = pA^3 * pC^4 * pG^3 * pT = 1.9e{-}7$ → **LLR = -23.1**

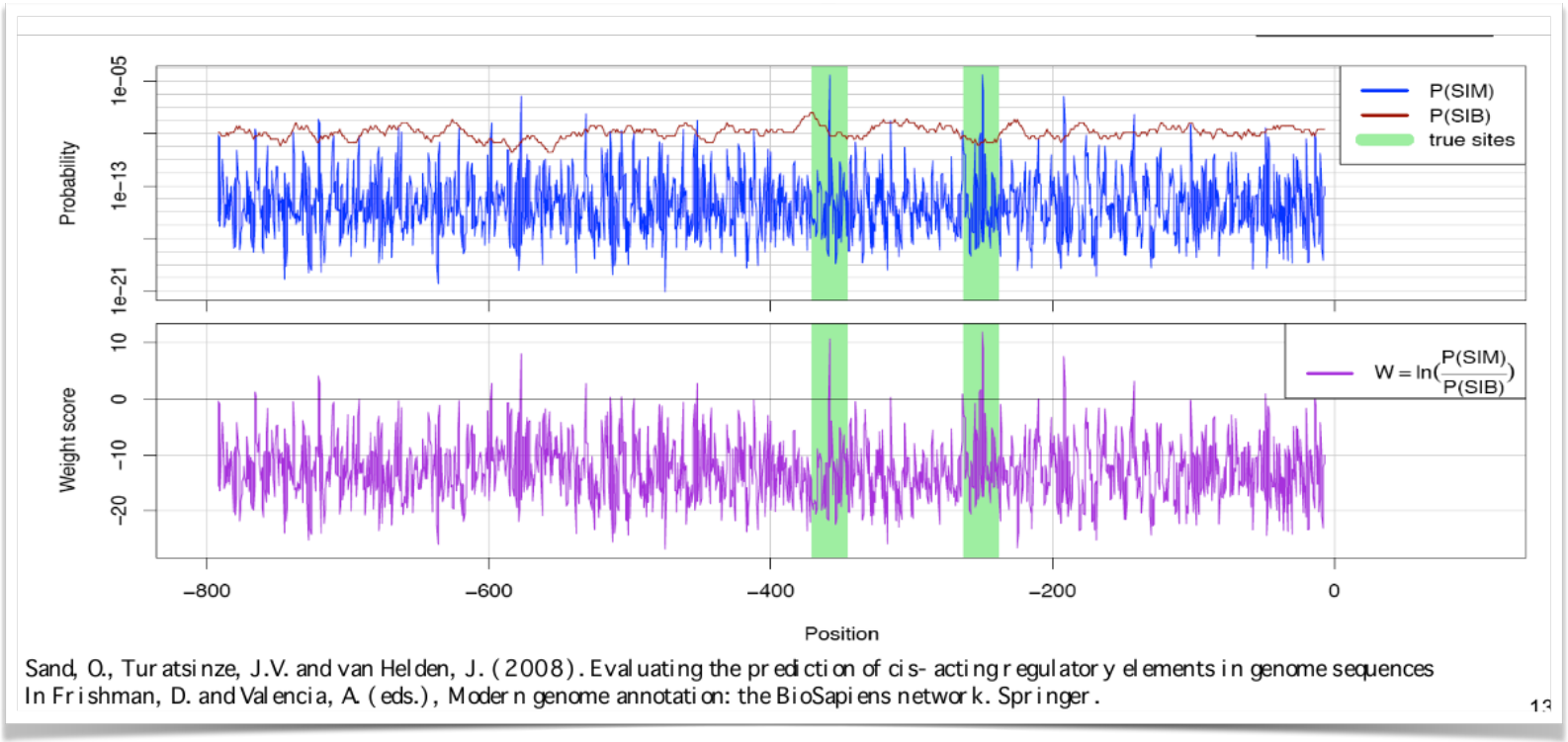
Predicting binding sites in sequences



Sand, O., Turatsinze, J.V. and van Helden, J. (2008). Evaluating the prediction of cis-acting regulatory elements in genome sequences In Frishman, D. and Valencia, A. (eds.), Modern genome annotation: the BioSapiens network. Springer.

13

Predicting binding sites in sequences



Sand, O., Turatsinze, J.V. and van Helden, J. (2008). Evaluating the prediction of cis- acting regulatory elements in genome sequences In Frishman, D. and Valencia, A. (eds.), Modern genome annotation: the BioSapiens network. Springer.

13

Defining a realistic background model

- In computing p(S|B) we made the following assumptions :

$$P(S \mid B) = \prod_{j=1}^{w} p_{i(j)}$$

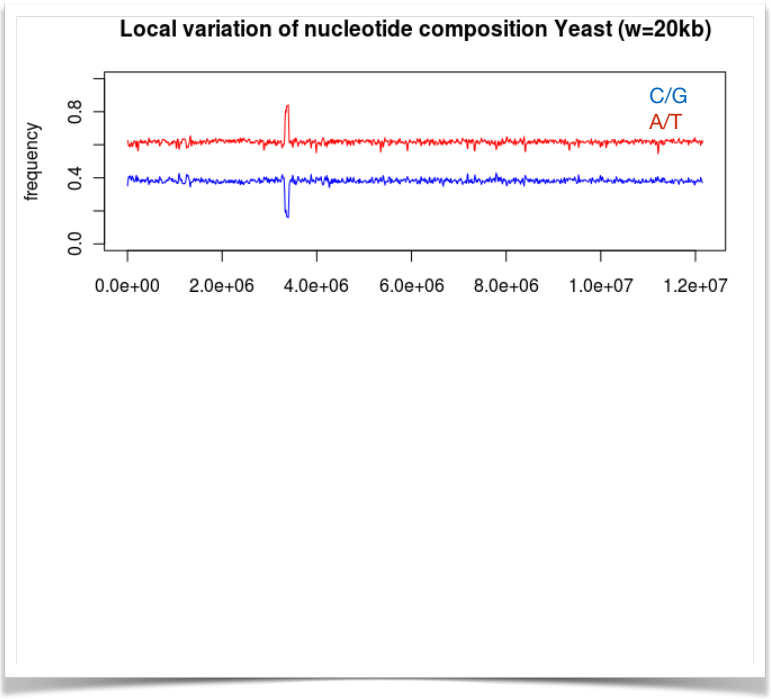
- **Homogeneity**: parameters remain identical everywhere in the genome

