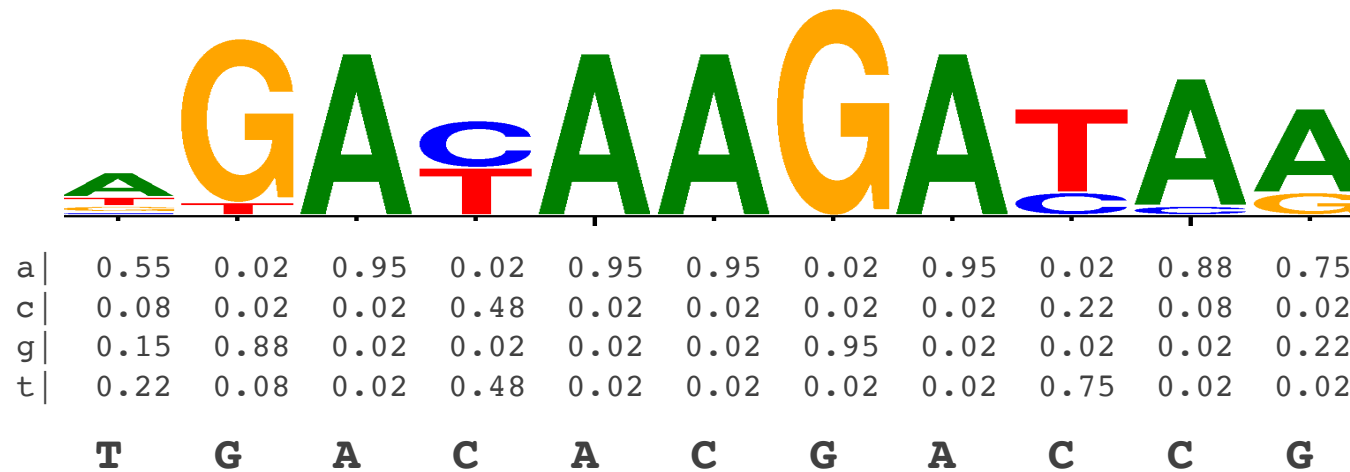


### 3. Predicting binding sites

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

# Predicting binding sites in sequences



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

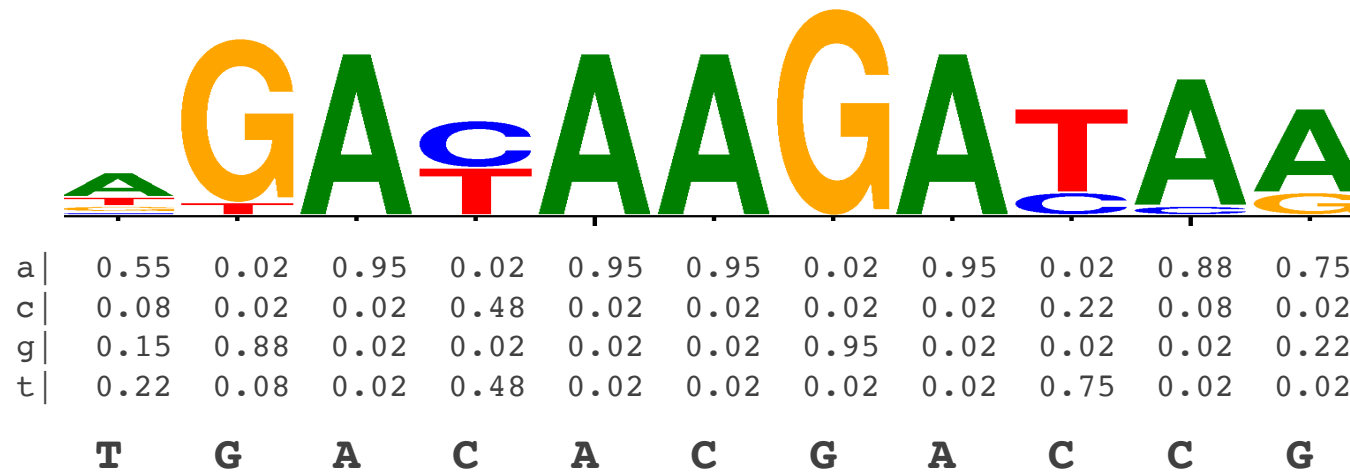
$$= \mathbf{5.8e-6}$$

$S$  = sequence  
 $M$  = TFBS model

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

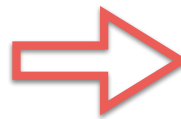
$i(j)$  = nucleotide  $i$  at position  $j$  of sequence

# Predicting binding sites in sequences



$$p(S|M) = 5.8e-6$$

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

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

$S$  = sequence  
 $M$  = TFBS model

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

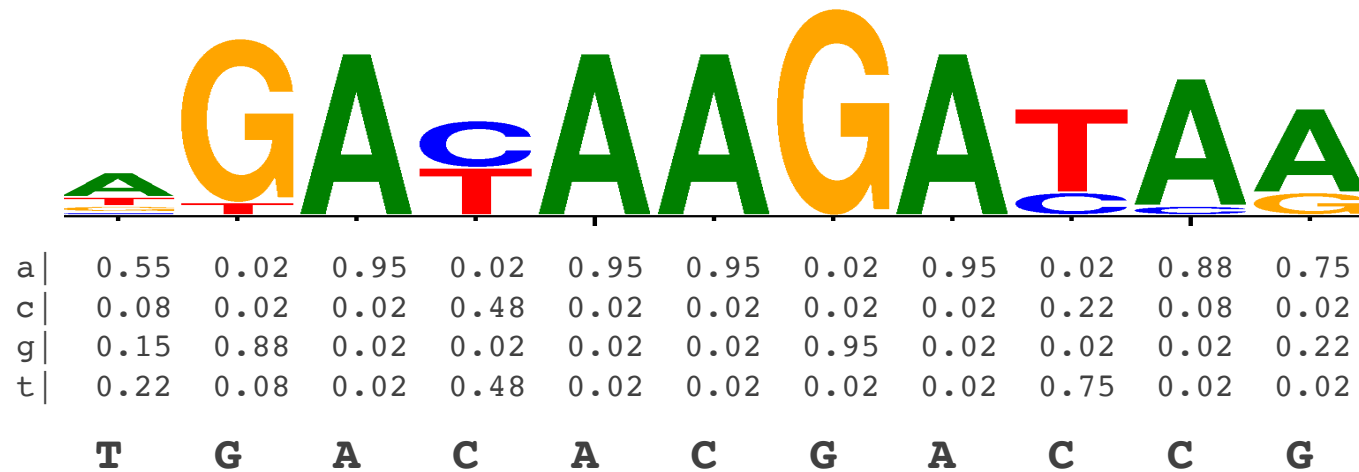
$S$  = sequence  
 $B$  = background model

