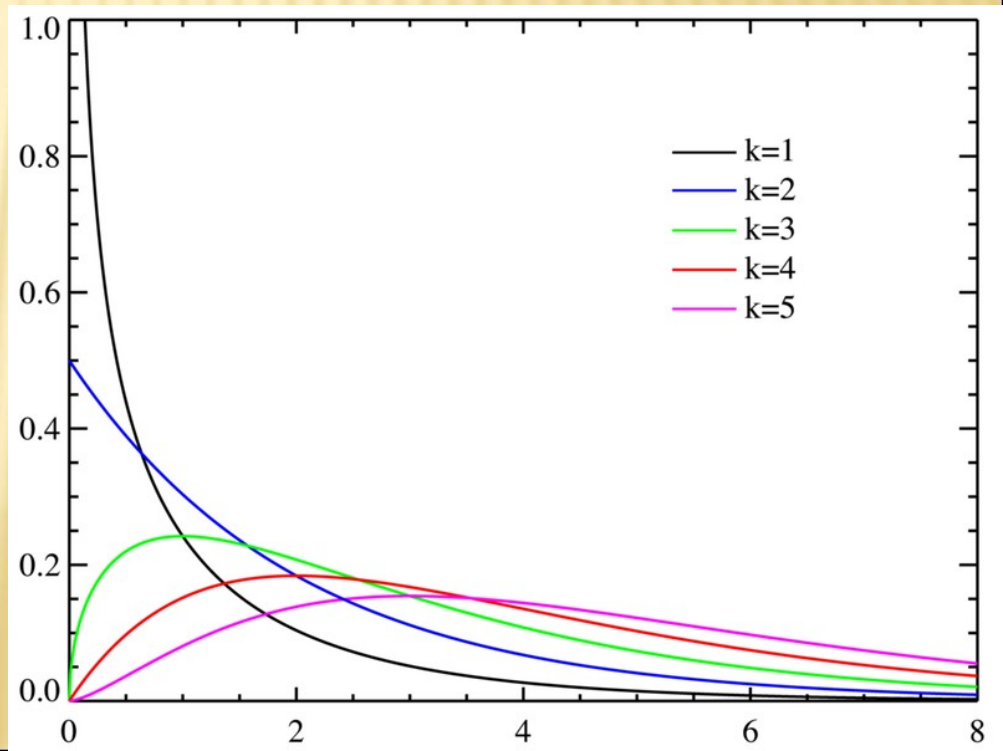
WHAT KINDS OF GENES CONTAIN TATA-BOXES?

- Are some classes of genes have TATA-boxes more frequently?
- Use the Chi-square test. Null hypothesis is that the frequency of TATA-boxes is the same for all classes of genes.

CHI-SQUARED TEST

O_i = an observed frequency; E_i = an expected frequency, n = the number of possible outcomes of each event. The chi-square statistic can then be used to calculate a p-value by comparing the value of the statistic to a chi-square distribution. The number of degrees of freedom K is equal to the number of possible outcomes, minus 1.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$



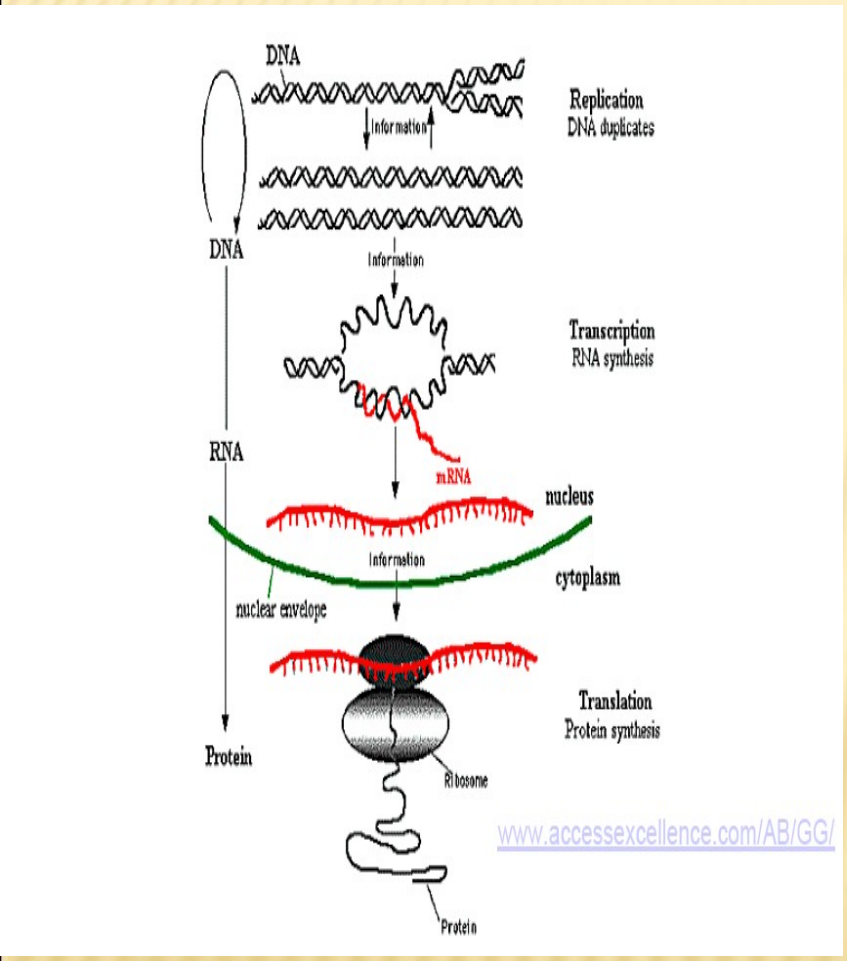
WHAT KINDS OF GENES CONTAIN TATA-BOXES?

