Mansoura University
Faculty of Computers and Information
Department of Computer Science
First Semester: 2020-2021

[MED121] Bioinformatics: Rabin-Karp Algorithm

Grade: Third Year (Medical Informatics Program)

Sara El-Metwally, Ph.D.

Faculty of Computers and Information,

Mansoura University,

Egypt.

AGENDA

- Rabin-Karp Algorithm Basic Idea
- Rabin-Karp Algorithm Trace
- Worst Case/Best Case Running Time

RABIN-KARP ALGORITHM

- Naïve algorithm slides the pattern **P** over the text **T** one by one.

- **The basic idea behind Rabin-Karp's algorithm is:**
  - **The pattern P and substrings of text T are represented as numbers**, called **hash values**.
  - Instead of comparing two strings, we **will compare two integers**.

NUMERIC REPRESENTATION OF STRINGS

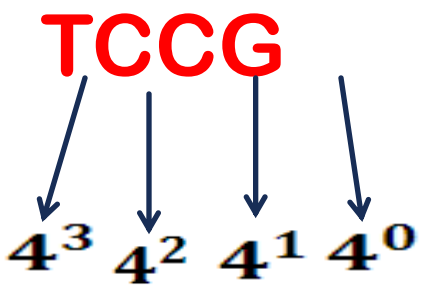
- Let $\Sigma = \{a_1, a_2, \ldots a_k\}$, we associate with each character a_i a digit $i-1$ in base k numeric system.

- Example: How to decode DNA alphabet ?

  DNA alphabet $\Sigma = \{A, C, G, T\}$
  k=4
  A=0,C=1,G=2,T=3

NUMERIC REPRESENTATION OF STRINGS

- Decode the string "**TCCG**"



$$= 214$$

$$T \times 4^3 + C \times 4^2 + C \times 4^1 + G \times 4^0$$

$$T \times 64 + C \times 16 + C \times 4 + G \times 1$$

$$3 \times 64 + 1 \times 16 + 1 \times 4 + 2 \times 1$$

M-SUBSTRINGS

- If you have a string = “ATTCCGTTGTTA”, Construct l-substrings and how many?

|T|=12, l=6

|T| - l +1

12 - 6 +1 = 7

ATTCCG

TTCCGT

TCCGTT

CCGTTG

CGTTGT

GTTGTT

TTGTTA

kmers, overlap by k-1

M-SUBSTRINGS

- If you have a string = "ATTCCGTTGTTA", Construct l-substrings and how many?

ATTCCG
TTCCGT
TCCGTT
CCGTTG
CGTTGT
GTTGTT
TTGTTA

kmers, overlap by k-1

ATTCCG
TTCCGT

k=6, overlap by k-1=5

RABIN-KARP ALGORITHM

- Compute hash value of P
- Compute hash value of first n-substring from T
- If both values matches, then go to compare characters.
- If not matches, go to the next n-substring from T

RABIN-KARP ALGORITHM

- Text = “ATTCCGT”
- Pattern =“ TCCG”

We need efficient calculations
O(1)

h(TCCG)= 214

h(ATTC)= 61 ✖

h(TTCC)= 245 ✖

h(TCCG)= 214 ✔ → Compare Letters

h(CCGT)= 91 ✖

RABIN-KARP ALGORITHM

$t_0 = h(ATTC) = 61$

$$A \times 4^3 + T \times 4^2 + T \times 4^1 + C \times 4^0 = 61$$

$t_1 = h(TTCC) = 245$

$$T \times 4^3 + T \times 4^2 + C \times 4^1 + C \times 4^0 = 245$$

$$t_{new} = K \times (t_{prev} - K^{l-1} \times S_{del}) + S_{add}$$

$$t_1 = 4(t_0 - 4^{4-1} \times A) + C$$

$$t_1 = 4(61 - 64 \times 0) + 1 = 245$$

RABIN-KARP ALGORITHM (PROBLEMS)

- Collision might occur when two different strings hash to the same value.
- When the length of Pattern n, and the size of alphabet k are big enough, the hash values become too large to fit into standard type integer.
- To over come this, instead of taking the hash value H, we will use its remainder when divided by a prime number q.
- When q is sufficiently large , q=101,13,11, it is less likely for collision to happen.

