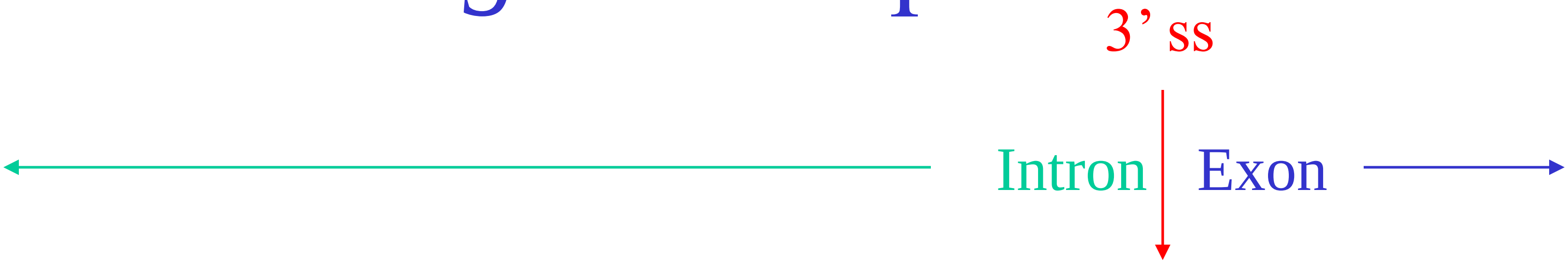
Today’s Lecture

- Weight matrices
- Score distributions
- Relative entropy and sequence logos

Nucleotide Counts for 8192 *C. elegans* 3’ Splice Sites



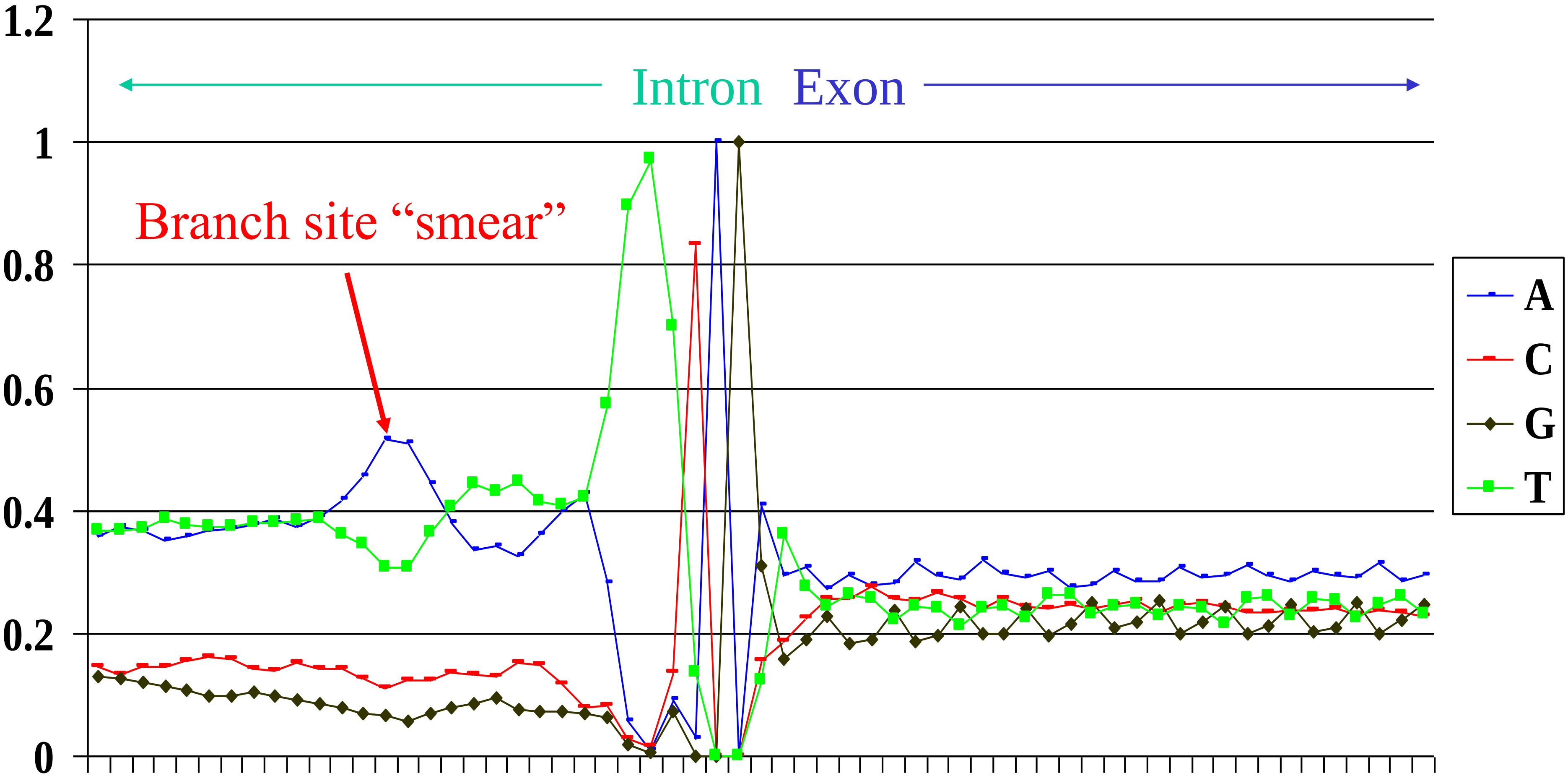
3’ ss

Intron

Exon

A	3276	3516	2313	476	67	757	240	8192	0	3359	2401	2514
C	970	648	664	236	129	1109	6830	0	0	1277	1533	1847
G	593	575	516	144	39	595	12	0	8192	2539	1301	1567
T	3353	3453	4699	7336	7957	5731	1110	0	0	1017	2957	2264
CONSENSUS	W	W	W	T	T	t	C	A	G	r	w	w
A	0.400	0.429	0.282	0.058	0.008	0.092	0.029	1.000	0.000	0.410	0.293	0.307
C	0.118	0.079	0.081	0.029	0.016	0.135	0.834	0.000	0.000	0.156	0.187	0.225
G	0.072	0.070	0.063	0.018	0.005	0.073	0.001	0.000	1.000	0.310	0.159	0.191
T	0.409	0.422	0.574	0.896	0.971	0.700	0.135	0.000	0.000	0.124	0.361	0.276

3’ Splice Sites – *C. elegans*



Weight Matrices for Site Models

- LR for sites: (prob under site model) / (prob under non-site (background) model)

$$\frac{P(s \mid M_{\text{site}})}{P(s \mid M_{\text{background}})} = \frac{\prod_{1 \le i \le n} P_i(s_i \mid M_{\text{site}})}{\prod_{1 \le i \le n} P_i(s_i \mid M_{\text{background}})}$$

- LLR = $\sum_{1 \le i \le n} \log(P_i(s_i \mid M_{\text{site}})) - \log(P_i(s_i \mid M_{\text{background}}))$
  - compute by reading from a *matrix* whose *i*-th column contains values $\log(P_i(r \mid M_{\text{site}}))$ - $\log(P_i(r \mid M_{\text{background}}))$ for each residue *r* (with *r* labelling the rows).
    - We use $\log_2$.

Example: 3’ splice sites in C. *elegans*

- For *background distribution* take
  - genomic residue freqs computed from C. *elegans* chrom. I:

A	4,575,132:	0.321
C	2,559,048:	0.179
G	2,555,862:	0.179
T	4,582,688:	0.321

