

Step1 Search instances of Kmers (length $K \in [12, 22]$) in CNEs allowing a given number (M) of mismatches.
 $M=0$ for $K \leq 13$; $M=1$ for $K \in [14, 16]$; $M=2$ for $K \in [17, 18]$; $M=3$ for $K \geq 19$

For instance, search TTCAGCACCGGACAGA ($K=18$) with $M=2$:

```
seq1 GCGATAGGAGTCCATTTCAGCACCTTGGACAGAGCCAACGGATTTGTCCGA
seq2 GCGGGCAGCGGGCGCCTCCTTCAGCACCGCGGACAGCGCCAGGCCAGTG
seq3 CCTCGGCCTTTCAGCACCGAGGACAGAGCCTCGCTCCCCGCCCGGAGCTA
.
.
.
seqC-1 CCACGCGATGGTAGCACCAACTGGGTCTTCAGCACCGTGGACAGAGCCAG
seqC AGCTAGATATGACGTTCAGCACCGGACAGCGCCAGCACACCTGCCAG
```

⇒ Total number of instances in CNEs:
 $C=83$

Step2 Collect Kmers with $C \geq 30$ and find the number of instances (N) in the entire human genome.

For instance, searching TTCAGCACCGGACAGA with $M=2$ in the human genome (size $G=3077\text{Mb}$) gives rise to $N=433$ instances.

Step3 Calculate enrichment score for each of the Kmers: a) Fold enrichment and b) Z-score.

For instance, given the total number of all possible 18-mers in CNEs $S=25\text{Mb}$, we estimate that the expected number of sites for TTCAGCACCGGACAGA in CNEs is $\mu=S/G*N=3.5$ sites.

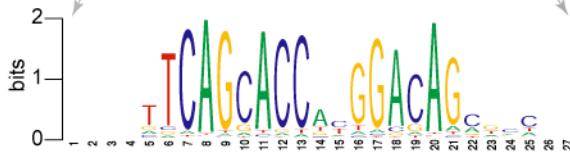
Fold enrichment: $SNR=C/\mu=23.4$

Z-score: $Z=(C-\mu)/\text{sqrt}(\mu)=42.4$

Step4 Keep Kmers with $F \geq 5$ and $Z \geq 10$. Cluster and align Kmers with similar nucleotide sequences.

Kmer	C	SNR	Z
. . . . TTCAGCACCGGACAGA	83	23.4	42.4
. . . . TTCAGCACCGGACAG	137	14.6	41.8
. . . . ATTCAGCACCGGACAGCG	76	21.0	38.2
. . . . TCAGCACCGGACAGCGAC	77	18.4	35.7
. . . . GCCTTCAGCACCGGACAG	61	17.1	30.6
. . . . TCAGCACCGGAGAGCGCC	62	16.9	30.6
. . . . TCAGCACCGGACAGCGCCC	58	16.7	29.4
. CAGCACCGGACAGCGCCCA	33	11.8	18.1
AGCGTTCAGCACCGGACA	30	21.1	24.0

Step5 Construct positional specific scoring matrix from the aligned Kmers.



Step6 Search the human genome using the positional specific scoring matrix and identify all instances.