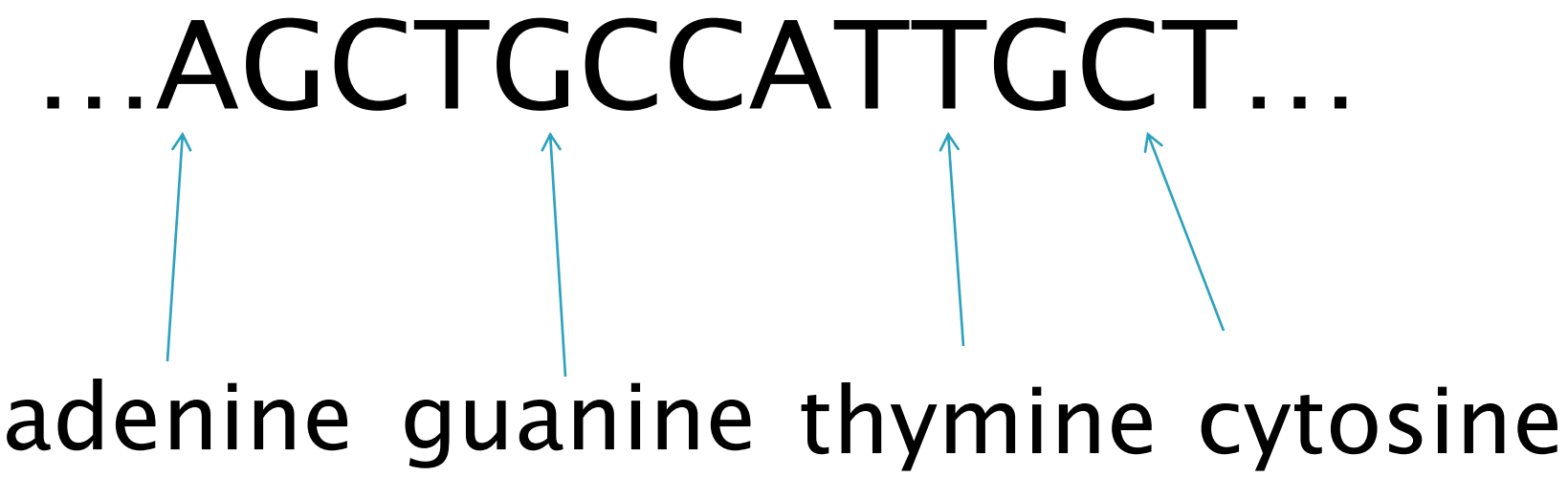
Relationship Between Edit Distance and Q-Gram Distance

Theresa Lye
Mentor: Alexander Schliep
Co-Mentor: Pavel Mahmud
July 19th, 2012

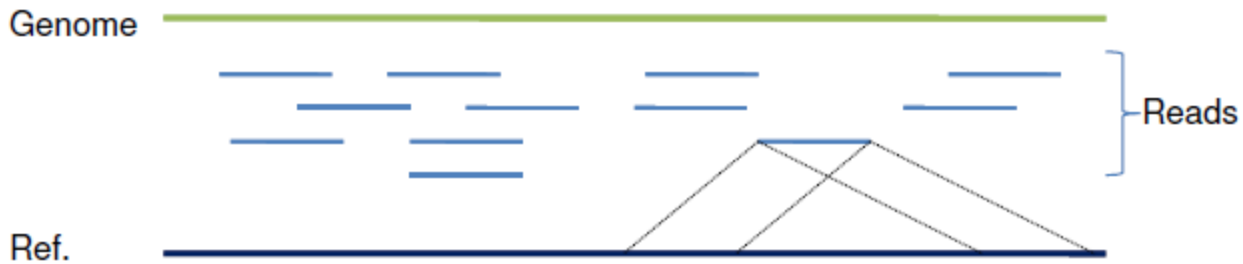
Motivation: Read Mapping to Reference Genomes