FUNCTION	$TATA^c+$	$TATA^c-$	R^c	P-value
response to oxidative stress	71	55	2.82	7.0E-10
response to abscisic acid stimulus	65	67	2.12	6.4E-06
response to auxin stimulus	69	76	1.98	1.6E-05
defense response	67	74	1.98	2.9E-05
response to cold	60	72	1.82	3.5E-04
protein folding	30	117	0.56	4.4E-03
protein amino acid phosphorylation	66	305	0.47	2.1E-08

TATA+ = 4,169 and TATA- 9,110 promoters in entire population.

$$R^c = \frac{TATA_+^c}{TATA_-^c} \frac{TATA_-}{TATA_+}$$

THE CENTRAL DOGMA
DNA MAKES RNA MAKES PROTEIN

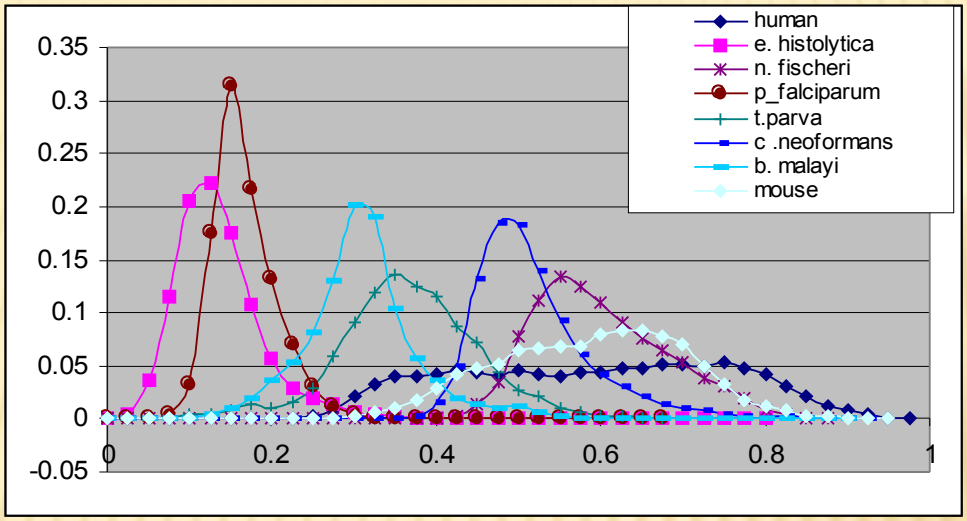


- There are 4^3 different codon combinations.
- There are 20 amino acids + START and STOP codons.

