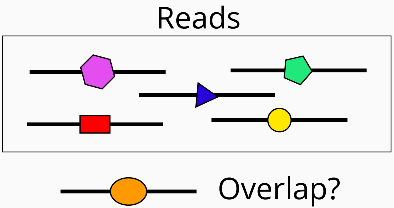
Locality sensitive hashing for the edit distance

Guillaume Marçais, Dan DeBlasio, Prashant Pandey, Carl Kingsford

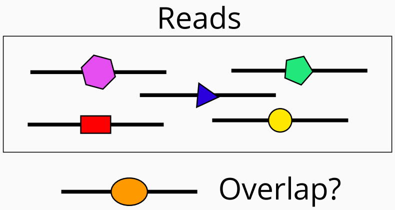
Carnegie Mellon University

Overlap computation

- Compute overlaps between reads (HMAP)
- Instance of "Nearest Neighbor Problem" for edit distance
- Use multiple hash tables
- Need meaningful hash collisions



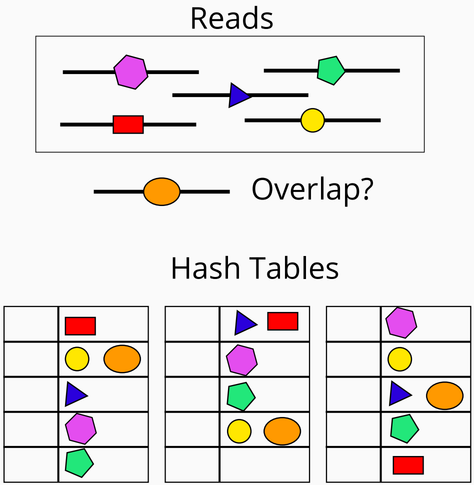
Overlap computation



- Compute overlaps between reads (HMAP)
- Instance of “Nearest Neighbor Problem” for edit distance
- Use multiple hash tables
- Need meaningful hash collisions

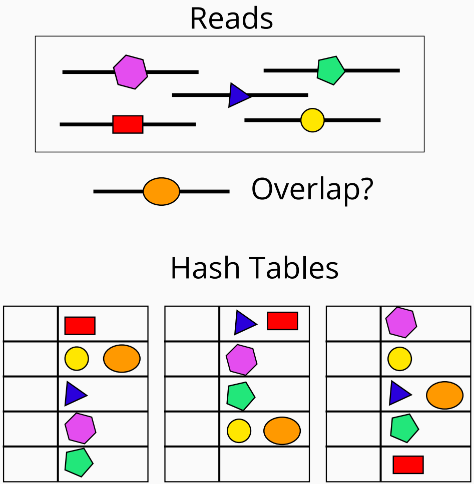
Overlap computation

- Compute overlaps between reads (HMAP)
- Instance of "Nearest Neighbor Problem" for edit distance
- Use multiple hash tables
- Need meaningful hash collisions



Overlap computation

- Compute overlaps between reads (HMAP)
- Instance of “Nearest Neighbor Problem” for edit distance
- Use multiple hash tables
- Need meaningful hash collisions



Locality Sensitive Hashing

Pick h at random from $\mathcal{H}$:

$$\Pr[h(\bigcirc) = h(\circ)] > \Pr[h(\bigcirc) = h(\blacktriangleright)]$$

Locality sensitive hash family

Family $\mathcal{H}$ of hash functions where similar elements are more likely to have the same value than distant elements.

Locality Sensitive Hashing

Pick h at random from $\mathcal{H}$:

$$\Pr[h(\bigcirc) = h(\bigcirc)] > \Pr[h(\bigcirc) = h(\blacktriangleright)]$$

$$\mathrm{Sketch}(\bigcirc) = \{h_1(\bigcirc), \ldots, h_m(\bigcirc)\}$$

Locality sensitive hash family

Family $\mathcal{H}$ of hash functions where similar elements are more likely to have the same value than distant elements.

Locality Sensitive Hashing

The family $\mathcal{H}$ is sensitive for distance D if there exists $d_1 < d_2$, $p_1 > p_2$ such that for all $x, y \in \mathcal{U}$

$$\mathrm{D}(x,y) \leq d_1 \implies \Pr_{h \in \mathcal{H}}[h(x) = h(y)] \geq p_1$$

$$\mathrm{D}(x,y) \geq d_2 \implies \Pr_{h \in \mathcal{H}}[h(x) = h(y)] \leq p_2$$

- **Low distance** $\iff$ **High collisions**
- **High distance** $\iff$ **Low collisions**

Locality sensitive hash family

Family $\mathcal{H}$ of hash functions where similar elements are more likely to have the same value than distant elements.

