

16 – 18 April

Readings and Synopsis

- Chapter 8 in text

Gene Modeling – How do you find the genes in raw DNA sequence?

Statistics: Part I

Slide 14, lecture Ap4C

Terminology

- **Positive/Negative** - label produced by a method, e.g. in a dotplot the points are positive, non-points are negative, in the example on the previous page, the boxed region is positive, everything else is a negative
- **True/False** - whether the labeling is correct. In a dotplot the **TRUE** points are the ones that are **CORRECTLY** assigned as positive (homologous) or **CORRECTLY** assigned as negative (not homologous). True or false depends on the **THRESHOLD**
- **Four possibilities: True positive, True Negative, False Positive, False Negative (P⁺, N⁺, P⁻, N⁻)**

		Correct Classification	
		Positive	Negative
Method Classification	Positive	P ⁺ True Positive	N ⁻ False Positive P ⁻
	Negative	P ⁻ False Negative N ⁻	N ⁺ True Negative

Gene Modeling

Basic Approaches

- **Gene modeling begins with an uncharacterized genomic sequence and predicts the transcriptional and translational products of each gene, including**
 - Gene location, direction, and/or frame
 - 5' and 3' untranslated regions
 - Introns and exons
 - Possibly includes regulatory elements
- **Gene modeling is notoriously difficult, especially in eukaryotes, but it is widely felt that current methods produce largely correct models, i.e. have errors in only 30% or so of eukaryotic genes and 10% of prokaryotic genes.**
 - Most common errors are in 5' end of gene and small exons
 - Difficult to distinguish errors from true genetic variation

Gene Modeling

Basic Approaches

- **Prokaryotic genes are obviously easier**
 - No introns, genes are (usually) a single open reading frame (ORF)
 - Simpler signals
 - Often better DNA sequence
- **Eukaryotic genes are very challenging**
 - Exons/introns may be very small (less than 10 bases)
 - Introns may be very large (greater than 1 Mbase)
 - Signals are poorly known and more complex
 - DNA sequence may be more poorly assembled

Gene Modeling

Basic Approaches

- **Staden identified two *ab initio* approaches in the 1980s**
 - Search by content - How does the fact that a sequence includes a gene affect its usage of the four bases?
 - *Constraints due to the encoded protein, the translation machinery, and the DNA itself*
 - Search by signal - Can you identify the important signals that delineate genes - promoters, terminators, splice sites, poly-A sites
- **New approaches: intergenomic comparison, ESTs, protein sequences**

Gene Modeling

Search by Content

- **Protein coding enforces several constraints on the underlying DNA sequence**
 - Amino acid residues are used unequally in proteins
 - Amino acid residues have unequal numbers of codons
 - Codons are used unequally

Gene Modeling

Search by Content

- **If reading frame 1 encodes a protein, there is an effect on**
 - The amino acid composition in both the coding frame and the two non-coding frames
 - The codon composition in all three frames
 - The frequencies of the four bases in the three positions of the codon (positional base frequency)
 - *asymmetry of bases in "codon" positions*
 - *preferences for certain bases in certain positions*

Gene Modeling

Search by Content

- **Methods**

- Usually measured using a sliding window approach. Window depends on the method but is often 50 - 200. A big window by dotplot standards.
- Differences are small and you have to average over a large window to get a relatively clear signal.
- Small exons are therefore hard to find

Gene Modeling

Search by Content

- The presence of a coding sequence in frame 1 causes a bias in the position specific base composition in all three frames
- This is the basis of the positional base preferences method

TABLE III
BASE COMPOSITIONS FOR FRAMES 1, 2, AND 3^a

Frame	T	C	A	G
1	17.68	21.08	27.67	33.57
2	27.07	23.78	30.97	18.18
3	25.06	25.06	23.96	25.92
Mean	23.27	23.30	27.53	25.89

^a Assuming an average amino acid composition in frame 1 and no codon preference.

Rodger Staden “Finding protein coding regions in genomic sequences”, *Methods in Enzymology* 183, 163-179, 1990.

Gene Modeling

Search by Content

- Uneven positional base frequencies
- Look for unequal use of the four bases in the three positions of the codon. We expect that the usage will be symmetric in non-coding regions but asymmetric in coding regions.
- for each base, count N_{ij} the number of times base i appears in the three positions j of the codon.

$$\text{Expected number} = E_{ij} = (N_{i1} + N_{i2} + N_{i3}) / 3$$

$$\text{Divergence} = D = \sum_{ij} |E_{ij} - N_{ij}|$$

- D is a windowed statistic usually calculated over 50 - 150 codons (150-450 bases)
- Does not predict frame
- Does not require training!