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

Sum of weights of the  
position weight matrix

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

# Predicting binding sites in sequences



$$p(S|M) = 5.8e-6$$

$$p(S|B) = p_A^3 p_C^4 p_G^3 p_T = 1.9e-7$$

$$LLR = 3.41$$

# 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			

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

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	0.02	0.75	0.02

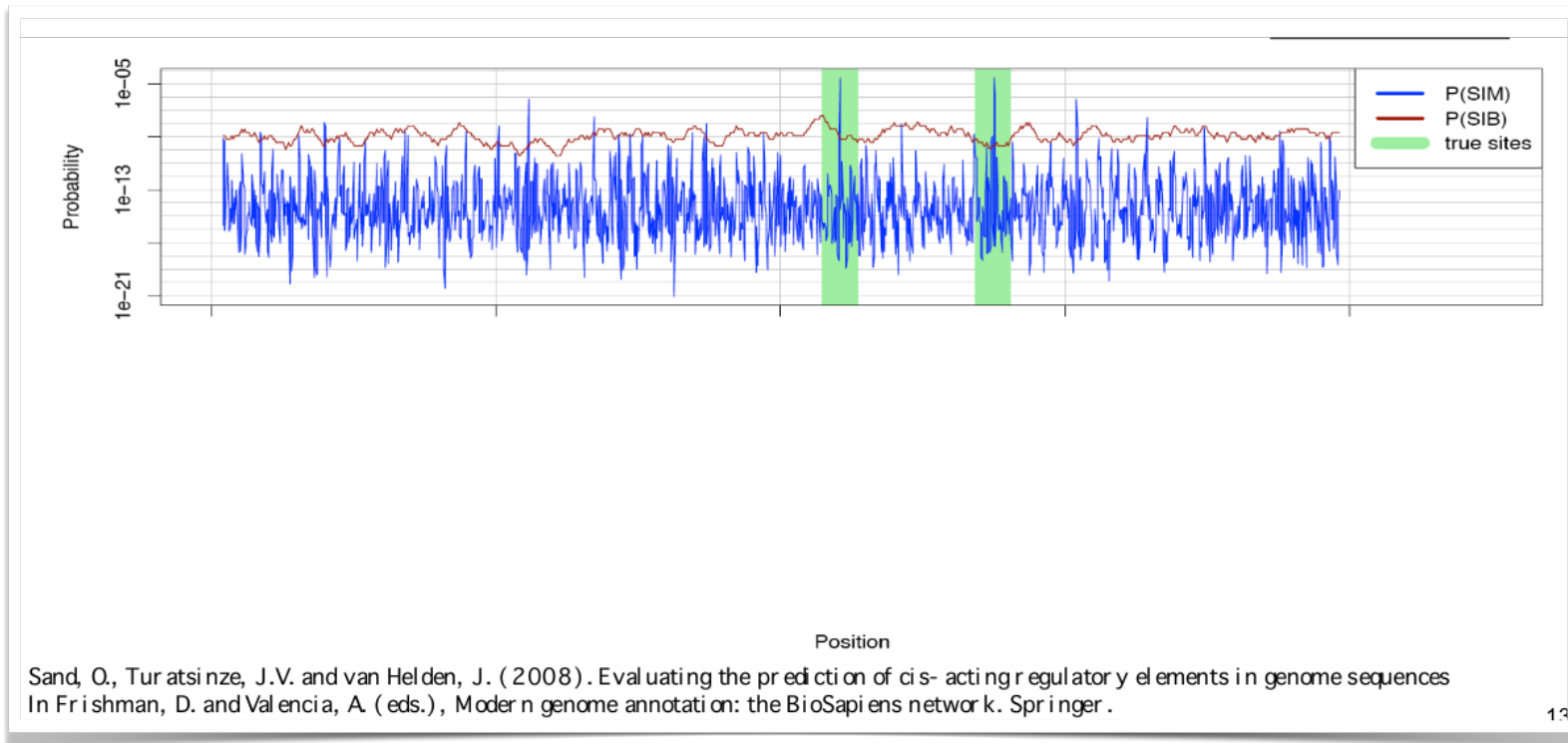
$$p(S|M) = 5.8e-6 ; ; P(S|B) = p_A^3 * p_C^4 * p_G^3 * p_T = 1.9e-7 \rightarrow \text{LLR} = 3.41$$

A T G A C A C G A C C G T ...

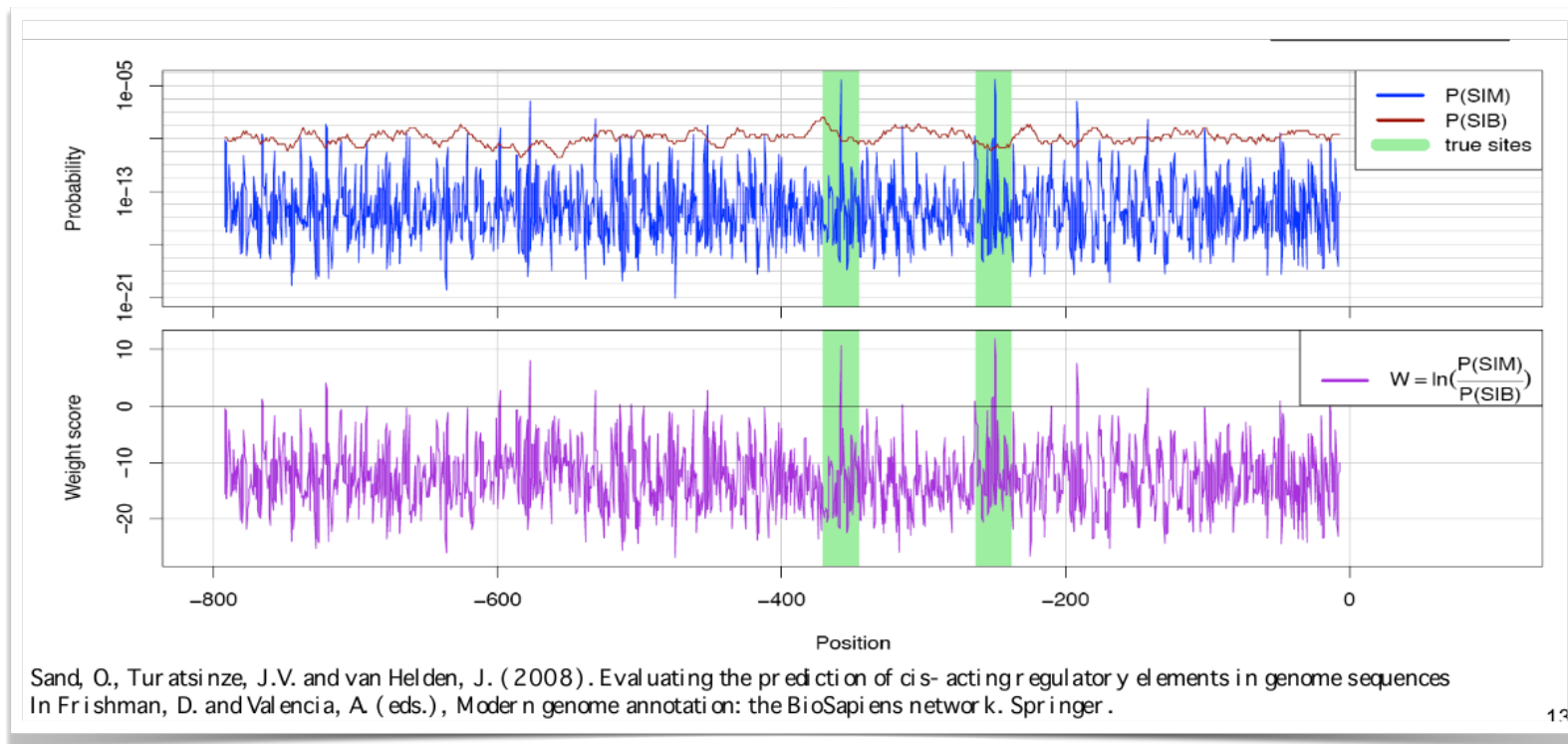
a	0.55	0.02	0.95	0.02	0.95	0.95	0.02	0.95	0.02	0.88	0.75
c	0.08	0.02	0.02	0.48	0.02	0.02	0.02	0.02	0.02	0.08	0.02
g	0.15	0.88	0.02	0.02	0.02	0.02	0.02	0.95	0.02	0.02	0.22
t	0.22	0.08	0.02	0.48	0.02	0.02	0.02	0.02	0.02	0.75	0.02

$$p(S|M) = 1.6e-17 ; ; P(S|B) = p_A^3 * p_C^4 * p_G^3 * p_T = 1.9e-7 \rightarrow \text{LLR} = -23.1$$

