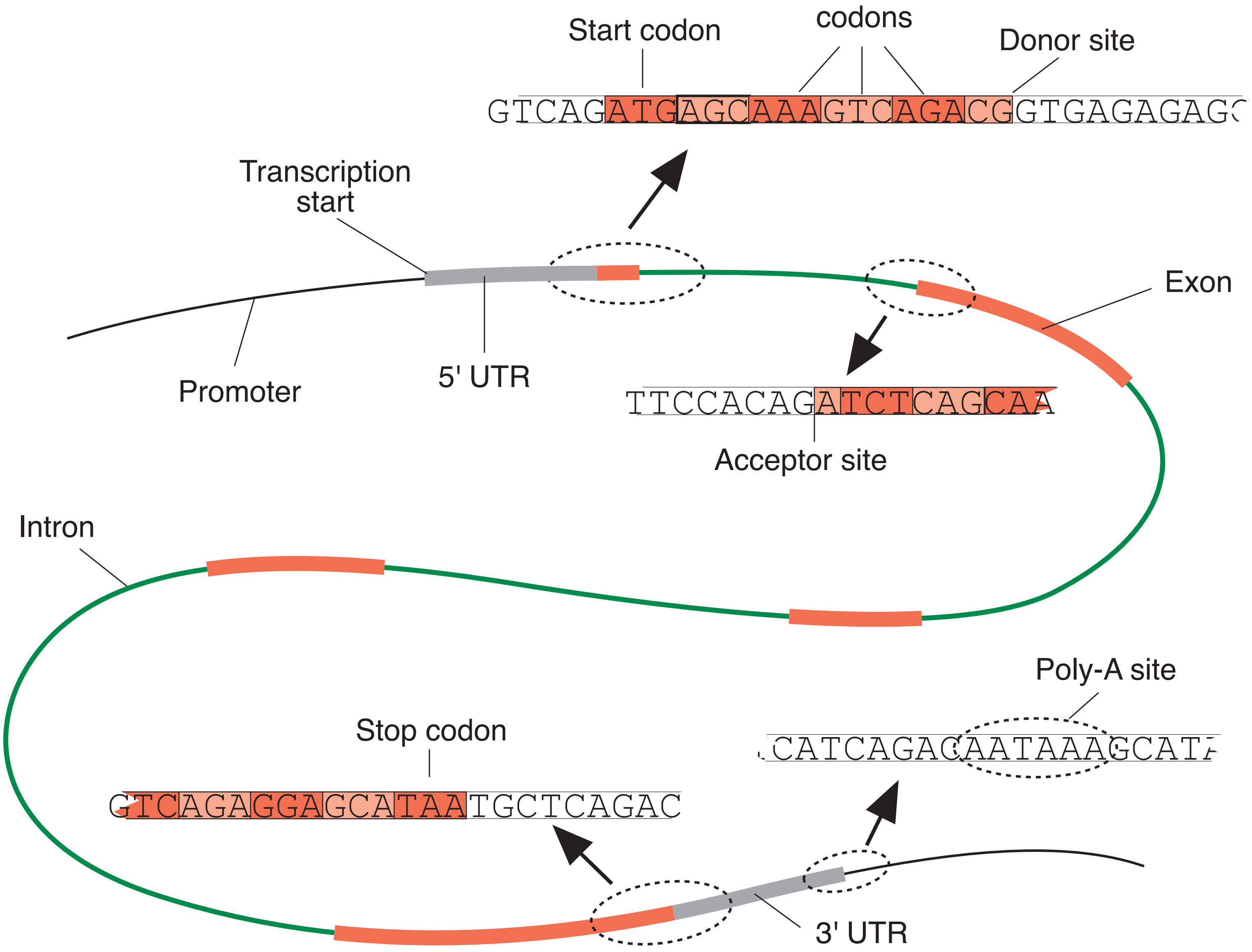
Weight matrices
- or -
Position Specific Scoring Matrices (PSSMs)

Introns and Exons



The most frequent donor sites with their number in the data set of 2068 donors.
Two bases upstream and 6 bases downstream are included.

AGGTGAGT	92	AGGTAAGA	43	AAGTGAGT	20	GGGTAAGT	16
GGGTGAGT	86	AGGTGGGT	40	CAGTGAGT	20	AGGTATGG	15
AGGTAAGC	67	ATGTGAGT	38	AGGTAGGC	19	GGGTGAGC	14
TGGTGAGT	62	TGGTAAGT	33	ATGTAAGT	19	TGGTATGT	14
AGGTGAGA	57	AAGTAAGT	26	AGGTCAGT	18	CTGTGAGT	14
AGGTGAGG	55	AGGTTGGT	26	TGGTAAGG	18	GGGTAAGA	14
AGGTAAGT	49	AGGTATGT	25	CTGTAAGT	18	GAGTAAGT	13
CGGTGAGT	48	AGGTAAAT	24	CAGTAAGT	18	AGGTAATT	13
AGGTGAGC	47	AGGTAGGT	24	TGGTGAGC	17	CGGTAAGT	13
AGGTAAGG	45	TGGTGAGA	21	TGGTAAGA	16	GAGTGAGT	13