Gene Modeling

Search by Content

- Uneven positional base frequencies
- Distribution of D scores (window=67 codons)

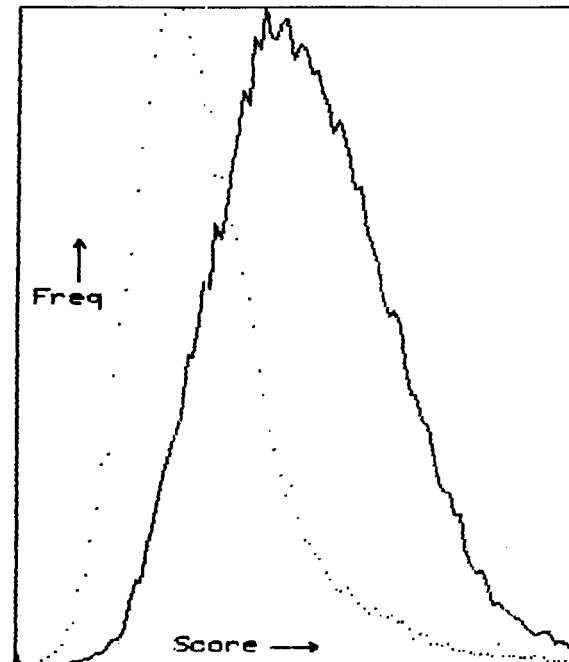


FIG. 1. Histograms of observed D values (see text) for all coding (solid line) and all noncoding (dotted line) sequences in the 1984 EMBL nucleotide library.

Rodger Staden “Finding protein coding regions in genomic sequences”, *Methods in Enzymology* 183, 163-179, 1990.

Gene Modeling

Search by Content

- Uneven positional base frequencies

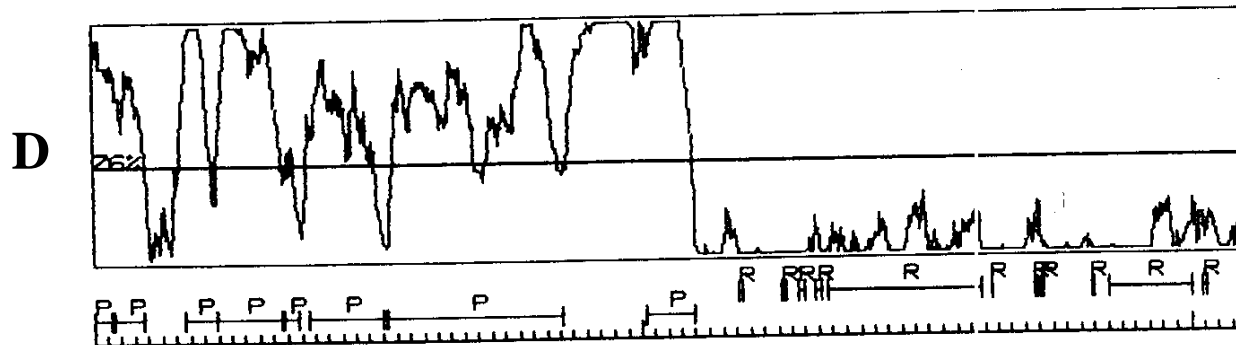


FIG. 2. Application of the uneven positional base frequencies method to bases 100,000 to 121,024 of the liverwort *Marchantia* chloroplast genome. The horizontal scale marks every 100th base, and the bars above indicate the extent of known protein coding segments (P) and known RNA genes (R).

Rodger Staden “Finding protein coding regions in genomic sequences”, *Methods in Enzymology* 183, 163-179, 1990.

Gene Modeling

Search by Content

- **Unequal use of amino acid residues**
- **Assume a protein of average composition (i.e. residue frequencies from the protein sequence database) Staden, Table I**
- **Assume all codons are used equally (so that there is no effect of codon preference) Staden, Table II**
- **Based on this we can calculate the codons that will appear in the other two reading frames, and thus their amino acid compositions. Staden, Table I**
- **Note the large differences in the amino acid compositions, and the fact that the protein sequence alone causes a use of 34% G bases in the first codon position. Staden, Table III**
- **These differences are due only to the encoded amino acid composition**

Gene Modeling

Search by Content

- The presence of a protein of average composition in frame 1 forces the codons in frames 2 and 3 away from the normal protein composition

TABLE I
AMINO ACID COMPOSITIONS FOR FRAMES 1, 2, AND 3^a

	A	C	D	E	F	G	H	I	K	L	
Frame 1	83	17	53	62	39	72	22	52	57	90	
Frame 2	48	27	14	23	27	50	23	50	49	101	
Frame 3	55	37	35	37	29	87	34	35	34	60	
	M	N	P	Q	R	S	T	V	W	Y	*
Frame 1	24	44	51	40	57	69	58	66	13	32	0
Frame 2	25	31	60	36	108	99	76	48	24	25	59
Frame 3	7	32	53	36	129	89	51	46	18	34	65

^a Assuming an average amino acid composition in frame 1 and no codon preference.