LSH for the edit distance

How to design an LSH for edit distance?

- LSH for Jaccard distance (minHash) used as proxy
- Jaccard distance is **significantly different** than edit distance

How to design an LSH for edit distance?

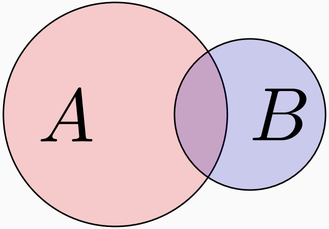
- LSH for Jaccard distance (minHash) used as proxy
- Jaccard distance is **significantly different** than edit distance

How to design an LSH for edit distance?

- LSH for Jaccard distance (minHash) used as proxy
- Jaccard distance is **significantly different** than edit distance

Jaccard distance

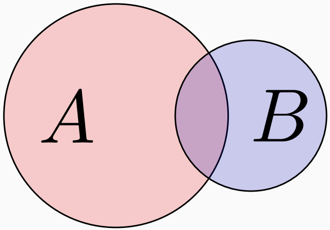
Jaccard distance between sets A, B:



$$\mathrm{J}(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

Jaccard distance

Jaccard distance between sets A, B:



$$\mathrm{J}(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

Jaccard between sequences x, y:
Jaccard distance of their k-mer sets

$$\mathrm{J}(x, y) = \mathrm{J}(\mathcal{K}(x), \mathcal{K}(y))$$

- Low $\mathrm{D}(x, y) \implies$ Low $\mathrm{J}(x, y)$
- High $\mathrm{D}(x, y) \not\Rightarrow$ High $\mathrm{J}(x, y)$

Jaccard ignores k-mer repetition

$$x = \overbrace{\texttt{AAAAAAAAAAAAAAA}}^{n-k}\overbrace{\texttt{CCCCC}}^{k}$$

$$y = \underbrace{\texttt{AAAAA}}_{k}\underbrace{\texttt{CCCCCCCCCCCCCCC}}_{n-k}$$

Jaccard ignores k-mer repetition

$$x = \overbrace{\texttt{AAAAAAAAAAAAAAA}}^{n-k}\overbrace{\texttt{CCCCC}}^{k} \quad \rightarrow \{\texttt{AAAAA}, \texttt{AAAAC}, \texttt{AAACC}, \texttt{AACCC}, \texttt{ACCCC}, \texttt{CCCCC}\}$$

$$y = \underbrace{\texttt{AAAAA}}_{k}\underbrace{\texttt{CCCCCCCCCCCCCCC}}_{n-k} \quad \rightarrow \{\texttt{AAAAA}, \texttt{AAAAC}, \texttt{AAACC}, \texttt{AACCC}, \texttt{ACCCC}, \texttt{CCCCC}\}$$

Jaccard ignores k-mer repetition

$$x = \overbrace{\texttt{AAAAAAAAAAAAAAA}}^{n-k}\overbrace{\texttt{CCCCC}}^{k} \rightarrow \{\texttt{AAAAA}, \texttt{AAAAC}, \texttt{AAACC}, \texttt{AACCC}, \texttt{ACCCC}, \texttt{CCCCC}\}$$

$$y = \underbrace{\texttt{AAAAA}}_{k}\underbrace{\texttt{CCCCCCCCCCCCCCC}}_{n-k} \rightarrow \{\texttt{AAAAA}, \texttt{AAAAC}, \texttt{AAACC}, \texttt{AACCC}, \texttt{ACCCC}, \texttt{CCCCC}\}$$

Jaccard distance $\mathrm{J}(x, y) = 0$ Edit distance $\mathrm{D}(x, y) \geq 1 - \frac{2k}{n}$

