Computational Genome Analysis

Lecture-4

Consider a DNA sequence:

ATGGTGGTCATGGCGCCCCGAACCCTCTTCCT
GCTGCTCTCGGGGGCCCTGACCCTGACCGAG
ACCTGGGCGGGTGAGTGCGGGGTCAGGAGGG
AAACAGCCCCTGCGCGGAGGAGGGAGGGGCC
GGCCCGGCGGGGTCTCAACCCCTCCTCGCCC
CCAGGCTCCCACTCCATGAGGTATTTCAGCGC
CGCCGTGTCCCGGCCCGGCCGCGGGGAGCCC
CGCTTCATCGCCATGGGCTACGTGGACGACAC
GCAGTTCGTGCGGTTC

Given a DNA sequence, there are a number of questions one might ask:

- What sort of statistics should be used to describe this sequence?
- What sort of organism this sequence came from based on sequence content?
- Do parameters describing this sequence differ from those describing bulk DNA in that organism?
- What sort of sequence it might be: Protein coding? Centromere? Telomere? Transposable element? Control sequence?

Let's attempt to address such Qs. by considering words:

- short strings of letters drawn from DNA alphabet, A, C, G, T.

What sort of information can we obtain from "*k*-tuple", or, *k*-mer analysis?

k = 1 (Base Composition)

***k*=1: Frequency of bases.**

DNA being a duplex,

No. of A's = No. of T's

No. of G's = No. of C's

What about on the same strand?

5'- GGATCGAAGCTAAGGGCT - 3'

3'- CCTAGCTTCGATTCCCGA - 5'

Top/Bottom Strand: No. of G = ?, C = ?

For the duplex molecule, G = ?, C = ?

$$N_A = N_T, \quad N_G = N_C, \quad N_A + N_G = N_T + N_C$$

- only for duplex molecule, not for a single strand.

Biological Words

For the duplex molecule,

$$fr(A+T) = 1 - fr(G+C)$$

$\Rightarrow$ only a single parameter, say fr(G+C), suffices in describing the base frequencies for duplex DNA.

i.e., there are 4 variables and 3 relations:

$$fr(A) = fr(T), \quad fr(G) = fr(C), \quad fr(A+T) = 1 - fr(G+C)$$

Base composition has been used as a descriptive statistic for genomes of various organisms since the early days of molecular biology.

Biological Words

If a genome is GC-rich, say, 60% GC, then find the fractions of A, T, G, and C.

The GC bond is stronger than the AT bond

- **higher temperatures are required for denaturation (opening the strands) if the genome is GC-rich**
- **Melting point of GC-rich genome is higher than a AT-rich genome**

So, what can you say about a bacterial species found in naturally occurring hot springs?

Base Compositions of Various Organisms

Organism	% G+C	Genome size (Mb)
Eubacteria		
Mycoplasma genitalium	31.6	0.585
Escherichia coli K-12	50.7	4.693
Pseudomonas aeruginosa PAO1	66.4	6.264
Archaebacteria		
Pyrococcus abyssi	44.6	1.765
Thermoplasma volcanium	39.9	1.585
Eukaryotes		
Caenorhabditis elegans (a nematode)	36	97
Arabidopsis thaliana (a flowering plant)	35	125
Homo sapiens (a bipedal tetrapod)	41	3080

Variation in GC content in different organisms – due to variation in selection, mutational bias & biased recombination-associated DNA repair

Biological Words

Distribution of individual bases within DNA molecule

- not uniform.

In prokaryotic genomes in particular, an excess of G over C is observed on the leading strand.

(strands whose 5' to 3' direction corresponds to the direction of replication fork movement).

This is described by "GC skew":

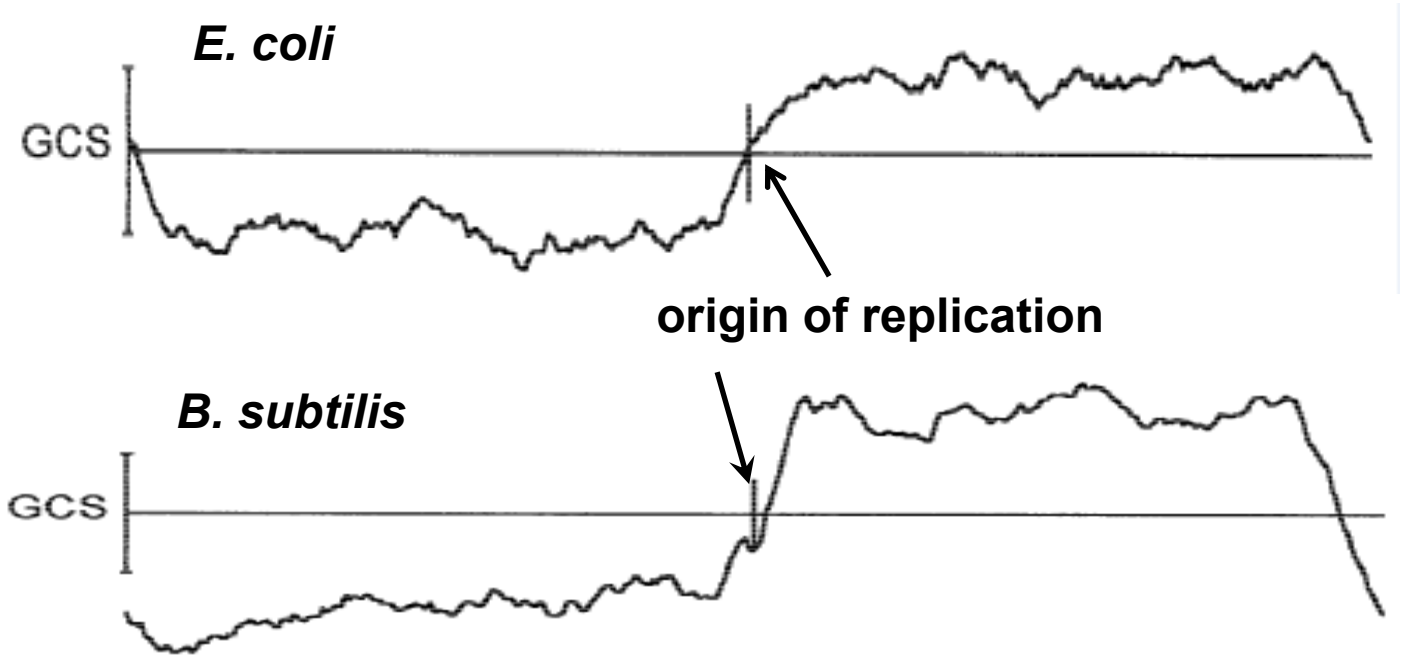
$$\text{GC skew} = (\#G - \#C)/ (\#G + \#C),$$

