7.91 / 7.36 / BE.490

Lecture #2

Feb. 26, 2004

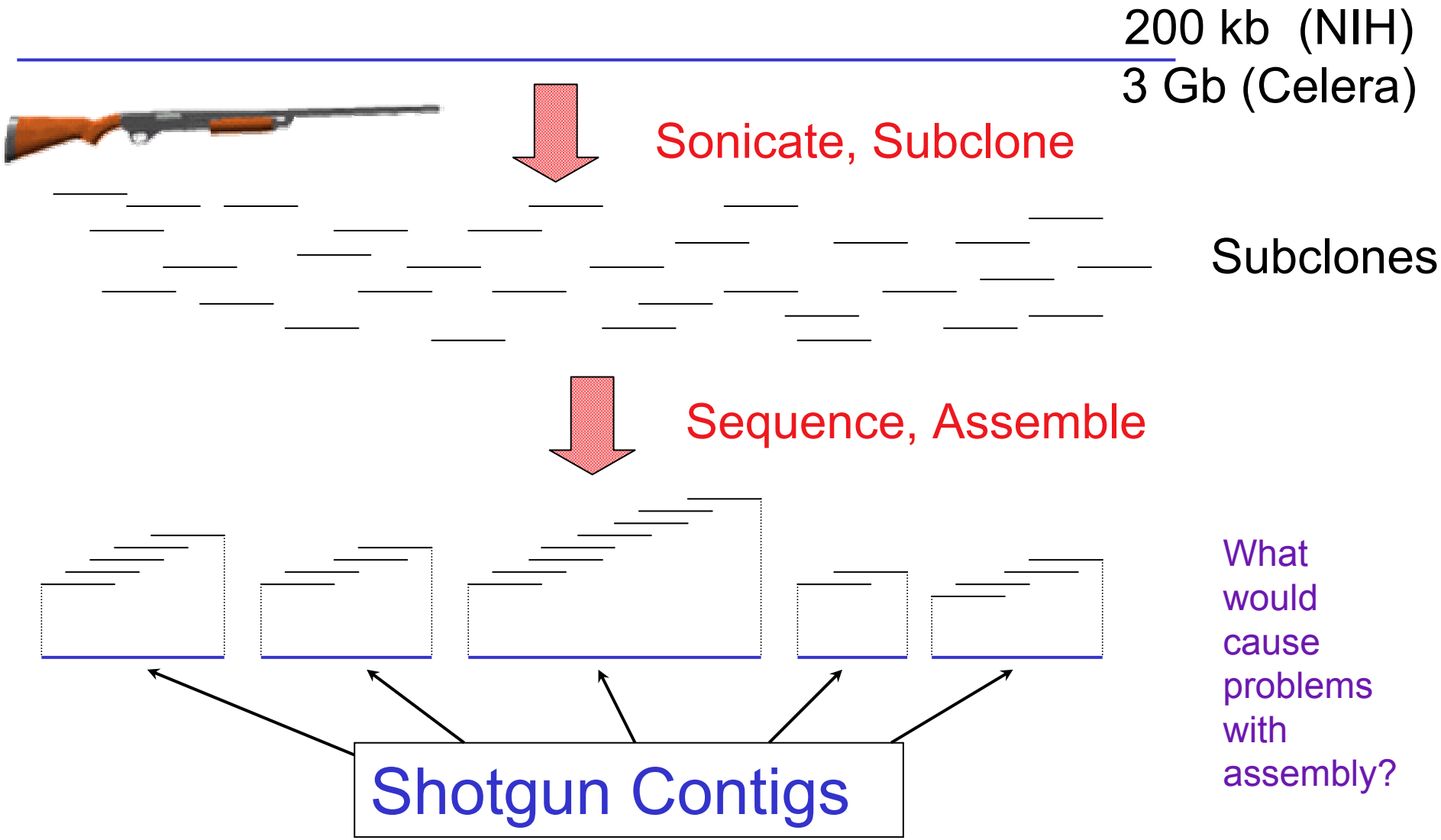
DNA Sequence Comparison
&
Alignment

Chris Burge

Review of Lecture 1: “Genome Sequencing & DNA Sequence Analysis”

- The Language of Genomics
  - cDNAs, ESTs, BACs, Alus, etc.
- Dideoxy Method / Shotgun Sequencing
  - The ‘shotgun coverage equation’ (Poisson)
- Flavors of BLAST
  - BLAST[PNX], TBLAST[NX]
- Statistics of High Scoring Segments

Shotgun Sequencing a BAC or a Genome



DNA Sequence Alignment IV

Which alignments are significant?

```
Q: 1   ttgacctagatgagatgtcgttcacttttactgagctacagaaaa 45
       |||| |||||||||||| | ||||||||||||||||||||||||||
S: 403 ttgatctagatgagatgccattcacttttactgagctacagaaaa 447
```

Identify high scoring segments whose score S exceeds a cutoff x using dynamic programming.