Identical k-mer content and high edit distance

Weighted Jaccard handles repetitions

$$x = \overbrace{\texttt{AAAAAAAAAAAAAAA}}^{n-k}\overbrace{\texttt{CCCCC}}^{k} \to \left\{ \begin{array}{c} (\texttt{AAAAA},1),(\texttt{AAAAA},2),\ldots,(\texttt{AAAAA},11) \\ (\texttt{AAAAC},1),(\texttt{AAACC},1),(\texttt{AACCC},1),(\texttt{ACCCC},1),(\texttt{CCCCC},1) \end{array} \right\}$$

$$y = \underbrace{\texttt{AAAAA}}_{k}\underbrace{\texttt{CCCCCCCCCCCCCCC}}_{n-k} \to \left\{ \begin{array}{c} (\texttt{AAAAA},1),(\texttt{AAAAC},1),(\texttt{AAACC},1),(\texttt{AACCC},1),(\texttt{ACCCC},1), \\ (\texttt{CCCCC},1),(\texttt{CCCCC},2),\ldots,(\texttt{CCCCC},11) \end{array} \right\}$$

Weighted Jaccard $\mathrm{J^w}(x,y) = 1 - \frac{k+2}{n}$ Edit distance $\mathrm{D}(x,y) \geq 1 - \frac{2k}{n}$

Weighted Jaccard $=$ Jaccard for multi-sets

Jaccard and weighted Jaccard ignore relative order

$$x = \texttt{CCCCACCAACACAAAACCC}$$

$$y = \texttt{AAAACACAACCCCACCAAA}$$

Jaccard and weighted Jaccard ignore relative order

$$x = \texttt{CCCCACCAACACAAAACCC} \quad \rightarrow \left\{ \begin{matrix} \texttt{AAAA,AAAC,AACA,AACC,ACAA,ACAC,ACCA,ACCC} \\ \texttt{CAAA,CAAC,CACA,CACC,CCAA,CCAC,CCCA,CCCC} \end{matrix} \right\}$$

$$y = \texttt{AAAACACAACCCCACCAAA} \quad \rightarrow \left\{ \begin{matrix} \texttt{AAAA,AAAC,AACA,AACC,ACAA,ACAC,ACCA,ACCC} \\ \texttt{CAAA,CAAC,CACA,CACC,CCAA,CCAC,CCCA,CCCC} \end{matrix} \right\}$$

x, y: de Bruijn sequences,
contain all 16 possible 4-mers once

Jaccard and weighted Jaccard ignore relative order

$$x = \texttt{CCCCACCAACACAAAACCC} \quad \rightarrow \left\{ \begin{matrix} \texttt{AAAA,AAAC,AACA,AACC,ACAA,ACAC,ACCA,ACCC} \\ \texttt{CAAA,CAAC,CACA,CACC,CCAA,CCAC,CCCA,CCCC} \end{matrix} \right\}$$

$$y = \texttt{AAAACACAACCCCACCAAA} \quad \rightarrow \left\{ \begin{matrix} \texttt{AAAA,AAAC,AACA,AACC,ACAA,ACAC,ACCA,ACCC} \\ \texttt{CAAA,CAAC,CACA,CACC,CCAA,CCAC,CCCA,CCCC} \end{matrix} \right\}$$

x, y: de Bruijn sequences,
contain all 16 possible 4-mers once

$$\mathrm{J}(x, y) = \mathrm{J}^{\mathrm{w}}(x, y) = 0 \qquad \mathrm{D}(x, y) = 0.63$$

Jaccard is different from edit distance

Unlike edit distance, Jaccard is insensitive to:

1. k-mer repetitions
2. relative positions of k-mers

OMH: Order Min Hash

- minHash is an LSH for Jaccard
- OMH is a refinement of minHash
- OMH is sensitive to
  - repeated k-mers
  - relative order of k-mers

minHash & OMH sketches

$x = \texttt{AGTTGAGCGGAAGGTG}$, $k = 2$

$$x = \texttt{AGTTGAGCGGAAGGTG},\ k = 2$$

Order: permutation of Σ^k

AT
AG
GT
CG
GT
TT
GG
CA
TG
TC
AA
AC
TA
GA
GC
CT

AG
GT
CG
TT
GG
TG
AA
GA
GC

minHash & OMH sketches

$$x = \texttt{AGTTGAGCGGAAGGTG},\ k = 2,\ m = 6$$

1	2	3	4	5	6
AG	GG	CG	AA	TG	TT
GT	GA	GA	AG	TT	GG
CG	CG	TG	TT	GG	AG
TT	AG	AG	GT	CG	GC
GG	GC	GC	TG	AA	GA
TG	GT	GG	GC	GT	AA
AA	AA	TT	GA	AG	CG
GA	TT	AA	CG	GC	TG
GC	TG	GT	GG	GA	GT