calculated in sliding "windows".

- It changes sign at positions of replication origins and termini in prokaryotic genomic sequences.

- another example of how a relatively simple statistic based on 1-tuple can be informative.

GC Skew



Third codon position GC skew (GCS), window size – 300Kb, step size – 10Kb.

Direction of skew switches at the genome's origin & terminus of replication, such that the leading strand in replication is always richer in G than C.

Biological Words

GC content is not uniformly constant throughout the genome - but is found to be markedly variable

- resulting in a mosaic-like formation with islet regions called isochores (GC-rich regions).

GC-rich isochores include many protein coding genes

- About 50% of humans are GC-rich

- another example of how a simple statistic based on *k* = 1 can be informative.

An isochore is a large region of DNA (> 300 KB) with a high degree of uniformity in G & C

k = 1 Analysis

To summarize:

- **G+C fraction gives the base composition of a genome**
- **Melting temperature of DNA sequence can be known**
- **GC skew on the leading strand can help in identifying the origin of replication in prokaryotes**
- **GC-rich regions – help in identifying protein-coding genes**
- **Since GC content of an organism is different, it is a useful measure in identifying horizontally transferred regions**

An interesting problem in genome analysis is in knowing whether certain patterns occur more often than by random chance in a given sequence;

- such sequences might be of biological significance

This can be addressed by using a probabilistic model of DNA sequence to identify if a pattern is over/under-represented in the genomic sequence compared to its expected frequency based on the iid model.

e.g., frequency of triplets is different in protein-coding regions compared to the genomic sequence.

Probabilistic model: a method for simulating observations from the model,

i.e., specify probabilistic rules to produce next letter in the simulated sequence, given the previous letters:

- **First base in the sequence is either an A, C, G, or T with probability p_A, p_C, p_G, p_T, respectively.**
- **Suppose the algorithm has generated *r* bases. To generate the base at position *r* +1, the algorithm pays no attention to what has been generated before and gives out A, C, G, or T with probabilities p_A, p_C, p_G, p_T.**

  - similar to coin tossing experiment.

χ = {A, C, G, or T} for DNA seq

Output of simulation at every step, *X* takes any one character from set χ with probability p_A, p_C, p_G, p_T

$$p_A, p_C, p_G, p_T \geq 0, \qquad p_A + p_C + p_G + p_T = 1$$

p_A, p_C, p_G, p_T is the probability distribution of *X*.

i.e., first base L_1 in our model for a DNA sequence has probability distribution

$$P(L_1 = A) = p_A, \quad P(L_1 = C) = p_C,$$
$$P(L_1 = G) = p_G, \quad P(L_1 = T) = p_T. \qquad (1)$$

To study probability distribution of no. of times a given pattern occurs in a random DNA sequence $L_1, L_2, \ldots, L_n$,

Consider patterns of length one, $k = 1$. To address this question, define a new sequence $X_1, X_2, \ldots, X_n$ by

$$X_i = \begin{cases} 1, & \text{if } L_i = \text{A} \\ 0, & \text{otherwise} \end{cases} \qquad (2)$$

No. of times that A appears is:

