Liubov Shilova and Richard Labri

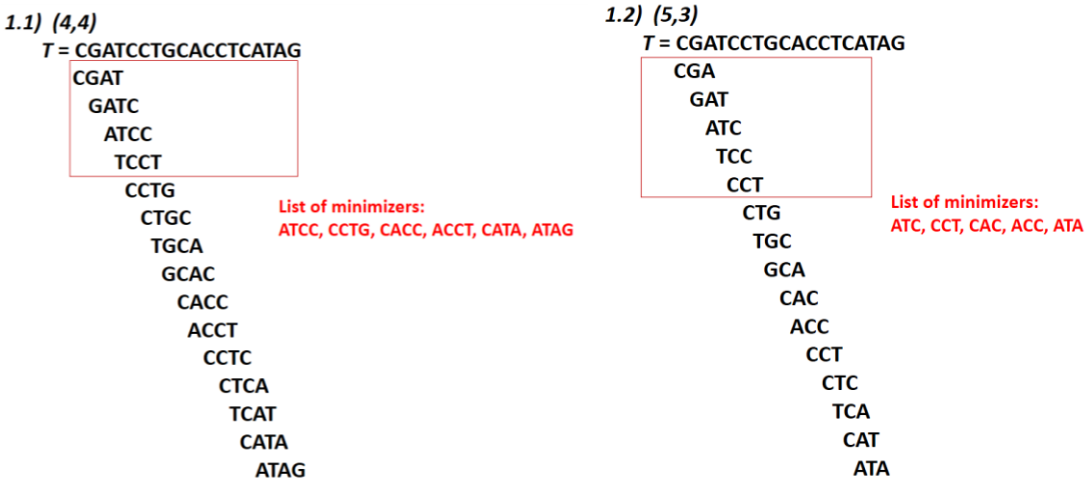
We both are new students and haven't received our matriculation numbers yet.

Ex. 1. Minimizers

The answers are:

1) ATCC, CCTG, CACC, ACCT, CATA, ATAG
2) ATC, CCT, CAC, ACC, ATA

Which is also illustrated below:



3) With a fixed k we need to have a window of $w <= k$ in order to be sure, that we cover the whole sequence.
   It can be easily seen on example from exercise (1.1): The first minimizer that we have is ATTC. As the first letter is A, it stays the biggest the whole time it is in the window. Once the window passes it (is located right below) the next biggest k-mer may appear in the bottom of the window (w positions later). If the window is larger than k, this kmer will not include letters between it and previous k-mer.
   Thus, we need a window equal or smaller than k.

Ex. 2. Hamming distance

The Hamming distance is defined as follows:

$$\hat{d}(P_1, P_2) = \frac{d(P_1, P_2)}{||P||},$$

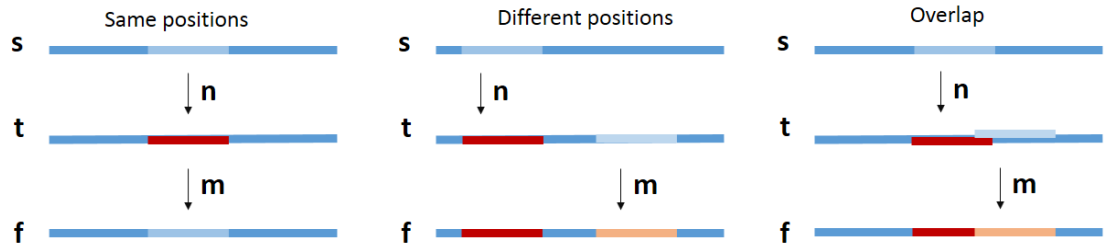
Where d(P1, P2) is a number of positions, where 2 strings of the same length differ and ||P|| is a length of a string.

1) Neither number of positions, nor length of a string can be negative. Thus, $d(x, y) >= 0$ (no negative distances)
2) If hamming distance equals to zero, it means number of changes is equal to zero: d(P1, P2) = 0. It can happen if and only if two strings are identical and there is no position i, where they differ. Thus, $d(x, y) = 0$ if and only if $x = y$ (distances are positive unless two points are identical)

3) Compared strings have the same length (from definition). If the string s is different from the string t in position i, it immediately means that t is also different from t in position i. Thus, $d(x, y) = d(y, x)$ (distance is *symmetric*)
4) Let there be 3 strings of the same length *l*: s, t, f. Let the hamming distance be:
   $d(s,t) = d_{st}$,
   $d(t,f) = d_{tf}$
   This means that the number of different positions, where s and t differ is d(s,t)*l. For t and f: d(t,f)*l.
   We make n changes in string s to get a string t. Then we make m changes to make a string f out of t.
   1. Suppose s and f are equal and two times we changed the string in the same positions (n = m). Thus, d(s,f) = 0 and since we proved that d is never negative, d(s,f)<= d(s,t)+d(t,s)
   2. Suppose we changed strings in different positions. And all n positions are different from all m positions. Then d(s,f) = d(s,t)+d(t,s)
   3. If part of the m positions, that we change, are from n, and part – new, it means, that the resulting number of changes from s to f is less than n+m.

   Thus, $d(x, y) <= d(x, z) + d(z, y)$ (the *triangle inequality*)

The proof of (4) can be clear from the picture below:



Ex. 3. Semi-global alignment, Ukkonen's trick

T = TGAATCAGG
query *S* =AATCG

match = 0, mismatch = 1, gap = 1

k = 1

		T	G	A	A	T	C	A	G	G
	0	0	0	0	0	0	0	0	0	0
A	1	1	1	0	0	1	1	0	1	1
A	2	2	2	1	0	1	2	0	1	2
T	3	2	3	2	1	0	1	1	2	2
C	4	3	3	3	2	2	0	1	2	3
G	5	4	3	4	3	3	1	1	1	1