minHash & OMH sketches

$$x = \texttt{AGTTGAGCGGAAGGTG},\ k = 2,\ m = 6$$

1	2	3	4	5	6
AG	GG	CG	AA	TG	TT
GT	GA	GA	AG	TT	GG
CG	CG	TG	TT	GG	AG
TT	AG	AG	GT	CG	GC
GG	GC	GC	TG	AA	GA
TG	GT	GG	GC	GT	AA
AA	AA	TT	GA	AG	CG
GA	TT	AA	CG	GC	TG
GC	TG	GT	GG	GA	GT

minHash & OMH sketches

$x = \texttt{AGTTGAGCGGAAGGTG}, k = 2, m = 6$

Order: permutation of $\Sigma^k \times \{1, \ldots, n\}$

1	2	3	4	5	6
AG	GG	CG	AA	TG	TT
GT	GA	GA	AG	TT	GG
CG	CG	TG	TT	GG	AG
TT	AG	AG	GT	CG	GC
GG	GC	GC	TG	AA	GA
TG	GT	GG	GC	GT	AA
AA	AA	TT	GA	AG	CG
GA	TT	AA	CG	GC	TG
$\boxed{\text{GC}}$	$\boxed{\text{TG}}$	$\boxed{\text{GT}}$	$\boxed{\text{GG}}$	$\boxed{\text{GA}}$	$\boxed{\text{GT}}$

GA, 4
TG, 3
AG, 5
GT, 1
GT, 13
AA, 10
AG, 11
TT, 2
AG, 0
CG, 7
GG, 12
GC, 6
TG, 14
GG, 8
GA, 9

minHash & OMH sketches

$x = \texttt{AGTTGAGCGGAAGGTG}, k = 2, m = 6$

1	2	3	4	5	6
AG	GG	CG	AA	TG	TT
GT	GA	GA	AG	TT	GG
CG	CG	TG	TT	GG	AG
TT	AG	AG	GT	CG	GC
GG	GC	GC	TG	AA	GA
TG	GT	GG	GC	GT	AA
AA	AA	TT	GA	AG	CG
GA	TT	AA	CG	GC	TG
GC	TG	GT	GG	GA	GT

1	2	3	4	5	6
GA, 4	CG, 7	GT,13	AG, 0	AA,10	GA, 9
TG, 3	TG,14	GA, 4	TT, 2	GT,13	GG, 8
AG, 5	AG, 0	GA, 9	AG,11	GA, 9	GC, 6
GT, 1	GA, 9	TG, 3	AG, 5	GT, 1	TG,14
GT,13	AG, 5	AG, 5	AA,10	AG, 5	GT,13
AA,10	AG,11	CG, 7	GT,13	TT, 2	TT, 2
AG,11	GA, 4	TT, 2	CG, 7	GA, 4	AA,10
TT, 2	GT,13	AA,10	GG, 8	CG, 7	AG, 0
AG, 0	TT, 2	GG,12	GA, 4	AG, 0	CG, 7
CG, 7	TG, 3	GG, 8	GA, 9	TG, 3	GG,12
GG,12	GG, 8	TG,14	TG,14	GG, 8	AG,11
GC, 6	AA,10	GT, 1	TG, 3	GG,12	TG, 3
TG,14	GG,12	AG,11	GC, 6	GC, 6	GT, 1
GG, 8	GT, 1	GC, 6	GT, 1	AG,11	GA, 4
GA, 9	GC, 6	AG, 0	GG,12	TG,14	AG, 5

minHash & OMH sketches

$x = \texttt{AGTTGAGCGGAAGGTG}, k = 2, m = 6$

1	2	3	4	5	6
AG	GG	CG	AA	TG	TT
GT	GA	GA	AG	TT	GG
CG	CG	TG	TT	GG	AG
TT	AG	AG	GT	CG	GC
GG	GC	GC	TG	AA	GA
TG	GT	GG	GC	GT	AA
AA	AA	TT	GA	AG	CG
GA	TT	AA	CG	GC	TG
GC	TG	GT	GG	GA	GT