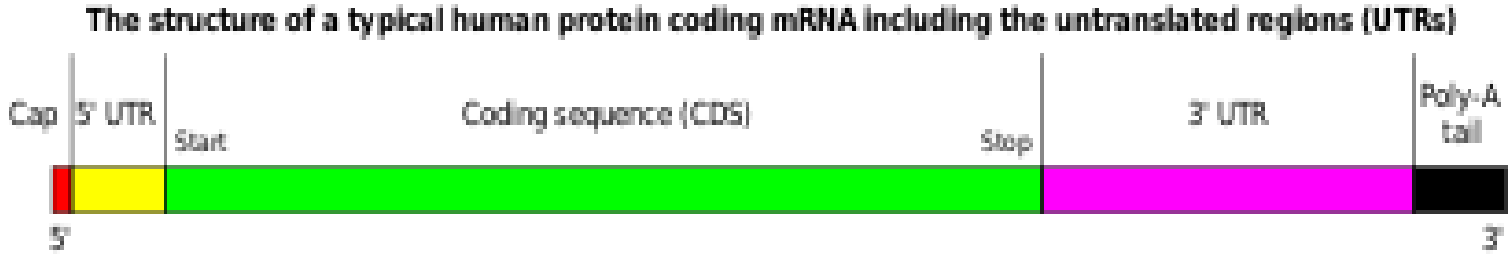
$$N = X_1 + X_2 + \ldots + X_n \qquad (3)$$

What is the probability of getting runs of A's of certain length?

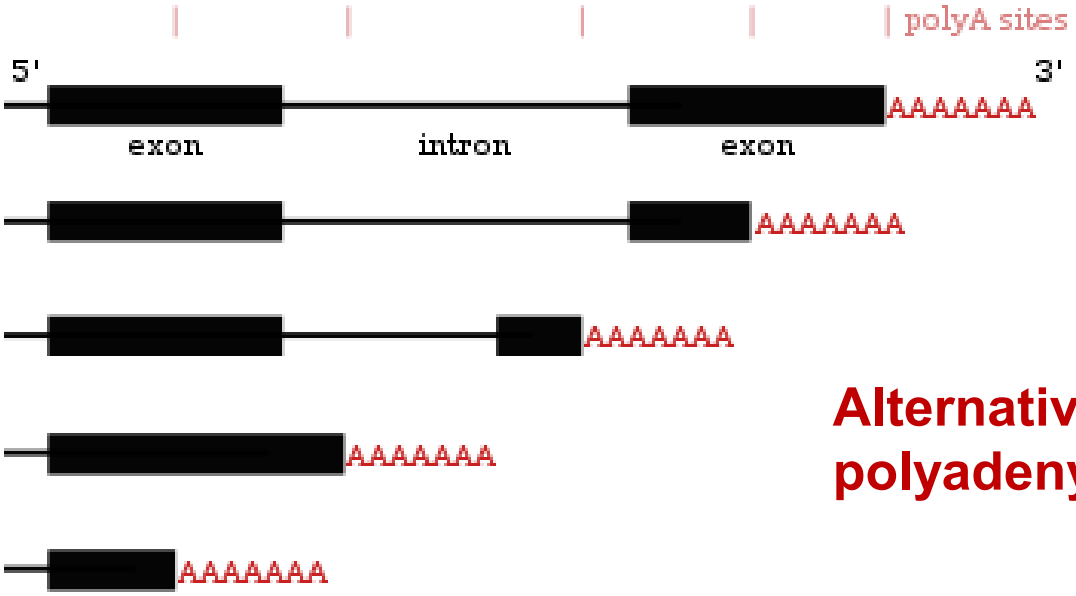
Is a homopolymer A region of any significance?

Polyadenylation signal – the site at which cleavage of 3'end of mature RNA takes place: AAUAAA

Typical structure of a mature eukaryotic mRNA



PolyA tail is important for nuclear export, translation and stability of mRNA



Alternative polyadenylation

Results of using different polyadenylation sites on the same gene

Starting from the initial probability distribution of L_i, the probability distribution of each X_i is:

$$P(X_i = 1) = P(L_i = A) = p_A,$$

$$P(X_i = 0) = P(L_i = C \text{ or } G \text{ or } T)$$

$$= p_C + p_G + p_T = 1 - p_A \qquad (4)$$

Different "runs" of the simulation would produce strings having different values of *N*, i.e., no.'s of A's.

- like in the case of a coin-tossing expt.

Then what is a "typical" value of *N*, i.e., its probability distribution?

To find the probability distribution of N is complicated because we need to know how individual outputs from our simulation are related to each other.

- it is not easy to find whether the random variables are independent,

Assume independence and then use the multiplicative rule to calculate the probabilities of different outcomes.

If the probability of a given pattern, assuming independence is similar to that observed, then the pattern is a random independent pattern, else the pattern is unique

Is this likely to be true in a DNA sequence?

Multiplicative Rule for Independent Events:

If events E & F are independent, probability of the event "E and F":

$$p(E \cap F) = p(E) \cdot p(F)$$

For a DNA sequence model, L_i are independent, so probability of obtaining the sequence l_1, l_2, …, l_n is

$$\mathbb{P}(L_1 = l_1,\ L_2 = l_2,\ \ldots, L_n = l_n) = \mathbb{P}(L_1 = l_1)\,\mathbb{P}(L_2 = l_2)\,\mathbb{P}(L_n = l_n)$$

For S = ATTGCGTGAG, what is the probability of observing the pattern, P(ATTG) ?

$$P(ATTG) = p_A p_T p_T p_G$$

A distribution is defined by its mean & variance

Measures of location (central tendency) & spread for a sequence $X_1, X_2, \ldots, X_n$ given by

1100101000110110101000111011010

with probability distribution

$$P(X_i = 1) = p_A, \qquad P(X_i = 0) = p_C + p_G + p_T = 1 - p_A$$

$$X_i = \begin{cases} 1, & \text{if } L_i = \text{A} \\ 0, & \text{otherwise} \end{cases}$$

Define the expected value (or mean of expectation) of X

- to see if X is closely concentrated about its expected value, or is spread out, compute variance.