Scores follow an extreme value distribution:

$$P(S > x) = 1 - \exp[-Kmn\, e^{-\lambda x}]$$

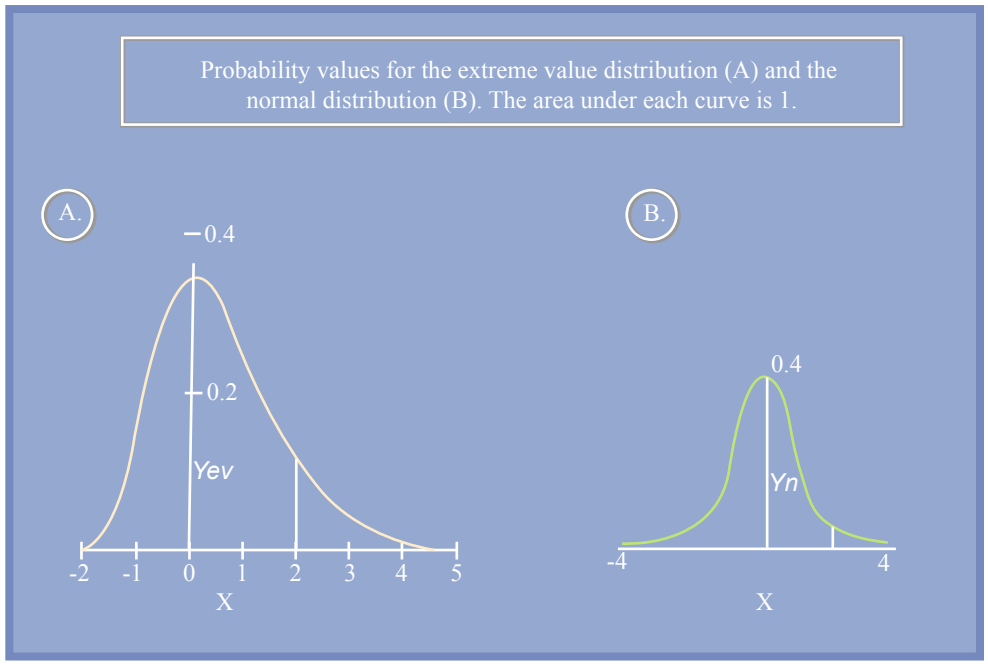
For sequences of length m, n where K, λ depend on the score matrix and the composition of the sequences being compared

(Same theory as for protein sequence alignments)

From M. Yaffe Lecture #2

Notes (cont)

- The random sequence alignment scores would give rise to an "extreme value" distribution – like a skewed gaussian.
- Called Gumbel extreme value distribution



For a normal distribution with a mean m and a variance σ, the height of the curve is described by $Y=1/(\sigma\sqrt{2\pi})\ \exp[-(x-m)^2/2\sigma^2]$

For an extreme value distribution, the height of the curve is described by $Y=\exp[-x-e^{-x}]$ …and $P(S\geq x) = 1-\exp[-e^{-\lambda(x-u)}]$ where $u=(\ln K mn)/\lambda$

Can show that mean extreme score is ~ $\log_2(nm)$, and the probability of getting a score that exceeds some number of "standard deviations" x is: $P(S\geq x)\sim K mn e^{-\lambda x}$. *$K$ and λ are tabulated for different matrices ******

For the less statistically inclined: $E\sim K mn e^{-\lambda S}$

DNA Sequence Comparison & Alignment

- Target frequencies and mismatch penalties
- Eukaryotic gene structure
- Comparative genomics applications:
  - Pipmaker (2 species comparison)
  - Phylogenetic Shadowing (many species)
- Intro to DNA sequence motifs

See Ch. 7 of Mount

DNA Sequence Alignment V

How is λ related to the score matrix?

λ is the unique positive solution to the equation*:

$$\sum_{i,j} p_i p_j e^{\lambda s_{ij}} = 1$$

p_i = frequency of nt i, s_{ij} = score for aligning an i,j pair

What kind of an equation is this?

What would happen to λ if we doubled all the scores?

What does this tell us about the nature of λ?

*Karlin & Altschul, 1990

DNA Sequence Alignment VI

What scoring matrix to use for DNA?

Usually use simple match-mismatch matrices:

$s_{i,j}$:

i j:	A	C	G	T
A	1	m	m	m
C	m	1	m	m
G	m	m	1	m
T	m	m	m	1

m = “mismatch penalty” (must be negative)

DNA Sequence Alignment VII

How to choose the mismatch penalty?

Use theory of High Scoring Segment composition*

High scoring alignments will have composition:

$$q_{ij} = p_i p_j e^{\lambda s_{ij}}$$

where q_{ij} = frequency of i,j pairs ("target frequencies")

p_i, p_j = freq of i, j bases in sequences being compared

What would happen to the target frequencies if we doubled all of the scores?