# Predicting binding sites in sequences



# Predicting binding sites in sequences





# Defining a realistic background model

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

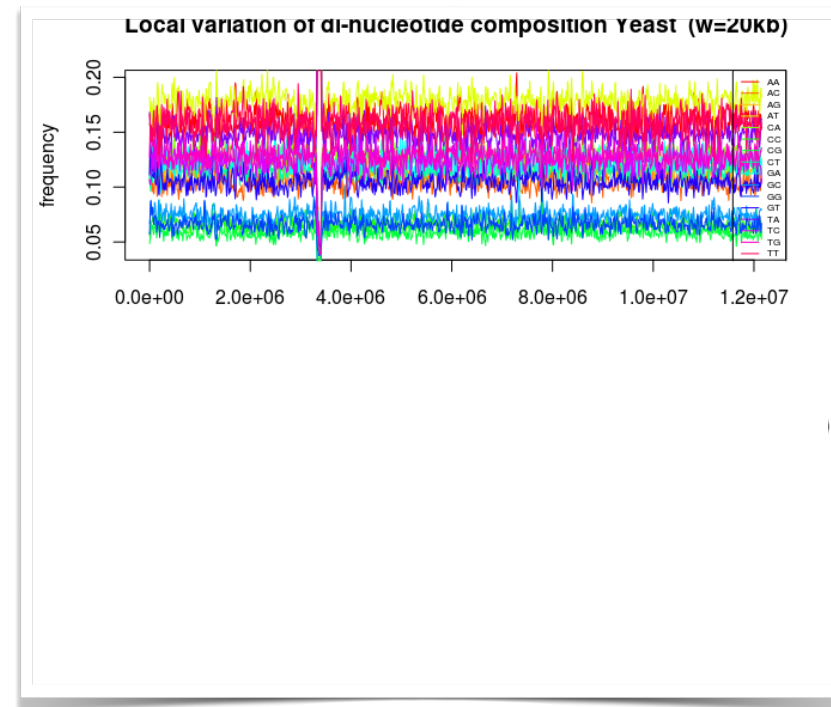
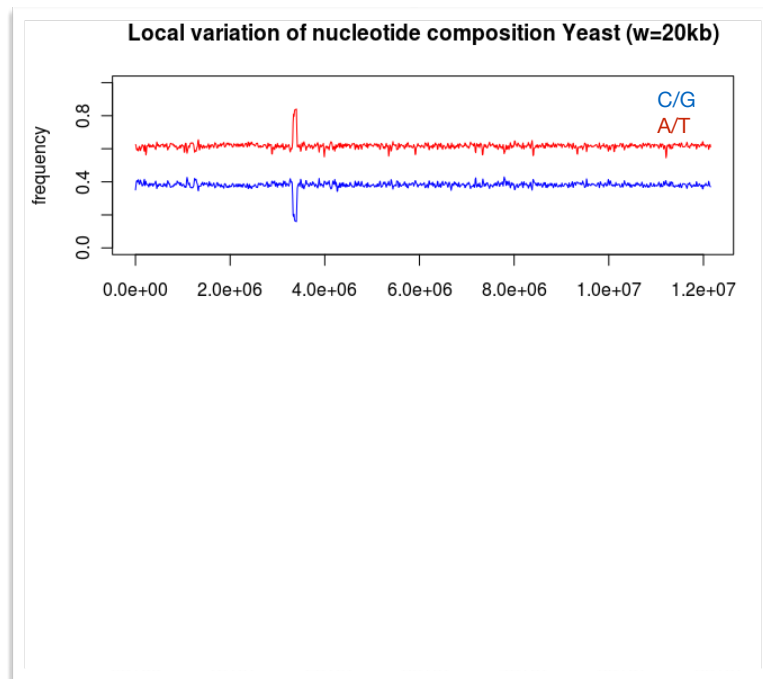
$$P(S|B) = \prod_{j=1}^w p_{i(j)}$$

- **Homogeneity**: parameters remain identical everywhere in the genome
- **Independance**: the probability of a nucleotide at each position is independant of its neighbors (Markov model of order 0)

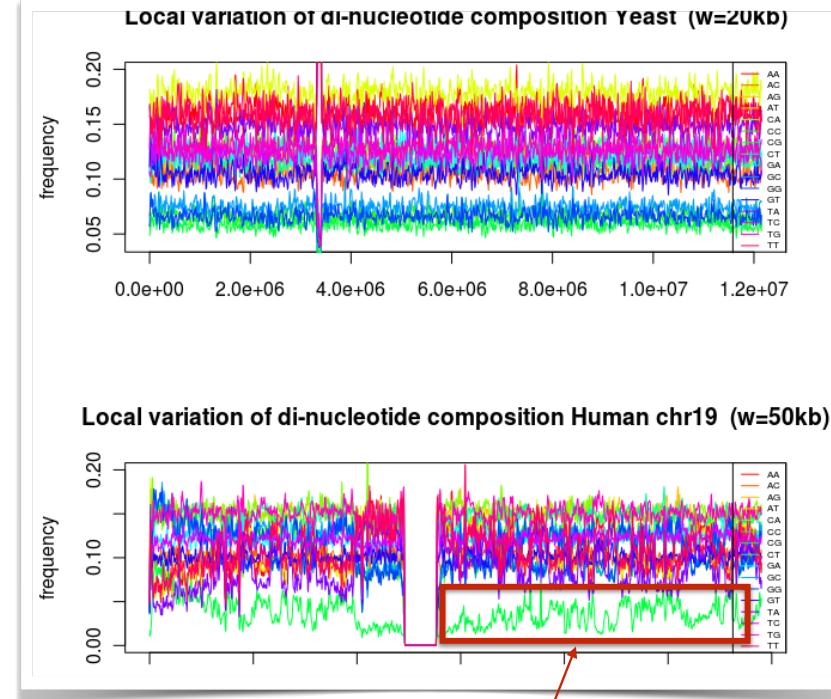
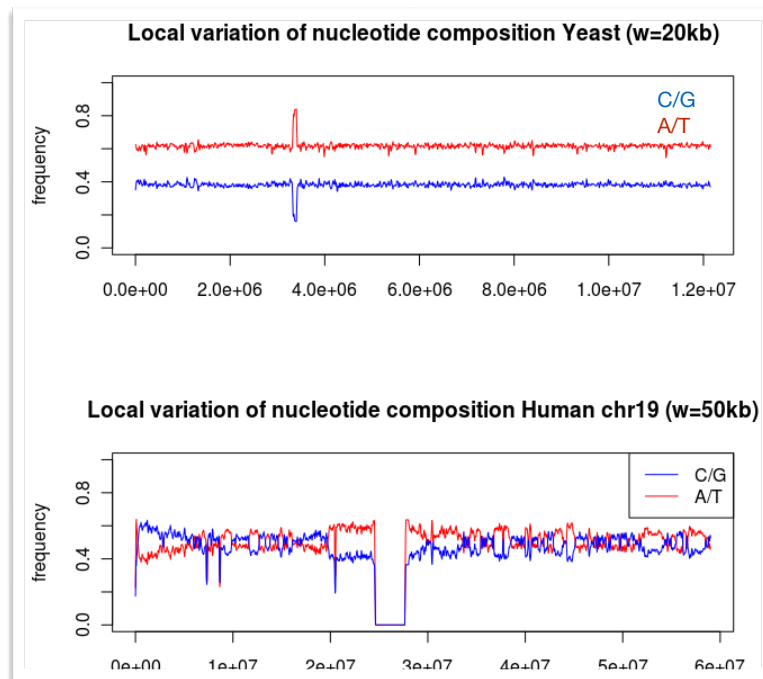
$$p(N) = p_{global}(N) \quad (N = \text{nucleotide})$$