Mathematical Expectation

If X assumes discrete values $X_1, X_2, \ldots, X_k$ with respective probabilities $p_1, p_2, \ldots, p_k$, the expectation of X is defined as

$$E(X) = p_1X_1 + p_2X_2 + \cdots + p_kX_k = \sum_{i=1}^{k} p_iX_i = \sum pX$$

If probabilities p_i are replaced by relative frequencies

$$f_i/N, \quad N = \sum f_i$$

For large N

the expectation reduces to $\left(\sum fX\right)/N$

- arithmetic mean of a sample of size N

We interpret that $E(X)$ represents the mean of the population from which the sample is drawn.

Expected Values for Random Variables:

Using the fact that mean of the sum of X_i is the sum of the mean of the X_i:

$$\mathbb{E}\,(X_1 + X_2 + \ldots + X_n) = \mathbb{E}\,X_1 + \mathbb{E}\,X_2 + \ldots\ \mathbb{E}\,X_n$$

the expected No. of times we see an *A* in our *n* bp sequence is

$$\mathbb{E}\,N = \mathbb{E}\,X_1 + \mathbb{E}\,X_2 + \ldots\ \mathbb{E}\,X_n = n\,\mathbb{E}\,X_1 = n\,p_{\mathrm{A}}$$

where $N = X_1 + X_2 + \ldots + X_n$

– no. of times that A appears is the sum of X's, and

$$EX_i = 1 \times p_A + 0 \times (1-p_A) = p_A$$

$$X_i = \begin{cases} 1, & \text{if } L_i = \mathrm{A} \\ 0, & \text{otherwise} \end{cases}$$

$P(X_i = 1) = P(L_i = A) = p_A,$

$P(X_i = 0) = P(L_i = C \text{ or } G \text{ or } T)$

$= p_C + p_G + p_T = 1 - p_A$

Expected Values for Random Variables:

Expected value of a random variable *X* gives a measure of its location - values of X tend to be scattered around this value,

$$\mu = \mathbb{E}\, X$$

So if we simulate this sequence a large no. of times, the no. of A's will take values around this mean value np_A.

The measure of spread is given by the variance.

Variance for Random Variables:

Measure of spread: Is *X* closely concentrated about its expected value, or is it spread out?

$$\text{Var}\, X = \mathbb{E}\,(X - \mu)^2 = \sum_{i=1}^{J} (x_j - \mu)^2 p_j$$

$$\text{Var}\, X = \mathbb{E}\, X^2 - \mu^2 = \sum_{j=1}^{J} x_j^2 p_j - \mu^2$$

For random variable X_i,

$$\text{Var}\, X_i = [1^2 \text{ x } p_A + 0^2 \text{ x } (1 - p_A)] - p_A{}^2 = p_A(1 - p_A)$$

+ve sq. root of variance is the standard deviation

Variance for Random Variables:

To calculate the variance of the no. *N* of A's in our simulated DNA sequences, we exploit the fact that

the variance of a sum of independent random variables is the sum of the individual variances,

$$\mathrm{Var}(X_1 + X_2 + \ldots + X_n) = \mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n)$$

$$\mathrm{Var}\, N = \mathrm{n}\, \mathrm{Var}\, X_1 = n\, p_{\mathrm{A}}(1 - p_{\mathrm{A}})$$

Expected value & Variance of the no. of times letter A occurs in the given DNA sequence

- are two statistics that describe its probability distribution

Ex: Find the no. of times letter A occurs in a sequence of length n = 1000 if the probability of occurrence of A, p_A = 0.15.

Expected value = n x p_A = .15 x 1000 = 150,

Variance = n x p_A (1 - p_A) = 1000 x .15 x .85 = 127.5,

Std. deviation = 11.29

No. of times A would occur in a sequence of length 1000, if p_A = 0.15 will lie between

150 ± 11.29, i.e., 139 – 161

Applications: finding low-complexity regions, polyA tail, occurrence of GC boxes in the vicinity of a gene, etc.

Ex: Simulate observations having the binomial distribution with $p = 0.25$ and $n = 1000$.

Suppose that for a sequence of length $n = 1000$ bp and assuming that each base is equally likely, what is the probability of observing 280 A's or more in such a sequence?

There are 3 ways of doing this!

- **by using the Binomial distribution formula for the probability of observing j A's,**
- **by simulation (~ 10,000 runs),**
- **by Central Limit Theorem**

Binomial Distribution:

Since X_i are iid, the probability of the sequence that has j 1's in $x_1, x_2, \ldots, x_n$, is $p^j(1 - p)^{n-j}$

To compute the probability that the sequence contains j A's (i.e., $N = j$), we need to know

how many different realizations of the sequence $x_1, x_2, \ldots, x_n$ have j 1's (and $n - j$ 0's).