*Karlin & Altschul, 1990

DNA Sequence Alignment VIII

Still figuring out how to choose the mismatch penalty m

Target frequencies: $q_{ij} = p_i p_j e^{\lambda s_{ij}} \Rightarrow$ $\boxed{s_{ij} = \ln(q_{ij} / p_i p_j)/\lambda}$

If you want to find regions with R% identities:

$r = R/100$ $\quad q_{ii} = r/4$ $\quad q_{ij} = (1-r)/12$ (i,j) $\quad$ Set $s_{ii} = 1$

Then $m = s_{ij} = s_{ij}/s_{ii} = \ln(q_{ij} / p_i p_j)/\lambda) / (\ln(q_{ii} / p_i p_i)/\lambda$ $\quad (i \neq j)$

$\Rightarrow$ $\boxed{m = \ln(4(1-r)/3)/\ln(4r)}$

DNA Sequence Alignment IX

The single most useful thing there is to know about mismatch penalties:

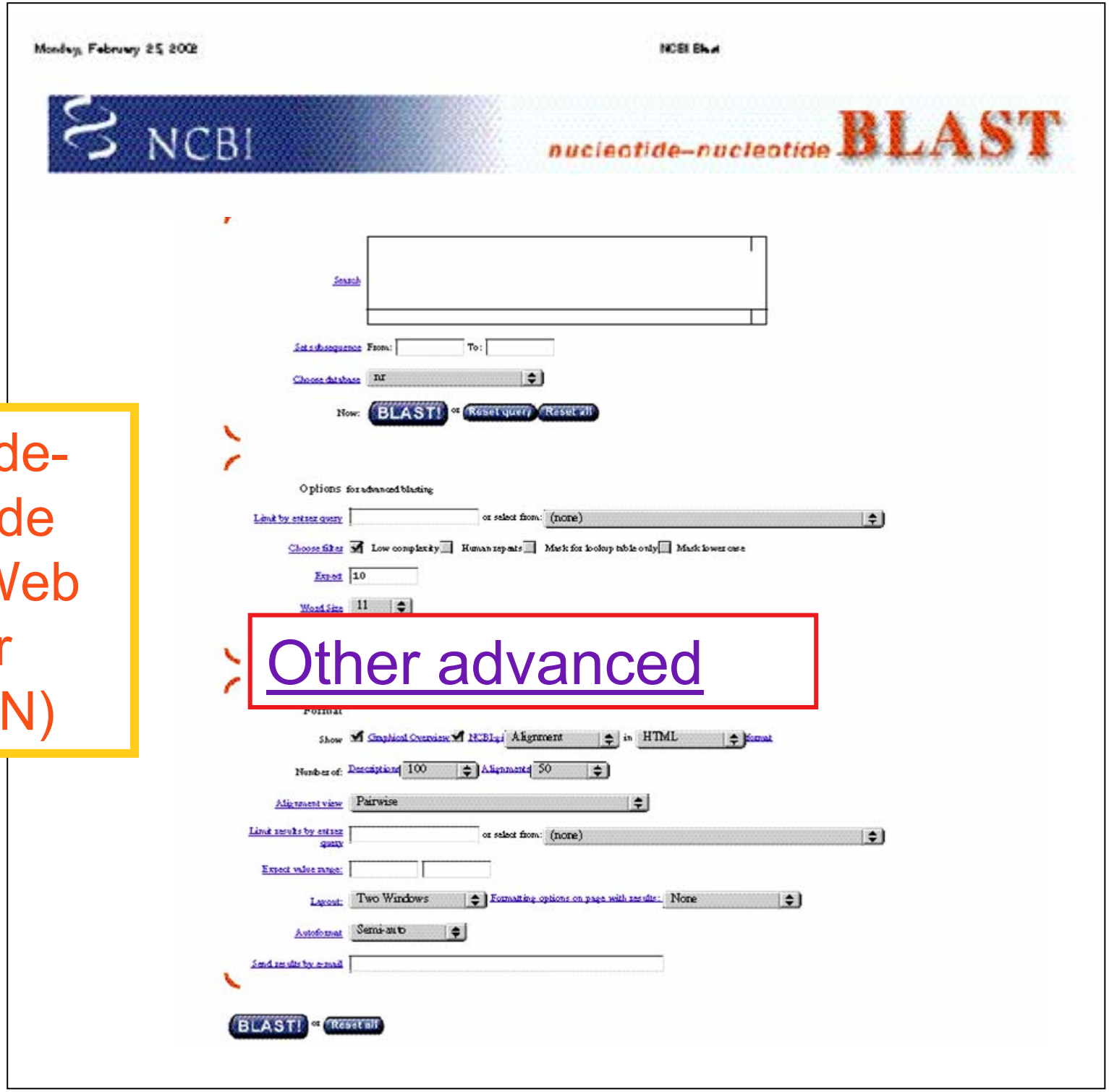
$$m = \ln(4(1-r)/3)/\ln(4r)$$

Examples:

r	0.75	0.95	0.99
m	-1	-2	-3

r = desired fraction of identities in BLAST hits

Nucleotide-nucleotide BLAST Web Server (BLASTN)