		2nd base			
		U	C	A	G
1st base	U	UUU (Phe/F)Phenylalanine UUC (Phe/F)Phenylalanine UUA (Leu/L)Leucine UUG (Leu/L)Leucine	UCU (Ser/S)Serine UCC (Ser/S)Serine UCA (Ser/S)Serine UCG (Ser/S)Serine	UAU (Tyr/Y)Tyrosine UAC (Tyr/Y)Tyrosine UAA Ochre (*Stop*) UAG Amber (*Stop*)	UGU (Cys/C)Cysteine UGC (Cys/C)Cysteine UGA Opal (*Stop*) UGG (Trp/W)Tryptophan
	C	CUU (Leu/L)Leucine CUC (Leu/L)Leucine CUA (Leu/L)Leucine CUG (Leu/L)Leucine	CCU (Pro/P)Proline CCC (Pro/P)Proline CCA (Pro/P)Proline CCG (Pro/P)Proline	CAU (His/H)Histidine CAC (His/H)Histidine CAA (Gln/Q)Glutamine CAG (Gln/Q)Glutamine	CGU (Arg/R)Arginine CGC (Arg/R)Arginine CGA (Arg/R)Arginine CGG (Arg/R)Arginine
	A	AUU (Ile/I)Isoleucine AUC (Ile/I)Isoleucine AUA (Ile/I)Isoleucine AUG (Met/M)Methionine, *Start*[1]	ACU (Thr/T)Threonine ACC (Thr/T)Threonine ACA (Thr/T)Threonine ACG (Thr/T)Threonine	AAU (Asn/N)Asparagine AAC (Asn/N)Asparagine AAA (Lys/K)Lysine AAG (Lys/K)Lysine	AGU (Ser/S)Serine AGC (Ser/S)Serine AGA (Arg/R)Arginine AGG (Arg/R)Arginine
	G	GUU (Val/V)Valine GUC (Val/V)Valine GUA (Val/V)Valine GUG (Val/V)Valine	GCU (Ala/A)Alanine GCC (Ala/A)Alanine GCA (Ala/A)Alanine GCG (Ala/A)Alanine	GAU (Asp/D)Aspartic acid GAC (Asp/D)Aspartic acid GAA (Glu/E)Glutamic acid GAG (Glu/E)Glutamic acid	GGU (Gly/G)Glycine GGC (Gly/G)Glycine GGA (Gly/G)Glycine GGG (Gly/G)Glycine

Is third nucleotide significant?

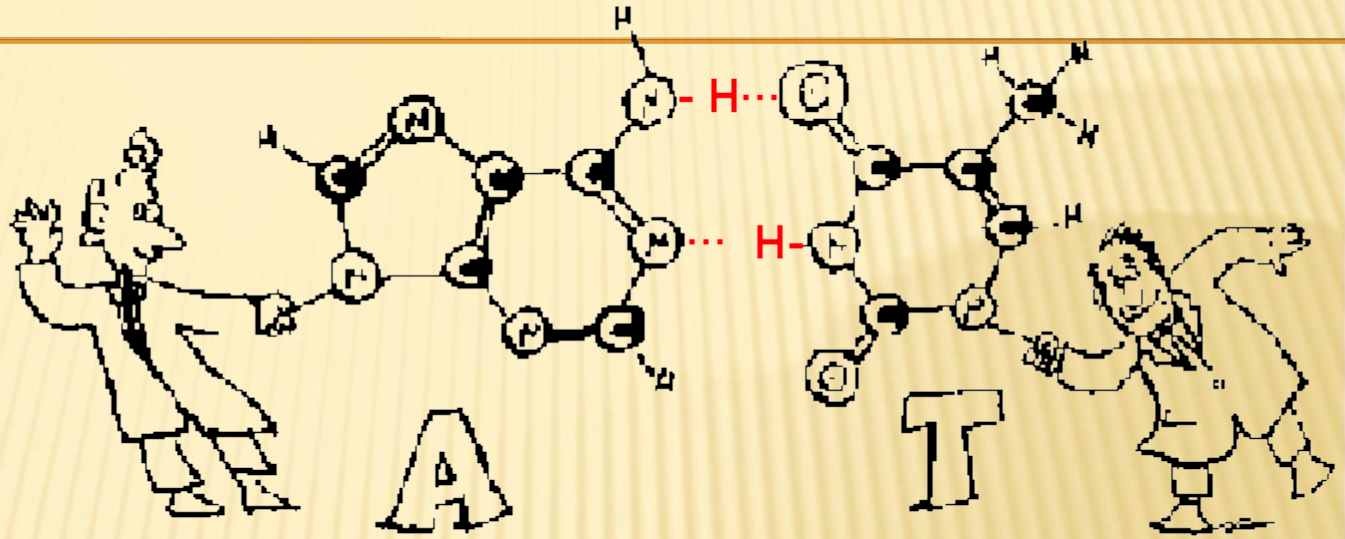
DISTRIBUTION OF C+G IN THE THIRD POSITION