This is given by the binomial coefficient ${}^nC_j$, defined as

$$ {}^nC_j = \frac{n!}{j!(n-j)!} $$

Thus, the probability of observing j A's is

$$ \mathbb{P}(N = j) = {}^nC_j\, p^j (1-p)^{n-j}, \quad j = 0, 1, 2, \ldots, n $$

Simulating from Probability Distributions:

To understand the behaviour of random variables such as, the no. of A's (*N*), simulate no. of sequences and obtain the no. of A's: N_1, N_2, …, N_n, having the same probability distribution as *N*.

We can use them to study the properties of this distribution. e.g., the sample mean

$$\overline{N} = (N_1 + N_2 + ... + N_n)/n$$

can be used to estimate the expected value μ of *N*.

We can use the sample variance to estimate the variance σ^2 of *N*:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(N_i - \overline{N})^2$$

And use the histogram of the values of N_1, N_2, …, N_n to estimate the probability of different outcomes for *N*

Simulating from Probability Distributions:

The pseudo-random number generators can be used to produce a sequence of random nos. U_1, U_2, …, U_n

We can thus simulate an observation with the distribution of X_1 given by

$$X_i = \begin{cases} 1, & \text{if} \quad L_1 = \text{A} \\ 0, & \text{otherwise} \end{cases}$$

by taking a uniform random number U and setting $X_1 = 1$ if $U_1 \leq p \equiv p_A$ and 0 otherwise.

Repeating this procedure n times (with a new U each time) results in a sequence X_1, X_2, …, X_n from which N can be computed by adding up the X's.

Simulating from Probability Distributions:

We can simulate the sequence of bases $L_1, L_2, \ldots, L_n$, using random-number generator.

This is done by diving the interval (0, 1) into four intervals with endpoints at

$$p_A, \quad p_A + p_C, \quad p_A + p_C + p_G, \quad p_A + p_C + p_G + p_T = 1$$

If the random no. called lies in the leftmost interval, set L_1 = A;

if it is in the second interval, set L_1 = C;

if it is in the third interval, set L_1 = G;

otherwise set L_1 = T.

Repeating this procedure with a new series of random nos. each time to produce a no. of sequences and count the occurrence of A.

Using Binomial distribution:

$$P(N \geq 280) = \sum_{j=280 \text{ to } 1000} {}^{n}C_{j} (1/4)^{j}(1 - 1/4)^{n-j} = 0.016$$

By Simulation:

Simulate large no. of sequences, say, 10,000, and find out in how many cases, no. of A's is ≥ 280?

If 149 such instances observed, then

$$P(N \geq 280) = 149/10000 = 0.0149$$

By Central Limit Theorem:

$$\mu = Np = 250, \sigma = \sqrt{Npq} = 13.69,$$

$$z = (280\text{-}250)/13.69 \sim 2.1,$$

Probability corresponding to this z-score is

$$\sim 0.5 - 0.4857 = 0.0143$$

Biological Words: $k = 2$

A dinucleotide is $l_i l_{i+1}$, l_i is one of the 4 bases A, C, G, T

$\Rightarrow$ there are 16 different dinucleotides: AA, AC, AG, AT, …TG, TT; the sum of the dinucleotide frequencies is 1.

Importance:

- **Set of di-nucleotide relative abundance values constitutes a "genomic signature" of an organism**
- **Useful in identifying genomic islands (laterally transferred regions) devoid of coding.**

Biological Words: $k = 2$

Suppose we model the sequence $L_1, L_2, \ldots, L_n$ using iid model with base probabilities given by

$$\mathbb{P}(L_1 = \mathrm{A}) = p_{\mathrm{A}}, \quad \mathbb{P}(L_1 = \mathrm{C}) = p_{\mathrm{C}}, \quad \mathbb{P}(L_1 = \mathrm{G}) = p_{\mathrm{G}}, \quad \mathbb{P}(L_1 = \mathrm{T}) = p_{\mathrm{T}}$$

Assuming independence in the occurrence of bases, using multiplication rule, the probabilities of each dinucleotide $r_1 r_2$

$$\mathbb{P}(L_i = r_1, \ L_{i+1} = r_2) = p_{r1} p_{r2}$$

i.e., under the independence model, the chance of seeing dinucleotide AA is p_{A}^2, & chance of seeing CG is $p_{\mathrm{C}} p_{\mathrm{G}}$.

Biological Words: $k = 2$

To test if a given sequence has unusual dinucleotide frequencies compared with an iid model,

Compare the observed no. O of dinucleotide $r_1 r_2$

with the expected no. given by $E = (n - 1)\, p_{r1} p_{r2}$.

χ^2 – test:

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

If the observed no. is close to the expected no. (the model is doing a good job of predicting the dinucleotide frequencies), χ^2 will be small.

Biological Words: $k = 2$

To determine which values of χ^2 are unlikely if in fact the model is true can be estimated:

(a)Calculate the number c given by

$$c = \begin{cases} 1 + 2p_{r1} - 3p_{r1}^2, & if \quad r_1 = r_2 \\ 1 - 3p_{r1}p_{r2}, & if \quad r_1 \neq r_2 \end{cases}$$

(b) Calculate the ratio χ^2 /c

(c) If this ratio is >3.84, the iid model is not a good fit.

If the base frequencies are unknown, the same approach works if these are estimated from the data.

For 1 degrees of freedom, and probability 0.05, chi-square value is 3.84.

Observed values of χ^2/c for the first 1000 bp of *E. coli* and *M. genitalium* genomes.