- DNA sequencing



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

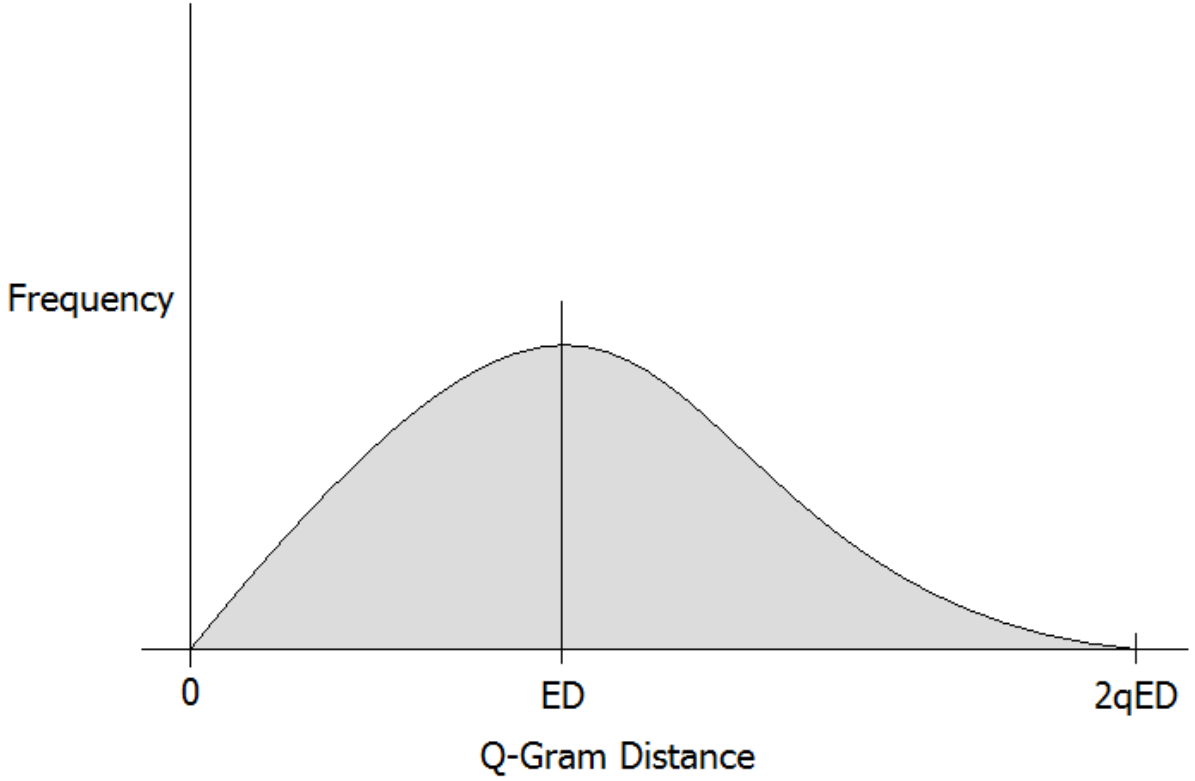
Edit Distance and Q-Gram Distance

- **Edit Distance**: Minimal number of edits (insertions, deletions, substitutions) required to transform one character string into another.