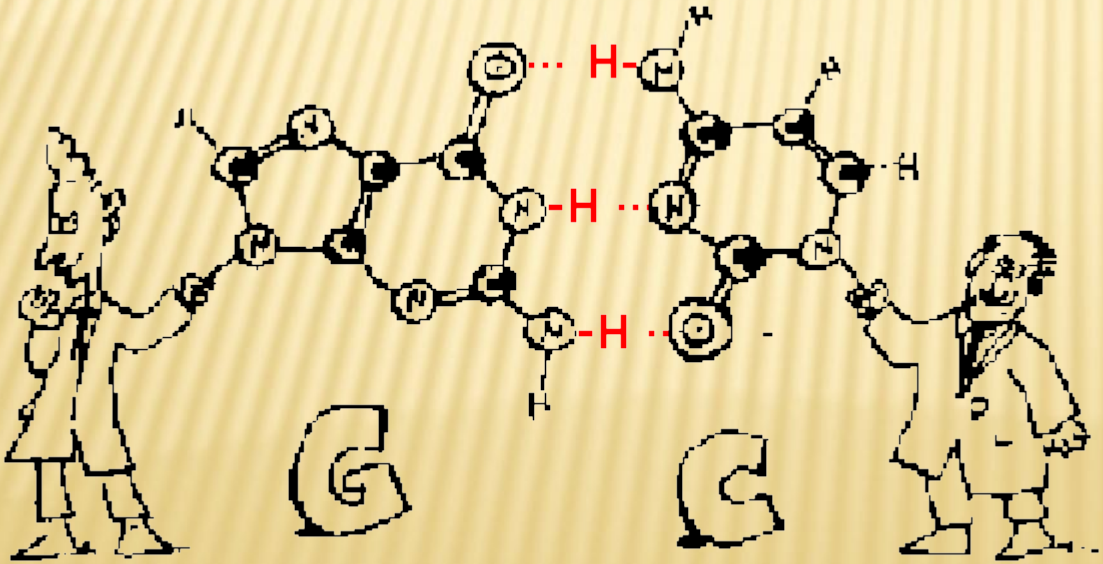
Possible explanations:

- Equilibrium is attained at low CG, hence older, stable genomes have higher AT3 content (first 2 positions are fixed). Other, newer genomes are under selective pressure
- Horizontal gene transfer

