AUG
Faculty of Computer and Information Sciences
Ain Shams University

Session 6
More to GC-content

GC skew

GC skew is when the nucleotides guanine & cytosine are over- or under-abundant in a particular region of DNA or RNA.

In most bacteria, nucleotide compositions are asymmetric between the leading strand & the lagging strand: the leading strand contains more guanine (G) & thymine (T), whereas the lagging strand contains more adenine (A) & cytosine (C).

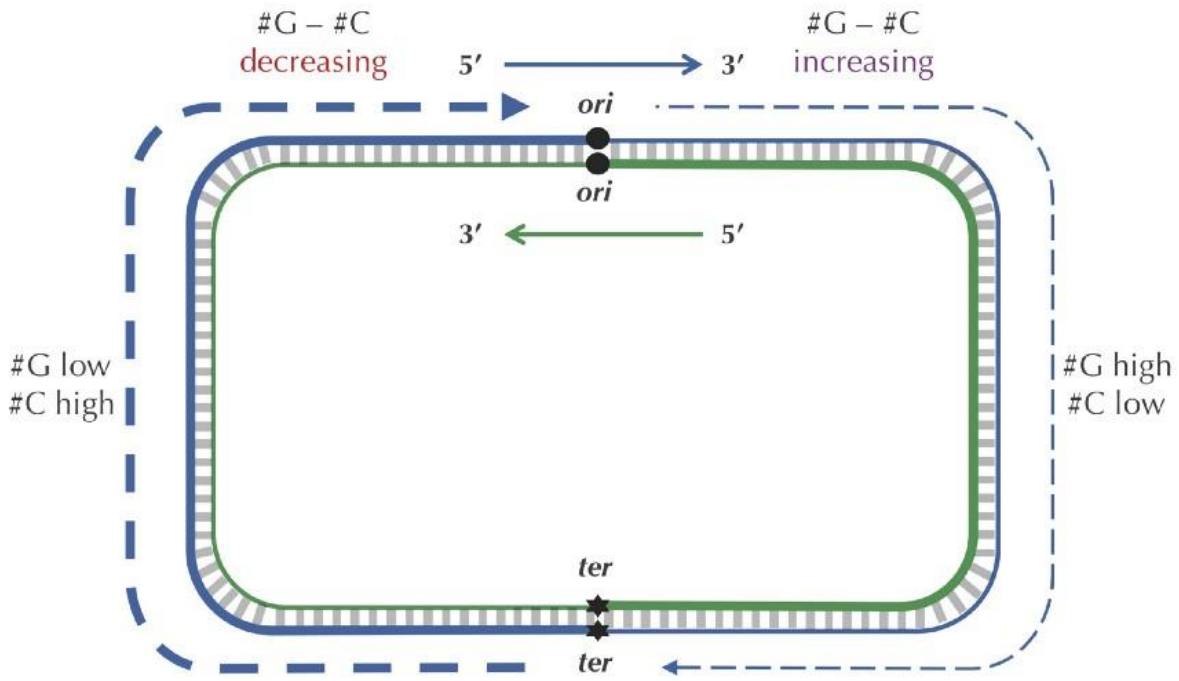
This phenomenon is referred to as GC and AT skew. It is represented mathematically as follows:

