Biological words: (in genomes)

1. $k=1$ (base composition): $\{A, T, G, C\}$

$fr(A)$, $fr(T)$, $fr(G)$, $fr(T) \Rightarrow$ Any one parameter is enough.

e.g., $fr.(A) = fr(T)$, & $fr.(C) = \left(\frac{1 - fr.(A+T)}{2}\right)$

$= fr.(G)$.

Org.	% G+C	Genome size (Mb)
Eubacteria:		
E - Coli	50.7	4.693
Archaebacteria		
Pyrococcus abyssi	44.6	1.765
Eukaryotes		
C-elegans (Nematodes)	36	97
Arabidopsis Thaliana (a flower's plant)	35	125
Homo Sapiens	41	3080.

In prokaryotic genomes, in particular, there is an excess G over C on the leading strands, (strands whose 5' to 3' direction corresponds to the direction of replication fork movement). $\Rightarrow$ known as "GC Skew" $= \frac{\#G - \#C}{\#G + \#C}$

2. $k=2$ $\{AA, AT, \ldots, CC\}$

χ^2 - Test for verifying the observation with the model (or expected value). (Here "dinucleotide strings").

Let $O =$ # dinucleotide (say "GC") in a seq. of length 'n'.

$E =$ expected no. of ($\simeq (n-1)\, p_G\, p_C$) (there are $(n-1)$ pairs)

$\therefore \chi^2 = \frac{(O-E)^2}{E}$

$\chi^2 \Rightarrow$ small if $O \simeq E$ else large.

Alg.: Let r_1 & r_2 be two nucleotides whose adjacent occurrence to be observed.

(a) Let $c = \begin{cases} 1 + 2p_{r_1} - 3p_{r_1}^2 & r_1 = r_2 \\ 1 - 3p_{r_1}p_{r_2} & r_1 \neq r_2 \end{cases}$

(b) If $\chi^2/c > 3.84^*$, conclude that the iid model (identically independent distribution) is not a good fit.

[from theory of Statistics] ~~to be~~

[Assignment to students]

(* accepted with 0.95 confidence level or ~~0.05 hypothesis~~ namely 5% level of significance error.)

$k = 3$ (Codons)

Codon Adaptation Index (for an amino acid):

$$CAI = \prod_{k=1}^{L} (p_k / q_k)^{\frac{1}{L}} ;$$

L = length of the amino acid.

p_k = Prob. of the kth codon's occurrence of the kth ~~amount~~ amino acid (among highly expressed gene.

q_k = max. prob. of a codon of that amino acid (among the same set of highly expressed gene)

$$\log(CAI) = \frac{1}{L} \sum_{k=1}^{L} \log(p_k / q_k)$$

$0 \leq CAI \leq 1$. In E-Coli, a sample of 500 - protein coding genes displayed CAI value in the range of 0.2 to 0.85.

CAI is +vely correlated with mRNA expression level.

k-word problem (a different approach).

Obj.: To find a particular 'k-word' occurs more often in a set of N strings (e.g. promoter regions) of length m (say)

Statistical Test:

N_w = No. of strings where the word "w" is observed.

Compute p_w = Prob. of occurrence of w at least once in a seq. of length m.

[Simulate by randomly generating string of length m for a large No of times (say 5000) using the prob. of letters (A, T, C or G) in a genome (computed from the seq.)].

∴ Following this random model given N strings:

Prob. of the string occurs for k times

$$= \binom{N}{k} p_w^k (1-p_w)^{N-k}.$$

$$E(k) = N \cdot p_w \quad \& \quad \text{Var}(k) = N \cdot p_w (1-p_w)$$

Hence the statistic

$$Z_w = \frac{N_w - N \cdot p_w}{\sqrt{N p_w (1-p_w)}} \approx N(0,1).$$

Compute p-value viz. $P(X > Z_w)$ for each word.

If $p(Z_w) < 0.001$ the string w is unusual.