* Stop codon.

Rodger Staden “Finding protein coding regions in genomic sequences”, *Methods in Enzymology* 183, 163-179, 1990.

Gene Modeling

Search by Content

- Average amino acid composition in frame 1, equal use of all codons, notice that codons in frame 2 and 3 are very uneven

TABLE II
CODON COMPOSITION FOR FRAMES 1, 2, AND 3^a

	Frame				Frame				Frame				Frame		
	1	2	3		1	2	3		1	2	3		1	2	3
F TTT	20	12	17	S TCT	12	12	15	Y TAT	16	11	19	C TGT	9	12	17
F TTC	20	14	12	S TCC	12	14	13	Y TAC	16	13	15	C TGC	9	15	21
L TTA	15	19	8	S TCA	12	19	16	* TAA	0	18	25	* TGA	0	20	29
L TTG	15	23	8	S TCG	12	23	10	* TAG	0	21	11	W TGG	13	24	18
L CTT	15	11	17	P CCT	13	11	15	H CAT	11	11	19	R CGT	10	11	17
L CTC	15	13	12	P CCC	13	13	13	H CAC	11	13	15	R CGC	10	13	21
L CTA	15	16	8	P CCA	13	16	16	Q CAA	20	16	25	R CGA	10	16	29
L CTG	15	20	8	P CCG	13	20	10	Q CAG	20	20	11	R CGG	10	20	18
I ATT	17	13	17	T ACT	15	13	14	N AAT	22	14	18	S AGT	12	14	16
I ATC	17	16	11	T ACC	15	16	12	N AAC	22	17	14	S AGC	12	17	20
I ATA	17	21	8	T ACA	15	21	15	K AAA	29	22	24	R AGA	10	22	28
M ATG	24	25	7	T ACG	15	25	9	K AAG	29	27	10	R AGG	10	27	17
V GTT	17	8	18	A GCT	21	8	16	D GAT	27	7	20	G GGT	18	9	17
V GTC	17	10	12	A GCC	21	10	13	D GAC	27	8	15	G GGC	18	11	22
V GTA	17	13	8	A GCA	21	13	16	E GAA	31	10	26	G GGA	18	14	30
V GTG	17	16	8	A GCG	21	16	10	E GAG	31	12	11	G GGG	18	17	19

^a Assuming an average amino acid composition in frame 1 and no codon preference.

* Stop codon.

Rodger Staden “Finding protein coding regions in genomic sequences”, *Methods in Enzymology* 183, 163-179, 1990.

Gene Modeling

Search by Content

- Positional base preferences
- How well do the positional base usages agree with the average protein model.
- Define E_{ij} as expected number of base i in position j , but take them for a real coding sequence, i.e. Table III
- count the number of observed bases, O_{if} , in frame f
- calculate the correlation for each choice of frame $C_f = \sum E_{ij} O_{if}$
- About 5% difference between coding from and non-coding frame
- Plot $C_f = \sum C_f$ for each frame

Gene Modeling

Search by Content

- Positional base preferences method on *Marchantia* chloroplast genome

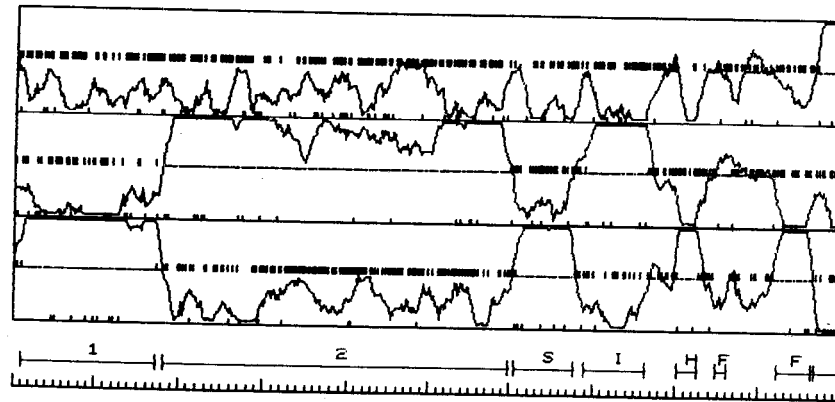


FIG. 3. Application of the positional base preferences method to bases 10,001 to 20,000 of the liverwort *Marchantia* chloroplast genome. The horizontal scale marks every 100th base, and the bars above indicate the extent of known protein coding segments. The three boxes above contain plots of the probability that each of the three reading frames is coding for a protein. The short vertical lines that bisect the mid-height of each box mark the positions of the stop codons in the corresponding reading frames.