  - other choices are possible, e.g. composition of *transcribed regions*
- For the *site distribution* we take
  - site residue freqs from 8192 sites:

Weight Matrix – 3’ Splice Sites

```
SITE FREQUENCIES:
A  0.400  0.429  0.282  0.058  0.008  0.092  0.029  1.000  0.000  0.410  0.293  0.307
C  0.118  0.079  0.081  0.029  0.016  0.135  0.834  0.000  0.000  0.156  0.187  0.225
G  0.072  0.070  0.063  0.018  0.005  0.073  0.001  0.000  1.000  0.310  0.159  0.191
T  0.409  0.422  0.574  0.896  0.971  0.700  0.135  0.000  0.000  0.124  0.361  0.276

BACKGROUND FREQUENCIES:
A  0.321  0.321  0.321  0.321  0.321  0.321  0.321  0.321  0.321  0.321  0.321  0.321
C  0.179  0.179  0.179  0.179  0.179  0.179  0.179  0.179  0.179  0.179  0.179  0.179
G  0.179  0.179  0.179  0.179  0.179  0.179  0.179  0.179  0.179  0.179  0.179  0.179
T  0.321  0.321  0.321  0.321  0.321  0.321  0.321  0.321  0.321  0.321  0.321  0.321



WEIGHTS:
A   0.32   0.42  -0.18  -2.46  -5.29  -1.79  -3.45   1.64 -99.00   0.36  -0.13  -0.06
C  -0.60  -1.18  -1.15  -2.64  -3.51  -0.41   2.22 -99.00 -99.00  -0.20   0.06   0.33
G  -1.31  -1.35  -1.51  -3.35  -5.23  -1.30  -6.93 -99.00   2.48   0.79  -0.17   0.10
T   0.35   0.39   0.84   1.48   1.60   1.12  -1.24 -99.00 -99.00  -1.37   0.17  -0.22
```

Scoring a Candidate 3’ Splice Site

A	0.32	0.42	-0.18	-2.46	-5.29	-1.79	-3.45	1.64	-99.00	0.36	-0.13	-0.06
C	-0.60	-1.18	-1.15	-2.64	-3.51	-0.41	2.22	-99.00	-99.00	-0.20	0.06	0.33
G	-1.31	-1.35	-1.51	-3.35	-5.23	-1.30	-6.93	-99.00	2.48	0.79	-0.17	0.10
T	0.35	0.39	0.84	1.48	1.60	1.12	-1.24	-99.00	-99.00	-1.37	0.17	-0.22

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

0.35 + 0.39 +-1.15 + 1.48 + 1.60 +-1.79 + 2.22 + 1.64 + 2.48 + 0.36 +-0.13 +-0.22 = 7.23