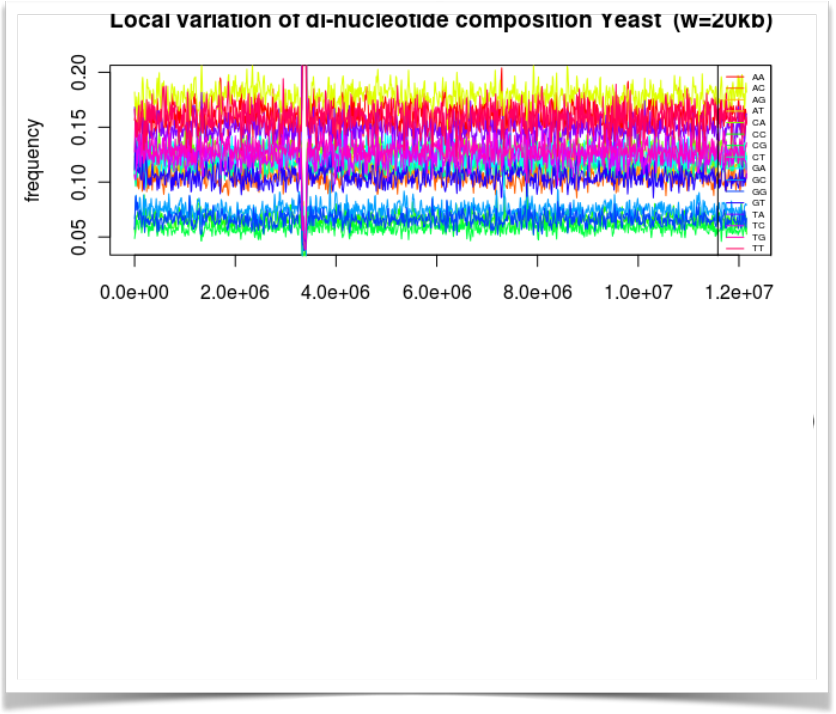
$p(N) = p_{global}(N)$ (N = nucleotide)

- **Independance**: the probability of a nucleotide at each position is independant of its neighbors (Markov model of order 0)

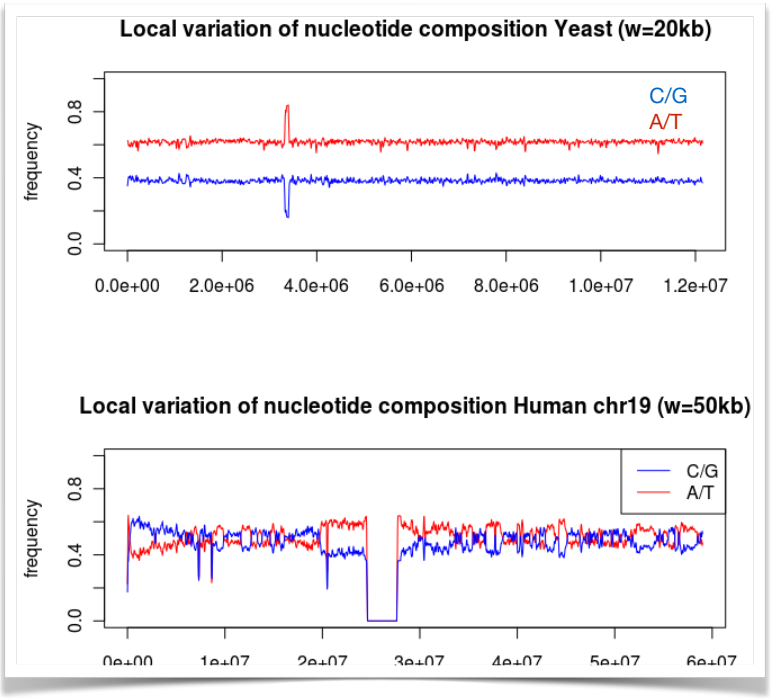
$p(CAGGCTAG|B) = p^2(A)\, p^2(C)\, p^3(G)\, p(T)$

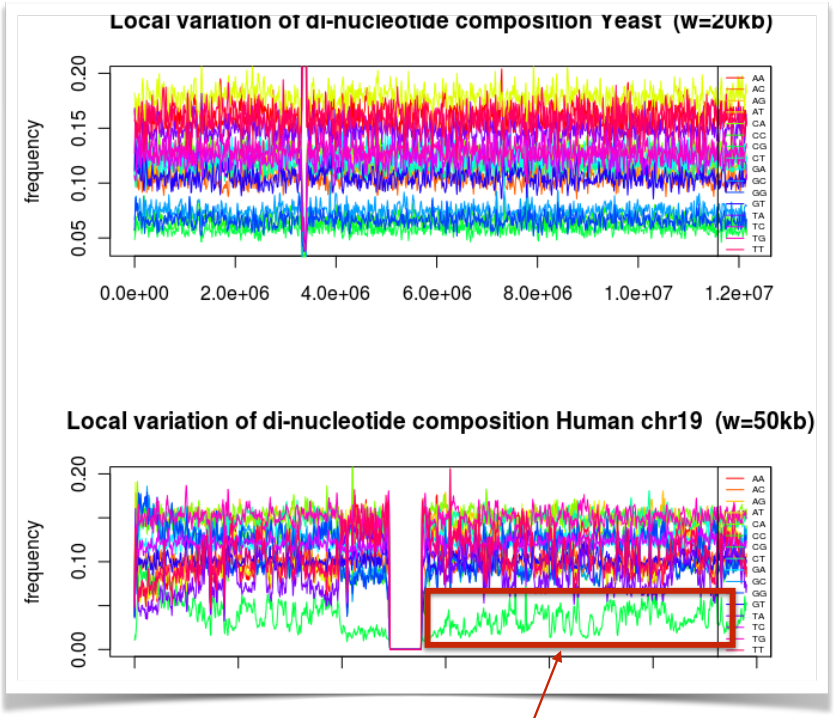
Homogeneity ?