Counts

Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	516	597	713	1230	162	0	0	1062	1469	135	318	548	391	443	421
C	495	596	725	261	49	0	0	59	167	106	322	469	623	593	516
G	523	542	399	291	1692	2068	0	895	250	1731	362	645	487	543	626
T	534	333	231	286	165	0	2068	52	182	96	1066	406	567	489	505

Frequencies

Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	0.25	0.29	0.34	0.59	0.08	0.00	0.00	0.51	0.71	0.07	0.15	0.26	0.19	0.21	0.20
C	0.24	0.29	0.35	0.13	0.02	0.00	0.00	0.03	0.08	0.05	0.16	0.23	0.30	0.29	0.25
G	0.25	0.26	0.19	0.14	0.82	1.00	0.00	0.43	0.12	0.84	0.18	0.31	0.24	0.26	0.30
T	0.26	0.16	0.11	0.14	0.08	0.00	1.00	0.03	0.09	0.05	0.52	0.20	0.27	0.24	0.24

Position Specific Score Matrix (PSSM)

$$\text{sequence } x = x_1, \ldots, x_l$$

$$P(x) = p_1(x_1) p_2(x_2) \cdots p_l(x_l)$$

$$\begin{aligned}\text{log-odds} &= \log \frac{P(x)}{Q(x)} \\ &= \log \frac{p_1(x_1)}{q(x_1)} + \log \frac{p_2(x_2)}{q(x_2)} + \ldots + \log \frac{p_l(x_l)}{q(x_l)}\end{aligned}$$

Sum of terms

$$s_i(a) = \log \frac{p_i(a)}{q(a)}$$

Counts

Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	516	597	713	1230	162	0	0	1062	1469	135	318	548	391	443	421
C	495	596	725	261	49	0	0	59	167	106	322	469	623	593	516
G	523	542	399	291	1692	2068	0	895	250	1731	362	645	487	543	626
T	534	333	231	286	165	0	2068	52	182	96	1066	406	567	489	505

Frequencies

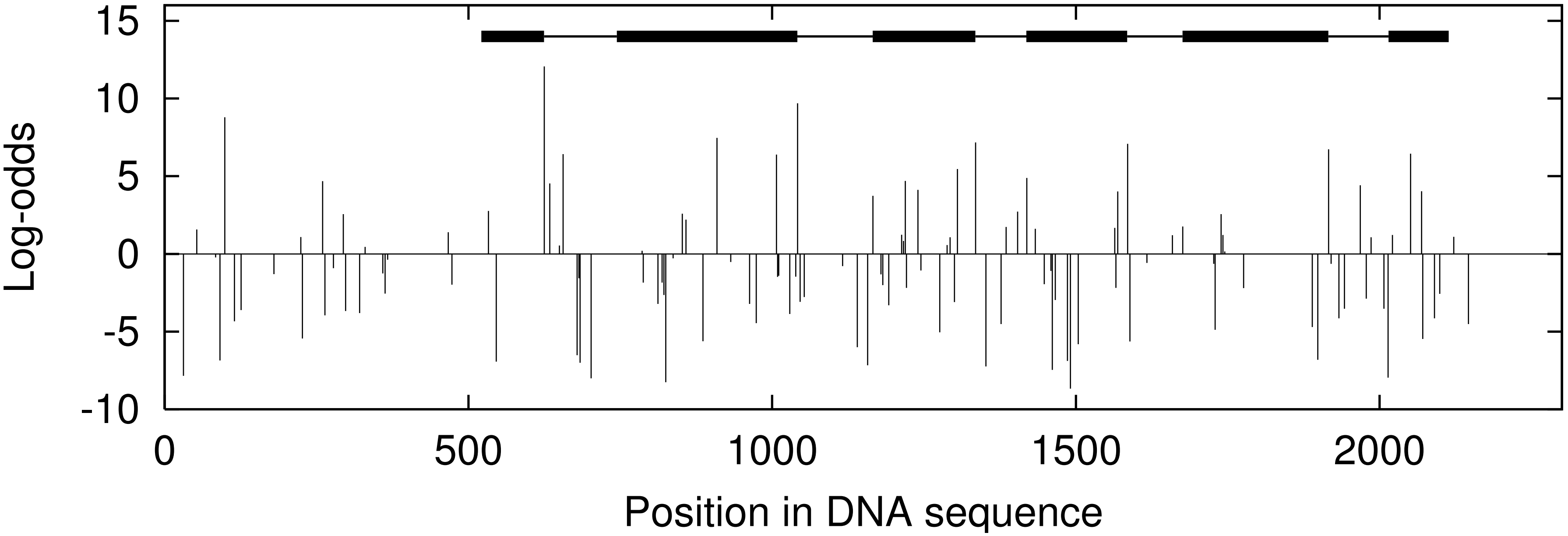
Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	0.25	0.29	0.34	0.59	0.08	0.00	0.00	0.51	0.71	0.07	0.15	0.26	0.19	0.21	0.20
C	0.24	0.29	0.35	0.13	0.02	0.00	0.00	0.03	0.08	0.05	0.16	0.23	0.30	0.29	0.25
G	0.25	0.26	0.19	0.14	0.82	1.00	0.00	0.43	0.12	0.84	0.18	0.31	0.24	0.26	0.30
T	0.26	0.16	0.11	0.14	0.08	0.00	1.00	0.03	0.09	0.05	0.52	0.20	0.27	0.24	0.24