Dinucleotide	Observed of χ^2/c for	
	E. coli	*M. genitalium*
AA	**6.78**	0.15
AC	0.05	1.20
AG	**5.99**	0.18
AT	0.01	0.01
CA	2.64	0.01
CC	0.03	0.39
CG	0.85	**4.70**
CT	**4.70**	1.10
GA	2.15	0.34
GC	**10.04**	1.07
GG	0.01	0.09
GT	1.76	0.61
TA	**5.99**	1.93
TC	**9.06**	2.28
TG	3.63	0.05
TT	1.12	0.13

***E. coli* base frequencies: (0.25, 0.25, 0.25, 0.25)**

***M. genitalium* frequencies: (0.45, 0.09, 0.09, 0.37)**

***E. coli* dinucl. frequencies not very well-described by the simple iid model**

Biological Words: *k* = 3

There are 61 codons that specify amino acids and three stop codons.

Since there are 20 AAs, most AAs are specified by more than one codon.

This has led to the use of a number of statistics to summarize "bias" in codon usage.

To see how these codon frequencies can vary, let's consider the specific example of the *E. coli* proteins.

Comparison of predicted & observed triplet frequencies in coding sequences for a subset of genes from *E. coli*.

	Codon	Predicted	Observed	
			Gene Class I (502)	Gene Class II (191)
Phe	TTT	0.493	0.551 ↔	0.291
	TTC	0.507	0.449 ↔	0.709
Ala	GCT	0.246	0.145	0.275
	GCC	0.254	0.276	0.164
	GCA	0.246	0.196	0.240
	GCG	0.254	0.382	0.323
Asn	AAT	0.493	0.409 ↔	0.172
	AAC	0.507	0.591 ↔	0.828

Class II genes ~ largely ribosomal proteins or translation factors
– genes expressed at high levels,
Class I genes ~ mostly those that are expressed at moderate levels

How are the predicted relative frequencies calculated?
Relative frequencies of two codons coding for the same A.A.?

Calculating the predicted relative frequencies:

For a sequence of independent bases L_1, L_2, …,L_n, the expected 3-tuple relative frequencies are computed as:

$$\mathbb{P}\left(L_i = r_1,\ L_{i+1} = r_2, L_{i+2} = r_3\right) = \mathbb{P}\left(L_i = r_1\right)\mathbb{P}\left(L_{i+1} = r_2\right)\mathbb{P}\left(L_{i+2} = r_3\right)$$

- provides the expected frequencies of particular codons

We first calculate the relative proportion of each of the codons making up a particular AA.

The relative proportion of each of the codons making up a particular AA.

$$\text{₱}\,(\text{TTT}) = 0.246 \times 0.246 \times 0.246 = 0.01489$$

$$\text{₱}\,(\text{TTC}) = 0.246 \times 0.246 \times 0.254 = 0.01537$$

using base frequencies. Among these codons making up the AA Phe, the expected proportion of TTT is

$$\frac{0.01489}{0.01489 + 0.01537} = 0.492$$

	Codon	Predicted	Observed	
			Gene Class I (502)	Gene Class II (191)
Phe	TTT	0.493	0.551	0.291
	TTC	0.507	0.449	0.709

Most widely used statistic for analyzing Codon usage bias in a protein is codon adaptation index (CAI)

- it compares the distribution of codons actually used in a particular protein with the preferred codons for highly expressed genes.

- **Highly expressed genes use mostly codons for tRNA species that are more abundant in the cell.**

⇒ CAI provides an indication of gene expression level under the assumption that there is translational selection to optimize gene sequences according to their expression levels

Consider a protein X = x_1, x_2, …, x_L with x_k representing the AA residue corresponding to codon k in the gene.

Here we are interested in comparing the actual codon usage with an alternative model:

- that the codons employed are the most probable codons for highly expressed genes.

p_k - probability that a particular codon k is used to code for the AA, and

q_k - probability for the most frequently used synonymous codon for that AA in highly expressed genes.

Codon Adaptation Index

CAI is defined as

$$\mathrm{CAI} = \left[\prod_{k=1}^{L} p_k / q_k \right]^{1/L}$$

i.e., the geometric mean of the ratios of the probabilities for the codons actually used to the probabilities for the codons most frequently used in highly expressed genes.

Alternatively,

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

There is a correlation between CAI and mRNA levels - if CAI is large, the expression level is also large, i.e., it provides a first approx. of its expression level.

Larger Words

No. and distributions of *k*-tuples > 3 have practical & biological significance.

Some important *k*-tuples corresponding to *k* = 4, 5, 6 and 8:

- determine the distribution of restriction endonuclease digest fragments.

- useful for constructing the physical maps of genomes and identifying structural variations

Restriction sites for restriction endonucelases

Enzyme	Recognition Sequence
EcoRI	G↓AATTC
HindIII	A↓AGCTT
BamHI	G↓GATCC
BglI	GCCNNNN↓NGGC
PvuI	CGATC↓G
HaeIII	GG↓CC
MboI	GAT↓C

RFLP – difference in the lengths & number of fragments as a result of mutations in the recognition sites, VNTRs, other genetic disorders such as insertions, deletions, translocations & inversions.

Larger Words

Any other example of larger *k*-words ($k \geq 4$)?