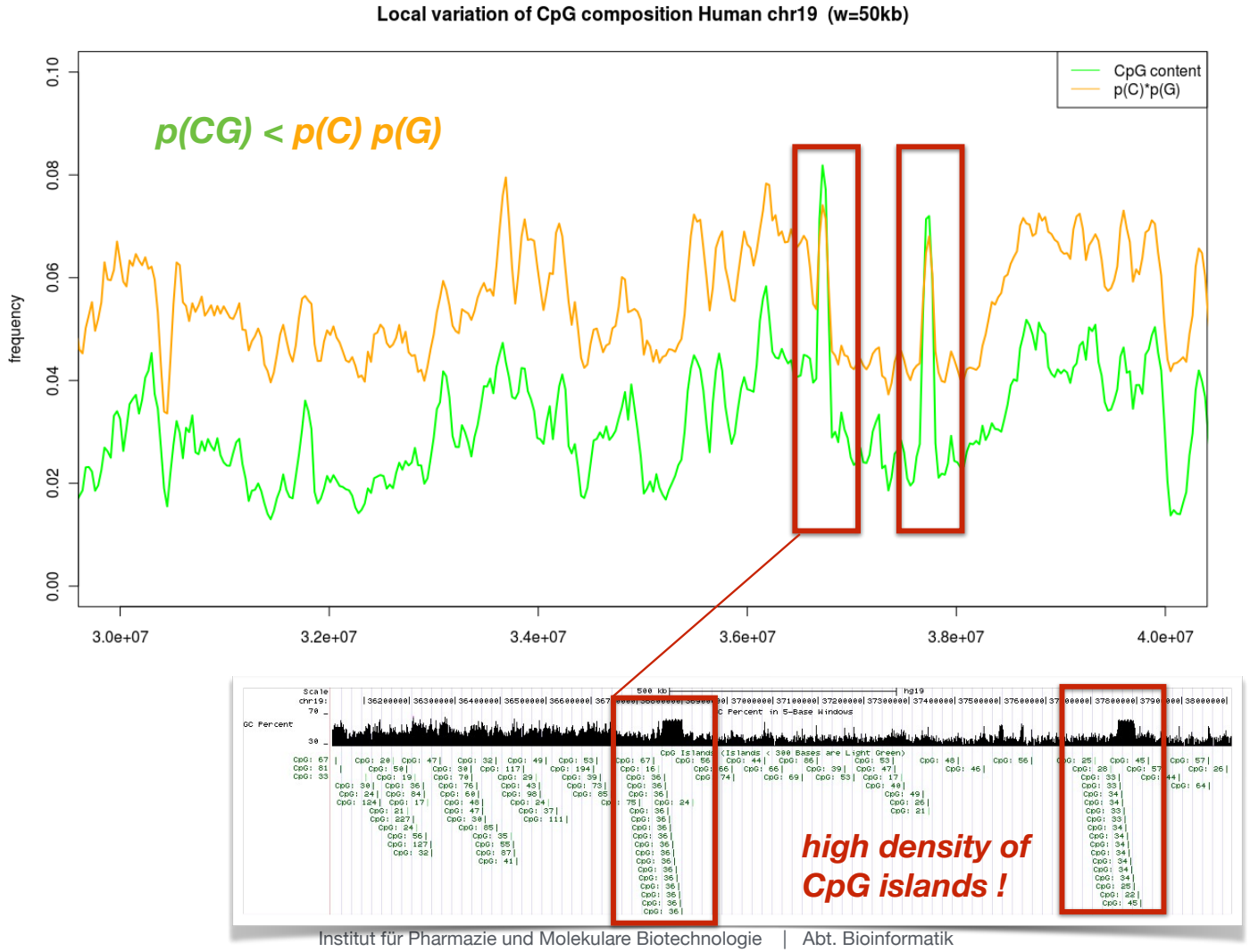
Homogeneity ?



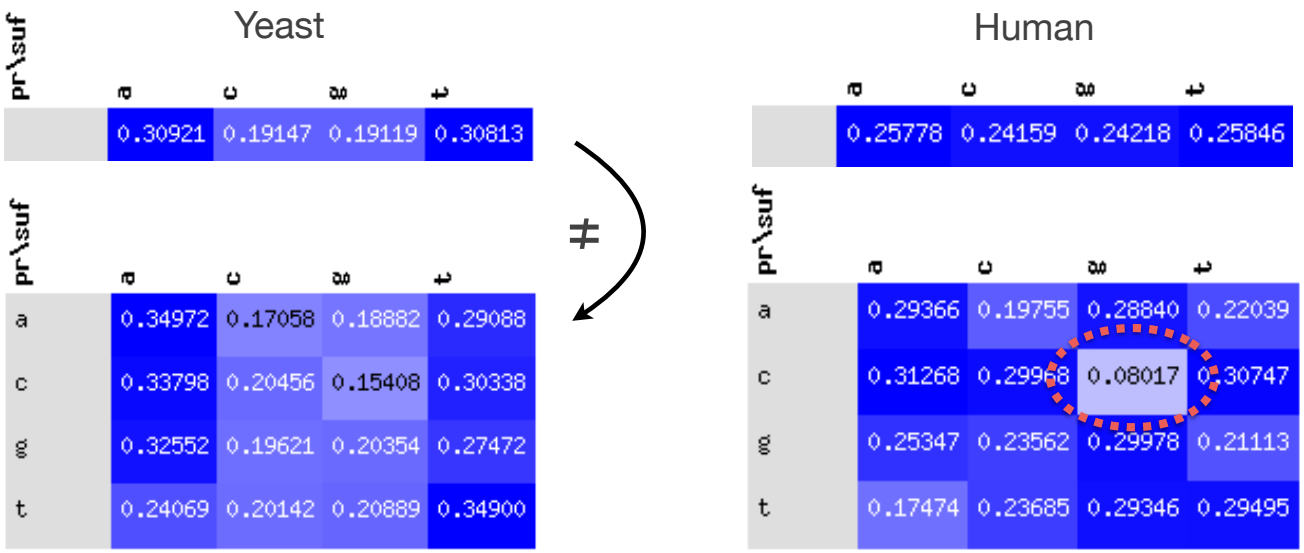


CpG avoidance in vertebrate genomes

Independance of nucleotides ?



Independance of nucleotides ?



$$p(G\ \ C) = 0.0817 \qquad p(G\ \ A) = 0.2884$$

Genomic sequences have a dependency between successive nucleotides

Markov Models

- Markov models of order k : nucleotide depends on its k predecessors

$$p(S_1 \ldots S_L) = p(S_1 \ldots S_k) \prod_{j=k}^{L-1} p(S_{j+1} \mid S_{j-k+1} \ldots S_j)$$

initial probability

next nucleotide

k predecessors

S = ACGTAGGCTACCGGATTAGCTTAGGCCATCGAGATCTAT

- the probability of **each nucleotide** depends on the sequence of the ***k* preceeding ones**
- Number of parameters
  - *k=0 :*
  - *k=1 :*
  - *k = n :*

Markov Models

- Markov models of order k : nucleotide depends on its k predecessors

$$p(S_1 \ldots S_L) = p(S_1 \ldots S_k) \prod_{j=k}^{L-1} p(S_{j+1} \mid S_{j-k+1} \ldots S_j)$$

initial probability

next nucleotide

k predecessors

S = ACGTAGGCTACCGGATTAGCTTAGGCCATCGAGATCTAT

- the probability of **each nucleotide** depends on the sequence of the ***k* preceeding ones**
- Number of parameters
  - *k=0 : p(A),p(C),p(G)* → 3 parameters
  - *k=1 : P(A|A),P(C|A),P(G|A), …* → 12 parameters
  - *k = n : 3 x 4n parameters*

Markov Models

Markov model oder 1: P(ACCGT) = P(A) x P(C|A) x P(C|C) x **P(G|C)** x P(T|G) = **2.4e-4**

Markov Model order 0: P(ACCGT) = P(A) x P(C) x P(C) x P(G) x P(T) = **9.4e-4**

Markov Models

- How are parameters learned ?

$$p(S_{k+1} \mid S_1 \ldots S_k) = \frac{\#S_1 \ldots S_k S_{k+1}}{\#S_1 \ldots S_k A + \#S_1 \ldots S_k C + \#S_1 \ldots S_k G + \#S_1 \ldots S_k T} = \frac{\#S_1 \ldots S_k S_{k+1}}{\#S_1 \ldots S_k}$$

- learning the parameters of a order *k* Markov model amounts to counting subsequences of length k+1
- Learning the parameters of the Markov model
  - using all intergenic sequences of the organism
  - using all promoter sequences
  - using the input sequences of the user (if sufficient!)