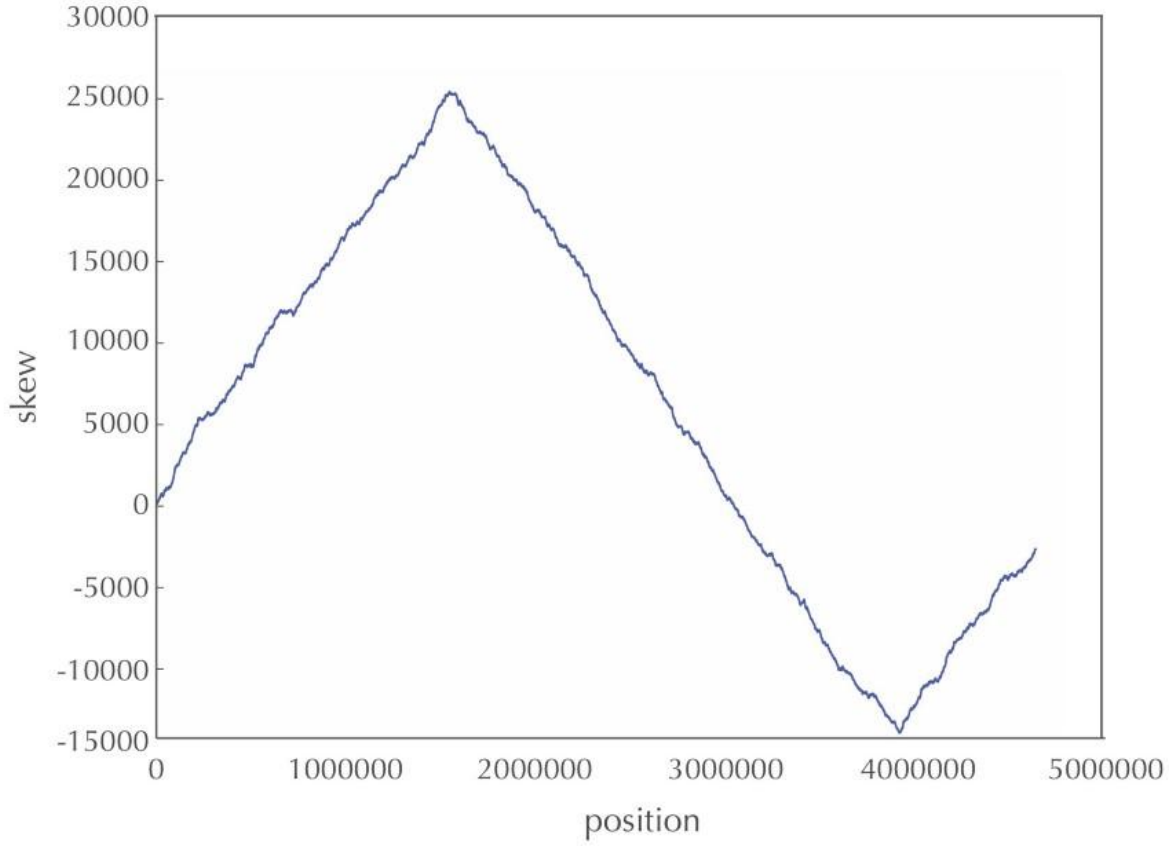
GC skew = (G − C)/(G + C)

AT skew = (A − T)/(A + T)

Why do we compute it?

The GC skew is useful as the indicator of the DNA leading strand, lagging strand, replication origin, & replication terminal.
The GC skew is +ve in the leading strand & -ve in the lagging strand. Therefore, it is expected to see a switch in GC skew sign at the point of DNA replication origin and terminus, where the maximum value of the GC skew corresponds to the terminal, and the minimum value corresponds to the origin of replication.





Maximum point: Replication terminal.
Minimum point: Replication origin.

K-mers

In bioinformatics, k-mers are substrings of length k contained within a biological sequence, composed of nucleotides (A, T, G, & C).
Usually, the term k-mer refers to all of a sequence's subsequences of length k, such that a sequence of length L will have L-K+1 k-mers.

How do they work?

e.g. Sequence 'AGAT' would have:
(L = 4 & K = 1, 2, 3, 4)

- Four monomers (4-1+1=4) -> (A, G, A, and T)
- Three 2-mers (4-2+1=3) -> (AG, GA, AT)
- Two 3-mers (4-3+1=2) -> (AGA and GAT)
- One 4-mer (4-4+1=1) -> (AGAT)

K-mer Composition

The k-mer composition of a string S encodes the number of times that each possible k-mer occurs in S.
To represent the k-mer composition of a string concisely, all possible k-mers (in the case of DNA strings, there will be 4^k total k-mers) are ordered lexicographically, and then an array A is created in which A[i] represents the number of times that the ith of these ordered k-mers appears in S.

e.g. The 2-mer composition of "TTGATTACCTTATTTGATCATTACACATTGTACGCTTGTGTCAAAATAT CACATGTGCCT" would be:
(4^k = 4^2 = 16 2-mers)

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
3	5	0	8	6	2	1	3	2	2	0	4	5	3	7	8

Why do we use them?

Primarily used within the context of computational genomics and sequence analysis, improve heterologous gene expression, identify species in metagenomic samples, & create vaccines.