- General def.: a *weight matrix* W has

  entries w_{rj} indexed by residues $r \in A$, and $1 \leq j \leq n$

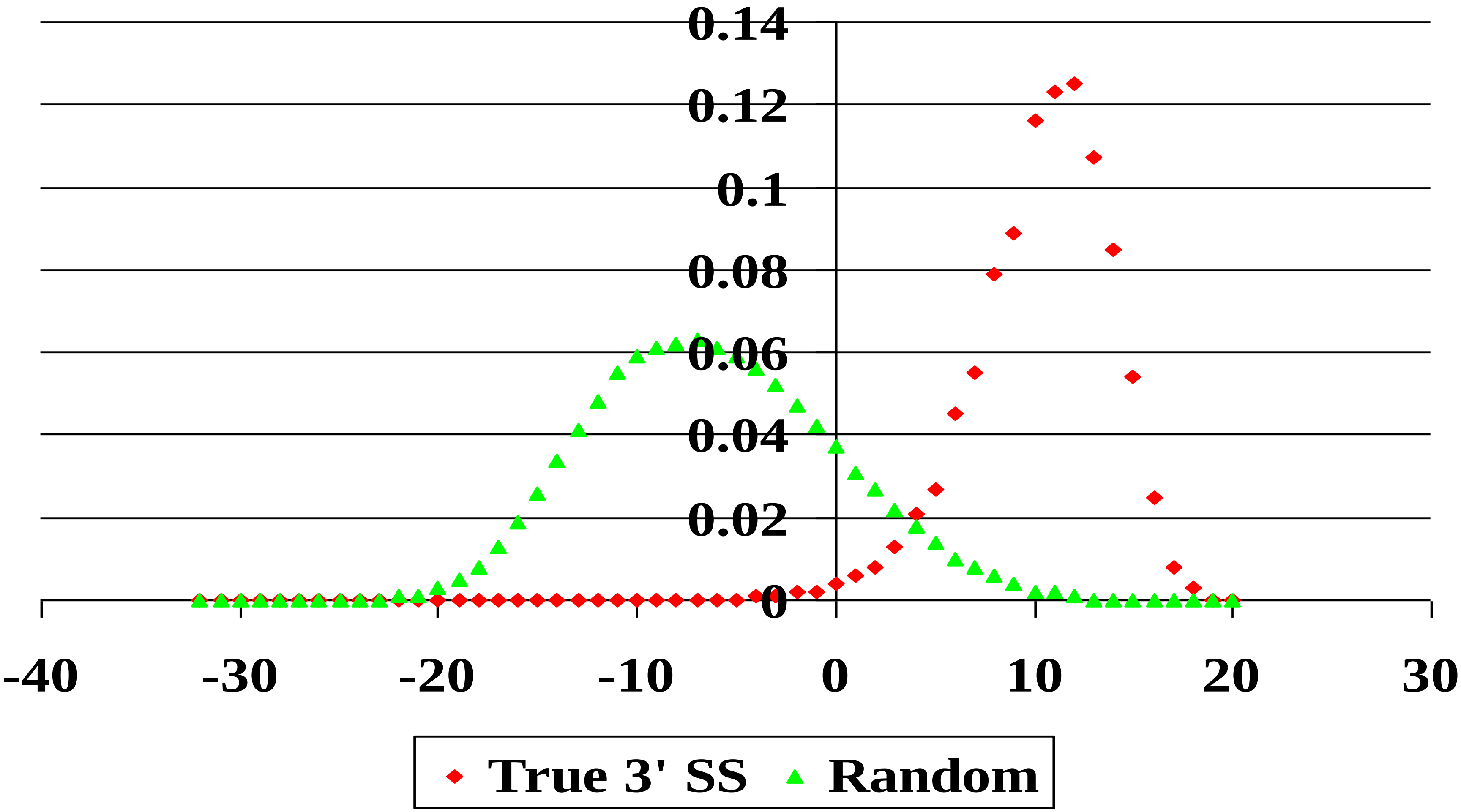
- *score* of a sequence $s = (s_1 s_2 \ldots s_n)$ is