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

# Homogeneity ?



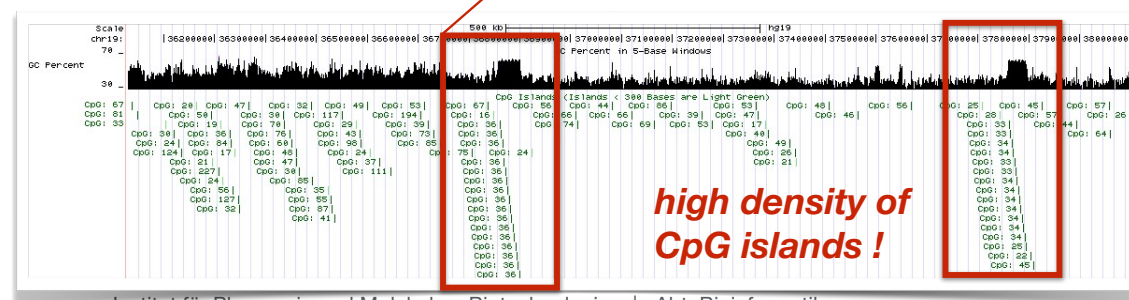
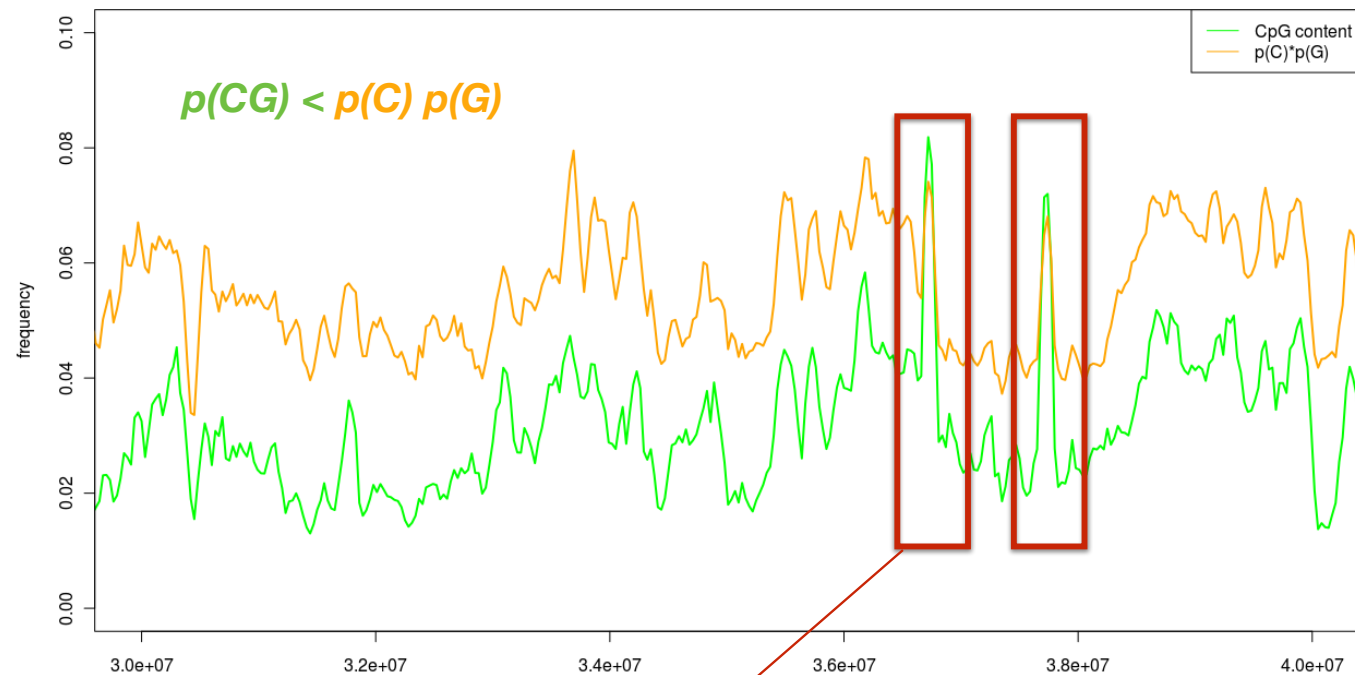
# Homogeneity ?



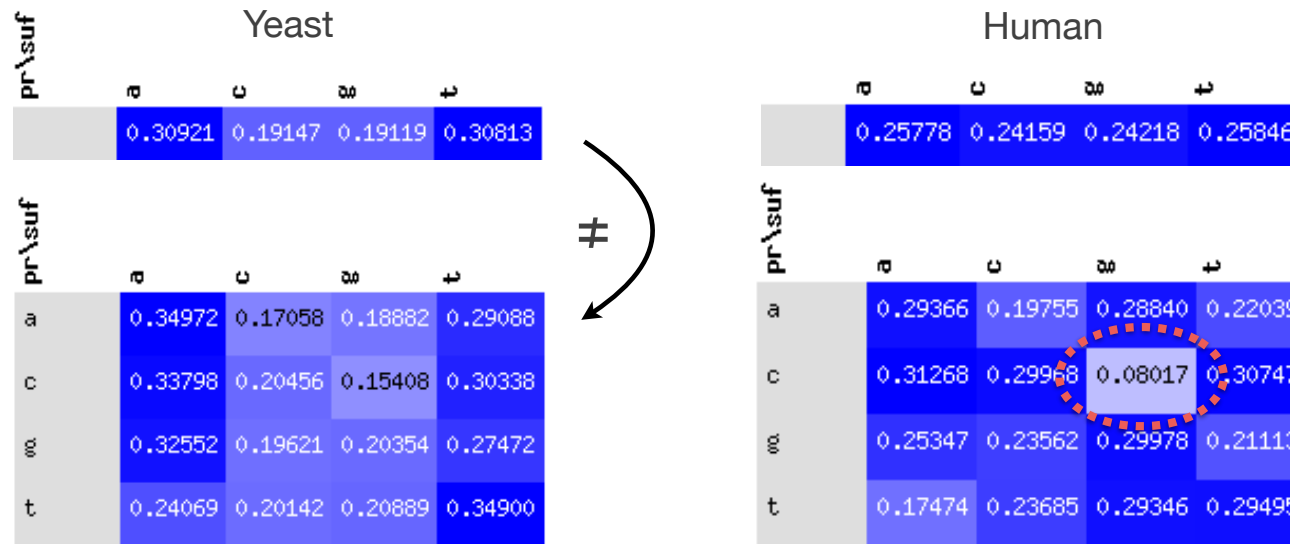
*CpG avoidance in vertebrate genomes*

# Independance of nucleotides ?

Local variation of CpG composition Human chr19 (w=50kb)



# Independence of nucleotides ?



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

**Genomic sequences have a dependency  
between successive nucleotides**

# Markov Models

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

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

initial probability                      next nucleotide                       $k$  predecessors

**S = ACGTAGGCTACCGGATTAGCTTAGGCCATCGAGATCTAT**

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

# Markov Models

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

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

initial probability      next nucleotide       $k$  predecessors

**S = ACGTAGGCTACCGGATTAGCTTAGGCCATCGAGATCTAT**

- the probability of **each nucleotide** depends on the sequence of the  **$k$  preceding ones**
- Number of parameters
  - $k=0$  :  $p(A), p(C), p(G) \rightarrow 3$  parameters
  - $k=1$  :  $P(A|A), P(C|A), P(G|A), \dots \rightarrow 12$  parameters
  - $k=n$  :  $3 \times 4^n$  parameters

# Markov Models

	a	c	g	t
	0.25778	0.24159	0.24218	0.25846
a	0.29590	0.19550	0.28357	0.22503
c	0.30990	0.27463	0.08988	0.32559
g	0.26673	0.22779	0.27765	0.22783
t	0.20207	0.23181	0.26906	0.29706

Markov model order 1:  $P(\text{ACCGT}) = P(A) \times P(C|A) \times P(C|C) \times P(G|C) \times P(T|G) = 2.4e-4$

Markov Model order 0:  $P(\text{ACCGT}) = P(A) \times P(C) \times P(C) \times P(G) \times P(T) = 9.4e-4$



# Markov Models

- How are parameters learned ?