AT AND CG

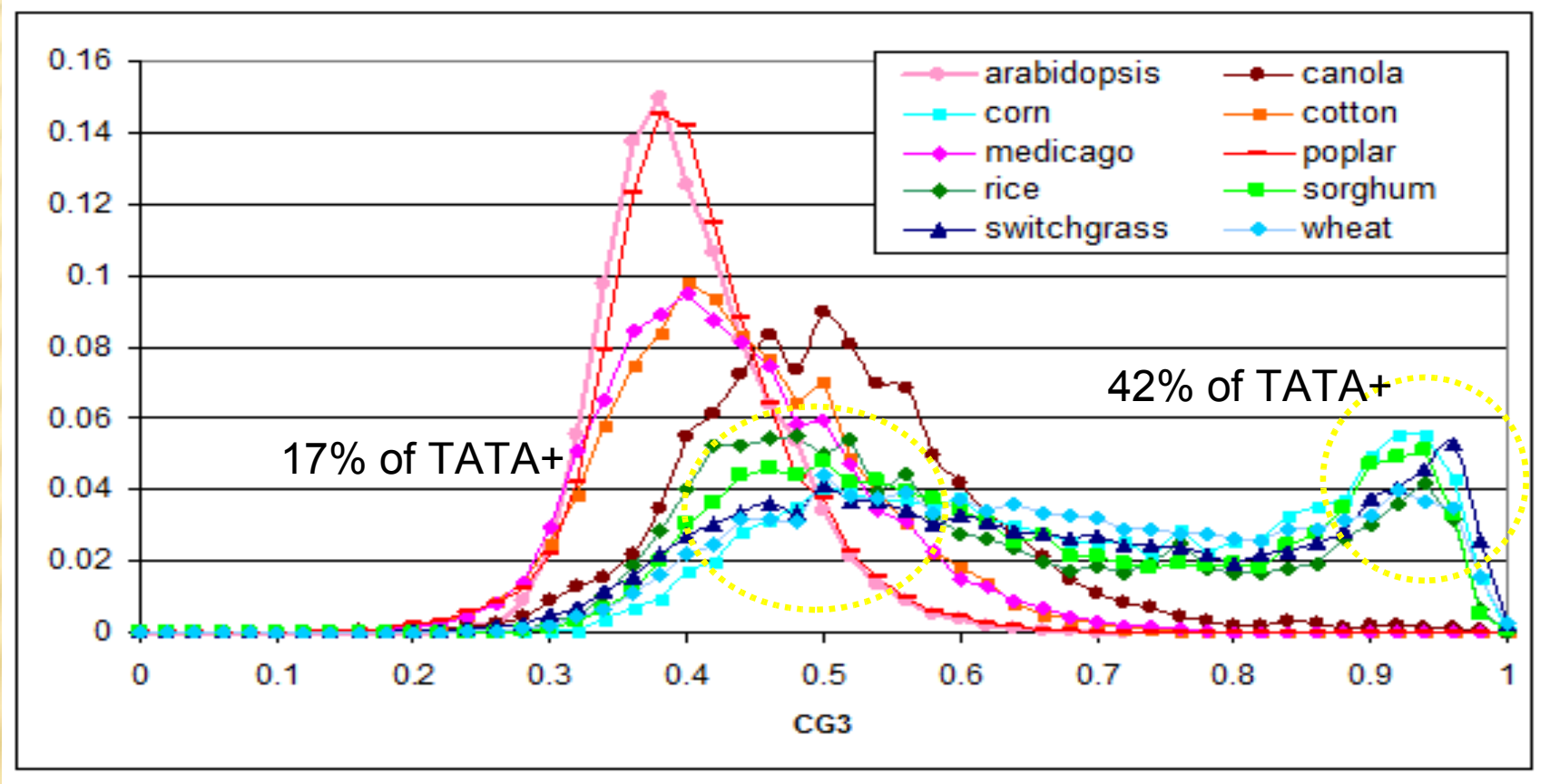


(Gonick & Whellis 1991)



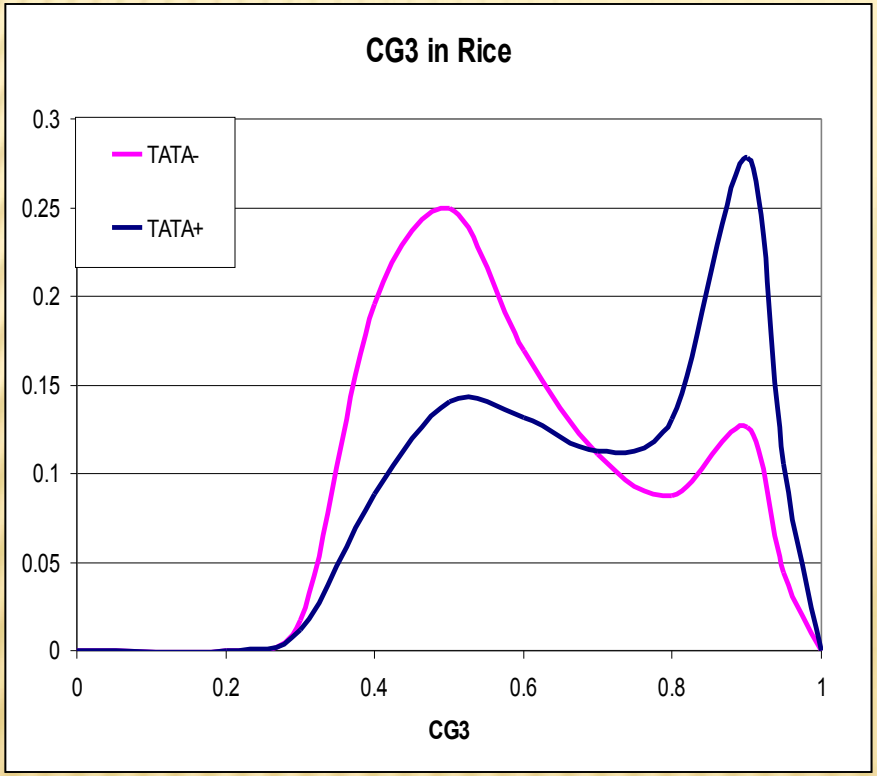
(Gonick & Whellis 1991)