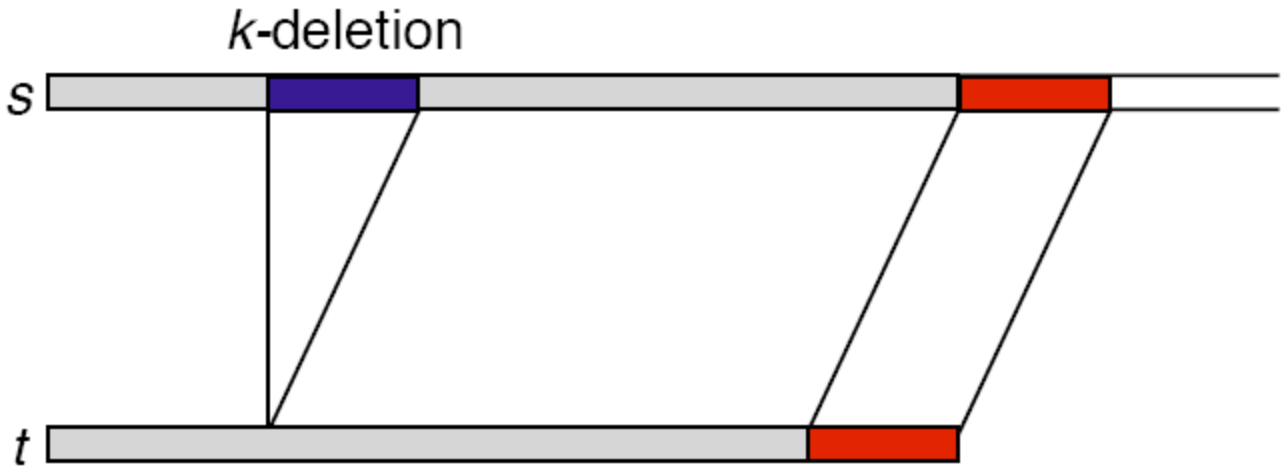
- **Q-Gram Distance**: L1 norm of the difference of the q-gram profiles
  - Example: AAA (f_{AA}:2) and AAG (f_{AA}: 1, f_{AG}: 1), q-gram distance of 2

Edit Distance and Q-Gram Distance

- $QD \leq 2qED$



K-Indel Problem



- L1(s,t) = q-gram distance between s and t
- Probability of L1(s,t) = 0? L1(s,t) = 2x?

1-Gram Case

$$P[L_1(s,t) = 0 | ED(s,t) = k, \text{one deletion}]$$
$$\leq \binom{k}{c_A, c_C, c_G, c_T} P_1^* \leq \frac{k!}{(\frac{k}{4}!)^4} (P_2^*)^k$$

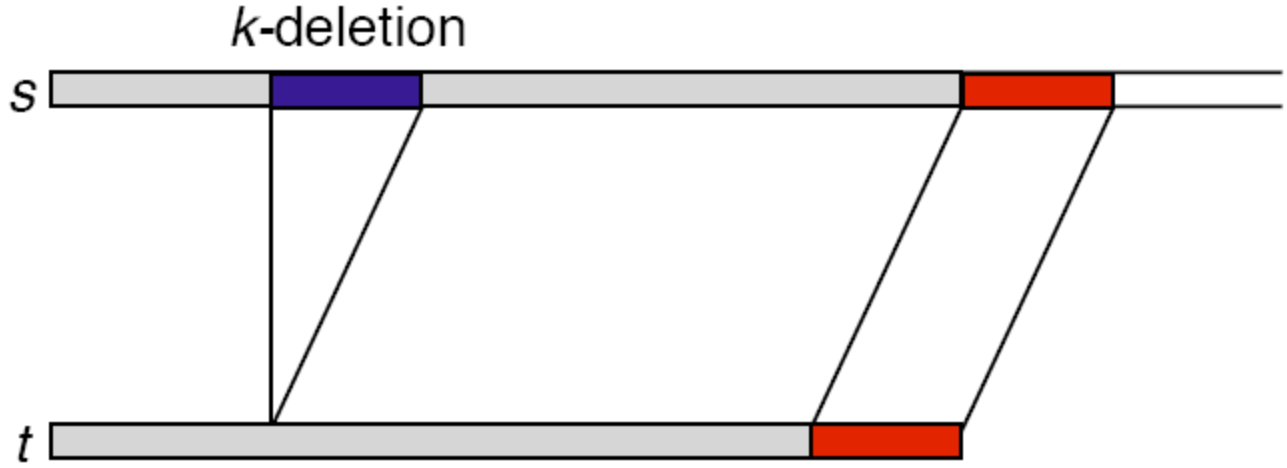
1-Gram Case

$$P[L_1(s,t) = 2x | ED(s,t) = k, onedeletion]$$
$$\leq \left[\binom{4}{1}\binom{x+2}{2} + \binom{4}{2}\binom{x-1}{1}\binom{x+1}{1} + \binom{4}{3}\binom{x-1}{2}\right] \frac{k!}{(\frac{k}{4}!)^4}(P_2^*)^k$$

