DNA Sequence Evaluation
Part II

More Codon Usage Bias

Scaled χ^2

χ^2 measure. In statistics, the χ^2 statstic computes how different the distribution of values is from a uniform distribution.

Let $d = c_1 \dots c_N$ be a DNA string in nucleotide alphabet, and let L be the total number of codons in d that are not Methinine or Tryptophan (the number does include the stop codons though).

Let a be a k-degenerate amino acid, and let L_a be the total number of codons coding for a in d. Let $O_1, \dots O_k$ be the number of occurrences of the k different codons for a in d ($O_1 + \dots + O_k = L_a$).

χ^2 statistic for codon usage bias. The χ^2 value of codon usage bias for the amino acid a in DNA string d is computed as follows [2, 1]:

$$\chi^2_a = \sum_{i=1}^{k} \frac{(O_i - E)^2}{E},$$

where

$$E = \frac{L_a}{k}$$

is the expected number of codon occurrences assuming no bias.

Scaled χ^2. A **scaled** χ^2 statistic for codon usage bias for the amino acid a in DNA string d is

$$\hat{\chi^2_a} = \frac{\chi^2_a}{L}.$$

Range. Smaller values of χ^2 and scaled χ^2 mean little or no codon usage bias. Larger values mean larger bias.

Other Means of Evaluating DNA.

Information Enthropy

Information Enthropy. Consider a set $S = S_1 \cup S_2 \cup \ldots \cup S_k$ of items, where $S_i \cap S_j = \emptyset$ when $i \neq j$.

As $Pr(s \in S_i)$ we denote the probability that a randomly chosen element $s \in S$ will be from set S_i.

$$Pr(s \in S_i) = \frac{|S_i|}{|S|}.$$

The **enthropy** of the set S w.r.t. the partition $S_1, \ldots, S_k$ is defined as follows:

$$enthropy(S) = -\sum_{i=1}^{k} Pr(s \in S_i) \cdot \log_2(Pr(s \in S_i)).$$

Enthropy is measured in **bits**.

(**Note:** In this computation, we assume that $0 \cdot \log_2(0) = 0$.)

Properties of enthropy. The enthropy of a *homogenous* dataset in which $|S_i| = \frac{|S|}{k}$ for all sets S_i $\log_2 k$, i.e., the number of bits necessary to represent k.

$$enthropy(S) = -\sum_{i=1}^{k} \frac{1}{k} \cdot \log_2\left(\frac{1}{k}\right) = -\log_2\left(\frac{1}{k}\right) \cdot \sum_{i=1}^{k} \frac{1}{k} = \log_2 k$$

The enthropy of a dataset where only one set S_i of the k sets is non-empty is 0.

$$enthropy(S) = -\sum_{i=1}^{k-1} 0 \cdot \log_2 0 - 1 \cdot \log_2 1 = 0.$$

Enthropy measures the impurity of data. The higher the *enthropy*, the more *impure* the data is.

For DNA sequence analysis, information enthropy is used in the following way. Let $d = c_1 \ldots c_N$ be a DNA sequence fragment in a nucleotide alphabet. Let N_A, N_T, N_C and N_G be the total numbers of occurrence of nucleotides A, T, C and G respectively in d ($N_A + N_T + N_C + N_G = N$). Then the enthropy of the sequence d is computed as follows:

$$enthropy(d) = -\left(\frac{N_A}{N} \cdot \log_2\left(\frac{N_A}{N}\right) + \frac{N_C}{N} \cdot \log_2\left(\frac{N_C}{N}\right) + \frac{N_T}{N} \cdot \log_2\left(\frac{N_T}{N}\right) + \frac{N_G}{N} \cdot \log_2\left(\frac{N_G}{N}\right)\right).$$

High enthropy in a DNA sequence means higher sequence complexity. Lower enthropy means lower sequence complexity. (The simplest DNA sequence is one that consists of a single nucleotide).

Proposition. Let d be a DNA sequence and $\hat{d}$ be its reverse compliment. Then

$$enthropy(d) = enthropy(\hat{d}).$$

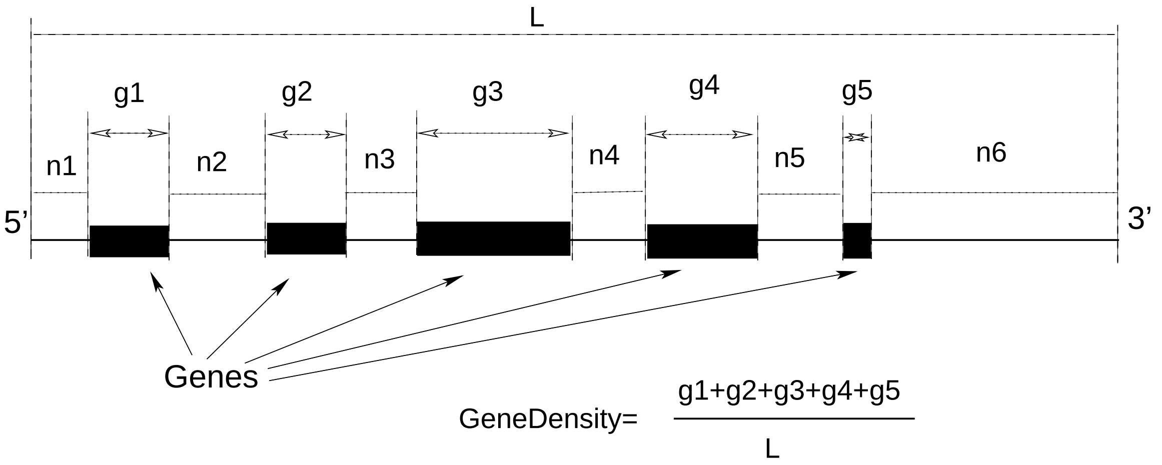

Figure 1: Gene Density.