Markov Models

Mouse 2kb promoter

Dinuc	Freq	Occurences
aa	0.07923	3483960
ac	0.05211	2291772
ag	0.07590	3337516
at	0.06009	2642346
ca	0.07165	3150994
cc	0.06343	2789452
cg	0.02036	895708
ct	0.07531	3311701
ga	0.06229	2739394
gc	0.05311	2335660
gg	0.06487	2852574
gt	0.05312	2336150
ta	0.05411	2379761
tc	0.06218	2734375
tg	0.07228	3178565
tt	0.07987	3512382

number of **AN** dinuc : 11755594
number of **AC** dinucl : 2291772
$P(C|A) = 2291772/11755594 = 0.195$

number of **CN** dinuc : 10147855
number of **CG** dinucl : 895708
$P(G|C) = 895708/10147855 = 0.09$

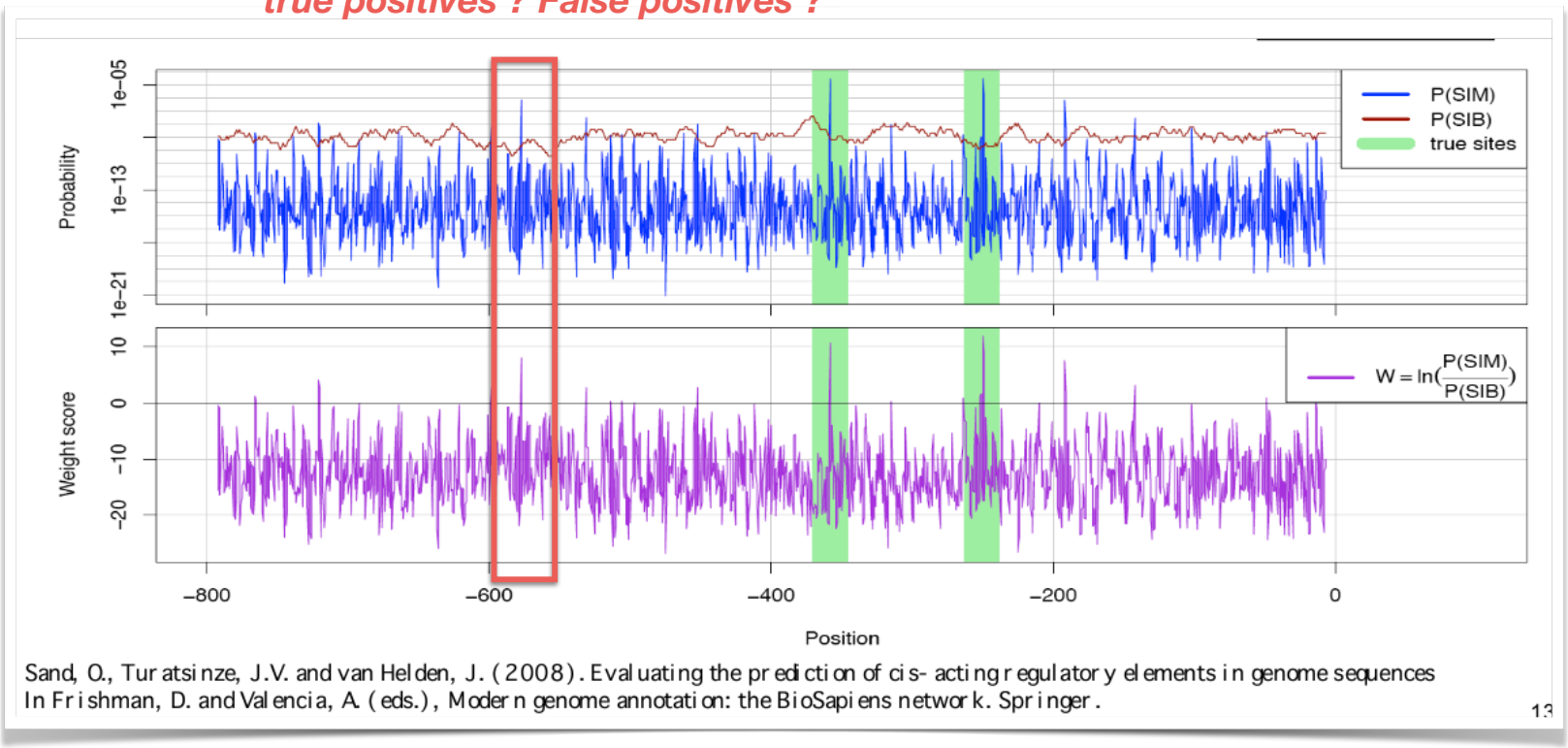
order 1 Markov model (MMo1)

$$P(ACCGT) = P(A) \times P(C|A) \times P(C|C) \times P(G|C) \times P(T|G)$$

Check : P(A|N) + P(C|N) + P(G|N) + P(T|N) = 1 for any nucleotide N

Predicting binding sites in sequences

true positives ? False positives ?



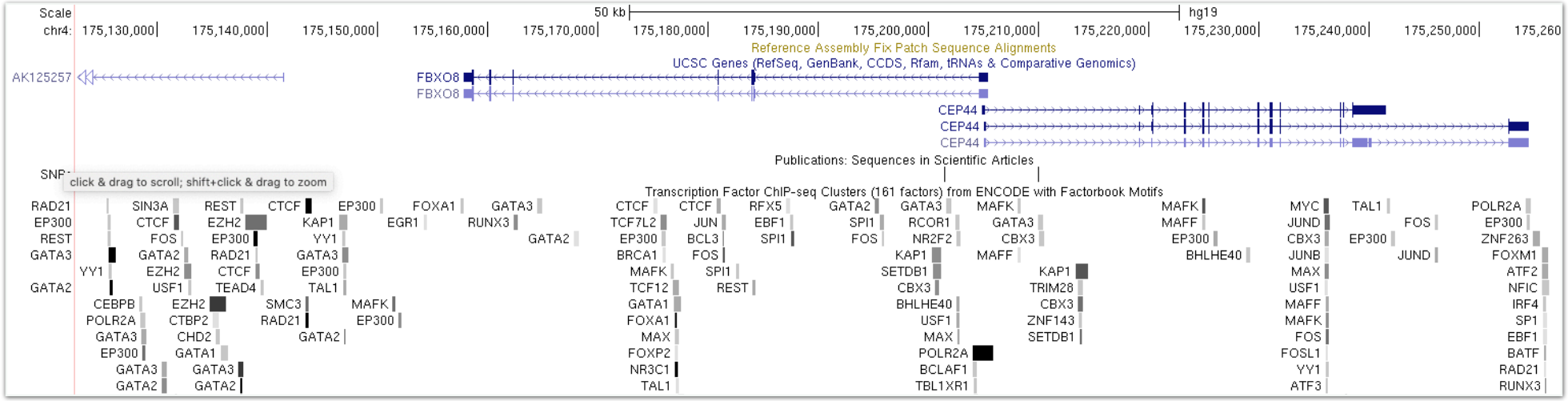
Sand, O., Turatsinze, J.V. and van Helden, J. (2008). Evaluating the prediction of cis- acting regulatory elements in genome sequences In Frishman, D. and Valencia, A. (eds.), Modern genome annotation: the BioSapiens network. Springer.