$$\sum_{1 \leq j \leq n} w_{s_j j}$$

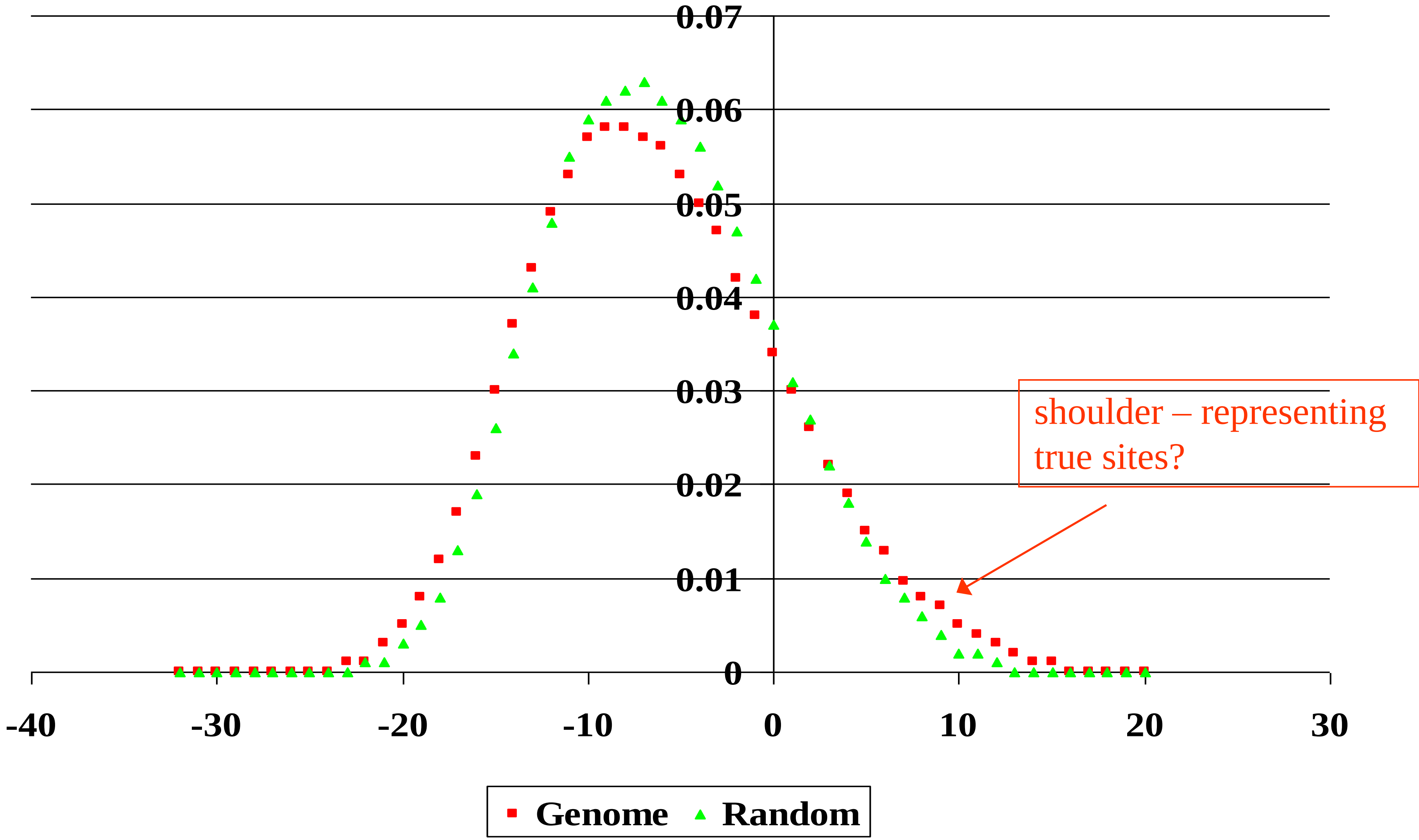
- In the site case,

$$w_{rj} = \log(P_j(r \mid M_{\text{site}})) - \log(P_j(r \mid M_{\text{background}}))$$

Score Distributions (AG sites)– 3’ SS Weight Matrix



Score Distributions (AG sites)– 3’ SS Weight Matrix



shoulder – representing true sites?

Some Issues for Site Weight Matrices (to be discussed later)

- Can derive *theoretical* probability distribution for scores, and compare with above *empirical* distributions
- Small sample correction to frequencies: pseudocounts
- Avoiding *overfitting* (e.g. using too large a window)

Relative Entropy

- The *relative entropy* or *Kullback-Leibler distance* for two dist'ns P and Q on S is

  $$D_b(P \parallel Q) \equiv \Sigma_{s \in S} P(s) \log_b(P(s) \,/\, Q(s))$$

  (the expected value of the loglikelihood ratio).
  - if $P(s) = 0$, set corresponding term = 0
  - if $P(s) \neq 0$ but $Q(s) = 0$, $D_b(P \parallel Q)$ is taken to be $+\infty$.
- By information inequality, $D_b(P \parallel Q) \geq 0$, with equality only if $P = Q$.
- In general

  $$D_b(P \parallel Q) \neq D_b(Q \parallel P)$$

- For site dist’n *P* and background dist’n *Q*,
  - $D(P \| Q)$ = the *mean* of site score distribution

  i.e. the sum, over sequences, of prob of seq times its LLR weight.

- Since $P(s) = \prod_{1 \le i \le n} P_i(s_i)$ and $Q(s) = \prod_{1 \le i \le n} Q_i(s_i)$,

$$D(P \| Q) = \sum_{s \in S} (\prod_{1 \le i \le n} P_i(s_i)) \sum_{1 \le j \le n} (\log(P_j(s_j)) - \log(Q_j(s_j)))$$

which simplifies to

$$\sum_{1 \le i \le n} (\sum_{r \in A} P_i(r)(\log(P_i(r)) - \log(Q_i(r)))) \quad = \sum_{1 \le i \le n} D(P_i \| Q_i)$$