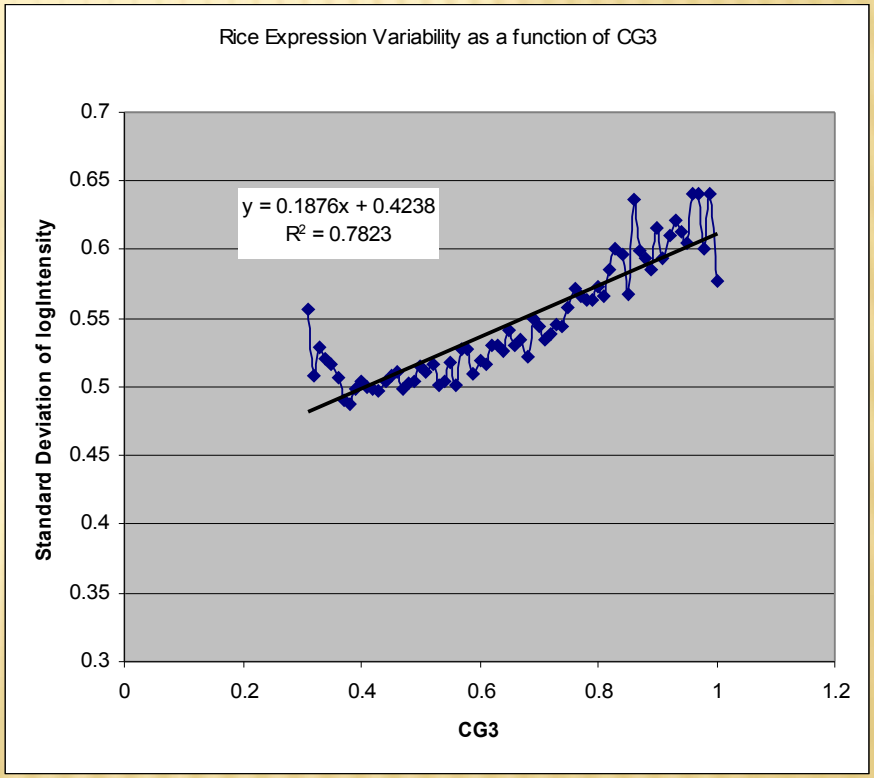
CG3 IS ASSOCIATED WITH TATA+



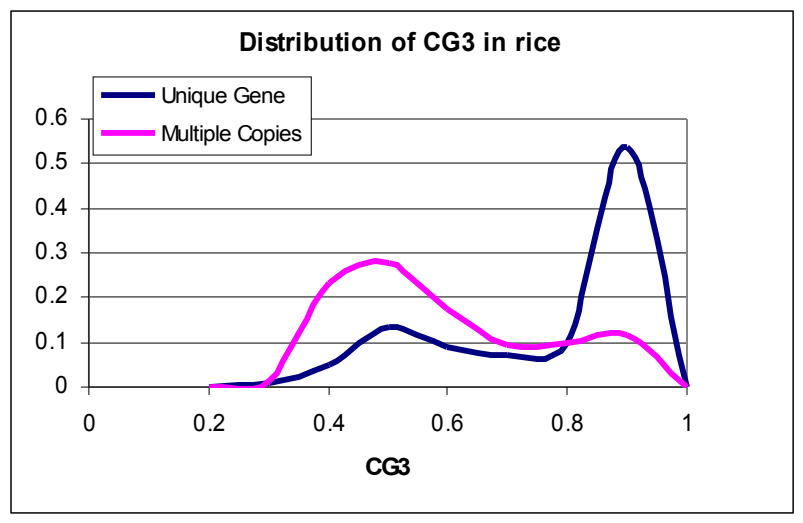
Grasses are relatively young and evolving

