Definition 1. *Number of occurences of string p in string s:*

$$occ(p, s) = \#\{i | s_i s_{i+1} ... s_{i+|p|-1} = p\}$$

where $|p|$ is length of string p
and $\#\{.\}$ is cardinality of this set.
/media/Storage/bioinf/BIM/bioalgo-task

Definition 2. *k-distance between two strings s_1 and s_2 is the difference in k-mers frequencies:*

$$d(s_1, s_2, k) = \sum_{p \in \{A,C,G,T\}^k} abs\left(\frac{occ(p, s_1)}{|s_1|} - \frac{occ(p, s_2)}{|s_2|}\right)$$

where $\{A, C, G, T\}^k$ is ACGT-string of length k,
$|s|$ is a number of valid (without N-nucleotide) k-mers in string s (if there is no N, it's $length(s) - k + 1$,
and $abs(.)$ is absolute value.

It's obvious that $0 \leq d(s_1, s_2, k) \leq 2$.

Problem 1. *Given genomic string $G = a_1 a_2 ... a_n$ where $\forall i : a_i \in \{A, C, G, T\}$.*
Find the set of indexes *$K = \{k_1, k_2, ...k_m\}$ such that:*

- *$m \geq 7$ is unknown (number of indexes in set K),*
- *$k_1 = 1$, $k_m = n + 1$,*
- *$\forall i < j : k_i < k_j$,*

I.e this indexes divide string G into $m - 1$ substrings g_1, g_2, ..., g_{m-1} — each from k_i to $k_{i+1} - 1$ positions (inclusive).

$$g_i = a_{k_i} a_{k_i+1} a_{k_i+2} ... a_{k_{i+1}-2} a_{k_{i+1}-1}$$

Maximize the following score among all possible sets K:

$$Score(K, G) = 2^{7-m} \cdot \sum_{i=1}^{m-2} d(g_i, g_{i+1}, 2) + d(g_i, g_{i+1}, 3)$$

Data: Homo sapiens (Reference Genome: GRCh37), Chromosome 1.

Motivation behind: find large genomic regions with high difference in dinucleotide and codon frequencies.