Biol 478/595 Intro to Bioinformatics

September					
	M 1		Labor Day		
4	W 3	MG	Database Searching		Ch. 6
5	F 5	MG	Database Searching	Hw1	
6	M 8	MG	Scoring Matrices		Ch 3 and Ch 4
7	W 10	MG	Pairwise Alignment		
8	F 12	MG	Pairwise Alignment	Hw2	
9	M 15	MG	Pairwise Alignment		Ch 9
10	W 17	MG	Genome Sequencing		
11	F 19	MG	Gene Finding/Annotation	Hw3	
12	M 22	MG	Gene Finding/Annotation		Ch. 11
13	W 24	MG	Quiz - Sequence Motifs		Ch. 7
14	F 26	MG	Sequence Motifs	No Hw	
X	M 29	Both	Exam		

- No new homework this week to allow time to study for exam (or not)
- Quiz Wednesday
- Midterm next Monday

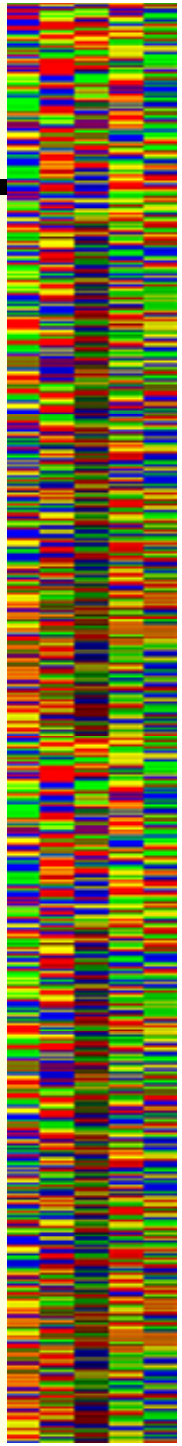
Biol478/595 Intro to Bioinformatics

Quiz – Wednesday 9/22

Covers through 9/15 (lecture 9)

- *database searching*
- *dynamic programming alignment*

- *Four-five mostly short answer questions*
- *Single sheet of notes for reference*
- *closed book, closed notes*
- *no calculators or computers.*

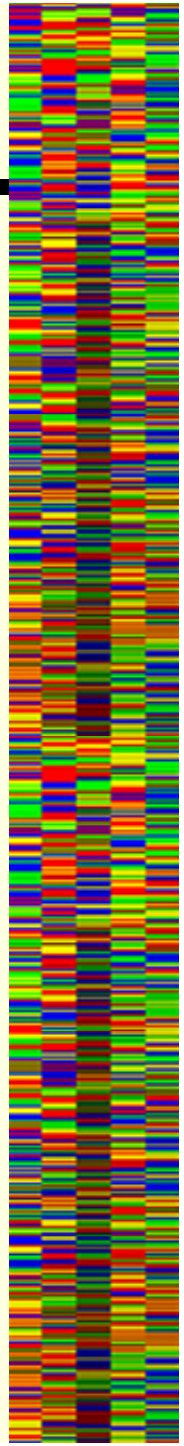


Biol478/595 Intro to Bioinformatics

Midterm exam

Monday September 29

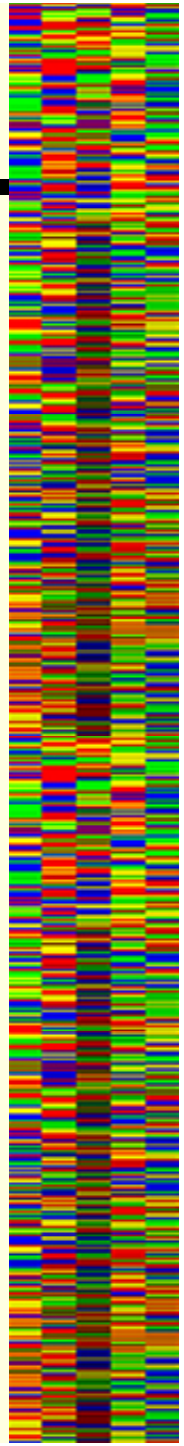
- ***Covers material through Friday 9/26***
- ***Example exams will be posted***
- ***Location TBA***