13

What scores would be obtained in a sequence which contains no binding sites ?

Positive controls

- **Positive controls** = sets of predicted binding sites, for which we know there is indeed binding of the transcription factor
- Best option: use experimental validated datasets, such as **ChIP-seq**

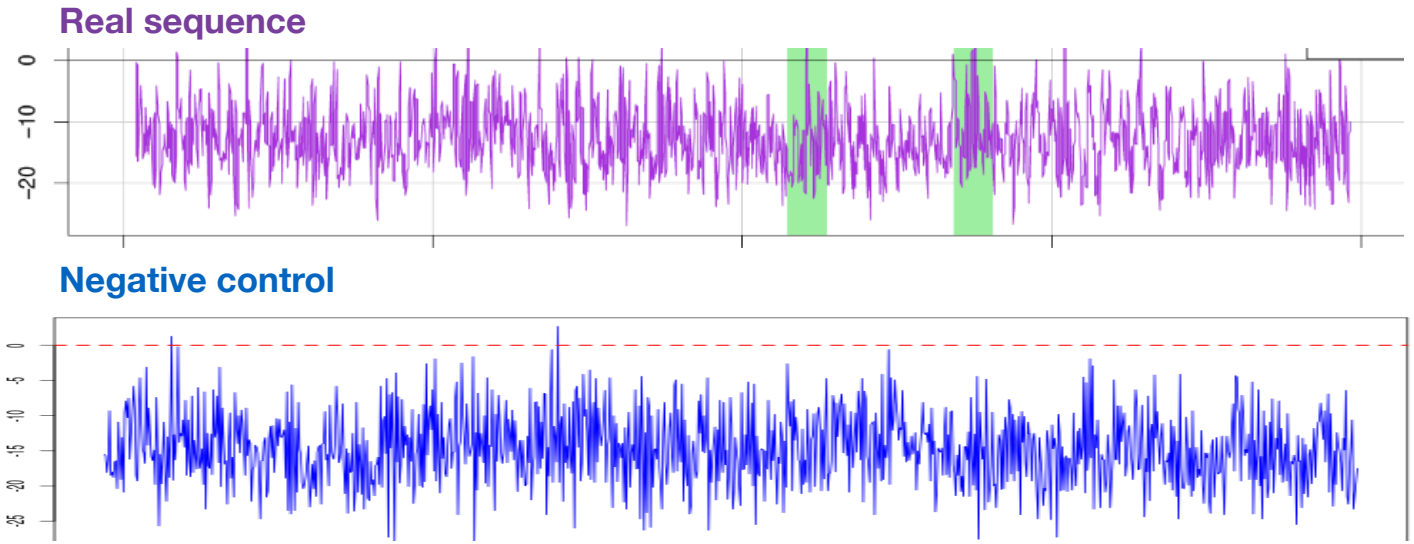


Negative controls

- **Negative controls** = test cases for which we know that the answer should be negative
- Here : sequences for which no binding site exists
- How to define such a sequence ?
  - real biological sequence : we cannot exclude the presence of a TFBS
  - generate **synthetic sequence** which is close enough to a real sequence e.g. using a Markov Model background

1. sample starting nucleotide
2. sample next nucleotides according to transition frequencies

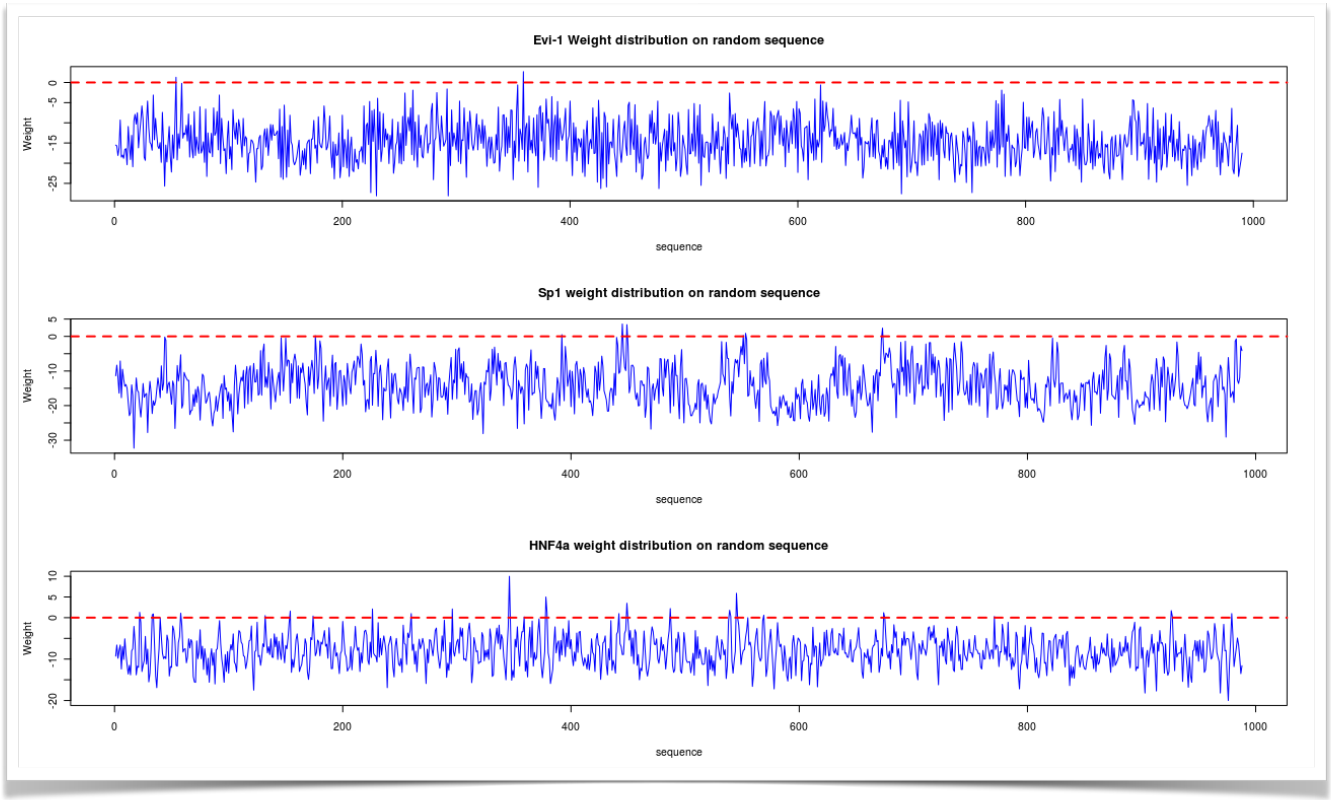
Distribution of weights

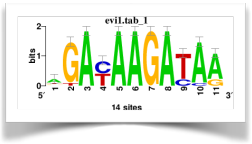


- Distribution of weights (=LLR) for Evi-1 matrix on a RANDOM sequence generated from mouse Markov model order = 3
  - most weights are negative (:-) but 2 sites have LLR>0 ***(false-positives)***
  - → false positive rate = 2 FP / 1000 negative binding sites = **2e-3**

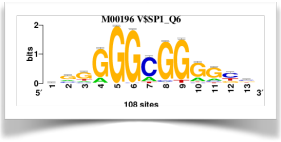
For a given LLR threshold, what is the rate of false positives that can be expected ?