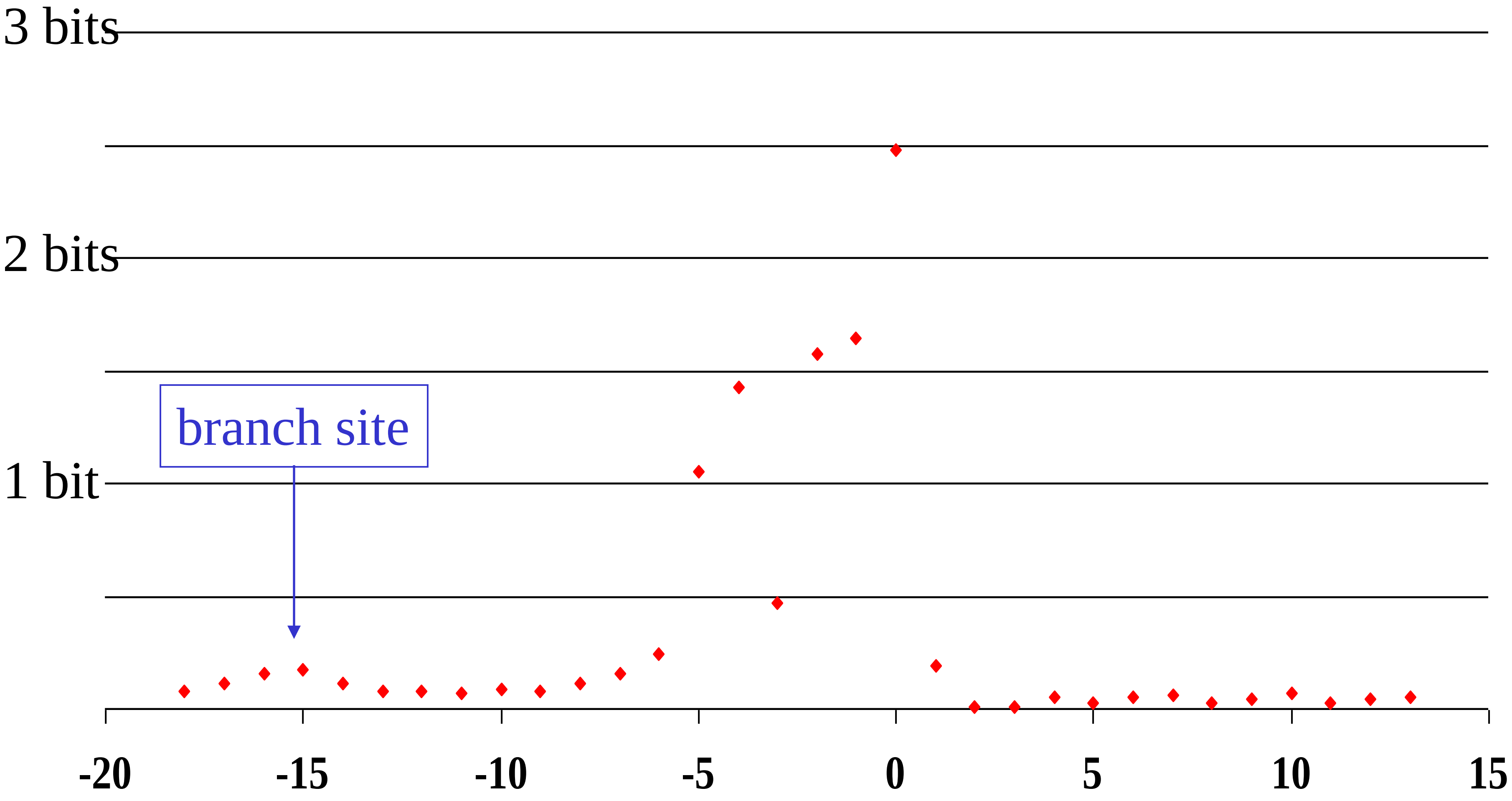
Weight Matrix – 3’ Splice Sites (*C. elegans*)

```
SITE FREQUENCIES:
A  0.400  0.429  0.282  0.058  0.008  0.092  0.029  1.000  0.000  0.410  0.293  0.307
C  0.118  0.079  0.081  0.029  0.016  0.135  0.834  0.000  0.000  0.156  0.187  0.225
G  0.072  0.070  0.063  0.018  0.005  0.073  0.001  0.000  1.000  0.310  0.159  0.191
T  0.409  0.422  0.574  0.896  0.971  0.700  0.135  0.000  0.000  0.124  0.361  0.276

BACKGROUND FREQUENCIES:
A  0.321  0.321  0.321  0.321  0.321  0.321  0.321  0.321  0.321  0.321  0.321  0.321
C  0.179  0.179  0.179  0.179  0.179  0.179  0.179  0.179  0.179  0.179  0.179  0.179
G  0.179  0.179  0.179  0.179  0.179  0.179  0.179  0.179  0.179  0.179  0.179  0.179
T  0.321  0.321  0.321  0.321  0.321  0.321  0.321  0.321  0.321  0.321  0.321  0.321



WEIGHTS:
A   0.32   0.42  -0.18  -2.46  -5.29  -1.79  -3.45   1.64 -99.00   0.36  -0.13  -0.06
C  -0.60  -1.18  -1.15  -2.64  -3.51  -0.41   2.22 -99.00 -99.00  -0.20   0.06   0.33
G  -1.31  -1.35  -1.51  -3.35  -5.23  -1.30  -6.93 -99.00   2.48   0.79  -0.17   0.10
T   0.35   0.39   0.84   1.48   1.60   1.12  -1.24 -99.00 -99.00  -1.37   0.17  -0.22
```