Rodger Staden “Finding protein coding regions in genomic sequences”, *Methods in Enzymology* 183, 163-179, 1990.

Gene Modeling

Search by Content

- Position base composition of a highly expressed gene. Note that the skew in the third position is even more extreme than for an average gene with no codon preference.

TABLE V
BASE COMPOSITION FOR *rpoC2* GENE OF
Marchantia CHLOROPLAST

Frame	T	C	A	G
1	24.66	14.20	40.74	20.40
2	32.73	16.80	39.08	11.39
3	44.20	4.61	45.78	5.41
Mean	33.86	11.87	41.86	12.40

Rodger Staden “Finding protein coding regions in genomic sequences”, *Methods in Enzymology* 183, 163-179, 1990.

Gene Modeling

Search by Content

- Codon usage/codon preference
- Organisms do not use the codons for each amino acid equally, this is called *codon preference*. The primary reason appears to be that the pools of isoaccepting tRNAs differ depending on gene number and expression.
- Highly expressed genes tend to use only codons corresponding to the most abundant tRNAs. This effect is stronger in prokaryotes. More weakly expressed genes use closer to equal usage and are therefore harder to detect.
- Rare codons (corresponding to low-level tRNAs) may also be used as a regulatory mechanism.
- The overall usage (number used) of each codon is determined by the codon preference and the amino acid preference
- In eukaryotes, codon usage/preference may be cell, developmental stage, or tissue specific

Gene Modeling

Search by Content

- Codon usage

- By Bayes Rule

$$P(F_i | f_{abc}) = P(f_{abc} | F_i) P(F_i) / \sum P(f_{abc} | F_j) P(F_j)$$

where $P(F_i)$ is the prior probability that frame i is the coding frame

$P(f_{abc} | F_i)$ is the preferred codon usage (tabulated from known genes)

- As usual, this is calculated over a window and the $P(f_{abc} | F_i)$ are the product of the values for the individual codons over the window. In real life one generally uses logs:

$$P(F_1 | f_{abc}) = P(F_1) e^{H_1} / (P(F_1) e^{H_1} + P(F_2) e^{H_2} + P(F_3) e^{H_3})$$

where $H_i = \sum \ln[P(F_i | f_{abc})]$

- This method matches to an expected codon usage so it is important to use an appropriate standard, both in terms of codon preference and amino acid composition.

- This can be difficult.

- Doesn't work well when amino acid composition is unusual

- Similar methods that use codon preference rather than codon usage work for genes with unusual amino acid choice

Gene Modeling

Search by Content

- Codon usage of a highly expressed gene

TABLE IV
CODON COMPOSITIONS FOR *rpoC2* GENE OF *Marchantia*
CHLOROPLAST

F TTT	77	S TCT	38	Y TAT	51	C TGT	9
F TTC	6	S TCC	6	Y TAC	3	C TGC	4
L TTA	98	S TCA	24	* TAA	1	* TGA	0
L TTG	8	S TCG	1	* TAG	0	W TGG	16
L CTT	32	P CCT	16	H CAT	27	R CGT	7
L CTC	0	P CCC	1	H CAC	4	R CGC	0
L CTA	5	P CCA	21	Q CAA	63	R CGA	14
L CTG	1	P CCG	2	Q CAG	2	R CGG	2
I ATT	84	T ACT	40	N AAT	113	S AGT	21
I ATC	5	T ACC	4	N AAC	12	S AGC	2
I ATA	58	T ACA	35	K AAA	154	R AGA	12
M ATG	18	T ACG	4	K AAG	2	R AGG	1
V GTT	29	A GCT	15	D GAT	33	G GGT	21
V GTC	3	A GCC	4	D GAC	3	G GGC	7
V GTA	26	A GCA	19	E GAA	68	G GGA	37
V GTG	4	A GCG	3	E GAG	6	G GGG	5

* Stop codon.

Rodger Staden “Finding protein coding regions in genomic sequences”, *Methods in Enzymology* 183, 163-179, 1990.

Gene Modeling

Search by Content

- **GC content of overall DNA also has a large effect**
- **Liverwort mitochondrial DNA is about 70% AT**
- **Effect is very strong on third codon position, Staden, Table V**
- **Third position changes do not usually change encoded amino acid sequence**

Gene Modeling

Search by Content

- **Hexamer in frame statistic**
- **Four models are used : coding frame 0, coding frame 1, coding frame 2, non-coding. Models are fifth order Markov model, i.e. given the last five bases, what is the next base.**
- **Score is sum of log-odds with appropriate coding frame as observed and non-coding as background model, e.g.,**

$$\text{Score} = \log (f_0 / f_n) + \log(f_1 / f_n) + \log(f_2 / f_n) + \log(f_n / f_n)$$