1	2	3	4	5	6
GA, 4	CG, 7	GT,13	AG, 0	AA,10	GA, 9
TG, 3	TG,14	GA, 4	TT, 2	GT,13	GG, 8
AG, 5	AG, 0	GA, 9	AG,11	GA, 9	GC, 6
GT, 1	GA, 9	TG, 3	AG, 5	GT, 1	TG,14
GT,13	AG, 5	AG, 5	AA,10	AG, 5	GT,13
AA,10	AG,11	CG, 7	GT,13	TT, 2	TT, 2
AG,11	GA, 4	TT, 2	CG, 7	GA, 4	AA,10
TT, 2	GT,13	AA,10	GG, 8	CG, 7	AG, 0
AG, 0	TT, 2	GG,12	GA, 4	AG, 0	CG, 7
CG, 7	TG, 3	GG, 8	GA, 9	TG, 3	GG,12
GG,12	GG, 8	TG,14	TG,14	GG, 8	AG,11
GC, 6	AA,10	GT, 1	TG, 3	GG,12	TG, 3
TG,14	GG,12	AG,11	GC, 6	GC, 6	GT, 1
GG, 8	GT, 1	GC, 6	GT, 1	AG,11	GA, 4
GA, 9	GC, 6	AG, 0	GG,12	TG,14	AG, 5

minHash & OMH sketches

$x =$ AGTTGAGCGGAAGGTG, $k = 2$, $m = 6$, $\ell = 2$

1	2	3	4	5	6
AG	GG	CG	AA	TG	TT
GT	GA	GA	AG	TT	GG
CG	CG	TG	TT	GG	AG
TT	AG	AG	GT	CG	GC
GG	GC	GC	TG	AA	GA
TG	GT	GG	GC	GT	AA
AA	AA	TT	GA	AG	CG
GA	TT	AA	CG	GC	TG
GC	TG	GT	GG	GA	GT

1	2	3	4	5	6
GA, 4	CG, 7	GT, 13	AG, 0	AA, 10	GA, 9
TG, 3	TG, 14	GA, 4	TT, 2	GT, 13	GG, 8
AG, 5	AG, 0	GA, 9	AG, 11	GA, 9	GC, 6
GT, 1	GA, 9	TG, 3	AG, 5	GT, 1	TG, 14
GT, 13	AG, 5	AG, 5	AA, 10	AG, 5	GT, 13
AA, 10	AG, 11	CG, 7	GT, 13	TT, 2	TT, 2
AG, 11	GA, 4	TT, 2	CG, 7	GA, 4	AA, 10
TT, 2	GT, 13	AA, 10	GG, 8	CG, 7	AG, 0
AG, 0	TT, 2	GG, 12	GA, 4	AG, 0	CG, 7
CG, 7	TG, 3	GG, 8	GA, 9	TG, 3	GG, 12
GG, 12	GG, 8	TG, 14	TG, 14	GG, 8	AG, 11
GC, 6	AA, 10	GT, 1	TG, 3	GG, 12	TG, 3
TG, 14	GG, 12	AG, 11	GC, 6	GC, 6	GT, 1
GG, 8	GT, 1	GC, 6	GT, 1	AG, 11	GA, 4
GA, 9	GC, 6	AG, 0	GG, 12	TG, 14	AG, 5

GC GA AG GG AG TG

OMH is a LSH for edit distance

Theorem: OMH is a LSH for edit distance

There exists (d_1, d_2, p_1, p_2) such that OMH is sensitive for the edit distance.

Conclusion

- OMH:
  - improvement on minHash
  - easy to compute
  - locality sensitive for edit distance
- LSH for other alignment scores?
- Smallest “gap” achievable?

Conclusion

- OMH:
  - improvement on minHash
  - easy to compute
  - locality sensitive for edit distance

- LSH for other alignment scores?
- Smallest “gap” achievable?

Conclusion

- OMH:
  - improvement on minHash
  - easy to compute
  - locality sensitive for edit distance
- LSH for other alignment scores?
- Smallest “gap” achievable?

Carnegie Mellon University

Natalie Sauerwald
Cong Ma
Hongyu Zheng
Laura Tung
Hongyu Zheng
Yutong Qiu
Yihang Shen
Minh Hoang
Mohsen Ferdosi
Shawn Baker
Yinjie Gao

Thank you

GORDON AND BETTY MOORE FOUNDATION
GBMF4554

NSF
CCF-1256087
CCF-1319998

NATIONAL INSTITUTES OF HEALTH
R01HG007104
R01GM122935

The Shurl and Kay Curci Foundation