RABIN-KARP ALGORITHM (PROBLEMS)

$$4^{50} = 1267650600228229401496703205$$

- short int x = value in 2 bytes
- value in 2 bytes = $0 \; to \; 2^{15} - 1$ 32767

- Modification: pick a number q, a prime number and perform the computations of p and ti values mod q

RABIN-KARP ALGORITHM

- Text = “ATTCCGT”
- Pattern =“ TCCG”
- q=11

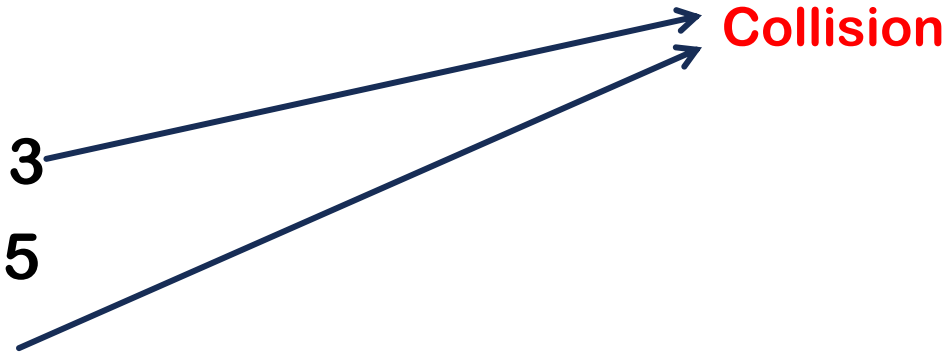
h(TCCG)= 214 mod 11 = 5

h(ATTC)= 61 mod 11= 6

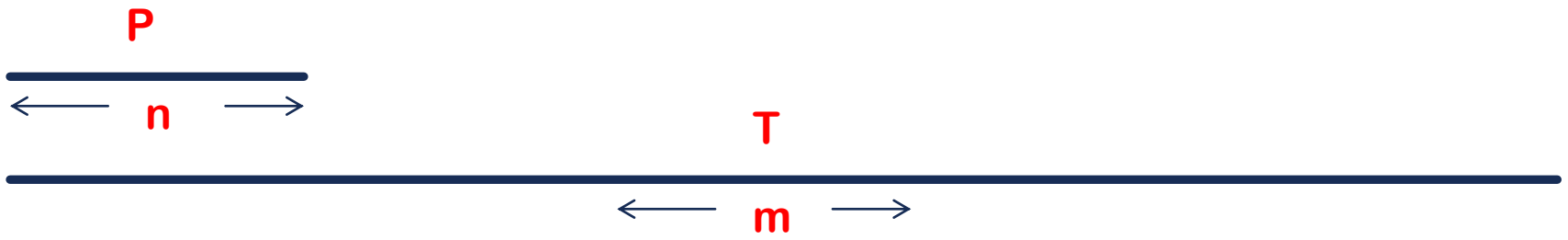
h(TTCC)= 245 mod 11 = 3

h(TCCG)= 214 mod 11= 5

h(CCGT)= 91 mod 11 =3



Q



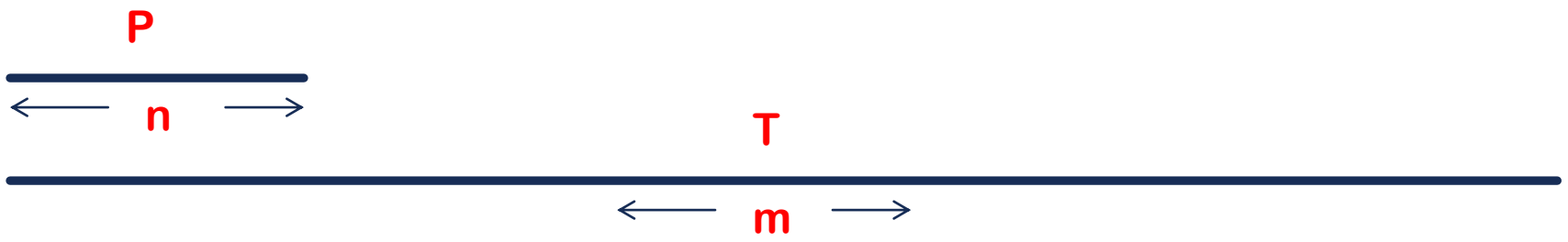
- What is the worst case running time?

  O(mn)

  T= aaaaaaaa

  P=aaaa

Q



- What is the best case running time?

  O(m)

  T= aaaaaaaa

  P=abbb

Thank you!