- Modeling the no. of restriction sites in DNA

Some of the Qs that we can answer are:

- **If we were to digest the DNA with a RE such as EcoRI, approx. how many fragments would be obtained, and what would be their size distribution?**
- **Suppose we observed 761 occurrences of the sequence GCTGGTGG in a genome that is 50% G+C and 4.6Mb in size. How does this number compare with the expected no.? How would one find the expected no.? Expected according to what model?**

Larger Words

The pattern GCTGGTGG is known as Chi sequence:

5′-GCTGGTGG-3′ in *E. coli* (*k*=8)

- are significantly over-represented in particular genomes or on one or the other strand of the genome

e.g., Chi sequences occur 761 times in *E. coli* genome compared with approx. 70 instances predicted using base frequencies under iid model

Chi sequences are more abundant on the leading strand than on the lagging strand

- these observations relate in part to the involvement of Chi sequences in generalized recombination

Larger Words

Another example is the uptake sequences that function in bacterial transformation, e.g., (*k*=10)

5'-GCCGTCTGAA-3' in *Neisseria gonorrhoeae*

Some sequences may be under-represented, e.g.,

5'-CATG-3' occurs in the *E. coli* K-12 genome 1/20th of the expected frequency.

***k*-words ($k \geq 4$) are also useful for analyzing particular genomic subsequences, e.g., 6-word frequencies can be used to quantify the differences between *E. coli* promoter sequences and "average" genomic DNA.**

Transformation - genetic alteration of a cell resulting from direct uptake, incorporation & expression of exogenous genetic material from its surroundings and taken up through the cell membrane(s).

Summary and Applications

For word sizes, *k* = 1, 2, and 3, we saw that

- frequencies of words or statistics derived from them (*viz*., GC skew for *k* = 1) were not as predicted from independent, identically distributed (i.i.d.) base model.

This is not surprising – since genomes code for biological information, and we would therefore not expect the iid model to provide an accurate description for real genomes.

Summary and Applications

The frequencies of k-tuples have a number of applications:

- **GC skew - to predict locations of replication origins and termini in prokaryotes.**
- **k-mers (k ≥ 2) – to identify regions having aberrant base compositions that may indicate genome segments acquired by lateral transfer.**

e.g., if there is gene transfer from organism 1 (GC content: 50%) to organism 2 (GC content: 70%), then a simple measure such as GC content can be computed in sliding windows to identify the laterally transferred regions.

Summary and Applications

Parametric Methods based on anomalous nucleotide compositions for identifying horizontally transferred regions:

- **GC content anomalies – defined as the ratio of GC content in sliding window by GC content of the whole genome**
- **Genomic signature - set of di-nucleotide relative abundance values constitutes a "genomic signature" of an organism**

$$\delta^*(f_w, g) = 1/16 \sum | \rho^*_{xy}(f_w) - \rho^*_{xy}(g) | \qquad \rho^*_{xy} = f^*_{xy} / f^*_x f^*_y$$

- ***k*-mer distribution - average *k*-mer difference is defined as**

$$\delta^*_k(w, g) = \frac{1}{n} \sum_{i=1}^{n} \left| f_i^w - f_i^g \right|$$

***n* = 4^k no. of distinct *k*-words (*k = 2 to 9*) are computed, both for the whole genome & the sliding window.**

Summary and Applications

Parametric Methods at the gene level are:

- **Codon Usage Bias** - i.e., unequal usage of synonymous codons
- **Amino Acid Usage Bias** - deviation in the frequency of usage of individual amino acids over the average usage of all 20 AAs
- **GC Content at Codon Positions** - frequency of occurrence of G & C at 3 codon positions, GC1, GC2 & GC3 for the set of genes compared to the whole genome/2nd set of genes

- All these are based on k-mer analysis

Summary and Applications

The frequencies of k-tuples have a number of applications:

- **For eukaryotes, gene regions, in general, have a different base composition than non-genic regions,**

  e.g., human genes are relatively GC-rich compared with the genome as a whole, triplet frequencies different in coding region compared to non-coding regions.

- **Different gene classes have different codon usage frequencies, e.g., highly expressed genes.**

- **Codon usage differs from organism to organism – useful in identifying horizontally transferred genes.**

Summary and Applications

- **Sometimes the observed frequencies of k-words can be used to make inferences about DNA sequences.**

  For e.g., suppose that we are given a sequence string that hypothetically could be a portion of a candidate exon or a prokaryotic gene:

GACGTTAGCTAGGCTTTAATCCGACTAAACCTTTGATGCATGCCTAGGCTG

Simply by noting the frequency of stop codons in all the three reading frames, and knowing that a typical bacterial gene contains, on average, more than 300 codons, or that the typical human exon, contains around 50 codons, one can make a reasonable inference that this string does not code for a protein.

Summary and Applications

- ***k*-tuple (*k* > 3) frequencies can also assist in predicting whether an unannotated sequence is coding or noncoding.**

  Because coding sequences commonly specify AA strings that are functionally constraint, the distribution of *k*-tuple frequencies differ from that of noncoding sequences, e.g., intergenic or intronic sequences.

  e.g., consider in-frame hexamers (*k* = 6) – there are 4096 6-mers. From known polypeptide sequences we can easily predict that some 6-tuples will be infrequent, *viz*., the pair of residues Trp-Trp in succession is not common in protein sequences, implying that the corresponding dicodon, TTGTTG is likely to be relatively infrequent in coding sequences.