Log-odds

Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	0.00	0.21	0.46	1.25	-1.67	$-\infty$	$-\infty$	1.04	1.51	-1.94	-0.70	0.08	-0.40	-0.22	-0.30
C	-0.06	0.21	0.49	-0.99	-3.40	$-\infty$	$-\infty$	-3.13	-1.63	-2.29	-0.68	-0.14	0.27	0.20	0.00
G	0.02	0.07	-0.37	-0.83	1.71	2.00	$-\infty$	0.79	-1.05	1.74	-0.51	0.32	-0.09	0.07	0.28
T	0.05	-0.63	-1.16	-0.85	-1.65	$-\infty$	2.00	-3.31	-1.51	-2.43	1.04	-0.35	0.13	-0.08	-0.03

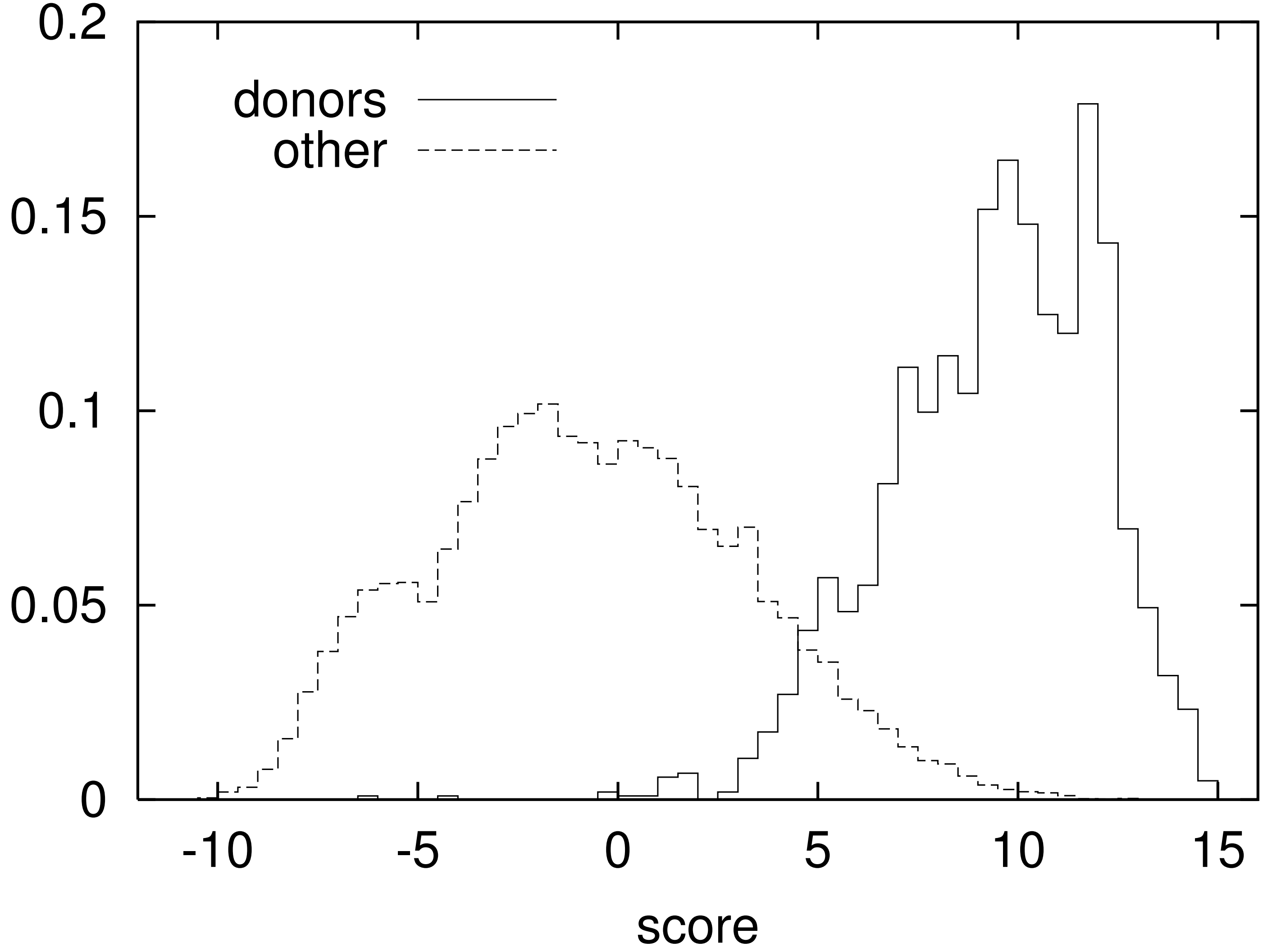