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

- 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

$$P(G|C) = 895708/10147855 = 0.09$$

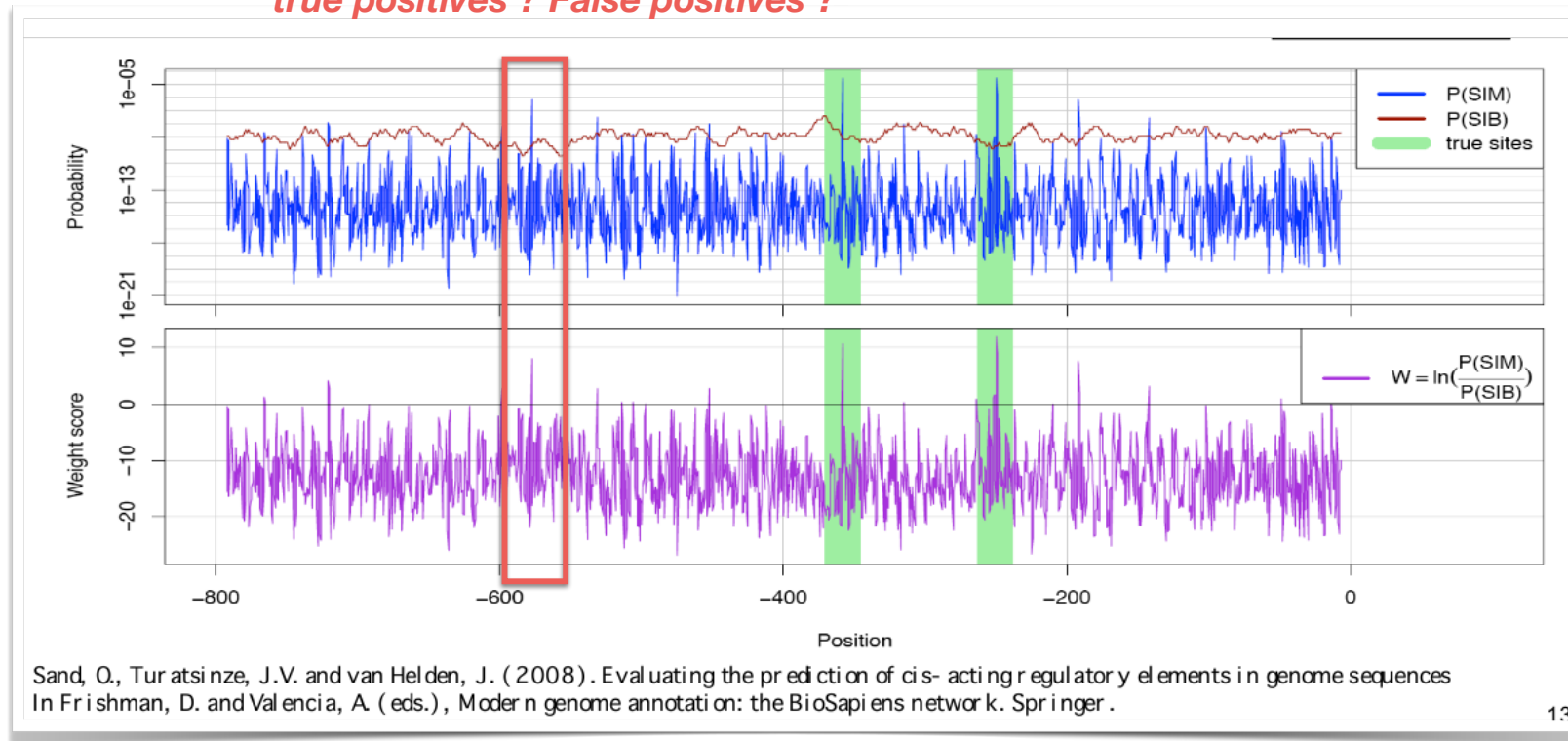
order 1 Markov model (MMo1)

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

**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 ?*



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



# 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



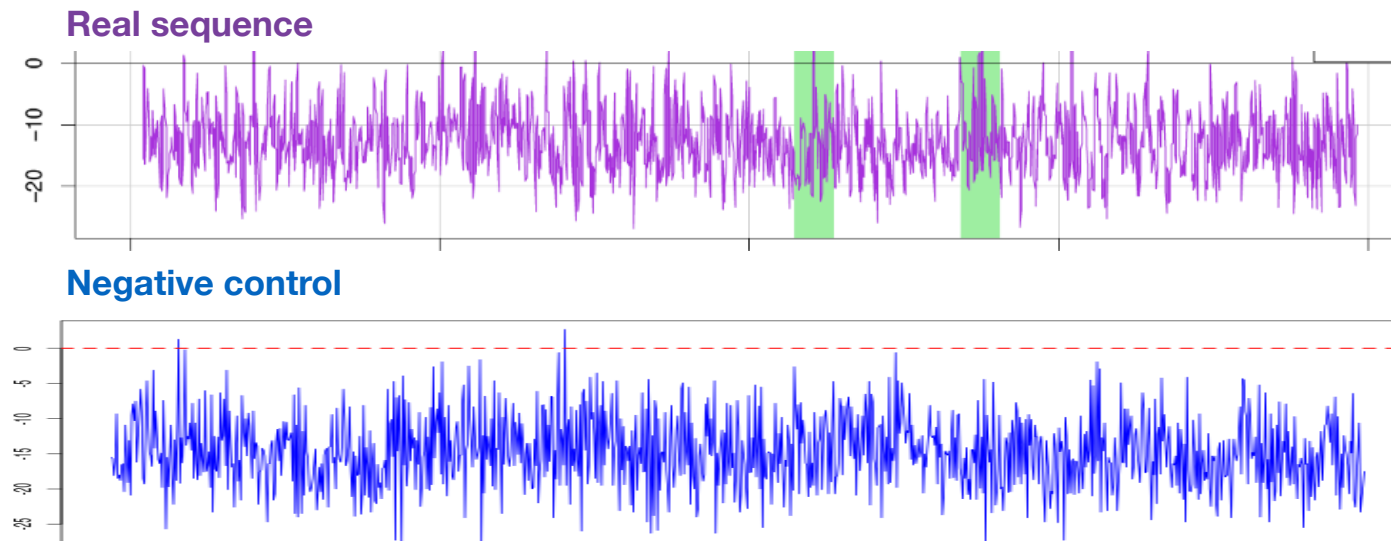
	a	c	g	t
a	0.25778	0.24159	0.24218	0.25846
c	0.29590	0.19550	0.28357	0.22503
g	0.30990	0.27463	0.08988	0.32559
t	0.26673	0.22779	0.27765	0.22783
	0.20207	0.23181	0.26906	0.29706