Summary and Applications

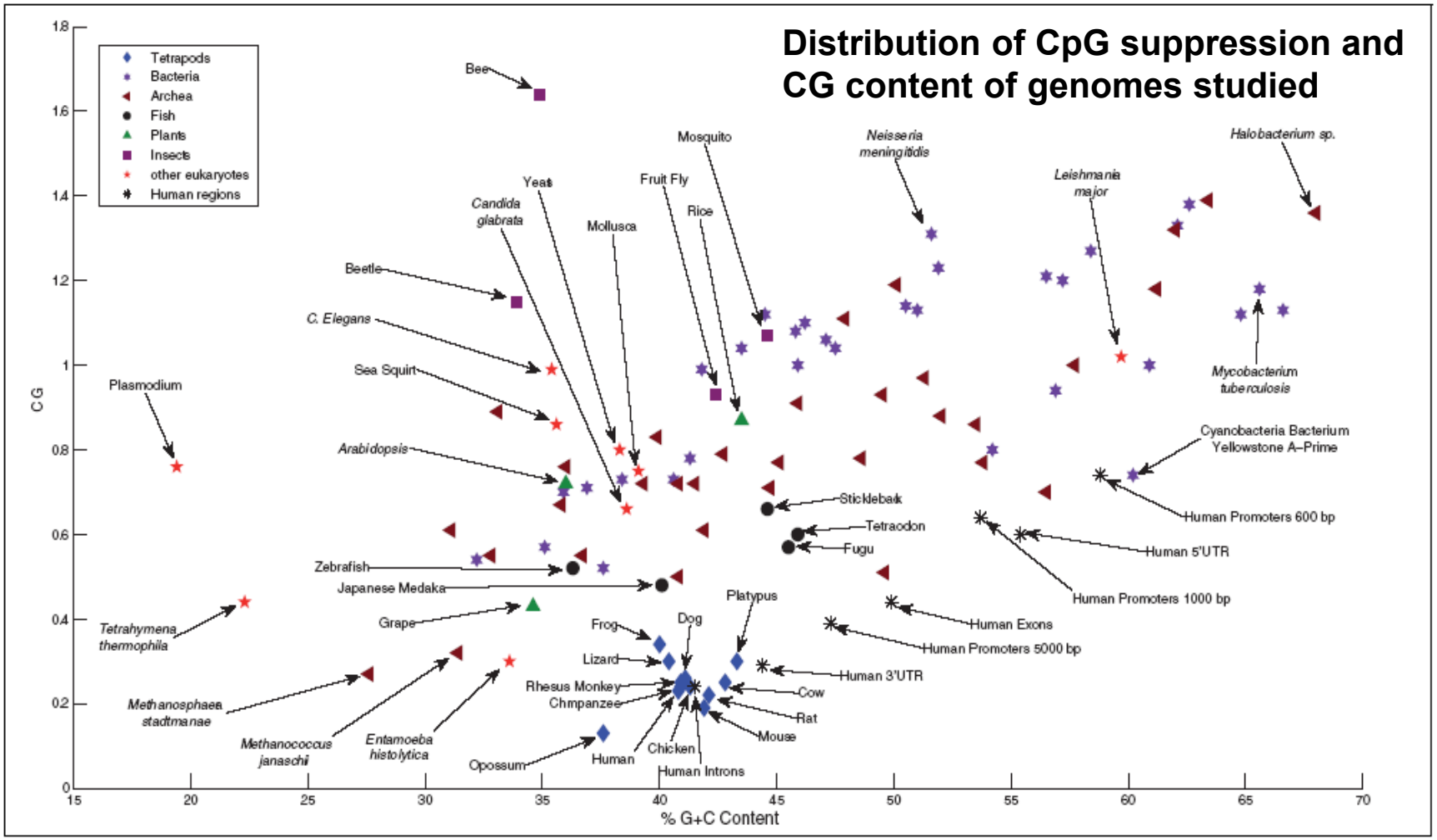
- **Predict gene expression:** using $k = 3$, compute CAI within ORFs to identify highly expressed genes (CAI ~1)
- ***k*-tuple frequencies and other content-based measures such as the presence of particular signals are among the statistical properties employed by computational gene finding tools.**

  Measures based on oligonucleotide counts:

  - Codon Usage
  - Amino Acid Usage
  - Codon Preference
  - Hexamer Usage

Summary and Applications

- ***k*-mer distributions are well-preserved among related strains/species - bacterial genomes can be clustered into natural groups according to *k*-mer distribution similarities.**



Reference

- **Computational Genome Analysis: An Introduction, R.C. Deonier, S. Tavare, M.C. Waterman (Chap-2).**

Mystery of the Chilean blob

- A 13-tonne blob containing no skin, bone or cells was washed ashore on a Chilean beach in July 2003
- Hypotheses ranged from remains of a giant squid, octopus, whale blubber, to some sea monster, alien
- DNA samples were extracted from the blob and sequenced (NADH2, control region)
- Search against the database unambiguously established the identity of the blob: sperm whale blubber

Pierce, S.K *et al.* 2004, *Biological Bulletinv*206, 125-133