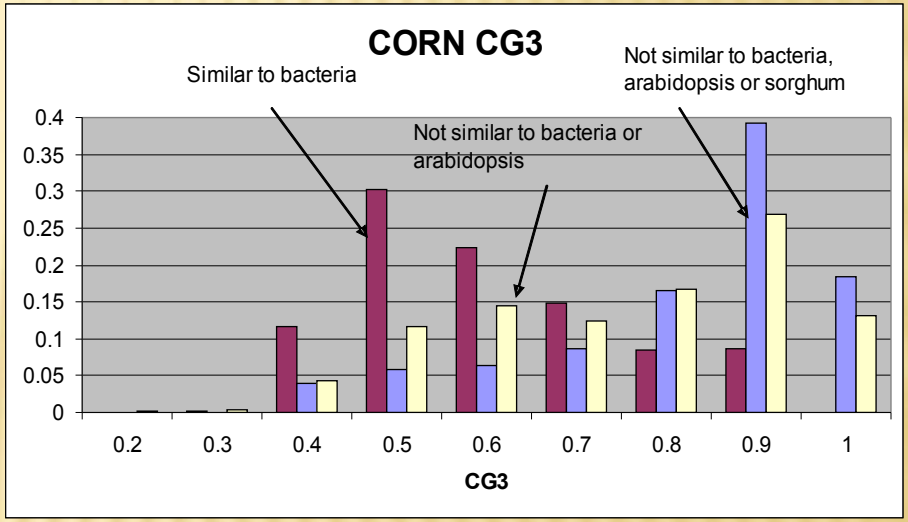
TATA/CG3/EXPRESSION VARIABILITY





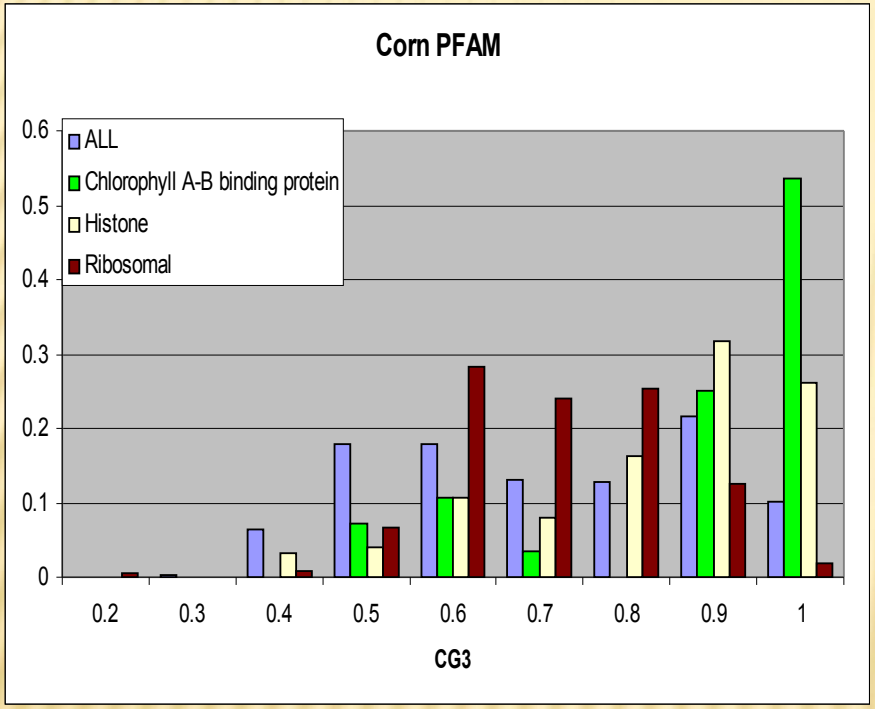
GENE UNIQUENESS AND CG3

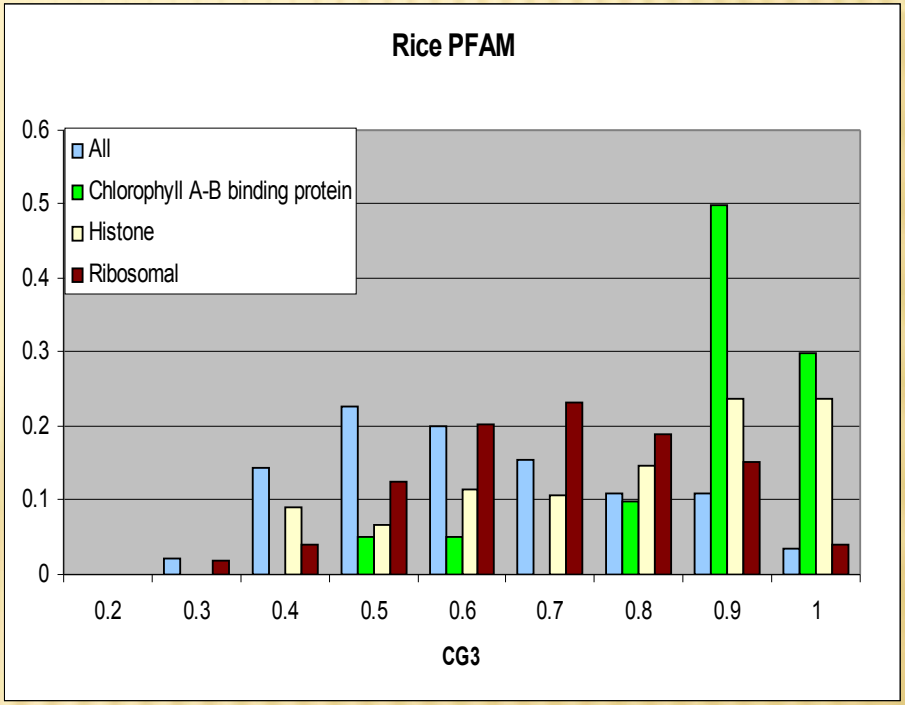




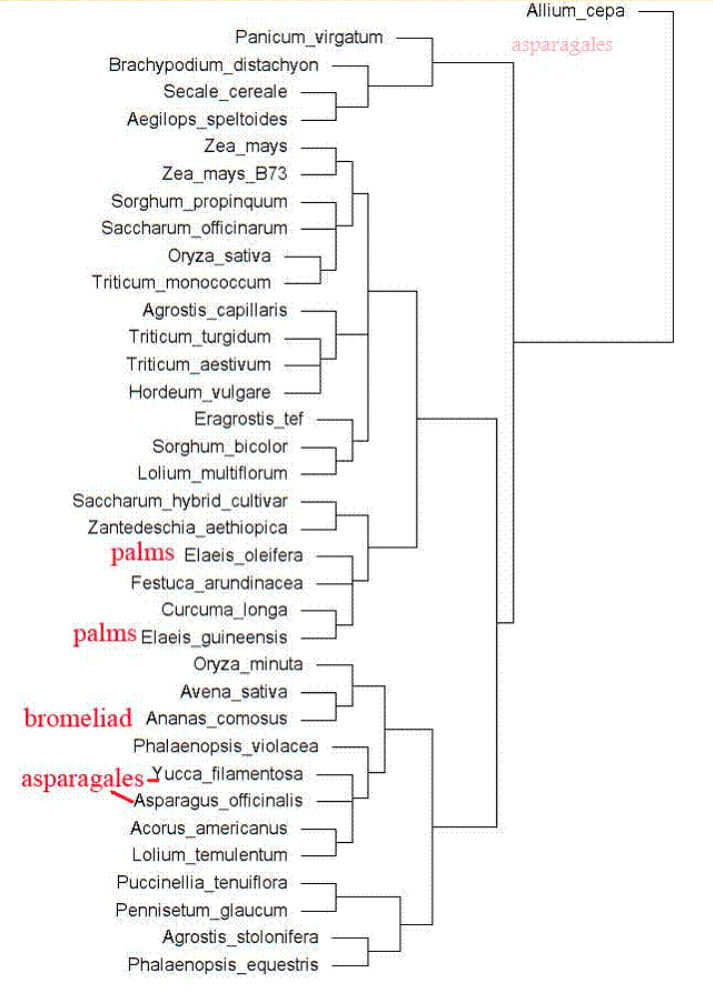
Genes from the higher GC peak group are usually intron-less

PROTEIN FAMILIES AND CG3



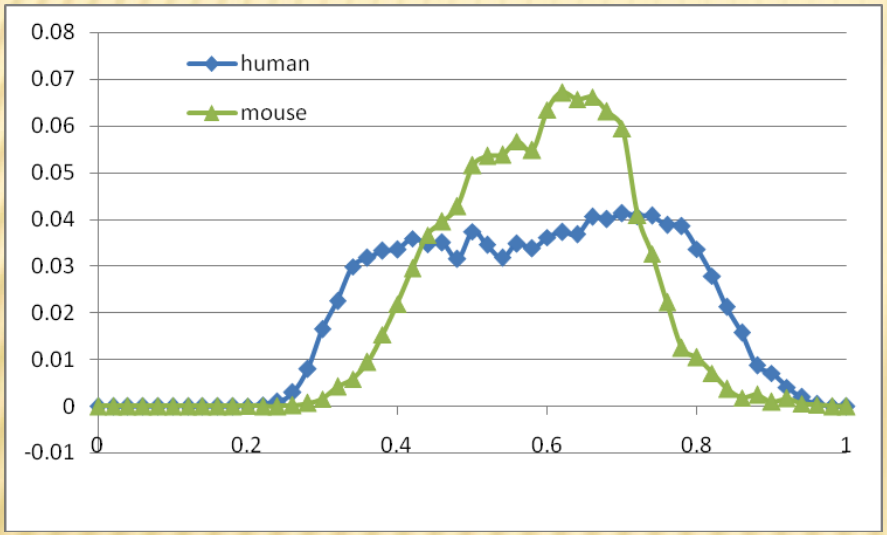


MONOCOT EVOLUTIONARY TREE



Distribution of CG3 is related to evolution. Since we observed that high-CG3 genes are related to stress response, presence of stress related genes in a particular genome may be an indication that the organism is evolving or adjusting to environment.

MAMMALS



CG3

- It appears that newer species and organisms under evolutionary pressure tend to develop high CG3 genes.
- To annotate genomes of such “camel” organisms, we should first classify proteins and then analyze their promoter region.