Genomics - Gene Modeling

Basic Approaches

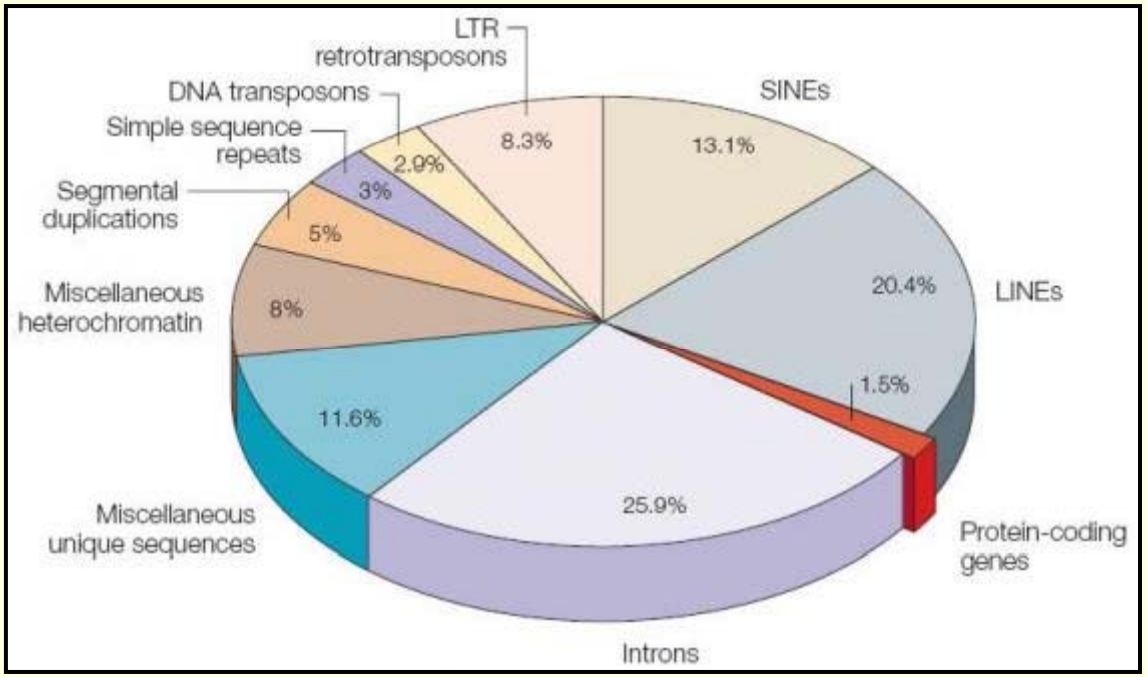
- ***Gene modeling begins with an uncharacterized genomic sequence and predicts the transcriptional and translational products of each gene, including***
  - Gene location, direction, and/or frame
  - 5’ and 3’ untranslated regions
  - Introns and exons
  - Possibly includes regulatory elements
  - Alleles and SNPs in population
- ***Gene modeling is relatively accurate in prokaryotes due to absence of introns***
- ***Gene modeling is notoriously difficult in eukaryotes***
  - Most common errors are in 5’ end of gene and small exons
  - Difficult to distinguish errors from true genetic variation, especially when there are paleoduplications and multiple haplotypes
  - junk DNA makes it hard to find genes



Genomics - Gene Modeling

Junk DNA

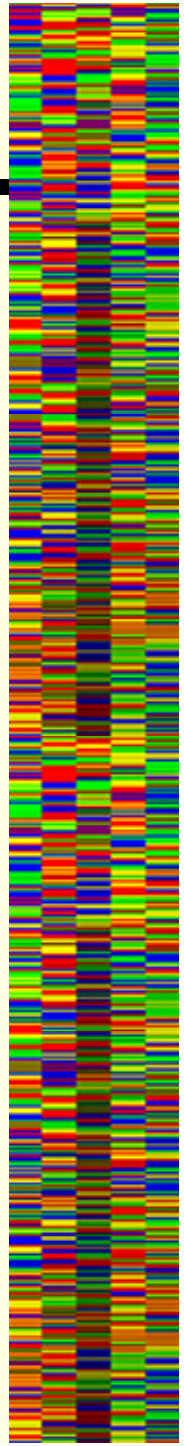
- ***Garbage you throw away***
- ***Junk you keep (but may not have an immediate need or use for)***
- ***Junk or garbage?***



Genomics - Gene Modeling

Basic Approaches

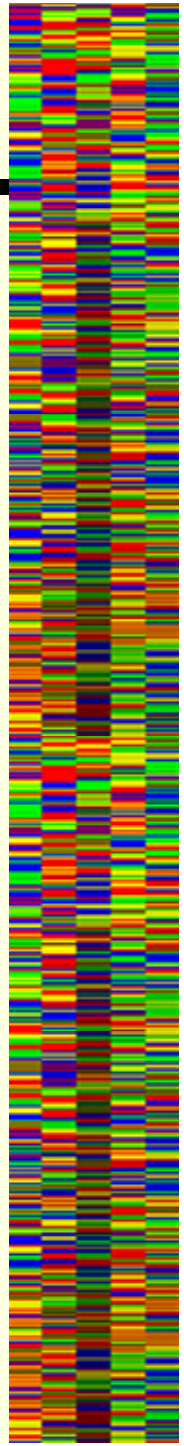
- ***Prokaryotic genes are obviously easier***
  - No introns
  - Simpler signals (promoters, etc)
  - Often better quality DNA sequence
    - higher fold coverage
    - fewer assembly errors
- ***Eukaryotic genes are very challenging***
  - Exons/introns may be very small (less than 10 bases)
  - Introns may be very large (greater than 1 Mbase)
  - Signals are poorly known and more complex
  - DNA sequence may be more poorly assembled
  - Genes may be silenced



