Definition

A short word or pattern can appear many times within a genomic sequence.

An example of a very short, one-letter pattern in molecular biology is the guanine cytosine content of a fragment of DNA, accounting for how often the patterns `G` and `C` occur within a given DNA sequence.

Further examples of patterns in DNA sequences are the TATA box found in the promoter region of most eukaryotic genes (an occurrence of the `TATAAAA` pattern within the sequence) and the Pribnow box found in the promoter region of most prokaryotic genes (an occurrence of the `TATAAT` pattern within the DNA sequence).

Definition

A nucleic acid or amino acid sequence can be seen as composed of a number of possibly overlapping k-mers or words of length k, for a certain $k \geqslant 1$.

The k-mer composition of a sequence is given by the frequency with which each possible k-mer occurs within the sequence.

The 1-mer composition is related to the GC content of a DNA sequence, and the 2-mer, 3-mer, and 4-mer compositions are also known as the di-nucleotide, tri-nucleotide, and tetra-nucleotide compositions of a DNA sequence.

Example

The fragment of DNA sequence

```
TTGATTACCTTATTTGATCATTACACATTGTACGCTTGTGTCAAAATATCACATGTGCCT
```

has the following 1-mer, 2-mer, and 3-mer compositions:

A	16
C	12
G	8
T	24

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

Example

AAA	2	AAC	0	AAG	0	AAT	1
ACA	3	ACC	1	ACG	1	ACT	0
AGA	0	AGC	0	AGG	0	AGT	0
ATA	1	ATC	2	ATG	1	ATT	4
CAA	1	CAC	2	CAG	0	CAT	3
CCA	0	CCC	0	CCG	0	CCT	2
CGA	0	CGC	1	CGG	0	CGT	0
CTA	0	CTC	0	CTG	0	CTT	2
GAA	0	GAC	0	GAG	0	GAT	2
GCA	0	GCC	1	GCG	0	GCT	1
GGA	0	GGC	0	GGG	0	GGT	0
GTA	1	GTC	1	GTG	2	GTT	0
TAA	0	TAC	3	TAG	0	TAT	2
TCA	3	TCC	0	TCG	0	TCT	0
TGA	2	TGC	1	TGG	0	TGT	4
TTA	3	TTC	0	TTG	4	TTT	1

Algorithm

The k-mer composition of a sequence can be computed by first obtaining all the sequences of length k over the alphabet Σ and then counting the number of occurrences of each possible k-mer within the given sequence.

The sequences of length k over the alphabet Σ are obtained using the words algorithm, and $S[i..i+k-1]$ denotes the word of length k starting at position i of sequence S.

Algorithm

```
function word_composition($S, k, \Sigma$)
    $L \leftarrow$ words$(k, \Sigma)$
    for each word $w$ of $L$ do
        freq$[w] \leftarrow 0$
    $n \leftarrow$ length$(S)$
    for $i \leftarrow 1$ to $n - k + 1$ do
        $w \leftarrow S[i..i + k - 1]$
        freq$[w] \leftarrow$ freq$[w] + 1$
    return freq
```

Definition

The similarities and differences between two biological sequences can be assessed by computing a distance measure between the two sequences.

The alignment free distance is based on the word composition of the sequences.

While similar DNA sequences have similar GC content, the *k*-mer frequencies for larger values of *k* reveal similarities and differences between two sequences.

Example

The fragments of DNA sequence

```
CCCCAATATGGGCGCGACCCCCCGGAATCTCTATTCACCAGCTT      (1)
CCCCAATATGGGCGCGACCCCCCGGAATCTGTCTCCGCCAGCCT      (2)
CCCCAATATGGGCGCTACTTTCACAATAACCCACTAGACAGCCT      (3)
```

have the following 1-mer and 2-mer frequencies:

word	A	C	G	T
(1)	9	18	8	9
(2)	7	20	10	7
(3)	13	16	6	9

word	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
(1)	2	2	1	4	3	9	3	3	2	3	3	0	2	3	1	2
(2)	2	1	1	3	2	11	4	3	2	4	3	1	1	3	2	0
(3)	3	5	2	3	5	6	1	4	1	3	2	0	4	1	1	2

Definition

The extent to which the *k*-mer frequencies of two sequences differ can be measured in a number of ways, such as by computing their covariance or correlation.

The alignment free distance between two sequences is given by the linear correlation coefficient of their *k*-mer frequencies.

Example

The alignment free distance between each pair of the DNA sequences

CCCCAATATGGGCGCGACCCCCCGGAATCTCTATTCACCAGCTT (1)

CCCCAATATGGGCGCGACCCCCCGGAATCTGTCTCCGCCAGCCT (2)

CCCCAATATGGGCGCTACTTTCACAATAACCCACTAGACAGCCT (3)

is as follows.

sequences		$k = 1$	$k = 2$	$k = 3$	$k = 4$
(1)	(2)	0.9453431	0.9205409	0.8260229	0.7631025
(1)	(3)	0.8081352	0.6153795	0.5388881	0.3561884
(2)	(3)	0.6148987	0.4210917	0.3995599	0.3561884

The first sequence is thus more similar to the second sequence than to the third sequence, no matter the word length *k* used to assess similarities.

Algorithm

Given the *k*-mer composition of two biological sequences, their alignment free distance can be obtained by computing the linear correlation coefficient of the *k*-mer frequencies, that is, by dividing the covariance of the *k*-mer frequencies by the product of their standard deviations.

```
function alignment_free_distance($S_1, S_2, k, \Sigma$)
    $F_1 \leftarrow$ word_composition($S_1, k, \Sigma$)
    $F_2 \leftarrow$ word_composition($S_2, k, \Sigma$)
    $cov \leftarrow$ covariance($F_1, F_2$)
    $sd_1 \leftarrow$ standard_deviation($F_1$)
    $sd_2 \leftarrow$ standard_deviation($F_2$)
    return $cov/(sd_1 sd_2)$
```