Distribution of weights





FPR = 2/1000



FPR = 7/1000

FPR = 29/1000

The rate of false positives depends on the matrix considered !
Different threshold on W changes the FPR

Distribution of weights

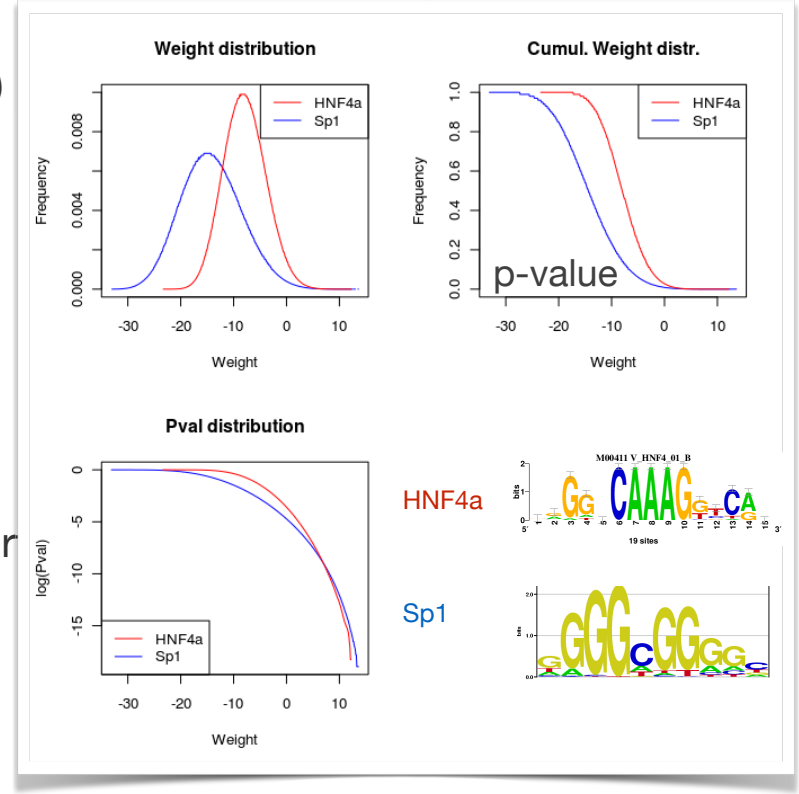
For a given matrix, we can exactly compute the **exact distribution** of scores under a particular background model:

1. for a matrix of width *k*, enumerate all possible *k*-mers S_k ($n = 4^k$)
2. compute the weight for a given background model as

$$W(S_k) = \log \frac{P(S_k|M)}{P(S_k|B)} = \log \frac{\prod_{i=1}^{L} f'_{ij(i)}}{P(S_k|B)}$$

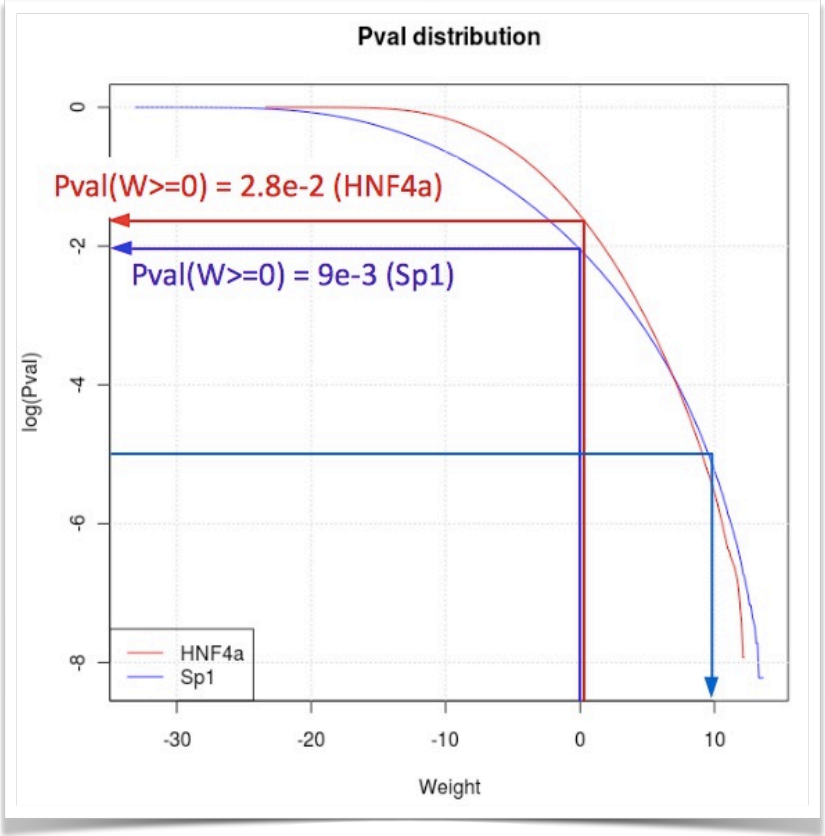
3. for a particular score W, the corresponding P-value is the propor
   higher score

   Pval(W) = freq. w >= W



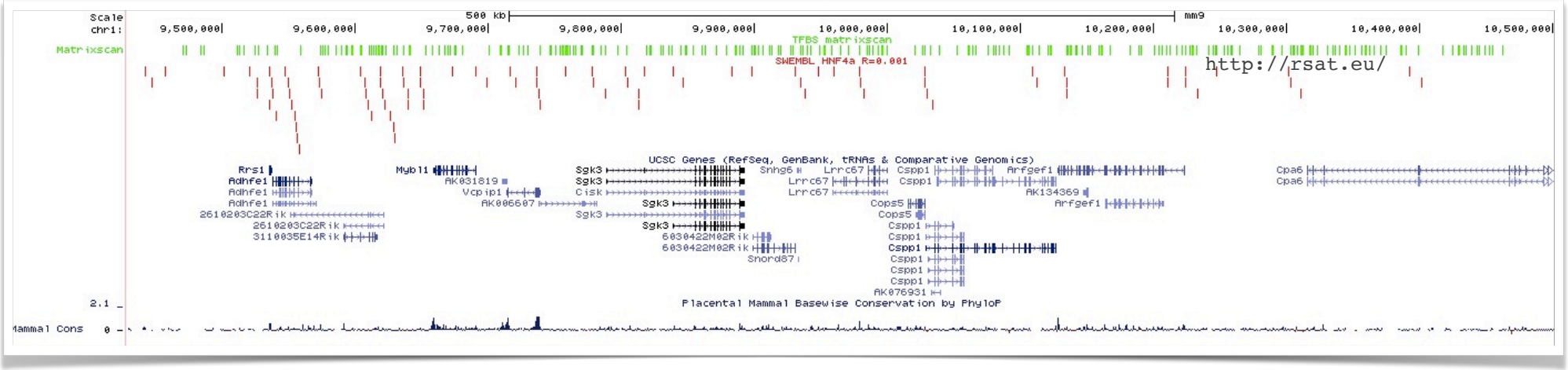
Distribution of weights

- the Pvalue indicates the **False Positive Rate (FPR)**
- if we set a threshold at $W \geq 0$, we expect
  - 1 FP every ~110 bp for Sp1
  - 1 FP every ~35 bp for HNF4a
- for a pvalue of 1e-5, we need to set a threshold of
  - 9.5 for Sp1
  - 9.1 for HNF4a