- Simplifies to…

$$[L_1(s,t) = 2x | ED(s,t) = k, \text{one deletion}] \leq (10x^2 + 2)\frac{k!}{(\frac{k}{4}!)^4}(P_2^*)^k$$

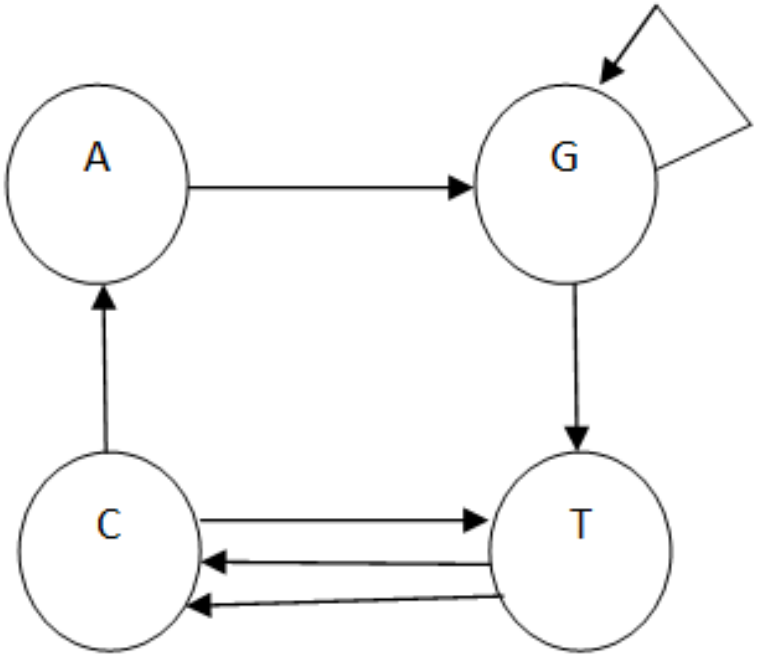
2-Gram Case: Previous Attempts



- 1-gram case as an upper bound
- 2-gram as independent
- 2-gram frequency vector as a state for a Markov Chain

2-Gram Case: Eulerian Graphs

- AGGTCTCA



2-Gram Case: Eulerian Graphs

- BEST Theorem

$$ec(G) = t_w(G) \prod_{v \epsilon V} (deg(v) - 1)!$$

- Requires some modification

$$ec^*(G) = \frac{f^* t_w(G) \prod_{v \epsilon V} (deg(v) - 1)!}{\prod_{a,b \epsilon \Sigma} f_{ab}!}$$

2-Gram Case: Eulerian Graphs

- Example: AGGTCTCA
- Using ec*(G), # of sequences with the same 2-gram composition = 2
- Using 1-gram upper bound, # of sequences with the same 1-gram composition = 2520

2-Gram Case: Eulerian Graphs

$$P[L_1(s,t) = 0 | ED(s,t) = k, \text{one deletion}]$$

$$\leq \frac{f^* t_w(G) \prod_{v \epsilon V} (deg(v) - 1)!}{\prod_{a,b \epsilon \Sigma} f_{ab}!} \left(\prod_{u \epsilon \Sigma^2} P_u^{f_u} \right)$$

Further Investigation

- Refine the Eulerian graph approach
- Generalize to $L1(s,t) \neq 0$
- Generalize to $q > 2$
- Generalize to other cases of modified strings (>1 indel, mismatches)

References

- *Circuits and Trees in Oriented Linear Graphs,* T. van Aardenne-Ehrenfest, N.G. de Bruijn
- *On Unicursal Paths in a Network of Degree 4*, W.T. Tutte, C.A.B. Smith
- *Approximate string-matching with q-grams and maximal matches*, Esko Ukkonen
- *Indel-tolerant Read Mapping with Trinucleotide Frequencies using Cache-Oblivious kd-Trees*, Md Pavel Mahmud, John Wiedenhoeft, Alexander Schliep