1.sample starting nucleotide

2.sample next nucleotides according  
to transition frequencies

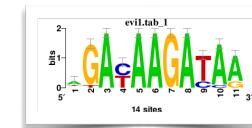
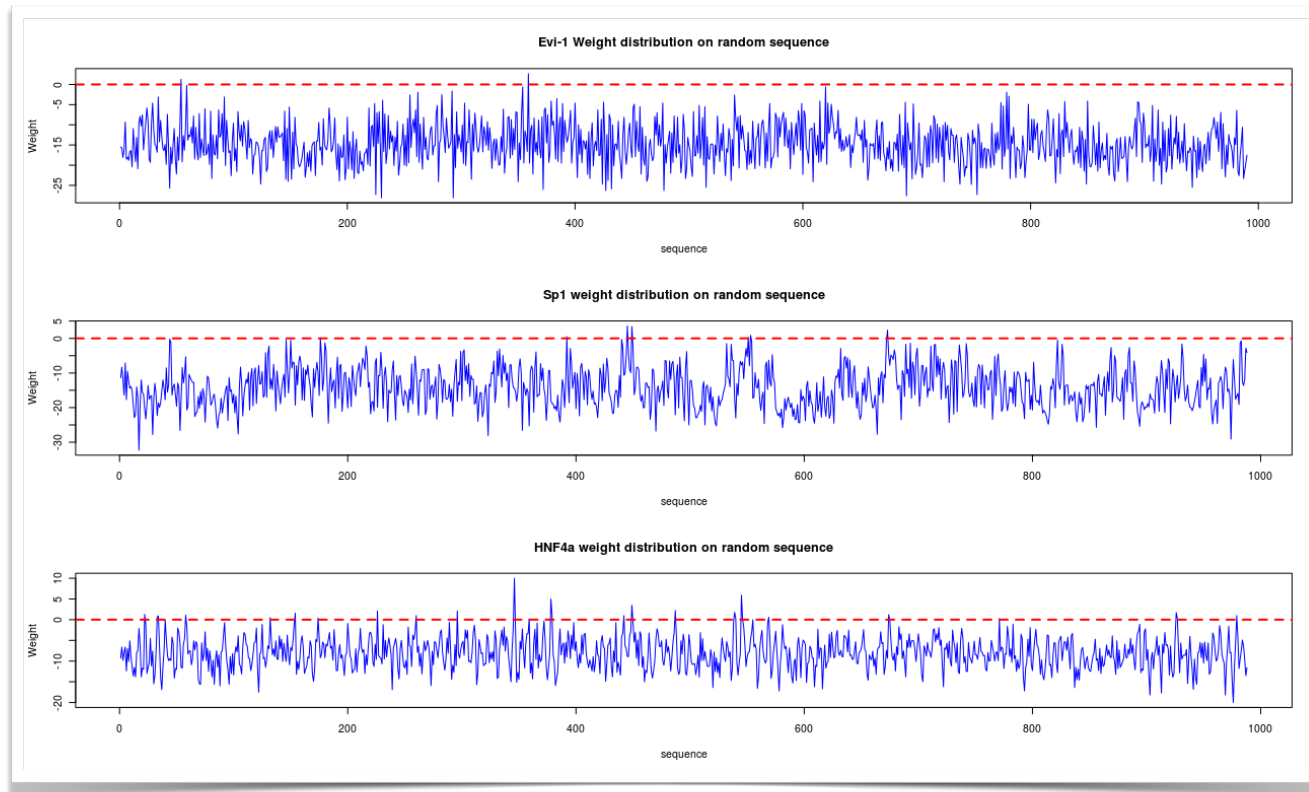
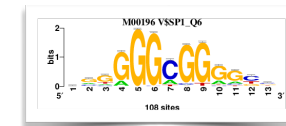
# Distribution of weights



- 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 ?***

1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025, 2026, 2027, 2028, 2029, 2030, 2031, 2032, 2033, 2034, 2035, 2036, 2037, 2038, 2039, 2040, 2041, 2042, 2043, 2044, 2045, 2046, 2047, 2048, 2049, 2050, 2051, 2052, 2053, 2054, 2055, 2056, 2057, 2058, 2059, 2060, 2061, 2062, 2063, 2064, 2065, 2066, 2067, 2068, 2069, 2070, 2071, 2072, 2073, 2074, 2075, 2076, 2077, 2078, 2079, 2080, 2081, 2082, 2083, 2084, 2085, 2086, 2087, 2088, 2089, 2090, 2091, 2092, 2093, 2094, 2095, 2096, 2097, 2098, 2099, 2100, 2101, 2102, 2103, 2104, 2105, 2106, 2107, 2108, 2109, 2110, 2111, 2112, 2113, 2114, 2115, 2116, 2117, 2118, 2119, 2120, 2121, 2122, 2123, 2124, 2125, 2126, 2127, 2128, 2129, 2130, 2131, 2132, 2133, 2134, 2135, 2136, 2137, 2138, 2139, 2140, 2141, 2142, 2143, 2144, 2145, 2146, 2147, 2148, 2149, 2150, 2151, 2152, 2153, 2154, 2155, 2156, 2157, 2158, 2159, 2160, 2161, 2162, 2163, 2164, 2165, 2166, 2167, 2168, 2169, 2170, 2171, 2172, 2173, 2174, 2175, 2176, 2177, 2178, 2179, 2180, 2181, 2182, 2183, 2184, 2185, 2186, 2187, 2188, 2189, 2190, 2191, 2192, 2193, 2194, 2195, 2196, 2197, 2198, 2199, 2200, 2201, 2202, 2203, 2204, 2205, 2206, 2207, 2208, 2209, 2210, 2211, 2212, 2213, 2214, 2215, 2216, 2217, 2218, 2219, 2220, 2221, 2222, 2223, 2224, 2225, 2226, 2227, 2228, 2229, 2230, 2231, 2232, 2233, 2234, 2235, 2236, 2237, 2238, 2239, 2240, 2241, 2242, 2243, 2244, 2245, 2246, 2247, 2248, 2249, 2250, 2251, 2252, 2253, 2254, 2255, 2256, 2257, 2258, 2259, 2260, 2261, 2262, 2263, 2264, 2265, 2266, 2267, 2268, 2269, 2270, 2271, 2272, 2273, 2274, 2275, 2276, 2277, 2278, 2279, 2280, 2281, 2282, 2283, 2284, 2285, 2286, 2287, 2288, 2289, 2290, 2291, 2292, 2293, 2294, 2295, 2296, 2297, 2298, 2299, 2300, 2301, 2302, 2303, 2304, 2305, 2306, 2307, 2308, 2309, 2310, 2311, 2312, 2313, 2314, 2315, 2316, 2317, 2318, 2319, 2320, 2321, 2322, 2323, 2324, 2325, 2326, 2327, 2328, 2329, 2330, 2331, 2332, 2333, 2334, 2335, 2336, 2337, 2338, 2339, 2340, 2341, 2342, 2343, 2344, 2345, 2346, 2347, 2348, 2349, 2350, 2351, 2352, 2353, 2354, 2355, 2356, 2357, 2358, 2359, 2360, 2361, 2362, 2363, 2364, 2365, 2366, 2367, 2368, 2369, 2370, 2371, 2372, 2373, 2374, 2375, 2376, 2377, 2378, 2379, 2380, 2381, 2382, 2383, 2384, 2385, 2386, 2387, 2388, 2389, 2390, 2391, 2392, 2393, 2394, 2395, 2396, 2397, 2398, 2399, 2400, 2401, 2402, 2403, 2404, 2405, 2406, 2407, 2408, 2409, 2410, 2411, 2412, 2413, 2414, 2415, 2416, 2417, 2418, 2419, 2420, 2421, 2422, 2423, 2424, 2425, 2426, 2427, 2428, 2429, 2430, 2431, 2432, 2433, 2434, 2435, 2436, 2437, 2438, 2439, 2440, 2441, 2442, 2443, 2444, 2445, 2446, 2447, 2448, 2449, 2450, 2451, 2452, 2453, 2454, 2455, 2456, 2457, 2458, 2459, 2460, 2461, 2462, 2463, 2464, 2465, 2466, 2467, 2468, 2469, 2470, 2471, 2472, 2473, 2474, 2475, 2476, 2477, 2478, 2479, 2480, 2481, 2482, 2483, 2484, 2485, 2486, 2487, 2488, 2489, 2490, 2491, 2492, 2493, 2494, 2495, 2496, 2497, 2498, 2499, 2500, 2501, 2502, 2503, 2504, 2505, 2506, 2507, 2508, 2509, 2510, 2511, 2512, 2513, 2514, 2515, 2516, 2517, 2518, 2519, 2520, 2521, 2522, 2523, 2524, 2525, 2526, 2527, 2528, 2529, 2530, 2531, 2532, 2533, 2534, 2535, 2536, 2537, 2538, 2539, 2540, 2541, 2542, 2543, 2544, 2545, 2546, 2547, 2548, 2549, 2550, 2551, 2552, 2553, 2554, 2555, 2556, 2557, 2558, 2559, 2560, 2561, 2562, 2563, 2564, 2565, 2566, 2567, 2568, 2569, 2570, 2571, 2572, 2573, 2574, 2575, 2576, 2577, 2578, 2579, 2580, 2581, 2582, 2583, 2584, 2585, 2586, 2587, 2588, 2589, 2590, 2591, 2592, 2593, 2594, 2595, 2596, 2597, 2598, 2599, 2600, 2601, 2602, 2603, 2604, 2605, 2606, 2607, 2608, 2609, 2610, 2611, 2612, 2613, 2614, 2615, 2616, 2617, 2618, 2619, 2620, 2621, 2622, 2623, 2624, 2625, 2626, 2627, 2628, 2629, 2630, 2631, 2632, 2633, 2634, 2635, 2636, 2637, 2638, 2639, 2640, 2641, 2642, 2643, 2644, 2645, 2646, 2647, 2648, 2649, 2650, 2651, 2652, 2653, 2654, 2655, 2656, 2657, 2658, 2659, 2660, 2661, 2662, 2663, 2664, 2665, 2666, 2667, 2668, 2669, 2670, 2671, 2672, 2673, 2674, 2675, 2676, 2677, 2678, 2679, 26