DNA Sequence Alignment X

NCBI BLAST Advanced Options

-G Cost to open gap [Integer]
default = 5 for nucleotides 11 proteins

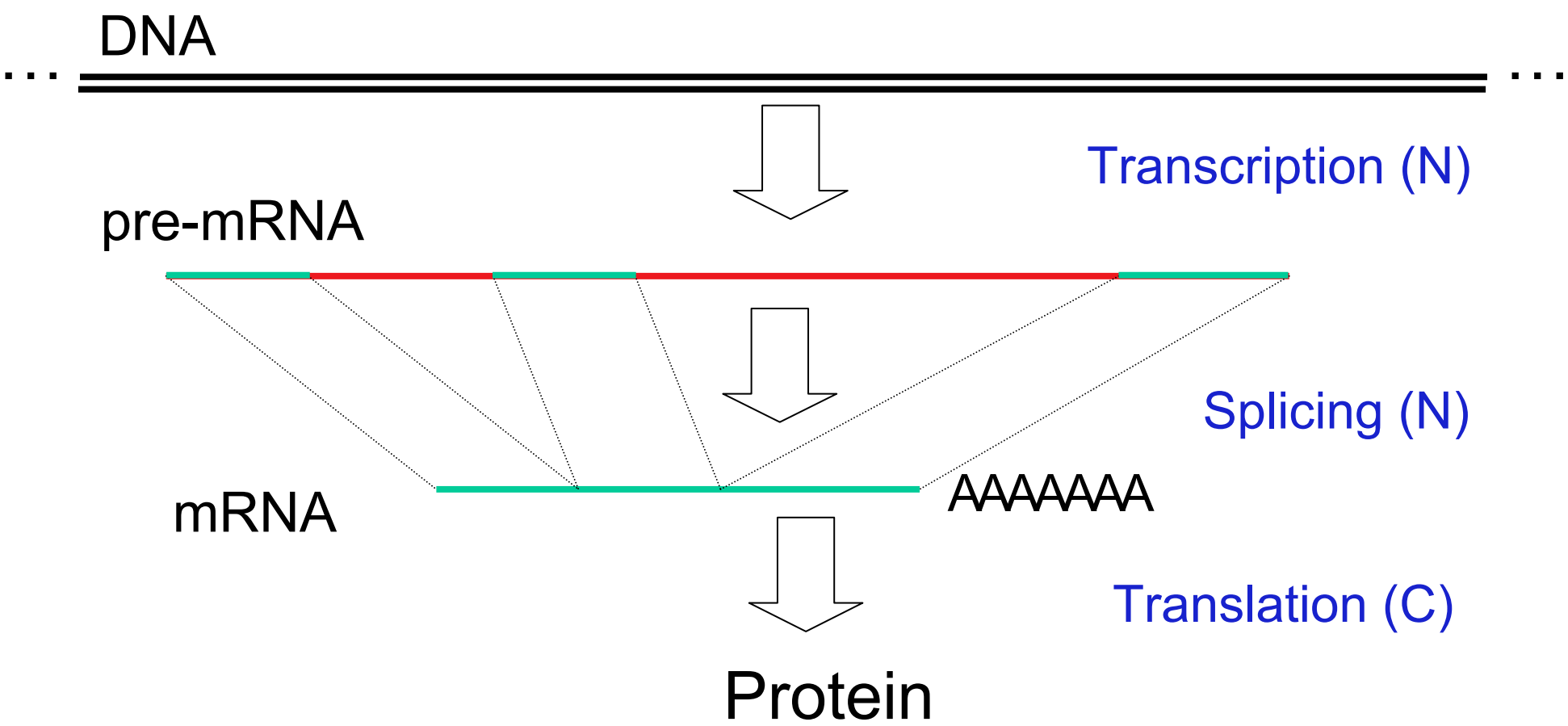
-E Cost to extend gap [Integer]
default = 2 nucleotides 1 proteins