Genomics - Gene Modeling

Information that can be used in gene modeling

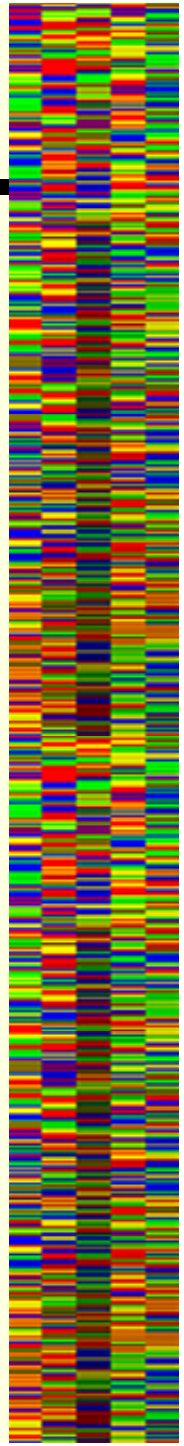
- ***Intrinsic methods - Search by content - How does the fact that a sequence encodes a gene affect its usage of the four bases?***
  - Constraints due to the encoded protein, the translation machinary, and the DNA itself
- ***Search by signal - Can you identify the important signals that delineate genes - promoters, terminators, splice sites, poly-A sites***
  - Constraints due to protein coding or global properties of transcribed regions
  - Signals
    - Promoter
    - Cap
    - Splice junction
    - Terminator
    - Poly-A
- ***Extrinsic methods - Matches to known genes/proteins***
  - Experimental evidence: EST and cDNA sequences
  - Computational: predicted coding and protein Sequences. ~70% of genes match well to other organisms
  - Known protein motifs



Genomics - Gene Modeling

Search by Content

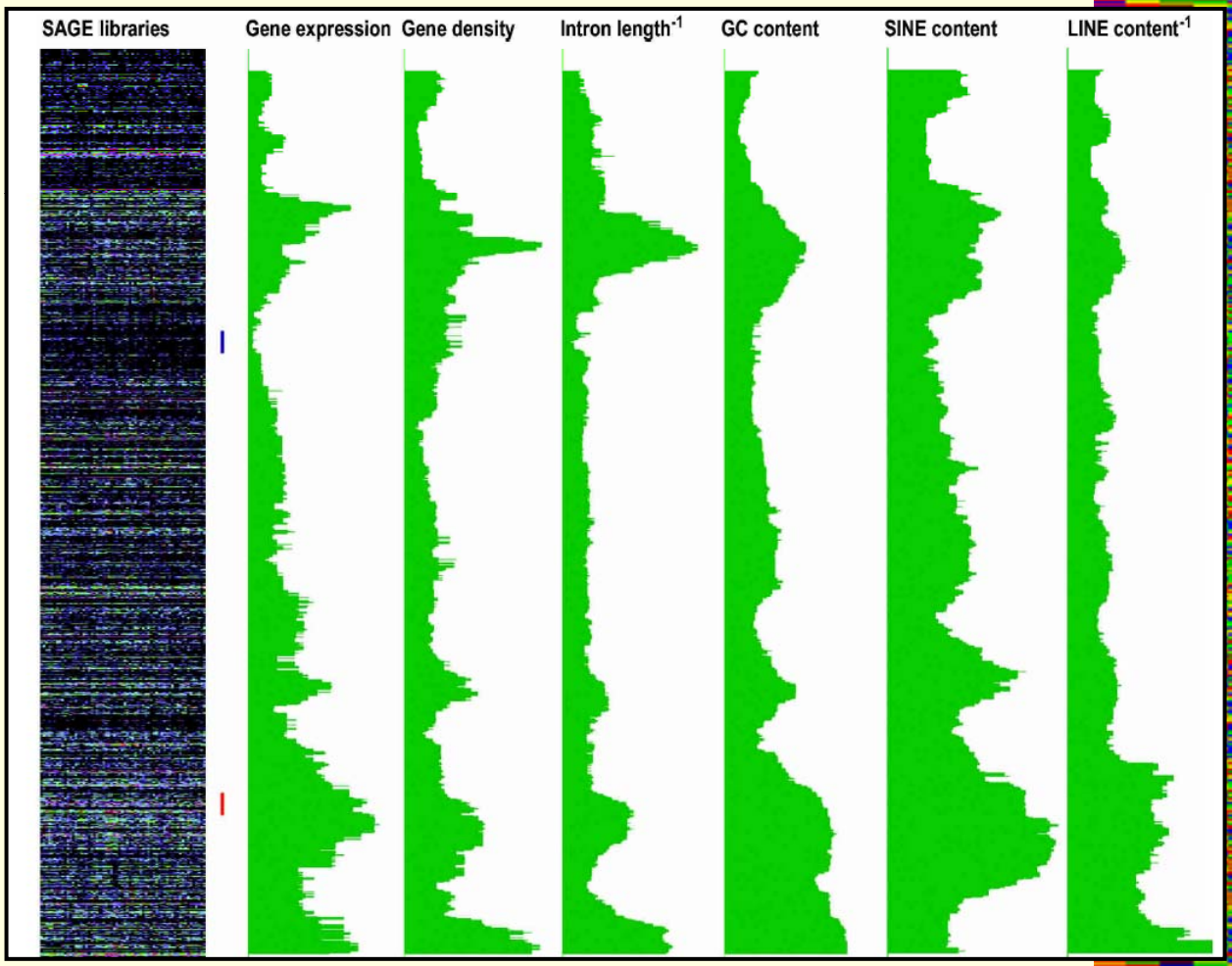
- ***Genes tend to have high GC content***
  - GC rich regions tend to be rich in genes
  - GC poor regions tend to be gene impoverished
    - Repeated elements tend to be common in low GC regions
  - GC content affects gene methylation, a major factor in gene silencing
  - "highly significant spatial autocorrelation of GC content, with most of the structure detectable at a relatively large (300-kb) scale" = isochores
  - high in SINES (e.g., alu elements)
  - low in LINES (contain TTTTA endonuclease site)
- ***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