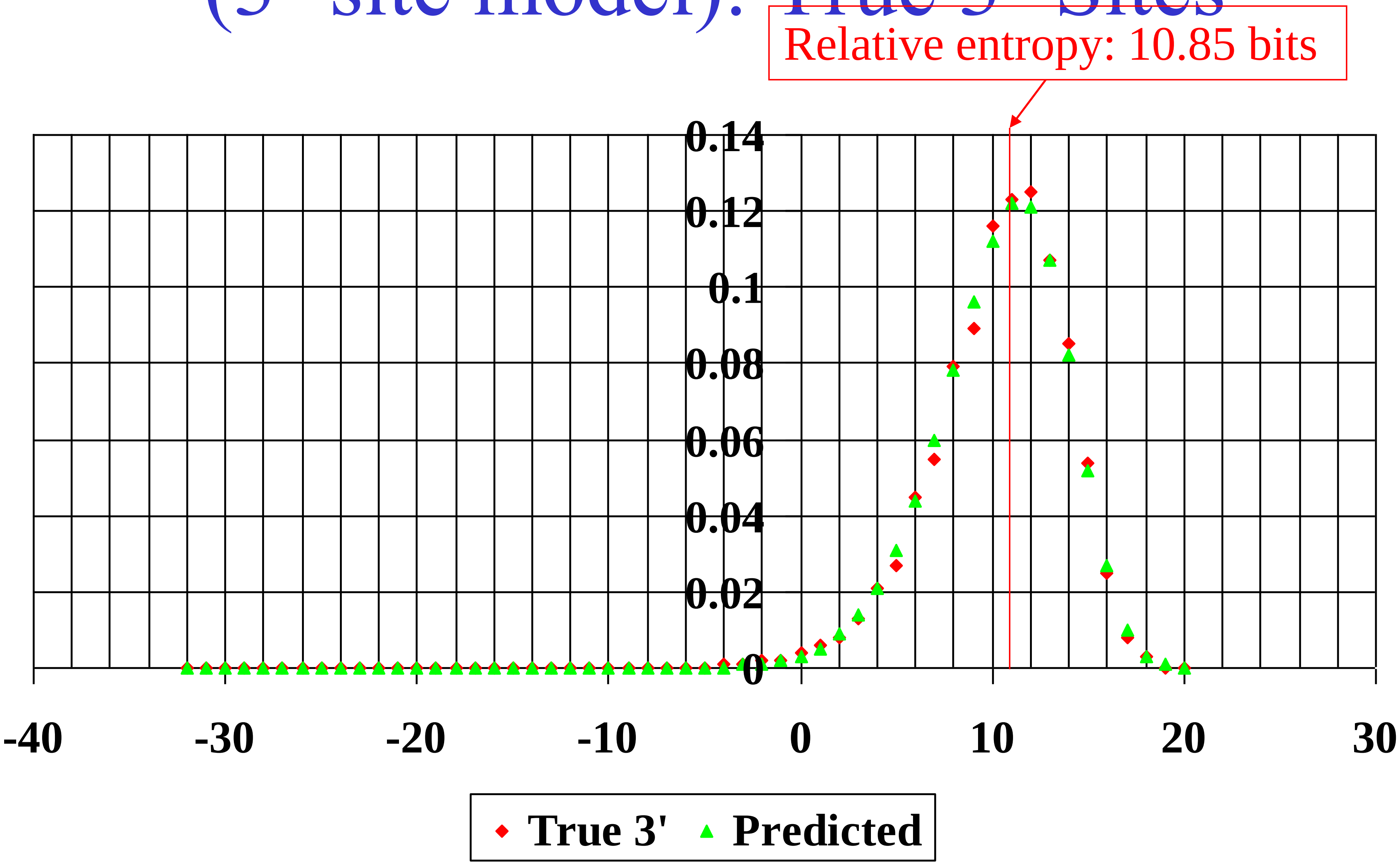
3’ Splice Sites

```
WEIGHTS:
A    0.32    0.42   -0.18   -2.46   -5.29   -1.79   -3.45    1.64 -99.00    0.36   -0.13   -0.06
C   -0.60   -1.18   -1.15   -2.64   -3.51   -0.41    2.22 -99.00 -99.00   -0.20    0.06    0.33
G   -1.31   -1.35   -1.51   -3.35   -5.23   -1.30   -6.93 -99.00    2.48    0.79   -0.17    0.10
T    0.35    0.39    0.84    1.48    1.60    1.12   -1.24 -99.00 -99.00   -1.37    0.17   -0.22

Position-specific Relative Entropy:
     0.11    0.16    0.24    1.05    1.43    0.47    1.57    1.64    2.48    0.19    0.01    0.01

e.g. 0.11 = .400 (.32) + .118 (-.60) + .072 (-1.31) + .409 (.35)

Total Relative Entropy (Sum of position-specific values) = 9.35
```