Searching for patterns

Score each possible position



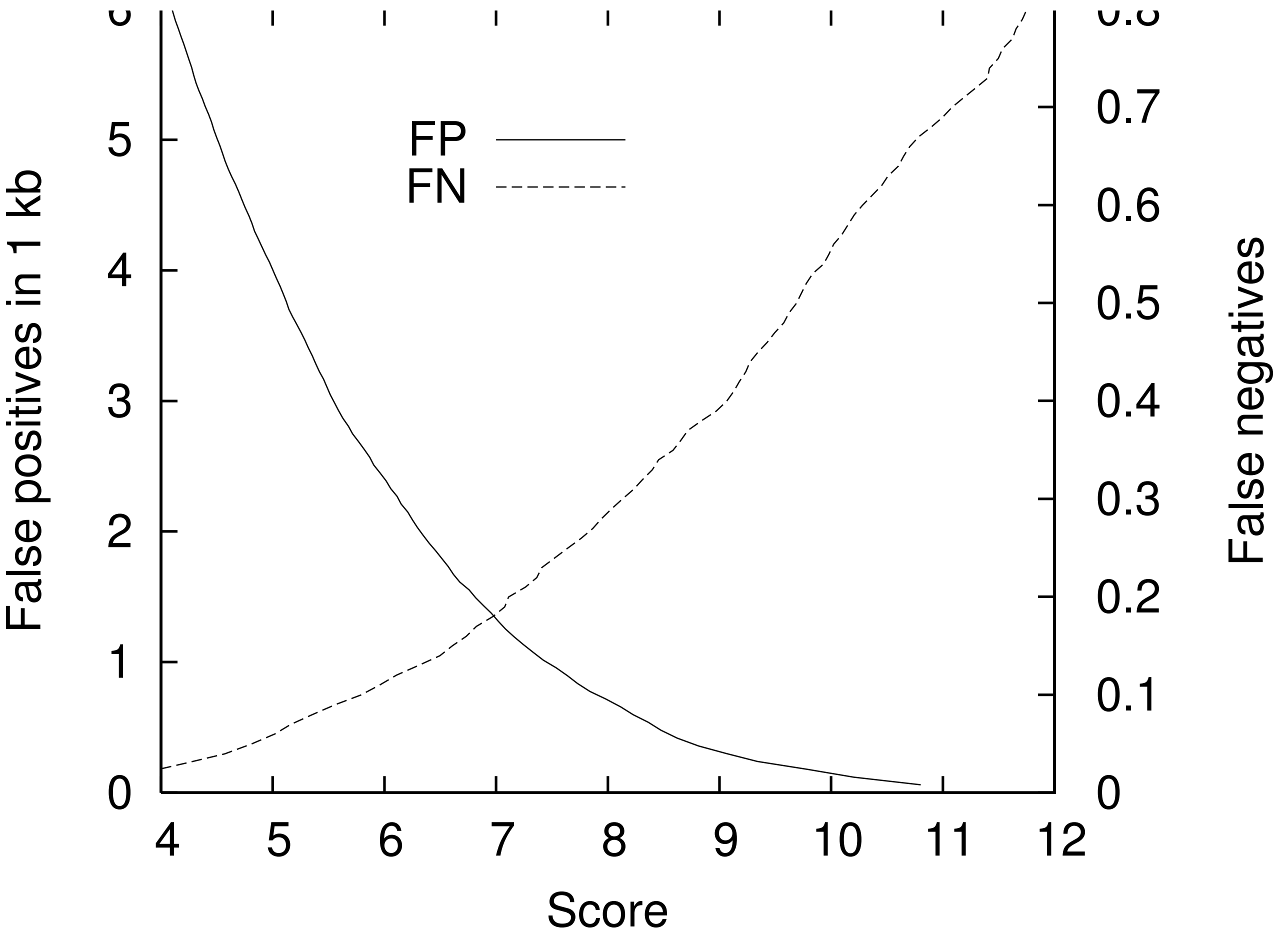
Discrimination

Compare real donors with a set of exons and introns

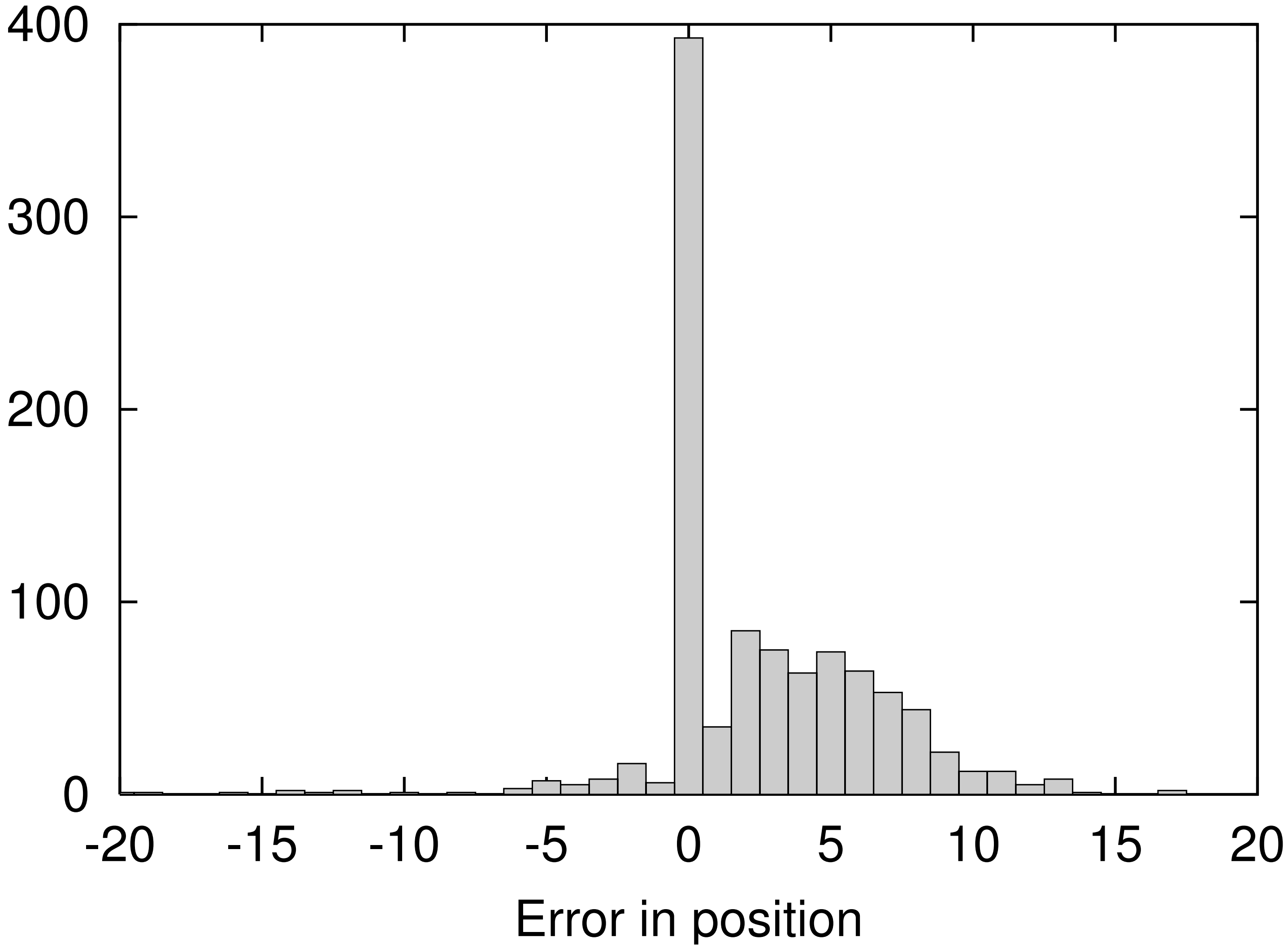


False positives and negatives