-q Penalty for nucleotide mismatch [Integer]
default = -3

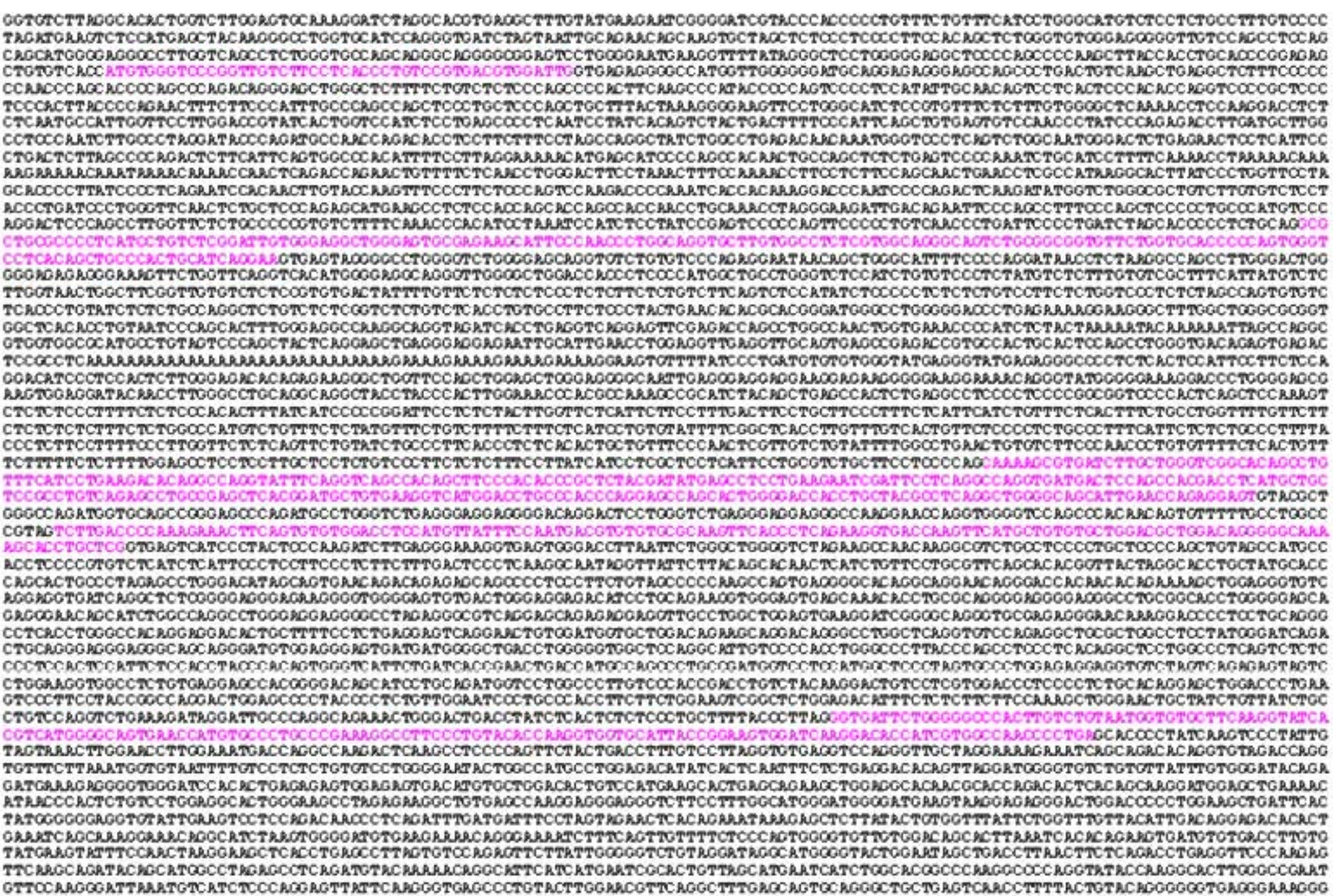
-r Reward for nucleotide match [Integer]
default = 1

-e Expect value [Real]
default = 10

Expression of a Eukaryotic Gene*



Structure of a Human Gene (PSA)



Typical Human Gene Statistics

Length of primary transcript:	~30,000 bp
Number of exons:	~8-10
Mean (internal) exon length:	~150 bp
Mean intron length:	~3,000 bp

Comparative Genomics - Examples

- PipMaker: applications to
  - human/mouse exon finding
  - human/mouse regulatory region finding
- “Phylogenetic Shadowing”: applications to
  - multi-genome exon finding
  - multi-genome regulatory region finding

PipMaker -
Percent Identity Plot (PIP) for two genomic sequences

For an illustration of pips, please see figure 1 of

Schwartz, Scott, Zheng Zhang, Kelly A. Frazer, Arian Smit, Cathy Riemer, John Bouck, Richard Gibbs, Ross Hardison, and Webb Miller. "PipMaker--A Web Server for Aligning Two Genomic DNA Sequences." *Genome Res*. 10 (April 2000): 577-586.

Application of PipMaker #1 - finding human/mouse exons

Please see figure 2 of

Schwartz, Scott, Zheng Zhang, Kelly A. Frazer, Arian Smit, Cathy Riemer, John Bouck, Richard Gibbs, Ross Hardison, and Webb Miller. "PipMaker--A Web Server for Aligning Two Genomic DNA Sequences." *Genome Res*. 10 (April 2000): 577-586.

A Computational Biology Paradigm for Finding Genomic Features of Interest

- Identify properties that feature of interest should have
- Develop an algorithm to find seq's with these properties
- Run algorithm on genome to predict features
- Test a subset of the predicted features experimentally to determine how well the method works

Application of PipMaker #2 - finding regulatory regions

Please see figure 2 of

Loots, GG, RM Locksley, CM Blankespoor, ZE Wang, W Miller, EM Rubin, and KA Frazer. "Identification of A Coordinate Regulator of Interleukins 4, 13, and 5 by Cross-species Sequence Comparisons." *Science* 288, no. 5463 (7 April 2000): 136-40.

