BLAST

Given a query $q = ACGACGAC$

$w = 2$

word length

$w = ?$

$O(m)$ word

$O(n)$

$q = m$

d = n

$(m-w+1) = O(m)$

$w$-len words

Match = 1

$M' \text{ match} = 0$

$T$: threshold for similarity

$\text{DNA}$

$\omega = \{A, C, G, T\}$

$\frac{y \cdots \cdots y}{y}$

$d$

AC

AC

AT

AA

CC

GG

TC

AC

AT

AA

$\begin{array}{c|c|c|c|c|}
\text{AC} & 1 & & & \\
\text{AC} & 1 & & & \\
\text{AT} & 1 & & & \\
\text{AA} & & & & \\
\text{CA} & 1 & & & \\
\end{array}$

$\begin{array}{c|c|c|c}
-1 & 0 & \\
1 & 1 & \\
\end{array}$

all $w$-len word except match
(1) seed/hit extension step.

worst case: $N = O(n \cdot m)$

Practically: $N = O(n \cdot m)$

Seed: 2 close hits on the same diagonal

$\Gamma \geq 2$

$O(n \cdot m)$

match = 1
mismatch = -1

drop threshold 1

$m = 1$

$V \leftarrow$ extended match

$\Gamma \leftarrow$ threshold 1
(4) \[ V \leftarrow \text{extended match} \]

\[ V \geq T \]

Keep only those extensions that score \( \geq V \)

Fill only those cells that score \( \geq T \)

\[ O \left( \frac{m \cdot q + N \cdot n + \log D P}{\text{neighbors}} \right) \]

Find \( w \) length words with at most \( k \) mismatches between \( q \) and \( t \)

\[ w = 100 \]

\[ k = 10 \]

Threshold for "hamming distance"
$O(n \cdot m)$

$1 \times 10^9$

human genome scale

$k$ possible mismatches

Threshold on the number of errors

w = 100

l = 10

$0 < w$

$w - l + 1$ l-len words inside $w$

Every one of the $k$
possible mismatches

ejectives $l$. $k$ l-len words

Remaining l-length word:

$(w - l + 1) - (k \cdot l)$
\[ w - x + 1 \in \text{exact matches of length } l \]

\[ w + 1 - (k+1) l \]

Any time there is a match of length \( w \), with at most \( k \) mismatches, then these two words of length \( w \) must have

\[ \geq \left[ w + 1 - (k+1) l \right] \] 

\( l \)-length word in common

\( l < w \)

1. Find all \( l \)-length exact matches between \( q \) and \( d \)

\[ l = 0 \]

\[ \text{hash} \]

\[ \text{ACR} \]

\[ 0(n^2) \]
Locality-sensitive Hashing (Randomized algorithm)

Find w-length word with at most 1ε mismatches

1. Define a "random" hash function

\[ h(x) = (\sum_{i=1}^{w} \theta_i) \mod \Gamma \]

\[ \Gamma = \begin{cases} 3 & \text{if } t=3 \\ 53,613 & \text{otherwise} \end{cases} \]

2. Now try w-length word & look for threshold.

\[ \frac{k}{w} \leq 16n \]

\[ \frac{2}{3} \leq 3 \]

\[ w+1 - (\text{last}) \cdot l \]

\[ l = 11.2 \]

\[ 101 - 33 \]

\[ 68 \]

In practice = \( O(n \cdot m) \)
Bound the probability of a false negative

- $k$: mismatches
- $w$: length of word
- $l$: hash length (number of positions)

$\left(1 - \frac{k}{w}\right)^l = \text{prob of hashing to the same cell}$

- prob of picking a matching position

$1 - \left(1 - \frac{k}{w}\right)^l = \text{prob of not hashing to the same cell}$

$\left(1 - \left(1 - \frac{k}{w}\right)^l\right)^m < \Theta = \frac{1}{1000}$

Prob $l = 41$
\[ O((|q| + |d|) \cdot m) \]

E.g.
\[ |q| = o(|d|) = o(n) \]

\[ n \log n \]

Sufix trees & Sufix Array

\[ O(n) \] space to store the database

Linear space

\[ O(n) \] to construct the sufix tree

Query: \[ O(|q|) \] to query!!!

+ enumerate the matches

1234567
Syn. 1: $ACAGCA\$ $ACAGCA$

Syn. 2: 3
   4
   5
   7

$O(n^2)$  $O(n)$

$q = CA$

$2, 5$

$AGC?$

$q = CC$

$k \in \Sigma$

terminal character

in suffix:
any substring of the form
$s[i \ldots \$]