Compare real donors with a set of exons and introns

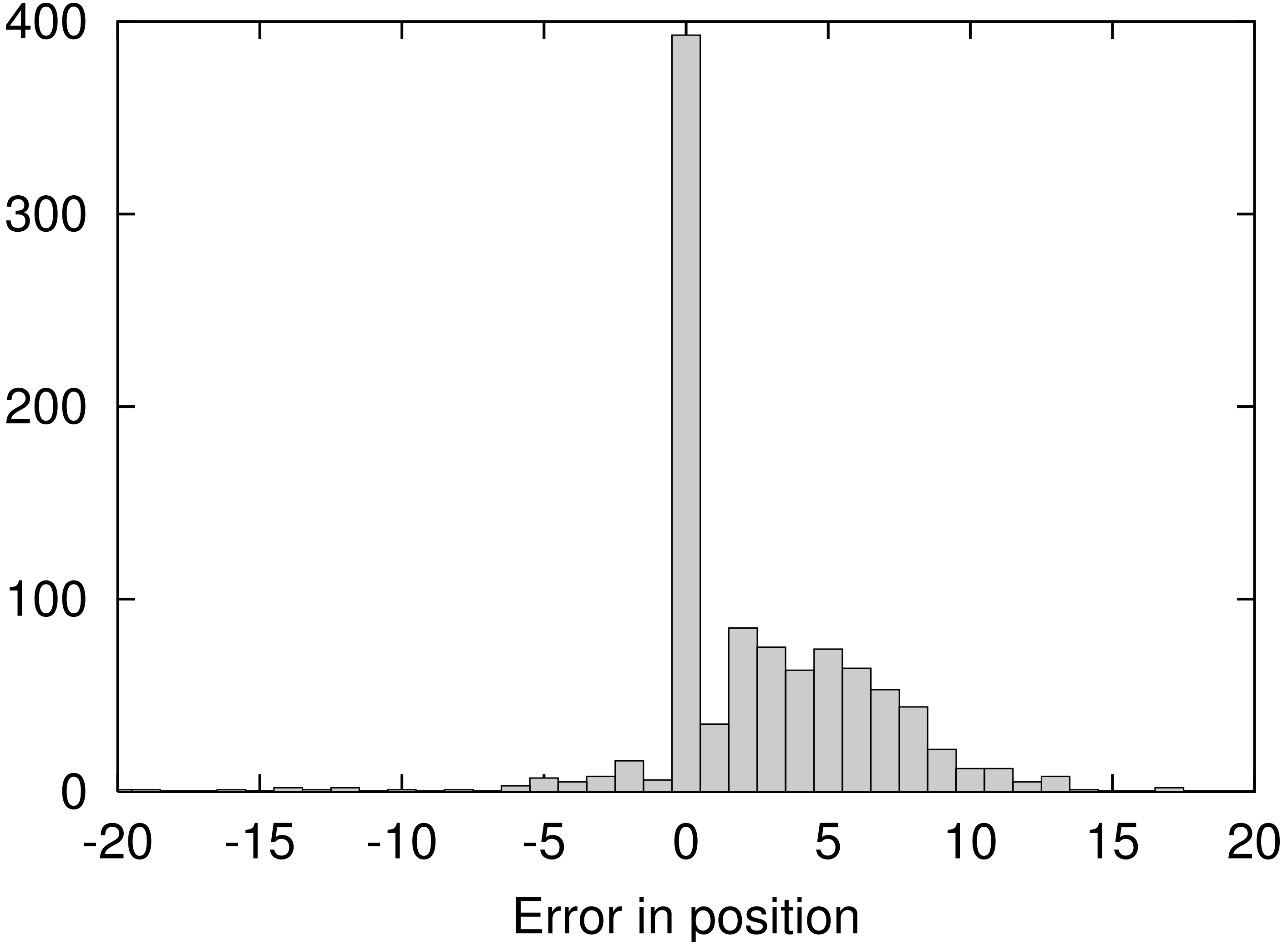


Signal Peptides



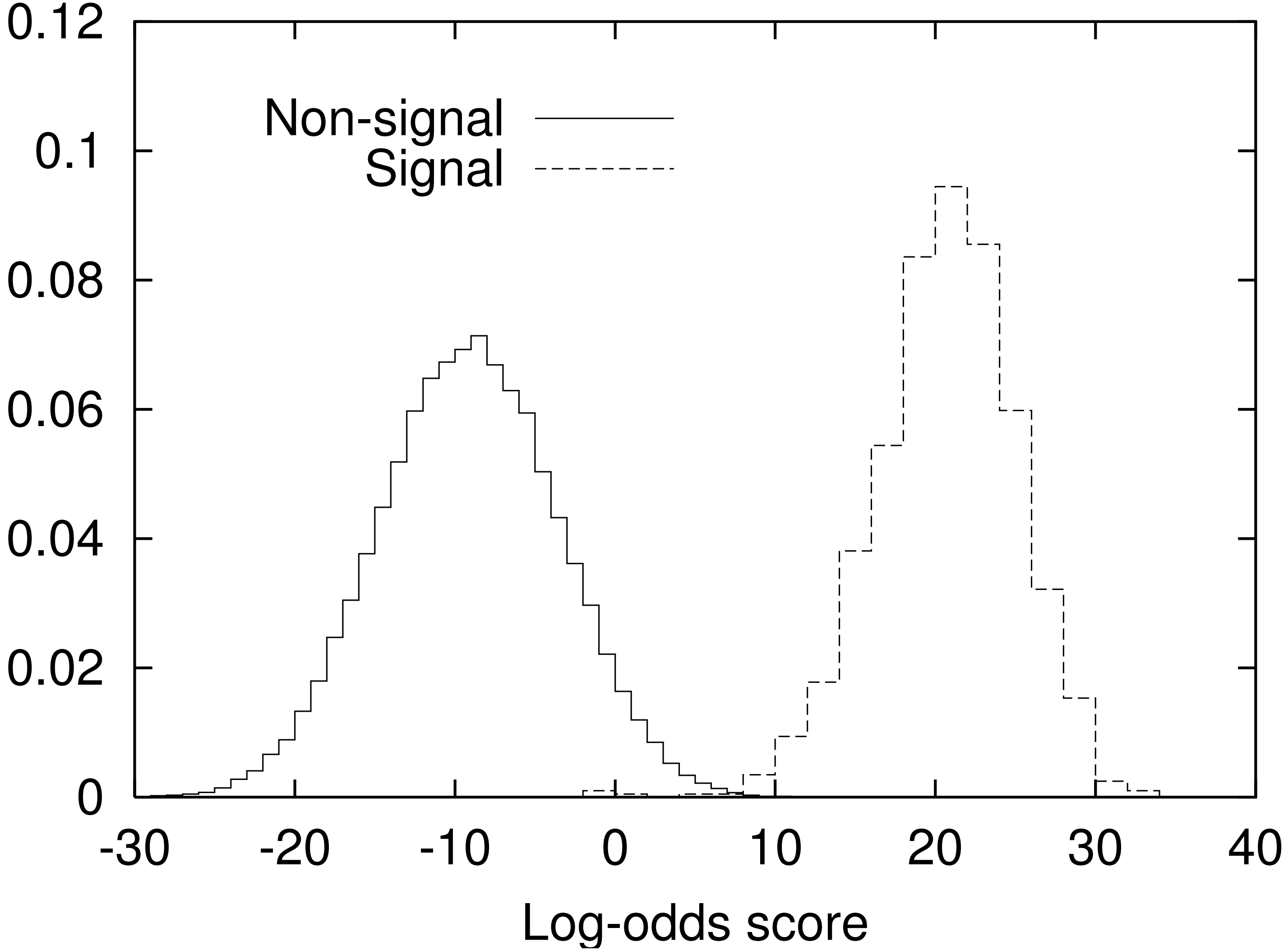
Discrimination

Cross-validation. Testing against proteins without signal peptides.