Genomics - Gene Modeling

Genes are unequally spaced on chromosomes

- ***Gene dense regions***
  - higher expression
  - high GC
  - short intron
  - more SINE
- ***gene sparse regions***
  - lower expression
  - low GC
  - longer introns
  - more LINES



Genomics - Gene Modeling

GC Content

Is high GC content of genes due to gene conversion?

- ***genes undergoing concerted evolution in mammals have high GC content, flanking regions do not***
  - ribosomal operons
  - transfer RNAs
  - histones
- ***GC content is high in recently translocated genes***
  - into mouse PAR gene, GC increased from 50 to 73% in <1 million years,
  - strongly suggeststhat recombination is the cause, not the consequence, of a high GC content.
- ***GC content is high in regions with high recombination***
  - Bird microchromosomes have a high recombination rate, at least one recombination event per generation.
    - very high GC content, probably even higher than mammalian GC-rich regions
    - Genomic analysis of drosophila, mice, yeast and others shows higher GC content in high recombination regions
  - significantly greater linkage disequilibrium (i.e., presumably less recombination) in GC-poor regions

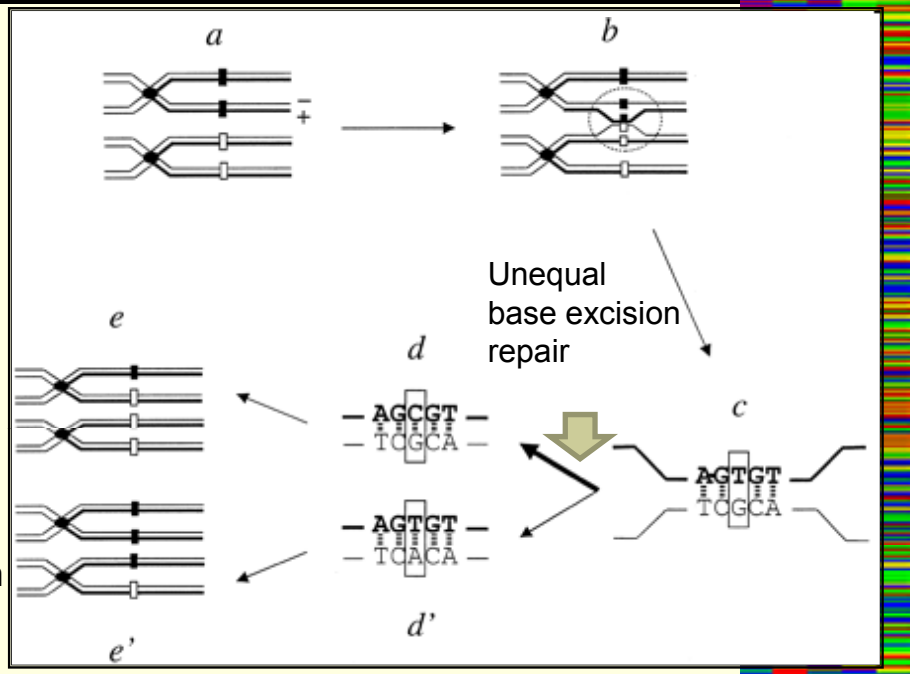


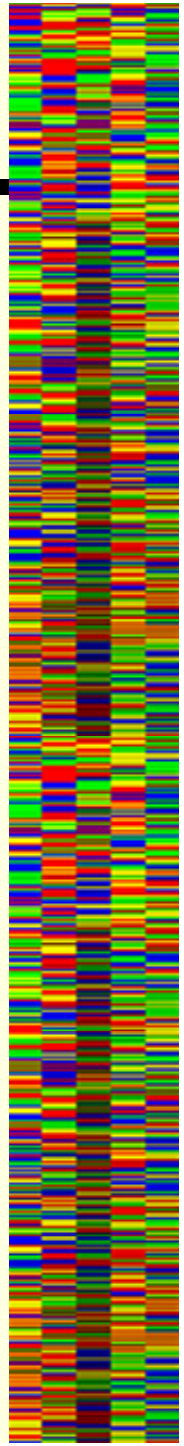
Table 1. GC content of ribosomal and transfer RNA in humans

Gene	Length	GC content (%)
5S rRNA	121	59.5
5' ETS	3656	79.2
18S rRNA	1871	56.1
ITS1	1095	79.6
5.8S rRNA	157	57.3
ITS2	1155	83.1
28S rRNA	5035	69.2
3' ETS	381	81.9
tRNA	65-88	58[a]

Genomics - Gene Modeling

Constraints on DNA

- ***Constraints modify the DNA sequence from random, letting us detect genes***
- ***If reading frame 1 encodes a protein, there is an effect on***
  - The base composition in both the coding frame and the two non-coding frames
  - The codon (trinucleotide) 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



Genomics - Gene Modeling

GC Content

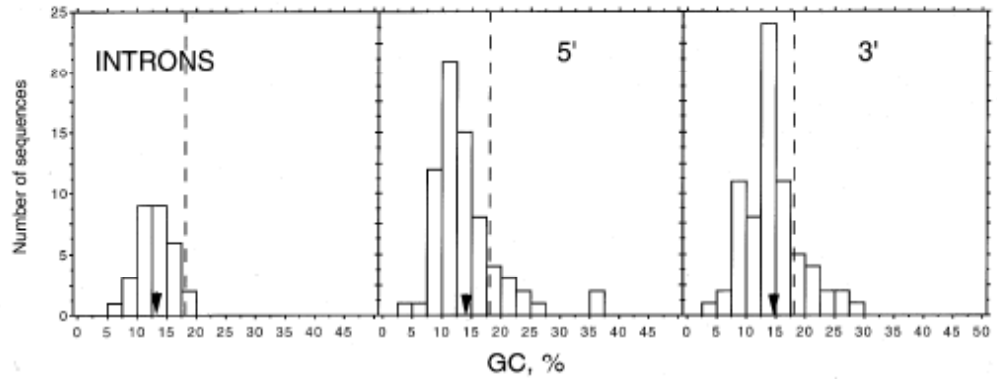


Fig. 2: compositional patterns of introns, 5'- UTR and 3'- UTR. Abscissae, ordinates, arrowheads and broken lines as in Fig. 1.

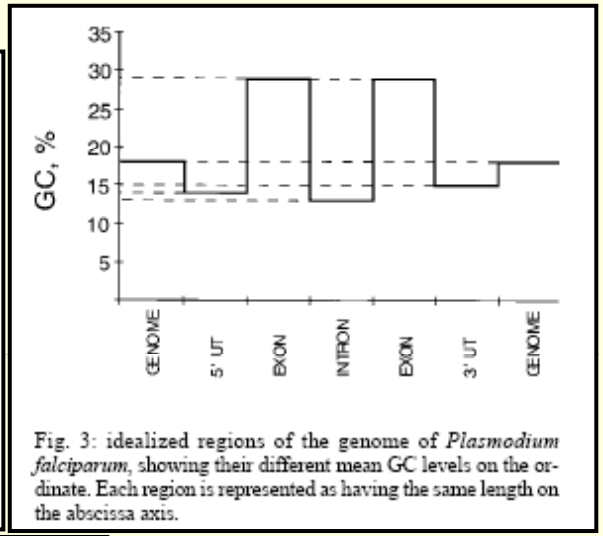


Fig. 3: idealized regions of the genome of *Plasmodium falciparum*, showing their different mean GC levels on the ordinate. Each region is represented as having the same length on the abscissa axis.

