AG Genominformatik

Sequence Analysis 3
Summer 2024

Exercises

Number 3, Discussion: 2024-05-02

For these exercises, do not consider the reverse complement of k-mers.

Notations:

- The i-th character of a string s is noted s[i].
- The concatenation of two strings is noted "·". Exemple: "A"·"C" = "AC"
- The concatenation of two integers is also noted "·".

Let **R** be the set of read {GTAGAGCTG, TCGAGCTGTG, GAGAGCTGT}.

- Compute the set of all 7-mers present in **R**. How many letters do you need to represent this set?
- Draw the associated de Bruijn graph.
- Compute the set of unitigs from the 7-mers of **R**. Let's call it **U.** How many letters do you need to represent **U**?

Say we are given a function **code**, that maps characters to integers, such that:
code(A) = 1
code(C) = 2
code(T) = 3
code(G) = 4

Let's define a hash function **h,** that hashes a string s (s = s[1]·s[2]·…·s[n]), by simply concatenating the code of each character of s (h(s) = code(s[1])·code(s[2])·…·code (s[n])). Exemple: h(AACTG) = 11234.

- Using BBHash, compute an MPHF on the 7-mers of **R**. Use an array of size 7. In case of collisions, add another array of size 6, then another of size 5, etc.
- Write down the hash of every 7-mer.
- Using these hashes, write an array that maps each 7-mer of **R** to the unitig in which it appears.
- Using the MPHF you computed, the array you just built, and **U,** search for the sequence **Q** = GGCGAGCTGTGGG in the sets of reads. What is the proportion of shared 7-mers between **Q** and **R**?