$$\text{FPR} = 2/1000$$

$$\text{FPR} = 7/1000$$

$$\text{FPR} = 29/1000$$

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

# Distribution of weights

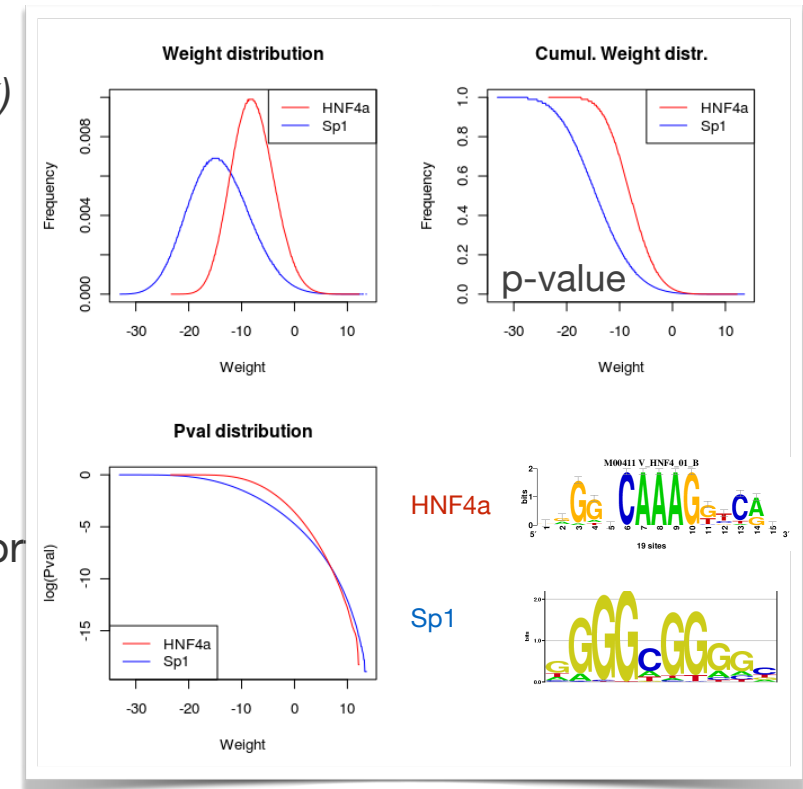
For a given matrix, we can exactly compute the **exact distribution** of scores under a particular background model:

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 proportion of higher score

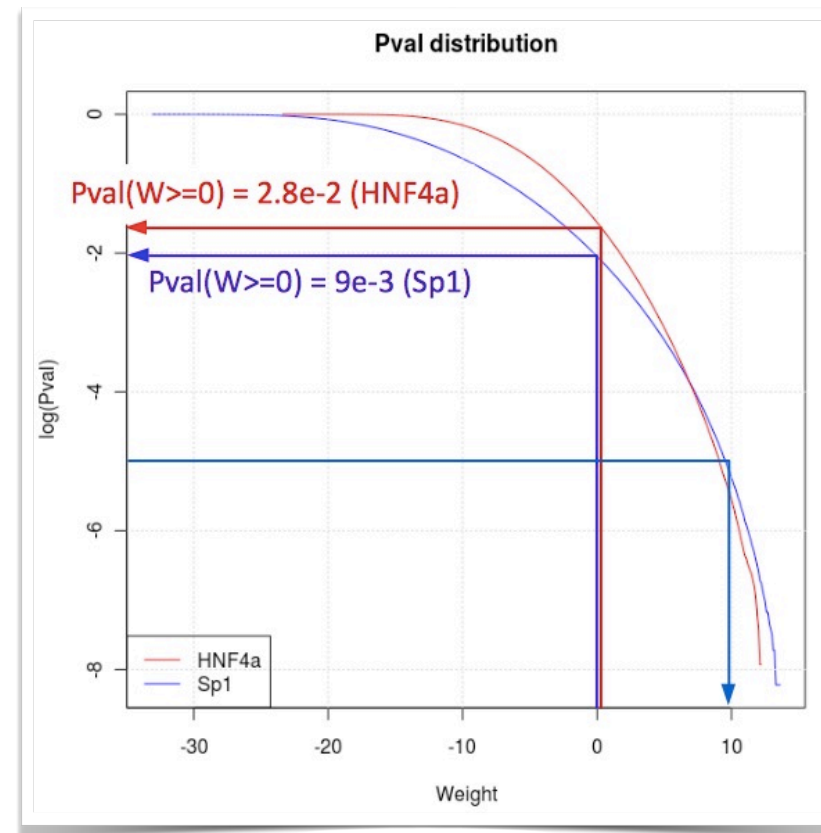
$$Pval(W) = \text{freq. } w \geq W$$





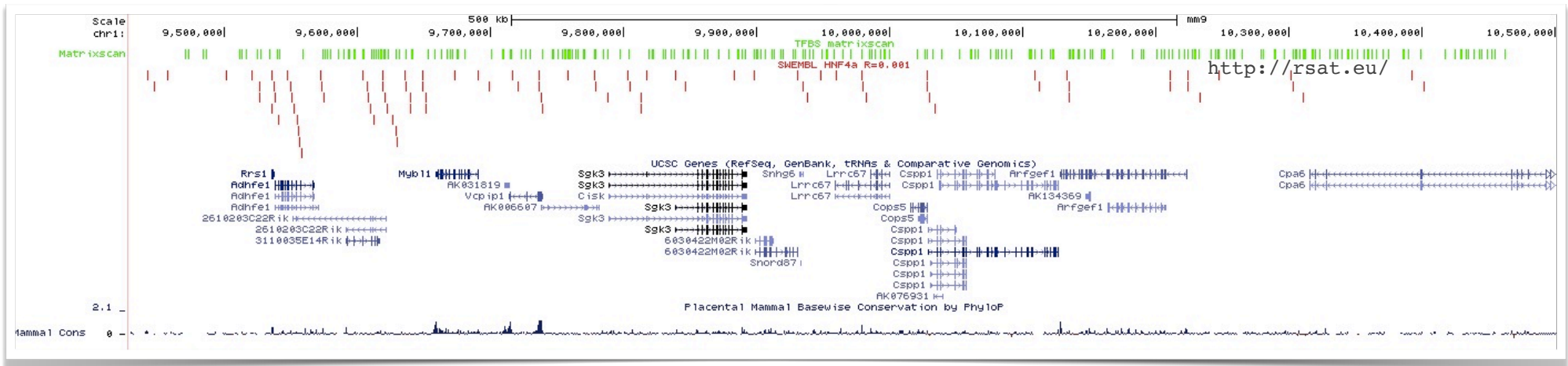
# Distribution of weights

- 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 :  $p \leq 1e-4$
- Background : Markov model order=3 estimated on input sequence
- Output : **259 predicted TFBS**
- **Control : HNF4a ChIP-seq peaks (red)**