Predicting TFBS on real sequences

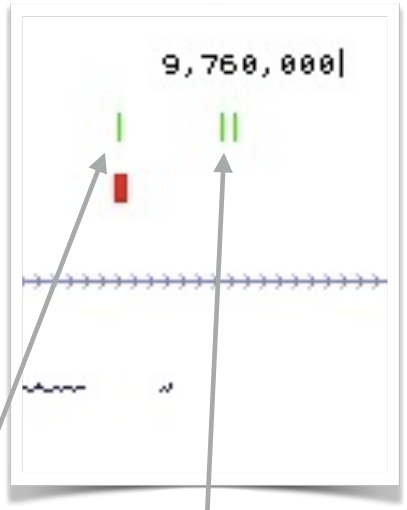
- Predicting TFBS on a 1 Mb portion of Mouse chromosome 1
- Software : Matrix-Scan ; Matrix : **HNF4a**
- Threshold to call TFBS : $p \leq 1e\text{-}4$
- Background : Markov model order=3 estimated on input sequence
- Output : **259 predicted TFBS**
- **Control : HNF4a ChIP-seq peaks (red)**



True/false positive/negative

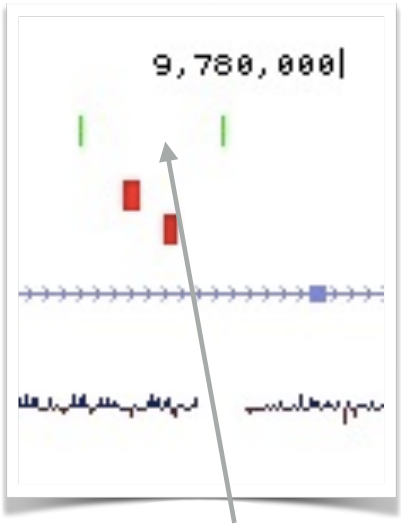
Predicted TFBS

ChIP-seq binding

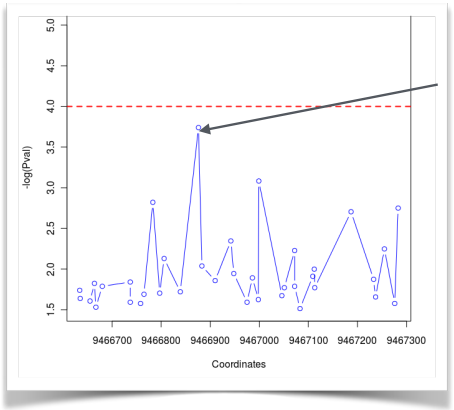


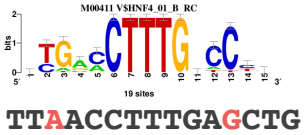
True positive

False positive
(TFBS predicted but no binding)



False negative
(binding but no TFBS predicted)

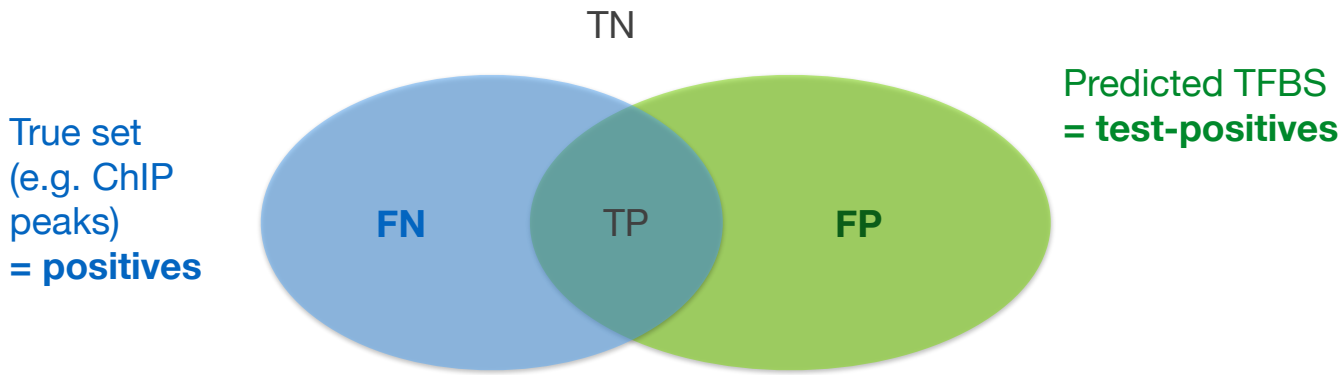




TTAACCTTTGAGCTG

no predicted TFBS passes the pval < 1e-4 threshold
→ low-affinity binding

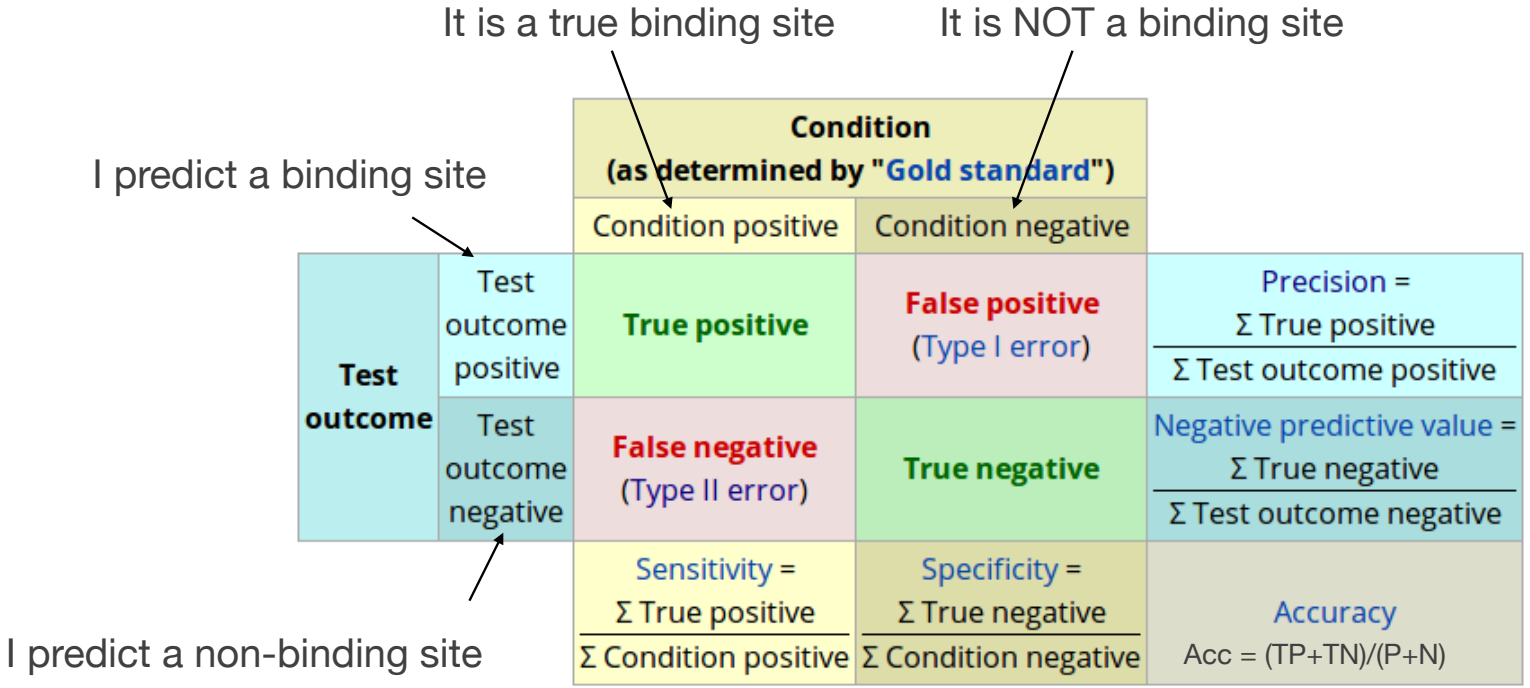
Evaluation of TFBS prediction performances