Position-Specific Relative Entropy: 3’ Splice Sites



Predicted vs. Observed Distributions (3’ site model): True 3’ Sites


Relative entropy: 10.85 bits

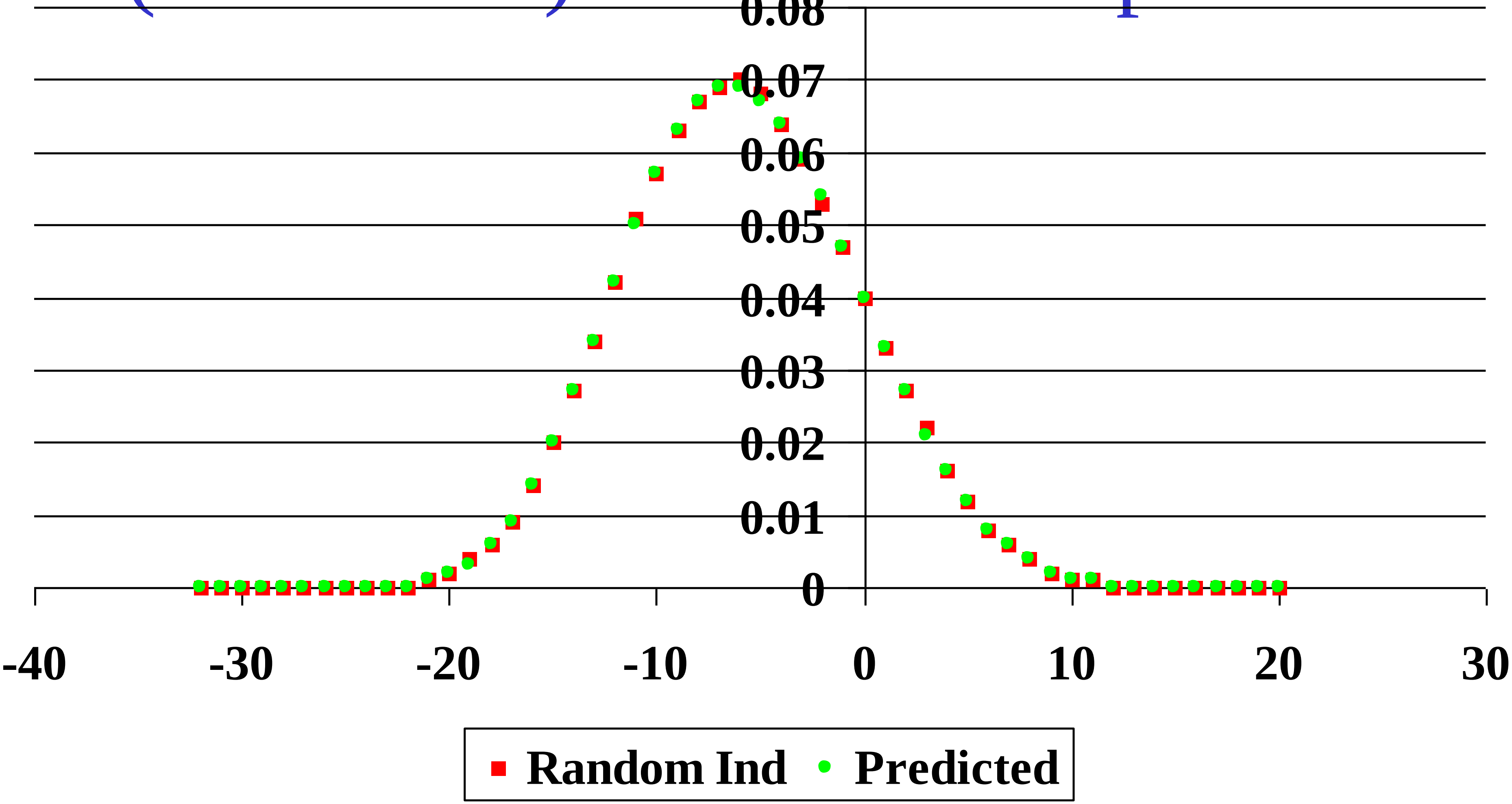
- Similarly,

$$D_b(Q \parallel P) = \Sigma_{s \in S} Q(s) \log_b(Q(s) \mathbin{/} P(s))$$
$$= - \Sigma_{s \in S} Q(s) \log_b(P(s) \mathbin{/} Q(s))$$

= *negative* of the mean of the dist'n of the LLR scores in background sequence (the "null distribution");

  – but must eliminate *s* for which $P(s) = 0$.

Predicted vs. Observed Distributions (3' site model): (Simulated) Random Independent