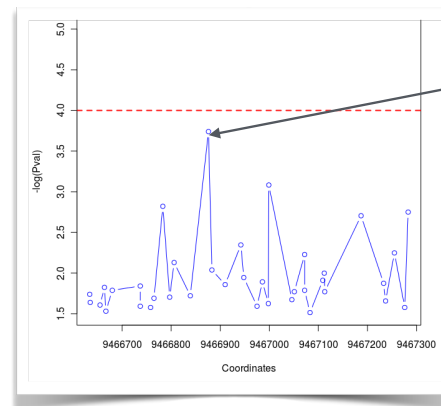
# True/false positive/negative

Predicted TFBS  
ChIP-seq binding

**True  
positive**

**False  
positive**  
(TFBS predicted  
but no binding)

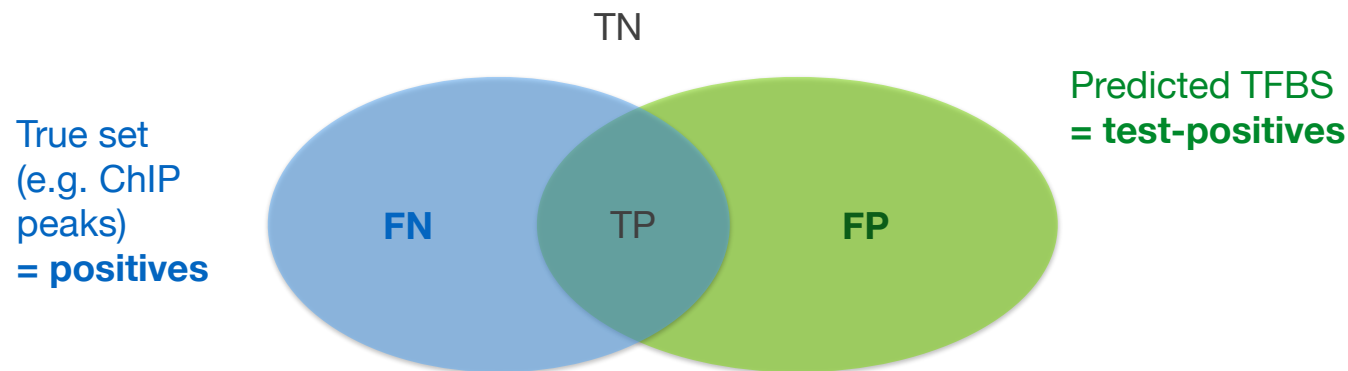
**False  
negative** (binding but  
no TFBS predicted)



MO0411 VSHNFA 01 B RC  
19 sites  
**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?*  
 $\text{sensitivity} = TP/P = TP / (TP+FN) \rightarrow \text{True Positive Rate (TPR)}$
- **specificity** : *how many of the non-binding sites did I identify correctly as negative ?*  
 $\text{specificity} = TN/N = TN / (TN + FP)$   
 $1\text{-specificity} = FP/N = FP / (TN+FP) \rightarrow \text{False Positive Rate (FPR)}$

# Performance measure

		Condition (as determined by "Gold standard")		
		Condition positive	Condition negative	
Test outcome	Test outcome positive	True positive	False positive (Type I error)	Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$
	Test outcome negative	False negative (Type II error)	True negative	Negative predictive value = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$
		Sensitivity = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	Specificity = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Accuracy $\text{Acc} = (TP+TN)/(P+N)$

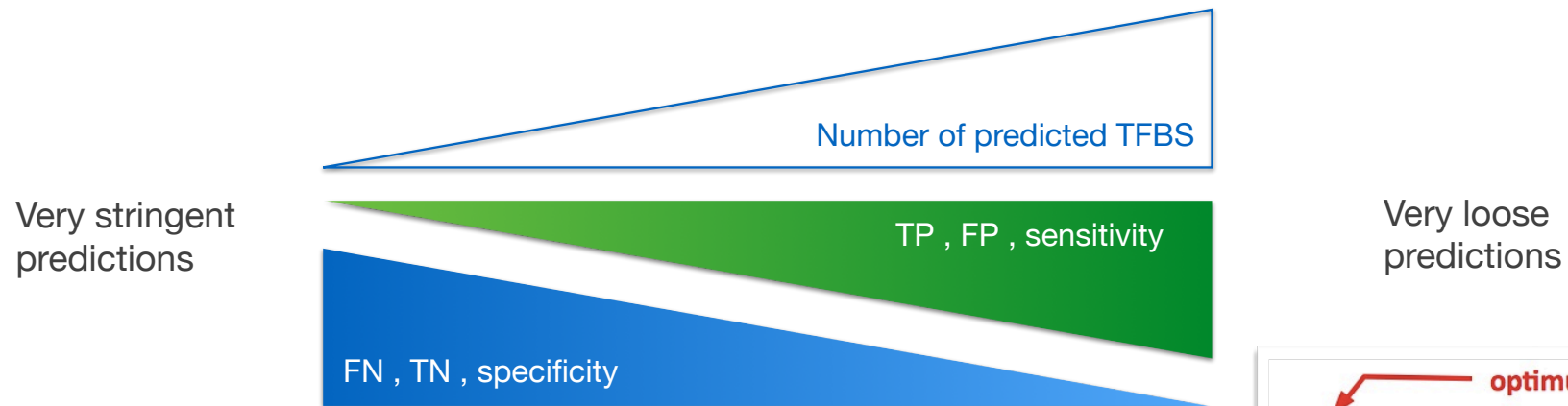
It is a true binding site

It is NOT a binding site

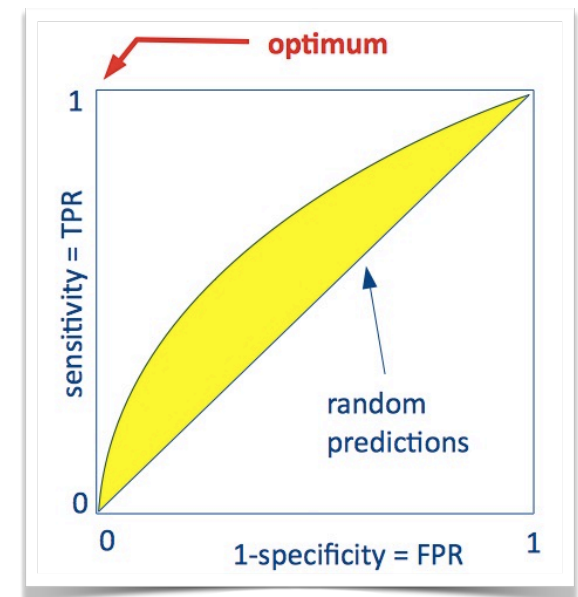
I predict a binding site

I predict a non-binding site

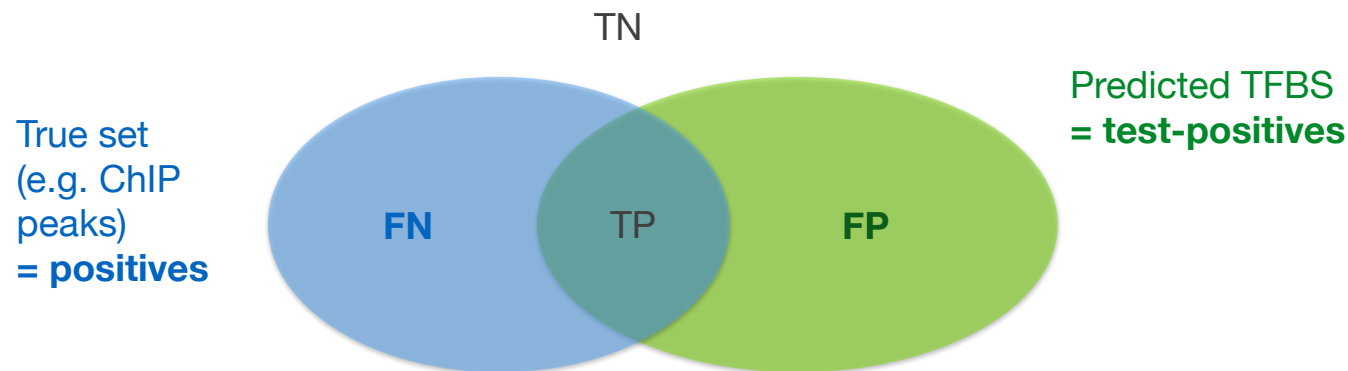
# Evaluation of TFBS prediction performances



- 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 than positives (as for ChIP data), then  $FPR = FP/N \ll 1$  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 =  $TP/P$**