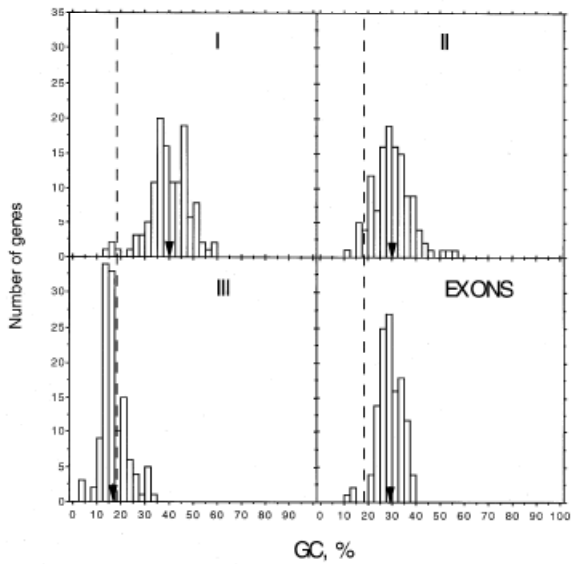
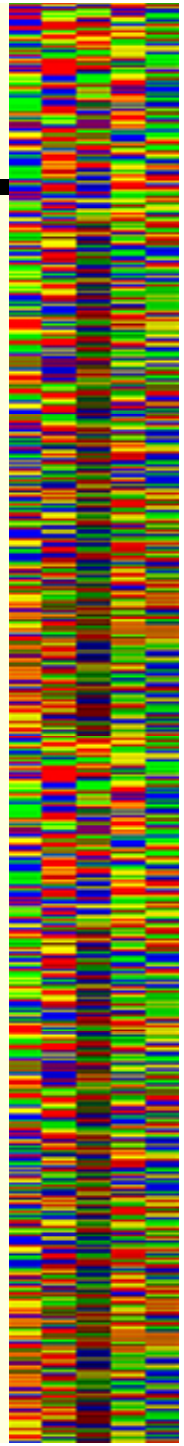


Fig. 1: compositional patterns of the three codon positions (denoted as I, II and III) and exons of *Plasmodium falciparum*. Abscissae and ordinates display the GC% and the number of sequences, respectively. Arrowheads indicate mean GC% values. The broken line corresponds to the genomic GC%.

Genomics - Gene Modeling

Search by Content

- *Methods*
- *Usually measured using a sliding window approach. Window depends on the method but is often 50 – 200 bp. 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*



Genomics – Gene Modeling

Nucleotide Frequencies

TABLE II

Base frequencies (%) in *Plasmodium falciparum*

	Nucleotide frequencies			
	A	C	G	T
I	38.9 (6.1)	11.0 (3.4)	28.8 (7.2)	21.3 (5.9)
II	42.9 (8.0)	17.0 (5.5)	13.2 (3.5)	27.0 (6.7)
III	43.0 (6.0)	8.4 (3.8)	8.6 (3.9)	40.0 (7.5)
tot	41.6 (4.4)	12.1 (3.1)	16.8 (3.1)	29.4 (4.9)
5′	40.3 (10.1)	7.3 (3.1)	6.4 (4.1)	45.9 (10.6)
Int	39.7 (5.3)	5.9 (2.0)	7.2 (1.5)	47.3 (6.4)
3′	45.2 (8.2)	6.7 (2.6)	7.5 (3.1)	40.6 (9.0)

I, II and III are first, second and third codon position, respectively. tot are total values for exons. 5′, Int and 3′ are 5′ UT. introns and 3′ UT, respectively. Standard deviations are given in parentheses.

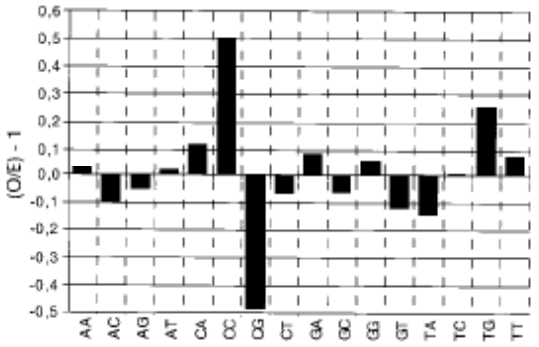


Fig. 4: weight-averaged (Observed/Expected)-1 dinucleotide frequencies for the total sum of exons.