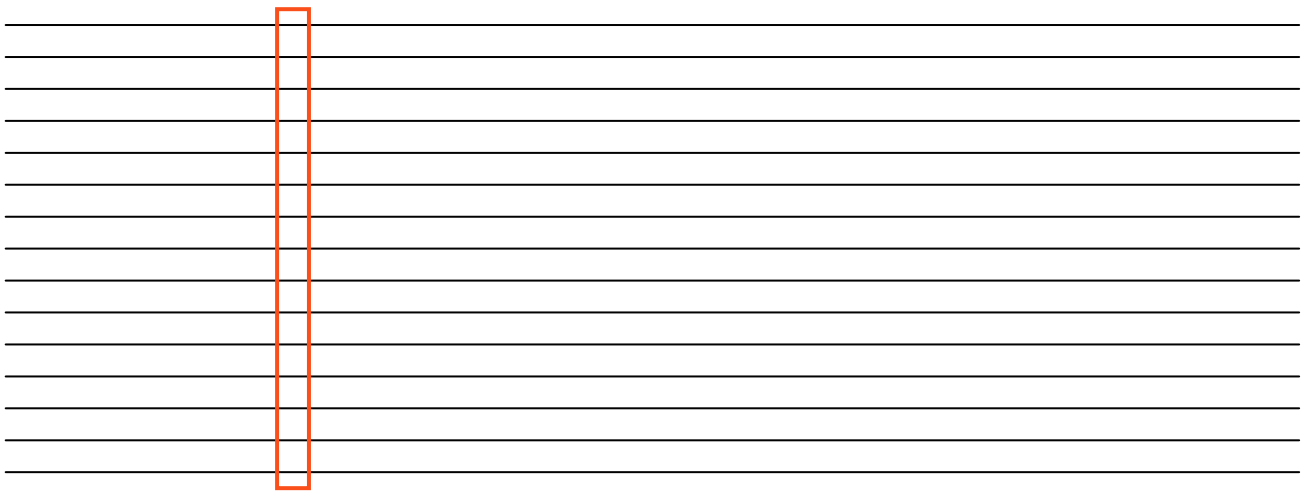
Effects on Transcription of Deleting CNS-1 region

Please see figure 1 of

Loots, GG, RM Locksley, CM Blankespoor, ZE Wang, W Miller, EM Rubin, and KA Frazer. "Identification of A Coordinate Regulator of Interleukins 4, 13, and 5 by Cross-species Sequence Comparisons." *Science* 288, no. 5463 (7 April 2000): 136-40.

“Phylogenetic Shadowing”
(the power of many genomes)

- Sequence orthologous region from 13-17 primates (~90+% identical)
- Do a multiple sequence alignment (MSA):



- Measure variability at each position
- Calculate P(data|fast evol.), P(data|slow evol.)

What are potential advantages of many close species versus fewer more distant organisms?

Please see figure 1 of

Boffelli, D, J McAuliffe, D Ovcharenko, KD Lewis, I Ovcharenko, L Pachter, and EM Rubin. "Phylogenetic Shadowing of Primate Sequences to Find Functional Regions of The Human Genome." *Science* 299, no. 5611 (28 February 2003): 1391-4.

Phylogenetic Shadowing of Regulatory Elements I

Please see figure 2 of

Boffelli, D, J McAuliffe, D Ovcharenko, KD Lewis, I Ovcharenko, L Pachter, and EM Rubin. "Phylogenetic Shadowing of Primate Sequences to Find Functional Regions of The Human Genome." *Science* 299, no. 5611 (28 February 2003): 1391-4.

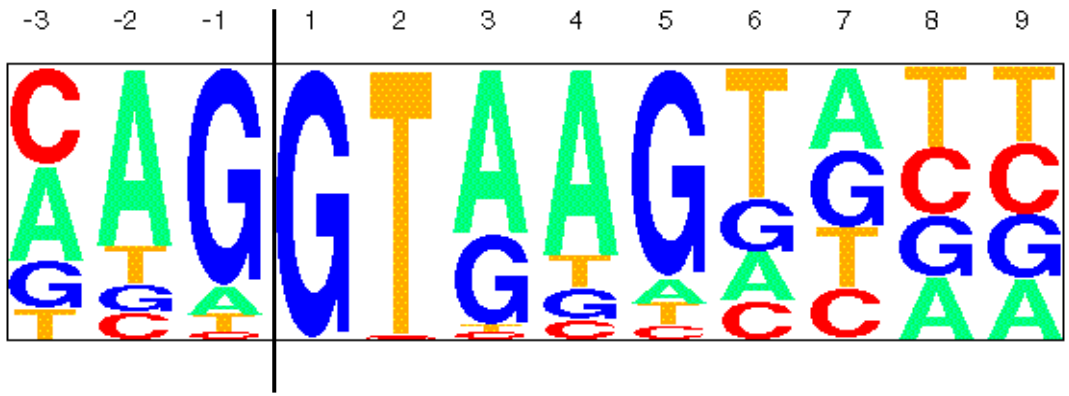
Phylogenetic Shadowing of Regulatory Elements II