Discrimination

Realistic background



Information content

Average score =

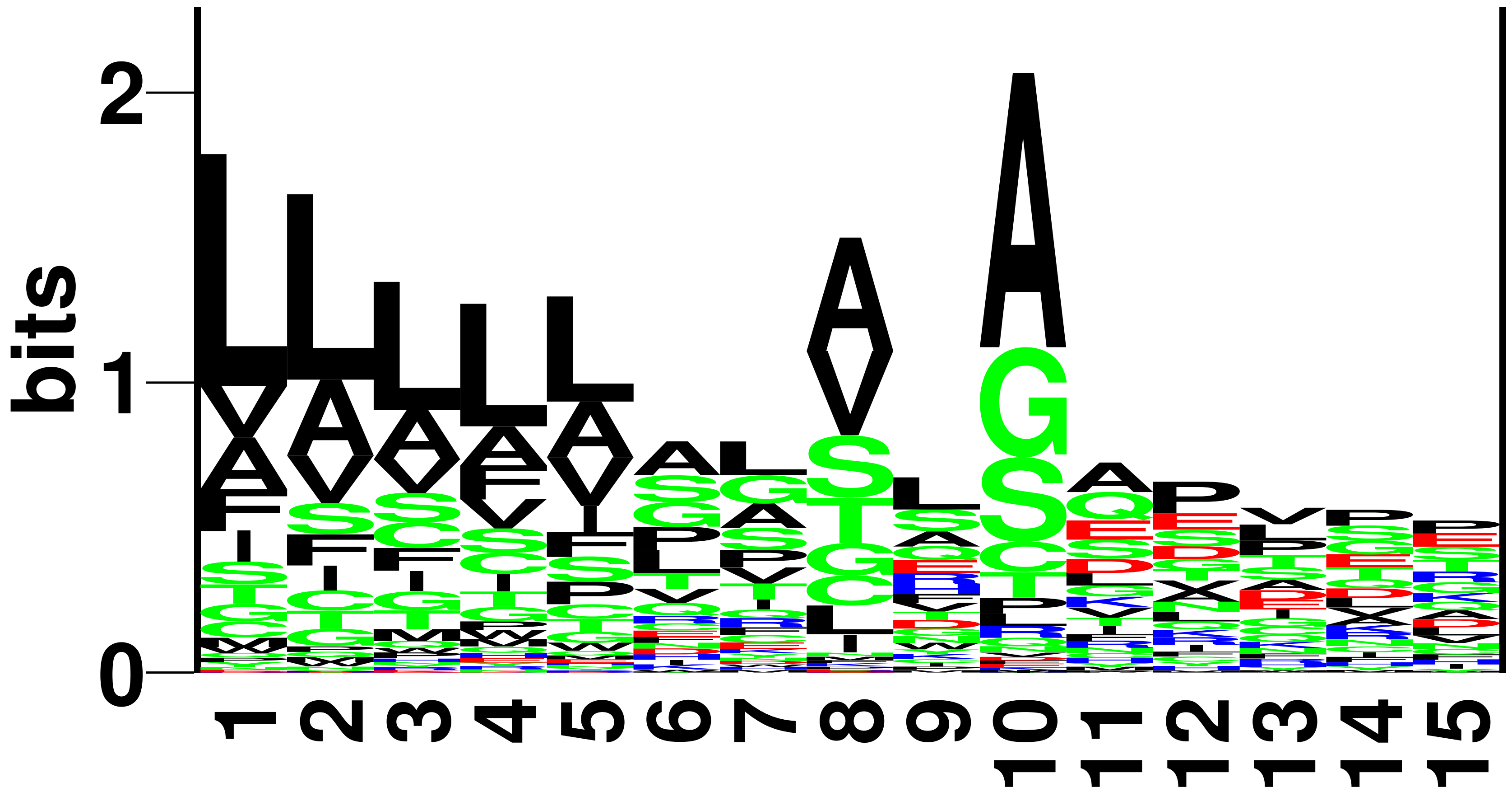
$$\sum_a p_1(a)\log\frac{p_1(a)}{q(a)}+\sum_a p_2(a)\log\frac{p_2(a)}{q(a)}+\ldots+\sum_a p_l(a)\log\frac{p_l(a)}{q(a)}$$