ences in translated regions implies a bias in amino

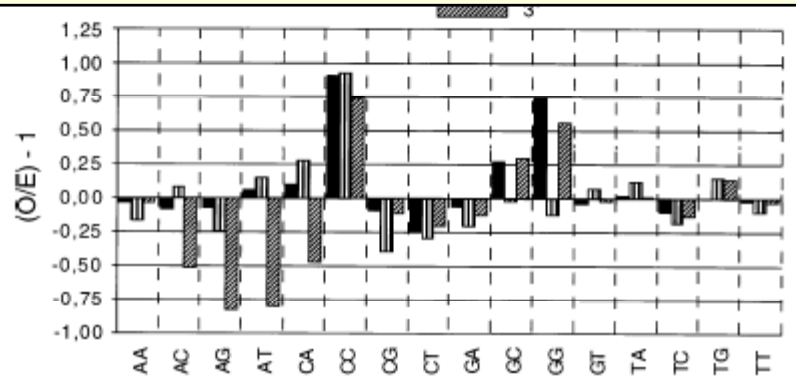


Fig. 5: weight-averaged (Observed/Expected)-1 dinucleotide frequencies for the total sum of 5'- UTR (5'), introns and 3'- UTR (3').

Genomics - Gene Modeling

Search by Content (intrinsic)

- *The presence of a coding sequence in frame 1 causes a bias in the position specific base composition in all three forward, and all three backward reading 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.

Genomics - Gene Modeling

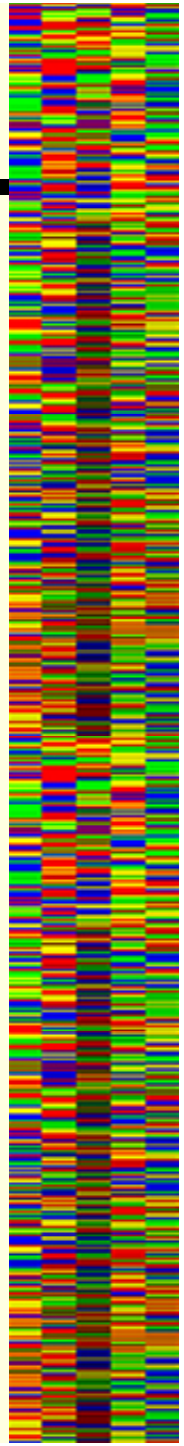
Search by Content

- *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.*

  Expected number = $E_{ij} = (N_{i1} + N_{i2} + N_{i3}) / 3$
  Divergence = $D_{ij} = S \mid E_{ij} - N_{ij} \mid$

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



Genomics - Gene Modeling

Search by Content

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

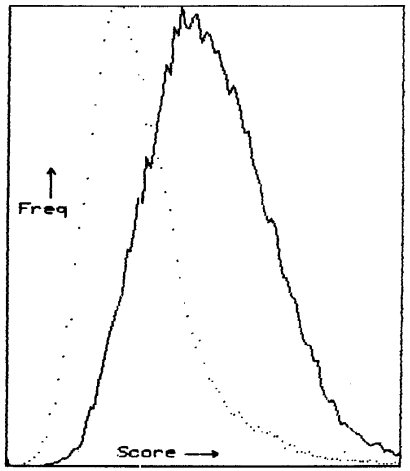


FIG. 1. Histograms of observed *D* values (see text) for all coding (solid line) and all noncoding (dotted line) sequences in the 1984 EMBL nucleotide library.