Please see figure 3 of

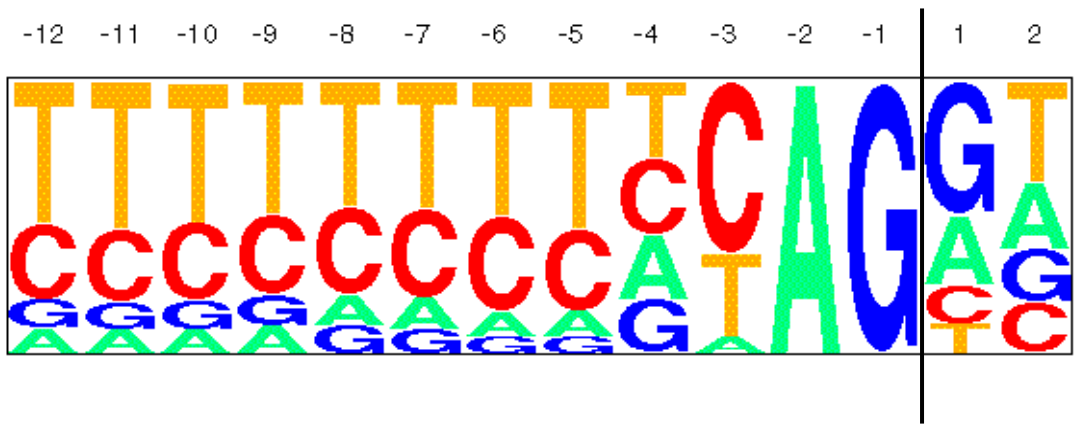
Boffelli, D, J McAuliffe, D Ovcharenko, KD Lewis, I Ovcharenko, L Pachter, and EM Rubin. "Phylogenetic Shadowing of Primate Sequences to Find Functional Regions of The Human Genome." *Science* 299, no. 5611 (28 February 2003): 1391-4.

Human Splice Signal Motif “Pictograms”

5' splice signal



3' splice signal



Binding Affinity of the Dog Transcription Factor

Site	Fraction bound
CAT	1/2
AAT	1/4
GAT	1/4
Others	0

Yeast Genome			
A	1/3	T	1/3
C	1/6	G	1/6

Search yeast promoters for potential Dog binding sites:

How should you prioritize the promoters for followup experiments?

Prioritizing Potential Dog Binding Sites

Site	Fraction bound		Odds Ratio (R)
CAT	1/2	(1/2) / (1/6)(1/3)(1/3)	27.0
AAT	1/4	(1/4) / (1/3)(1/3)(1/3)	6.75
GAT	1/4	(1/4) / (1/6)(1/3)(1/3)	13.5
Others	0	(0) / ( ) ( ) ( )	0.0

Yeast Genome

A 1/3 T 1/3

C 1/6 G 1/6

Neyman-Pearson Lemma:

Optimal decision rules are of the form R > C

(C, a chosen cutoff value)

Therefore: CAT > GAT > AAT > Others

Weight Matrix Model (WMM)

Pos	-3	-2	-1	+1	+2	+3	+4	+5	+6
A	**0.3**	**0.6**	0.1	0.0	0.0	**0.4**	**0.7**	0.1	0.1
C	**0.4**	0.1	0.0	0.0	0.0	0.1	0.1	0.1	0.2
G	0.2	0.2	**0.8**	**1.0**	0.0	**0.4**	0.1	**0.8**	0.2
T	0.1	0.1	0.1	0.0	**1.0**	0.1	0.1	0.0	**0.5**

$$S = S_1\, S_2\, S_3\, S_4\, S_5\, S_6\, S_7\, S_8\, S_9$$

$$P(S|+) = P_{-3}(S_1)P_{-2}(S_2)P_{-1}(S_3) \bullet\bullet\bullet P_5(S_8)P_6(S_9)$$

Inhomogeneous, assumes independence between positions

Statistical Independence

Two events A and B are said to be independent if:

$A \perp B$ if and only if $P(A,B) = P(A)\,P(B)$

In terms of conditional probabilities $P(B|A) = P(A,B)/P(A)$

$A \perp B \Rightarrow P(B|A) = P(B)$ and $P(A|B) = P(A)$

Example: E = { die roll an even number } (2,4,6)

R = { die roll a prime number } (2,3,5)

Are the events E and R independent?

P(A,B) indicates probability that both A and B occur

Weight Matrix Models II

5′ splice signal

Con:	C	A	G	…	G	T
Pos	-3	-2	-1	…	+5	+6
A	**0.3**	**0.6**	0.1	…	0.1	0.1
C	**0.4**	0.1	0.0	…	0.1	0.2
G	0.2	0.2	**0.8**	**…**	**0.8**	0.2
T	0.1	0.1	0.1	…	0.0	**0.5**

Background

Pos	Generic
A	0.25
C	0.25
G	0.25
T	0.25

$$S = S_1\, S_2\, S_3\, S_4\, S_5\, S_6\, S_7\, S_8\, S_9$$

$$\text{Odds Ratio: } R = \frac{P(S|+)}{P(S|-)} = \frac{P_{-3}(S_1)P_{-2}(S_2)P_{-1}(S_3) \bullet\bullet\bullet P_5(S_8)P_6(S_9)}{P_{bg}(S_1)P_{bg}(S_2)P_{bg}(S_3) \bullet\bullet\bullet P_{bg}(S_8)P_{bg}(S_9)}$$

