

Outline

1 n-grams (k-mers)

n-grams (k-tuples) are vectors of n characters derived from input sequence(s). They may form continuous sub-sequences or be discontinuous.

Important n-gram parameter is its position. Instead of just counting n-grams, one may want to count how many n-grams occur at a given position in multiple (e.g. related) sequences.

	P1	P2	P3	P4	P5	P6
S1	C	T	T	A	G	C
S2	C	A	G	A	C	G
S3	G	T	G	A	T	T

Sample sequences. S - sequence, P - position.

	A	C	G	T
S1	1	2	1	2
S2	2	2	2	0
S3	1	0	2	3

1-gram counts.

	P1	P2	P3	P4	P5	P6
S1	C	T	T	A	G	C
S2	C	A	G	A	C	G
S3	G	T	G	A	T	T

Sample sequences. S - sequence, P - position.

	AA	CA	GA	TA	AC	CC	GC	TC	AG	CG	GG	TG	AA
S1	0	0	0	1	0	0	1	0	1	0	0	0	0
S2	0	1	1	0	1	0	0	0	1	1	0	0	0
S3	0	0	1	0	0	0	0	0	0	0	0	1	0

2-gram counts.