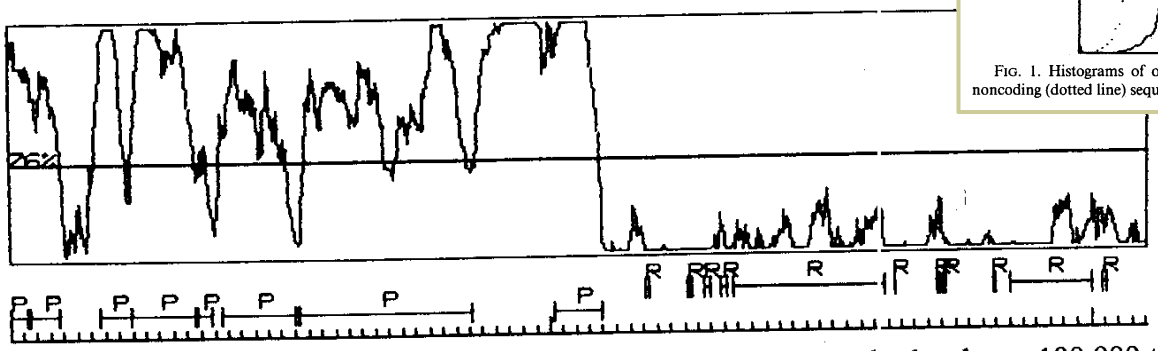


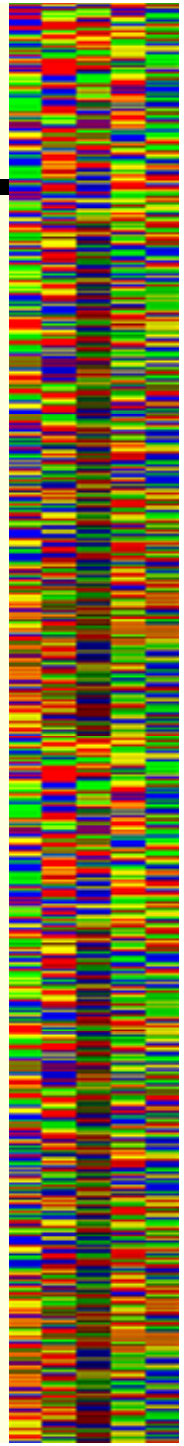
FIG. 2. Application of the uneven positional base frequencies method to bases 100,000 to 121,024 of the liverwort *Marchantia* chloroplast genome. The horizontal scale marks every 100th base, and the bars above indicate the extent of known protein coding segments (P) and known RNA genes (R).

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

Genomics - 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
  - Assume all codons are used equally (so that there is no effect of codon preference)
  - Based on this we can calculate the codons that will appear in the other two reading frames, and thus their amino acid compositions.
  - 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.
  - These differences are due only to the encoded amino acid composition



Genomics - 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.

Genomics - 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

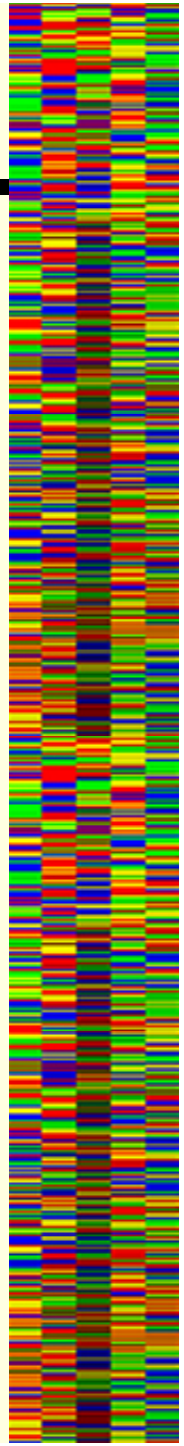
[a] Assuming an average amino acid composition in frame 1 and no codon preference.
* Stop codon.

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

Genomics - 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 $C_f = \Sigma E_{ij} O_{if}$
  - About 5% difference between coding from and non-coding frame
  - Plot $C_f = \Sigma C_f$ for each frame



Genomics - Gene Modeling

Search by Content

- *Positional base preferences method on Marchantia chloroplast genome*
- *Tells location AND coding frame*

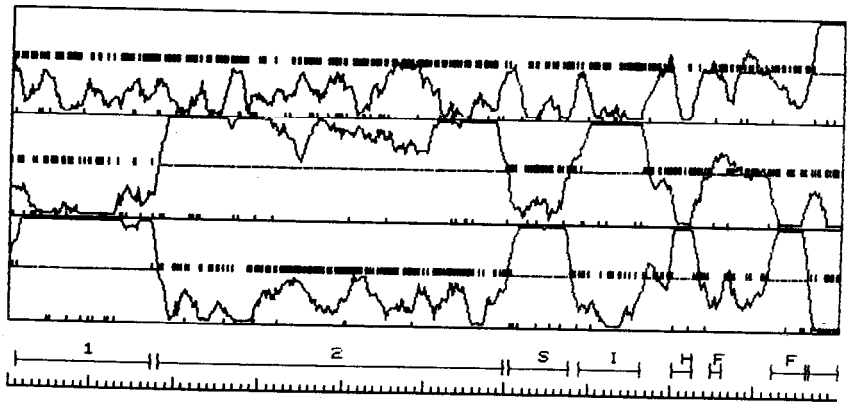


FIG. 3. Application of the positional base preferences method to bases 10,001 to 20,000 of the liverwort *Marchantia* chloroplast genome. The horizontal scale marks every 100th base, and the bars above indicate the extent of known protein coding segments. The three boxes above contain plots of the probability 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.

Genomics - Gene Modeling

Search by Content

- ***Positional base composition of a highly expressed gene. Note that the skew in the third position is even more extreme than for and 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.

Genomics - 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.
  - Rare codons (corresponding to low-level tRNAs) may also be used as a regulatory mechanism.
  - 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.
  - The overall codon 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