Proof. Let N_A, N_T, N_C, N_G be the numbers of occurrences of A,T,C and G in d, and $\hat{N_A}, \hat{N_B}, \hat{N_C}, \hat{N_G}$ be the numbers of occurrences of A,T,C,G in $\hat{d}$. Since $\hat{d}$ is a reverse complement of d:

$$N_A = \hat{N_T}; N_T = \hat{N_A};$$

$$N_C = \hat{N_G}; N_G = \hat{N_C}.$$

Using these equalities and plugging the values into the enthropy formula, we obtain the desired result.

Gene Content

Gene content of DNA is the name of a collection of measures that evaluate the frequency and the size of genes, their components (introns and exons) and the intergenic regions (regions of DNA between genes).

Notation. Let $d = c_1 \dots c_N$ be a DNA string in a nucleotide alphabet. Let $d = d_1 e_{11} i_{11} e_{12} i_{12} \dots e_{1s_1} d_2 \dots d_k g_{k1} \dots e_{ks_k} d_{k+1}$, where

1. d_is are non-coding intragenic regions, e_{lj} are *exons*, i.e., coding regions
2. e_{lj} is the jth exon of lth gene in d)
3. i_{lj} are *intron*, i.e., *non-coding DNA regions* separating exons of the same gene.

Let

$$N_e = |e_{11}| + |e_{12}| + \dots |e_{ks_k}|$$

be the number of base pairs in the exons,

$$N_o = |d_1| + |d_2| + \dots + |d_{k+1}|$$

be the number of base pairs in the non-coding intragenic regions,

$$N_i = |i_{11}| + |i_{12}| + \ldots |i_{ks_{k-1}}|$$

be the number of base pairs in all introns,

$$N_{nc} = N_o + N_i$$

be the total number of non-coding base pairs in the DNA fragment, and, finally,

$$N_g = N_e + N_i$$

be the total length of genes in the DNA fragment[1].

In density computations, **both** genes expressed on the top **and** the bottom strands are considered: so e_{lj}s in the notation above refer to **all** coding regions from both strands.

We let k represent the total number of genes in d and q be the total number of exons.

The following measures are used for tracking gene density.

Average gene size. The average gene size, $Avg_g(d)$ is computed as follows:

$$Avg_g = \frac{N_e + N_i}{k} = \frac{N_g}{k}.$$

Average coding DNA sequence (CDS) size. The average coding DNA sequence size, $Avg_c(d)$, only counts exon lengths in computations:

$$Avg_c(d) = \frac{N_e}{k}.$$

Average exon size. The average exon size, $Avg_e(d)$, uses q, the total number of exons, in the denominator:

$$Avg_e(d) = \frac{N_e}{q}.$$

Average intron size. A gene with l exons contains $l-1$ introns. A DNA sequence with k genes and q exons has $p = q - k$ introns. The average intron size, $Avg_i(d)$, is computed as follows:

$$Avg_i(d) = \frac{N_i}{p} = \frac{N_i}{q-k}.$$

Nucleotides to genes ratio. This measure, denoted $ratio(d)$ is defined as

$$ratio(d) = \frac{N}{k}.$$

[1] Note the difference between the notions of "length of a gene" and "length of all exons of a gene".

Usually, it is measured in Kilo-base pairs per gene, so the exact computation would be:

$$ratio(d) = \frac{N}{1000 \cdot k}.$$

$$Avg_g = \frac{N_e + N_i}{N_e + N_i + N_{nc}} = \frac{N_g}{N_g + N_{nc}} = \frac{N_g}{N}$$

Gene nucleotide fraction. The gene nucleotide fraction of d, denoted $density_g(d)$, is the percent of base pairs in the genes:

$$frac_g(d) = \frac{N_e + N_i}{N_e + N_i + N_{nc}} = \frac{N_g}{N_g + N_{nc}} = \frac{N_g}{N}$$

Coding nucleotide fraction. The coding nucleotide fraction of d is the percent of base pairs in the exons:

$$frac_e(d) = \frac{N_e}{N_e + N_i + N_{nc}} = \frac{N_e}{N}.$$

Exon density. The exon density of d is the total number of coding regions (exons) in d divided by the length of d:

$$\mathsf{density}_e(d) = \frac{q}{N}.$$

Gene density. The gene density of d is the total number of genes in d divided by the length of d:

$$\mathsf{density}_g(d) = \frac{k}{N}.$$

Since genes/coding regions typically span $1 - 10$ Kbp (Kilo-base pairs), density is usually expressed as the number of genes per 10Kbp, 100Kbp, or 1Mbp. Alternatively, a reciprocal value can quantify average gene spacing: DNA length in nucleotides divided by the number of genes present:

$$\mathsf{spacing}(d) = \frac{N}{k}.$$

Relative gene coverage. This measure, denoted $coverage_g(d)$, is the ratio of the average gene size and the toal length of the DNA sequence:

$$coverage_g(d) = \frac{Avg_g(d)}{N} = \frac{N_g}{k \cdot N}.$$

Relative exon coverage. Same as relative gene coverage, but for exons only:

$$coverage_e(d) = \frac{Avg_e(d)}{N} = \frac{N_e}{k \cdot N}.$$

References

[1] Estuko Moriyama, (2003), Codon Usage, in *Encyclopedia of the Human Genome*, Macmillan publishers, Ltd.

[2] D. Shields, P. Sharp, D. Higgins and F. Wright (1988) Silent sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. *Molecular Biology and Evolution*, Vol. 5, pp. 704716.