Background model homogenous, assumes independence

Weight Matrix Models III

$$S = S_1\, S_2\, S_3\, S_4\, S_5\, S_6\, S_7\, S_8\, S_9$$

$$\text{Odds Ratio: } R = \frac{P(S|+)}{P(S|-)} = \frac{P_{-3}(S_1)P_{-2}(S_2)P_{-1}(S_3)\,\bullet\bullet\bullet\, P_5(S_8)P_6(S_9)}{P_{bg}(S_1)P_{bg}(S_2)P_{bg}(S_3)\,\bullet\bullet\bullet\, P_{bg}(S_8)P_{bg}(S_9)}$$

$$= \prod_{k=1}^{k=9} P_{-4+k}(S_k)/\, P_{bg}(S_k)$$

$$\text{Score } \quad s = \log_2 R \;= \sum_{k=1}^{k=9} \log_2 \left(P_{-4+k}(S_k)/\, P_{bg}(S_k)\right)$$

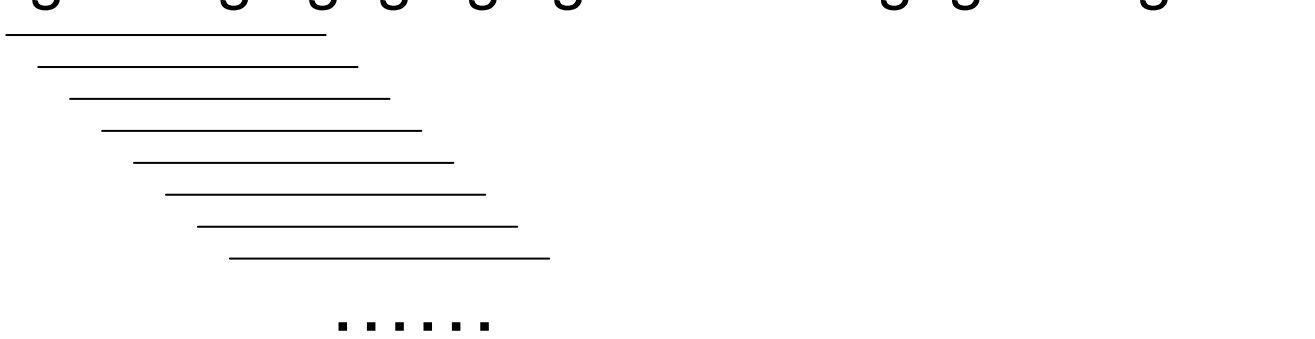
Neyman-Pearson Lemma:

Optimal decision rules are of the form $R > C$

Equiv.: $\log_2(R) > C'$ because log is a monotone function

Weight Matrix Models IV

Slide WMM along sequence:

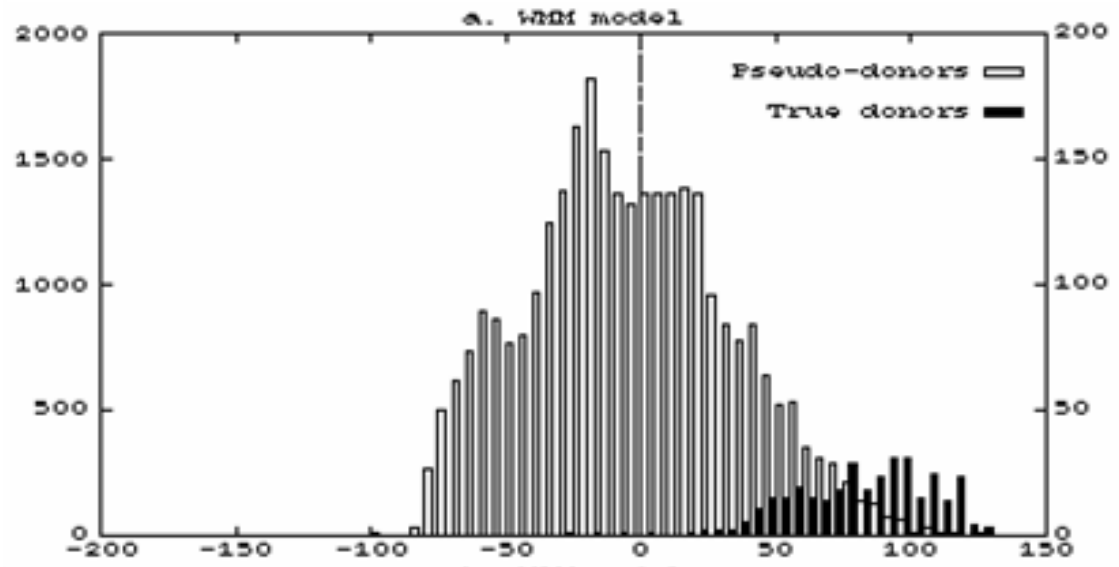


Assign score to each 9 base window.

Use score cutoff to predict potential 5′ splice sites

Histogram of 5′ss Scores

"Decoy" 5′ Splice Sites



True 5′ Splice Sites

Score (1/10 bit units)

Measuring Accuracy:
Sensitivity = % of true sites w/ score > cutoff
Specificity = % of sites w/ score > cutoff that are true sites

Sn:	20%	50%	90%
Sp:	50%	32%	7%

What does this result tell us?