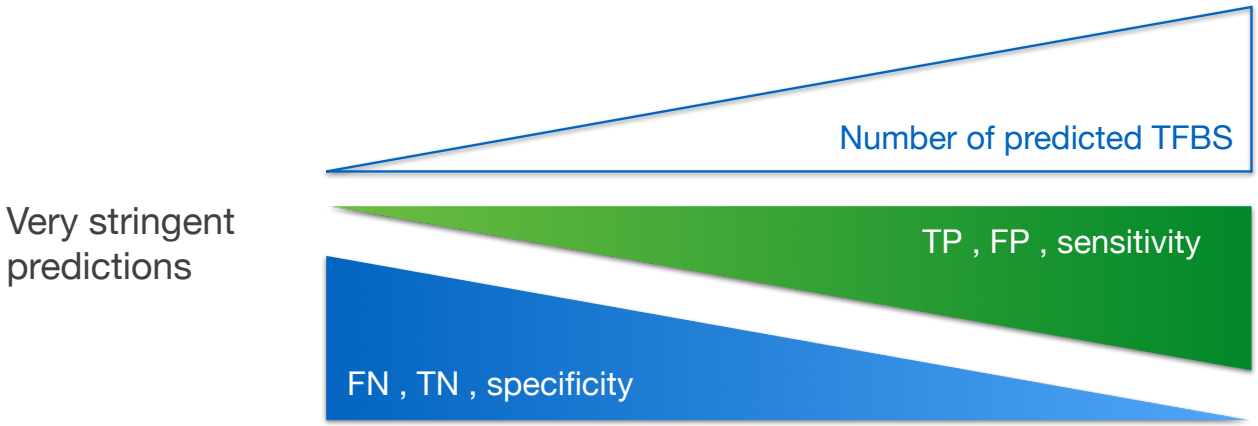
- **sensitivity** : *how many of the true binding sites did I find?*
  sensitivity = TP/P = TP / (TP+FN) → **True Positive Rate (TPR)**

- **specificity** : *how many of the non-binding sites did I identify correctly as negative ?*
  specificity = TN/N = TN / (TN + FP)
  1-specificity = FP/N = FP / (TN+FP) → **False Positive Rate (FPR)**

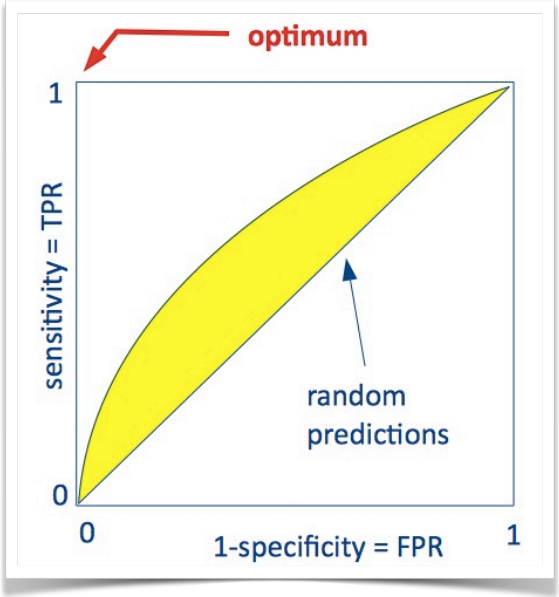
Performance measure



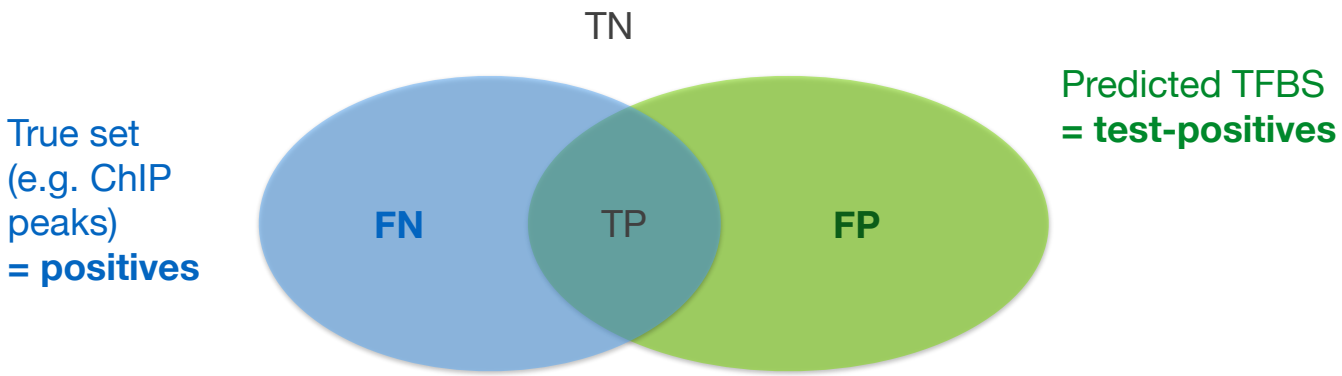
Evaluation of TFBS prediction performances



- The tradeoff specificity/sensitivity for a continuous range of parameters can be summarized as a **Receiver Operating Curve (ROC)**
- Performance of a model is determined by the **Area under the curve (AUC)**



Evaluation of TFBS prediction performances



- if the number of negatives is much larger that positives (as for ChIP data), then FPR =FP/N << 1 and is insensitive to variations in FP
  → we replace it with another measure
- **positive predictive value (PPV)**
  (a.k.a. precision) : how many true events among those predicted ?
  - **precision = PPV = TP / (TP+FP)**
  - **recall = sensitivity = TP/P**