

Relationship between Edit Distance and Q-Gram Distance

Theresa Lye

Mentor: Alexander Schliep

Co-mentor: Pavel Mahmud

June 8th, 2012

Motivation: Read Mapping to Reference Genomes

- ▶ DNA sequencing

...AGCTGCCATTGCT...

adenine guanine thymine cytosine



Motivation: Read Mapping to Reference Genomes

- ▶ Read mapping (aligning) to reference genome: matching shorter sequences (reads) to a longer reference sequence



TCAGTCGTAATATACG — TAAGCCATGCGCTCGTACGTTAC
ATGTGCTAA — AT

Motivation: Read Mapping to Reference Genomes

- ▶ Researching a possible faster method for read mapping that utilizes q-gram distance
- ▶ Want to analyze the difference between edit distance and q-gram distance

Edit Distance

- ▶ Minimal number of edits (insertions, deletions, substitutions) required to transform one character string into another.
- ▶ What is the edit distance of ACGTAAGCCAT to ATGTGCTAAAT?

ACG ——— TAAGCCAT
ATGTGCTAA ——— AT

Edit Distance:
7

Q-Gram Distance

- ▶ q-gram: substring of length q
- ▶ q-gram profile: vector containing the counts of the q-grams found in a string

Q-Gram Distance

- ▶ Consider a string in binary (only 0's and 1's)
- ▶ Possible 2-grams are: 00, 01, 10, and 11

01001 , 10110

the 2-gram profiles of these strings are:

(1, 2, 0, 0) , (0, 0, 2, 1)

Q-Gram Distance

- ▶ Q-gram distance between two strings = L1 norm of the difference of their q-gram profiles
- ▶ L1 norm: take the absolute difference between each column, and sum the result

Q-Gram Distance

- ▶ The two strings: 01001 and 10110
- ▶ Their q-gram profiles:
(1, 2, 0, 0) and (0, 0, 2, 1)

The q-gram distance is:

$$\begin{aligned} |1-0| + |2-0| + |0-2| + |0-1| &= \\ 1 + 2 + 2 + 1 &= \\ 6 \end{aligned}$$

Q-Gram Distance

- ▶ In terms of DNA sequences, there would be sixteen different possible 2-grams
- ▶ AA, AG, AT, AC, GA, GG, GT, GC, TA, TG, TT, TC, CA, CG, CT, CC
- ▶ Example: AAGTC
 - AA: 1, AG: 1, GT: 1, TC: 1

Q-Gram Distance vs Edit Distance

- ▶ Q-gram and edit distance can be similar...

- ▶ Consider two strings:

ATATATATAT

TATATATATA

- ▶ Edit distance = 2
- ▶ 2-gram distance = 2

Q-Gram Distance vs Edit Distance

- ▶ But q-gram distance can be small even when edit distance is large
- ▶ Consider two strings of length 10:
AAAAATTTTT
TTTTTAAAAA
- ▶ Edit distance = 10
- ▶ 2-gram distance = 2

Q-Gram Distance vs Edit Distance

- ▶ And q-gram distance can be large even when edit distance is small
- ▶ Consider two strings:
AATAATAA
AAAAAAAAA
- ▶ Edit distance = 2
- ▶ 2-gram distance = 8

Q-Gram Distance vs Edit Distance

Relationship between q-gram distance and edit distance:

$$QD \leq 2qED$$

Further Investigation

Investigate the relationship between q-gram distance and edit distance