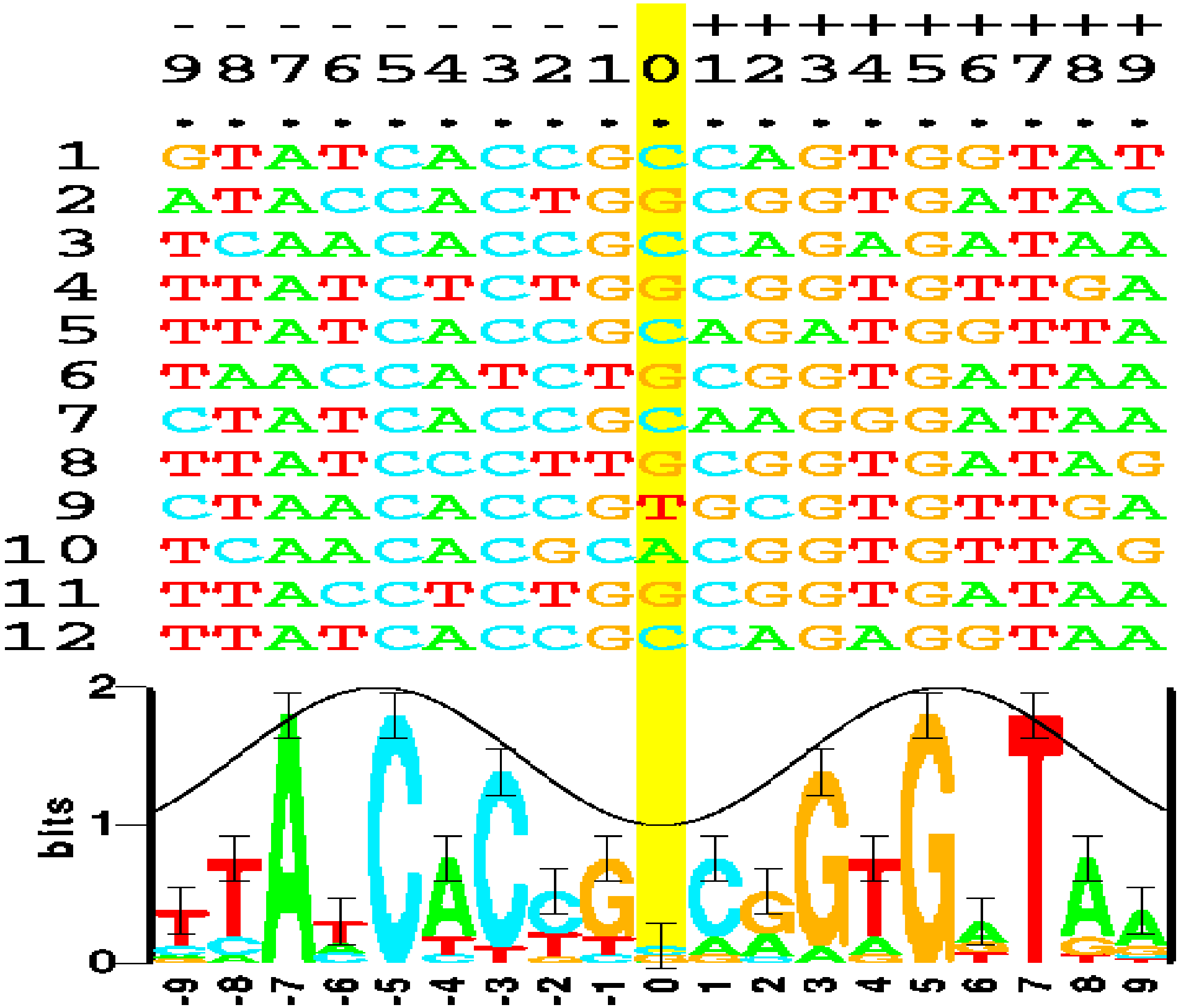


- Note pos-specific relative entropy always ≥ 0
  - = 0 only if site freqs *exactly* equal backgd freqs.
    - will rarely happen, even far from site (when we're in backgd).
- So rel entropy increases indefinitely as window size increases
  - – even when no biological information being added.
- For large enough window get spuriously clean score separation between training seqs and other seqs
  - – *overfitting*.

Sequence Logos

- Schneider and Stephens (NAR 18, 6097-6100, 1990)– see http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html
- At i^{th} position, each residue r gets height
  $P_i(r) D(P_i \,||\, Q_i)$
- Schneider
  - takes Q_i to be the equal-frequency model
  - subtracts small-sample correction from $D(P_i \,||\, Q_i)$
- Gorodkin, Heyer, Brunak and Stormo (CABIO 13, 583-586, 1997)
  - use unequal frequency Q_i
  - allow for gaps
  - take height either proportional to $P_i(r)$ (as above) or to $P_i(r)/\,Q_i(r)$, letter upside down if $P_i(r) < Q_i(r)$.

From http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html



```
   --------- +++++++++
   9876543210123456789
   ...................
 1 GTATCACCGCCAGTGGTAT
 2 ATACCACTGGCGGTGATAC
 3 TCAACACCGCCAGAGATAA
 4 TTATCTCTGGCGGTGTTGA
 5 TTATCACCGCAGATGGTTA
 6 TAACCATCTGCGGTGATAA
 7 CTATCACCGCAAGGGATAA
 8 TTATCCCTTGCGGTGATAG
 9 CTAACACCGTGCGTGTTGA
10 TCAACACGCACGGTGTTAG
11 TTACCTCTGGCGGTGATAA
12 TTATCACCGCCAGAGGTAA
```

12 Lambda cI and cro binding sites

Fig. 1. Some aligned sequences and their sequence logo. At the top of the figure are listed the 12 DNA sequences from the P_L and P_R control regions in bacteriophage lambda. These are bound by both the cI and cro proteins [16]. Each even numbered sequence is the complement of the preceding odd numbered sequence. The sequence logo, described in detail in the text, is at the bottom of the figure. The cosine wave is positioned to indicate that a minor groove faces the center of each symmetrical protein. Data which support this assignment are given in reference [17].