Relative entropy

$$H(p||q)=\sum_a p(a)\log\frac{p(a)}{q(a)}$$

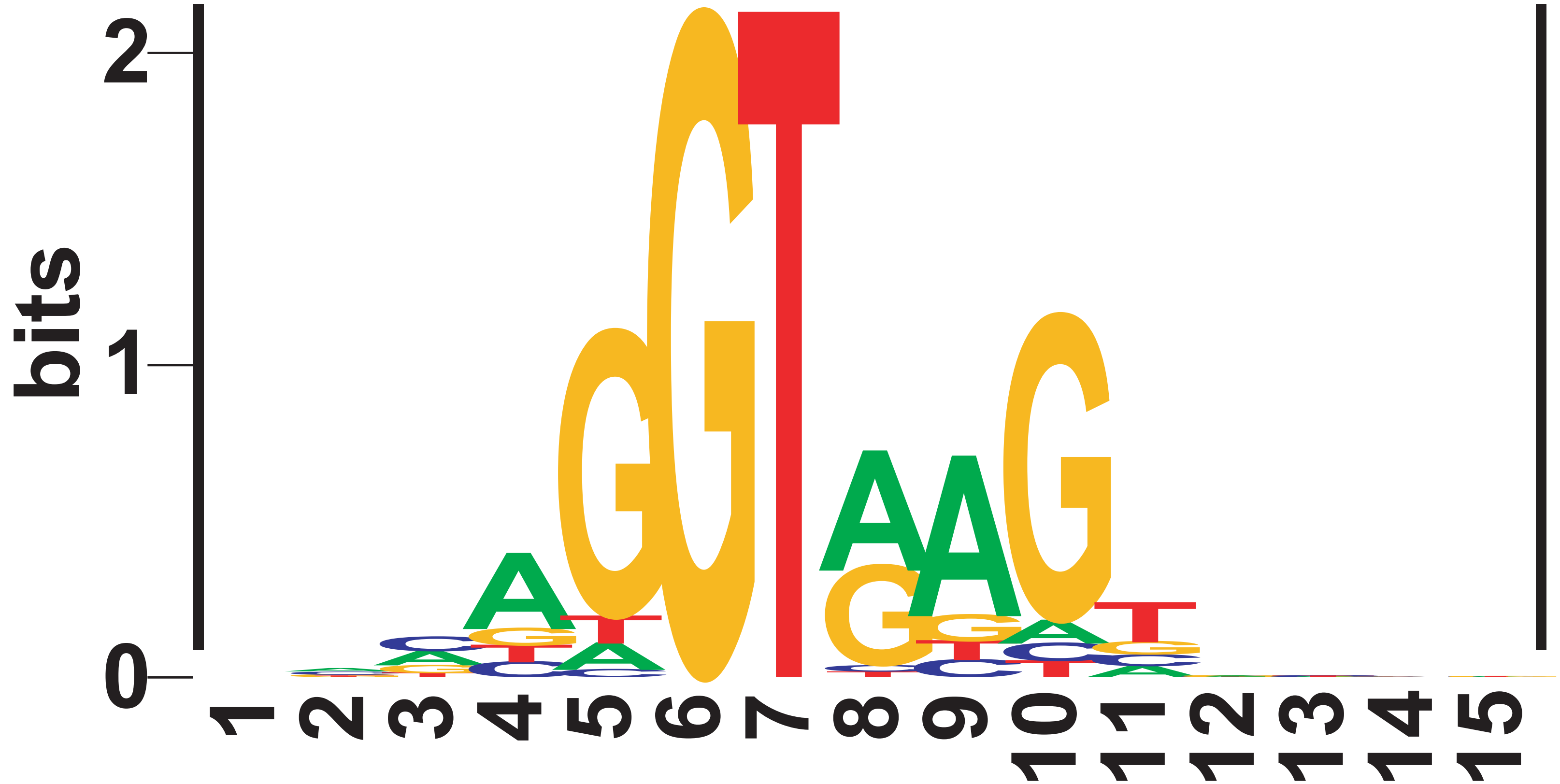
Logo

Signal peptide



Logo

Donor site


bits
A
G
G
T
A
A
G
T