The statistics of k-mers from a sequence undergoing a simple mutation process without spurious matches

Antonio Blanca, Robert S. Harris, David Koslicki, Paul Medvedev

Pennsylvania State University

Data structures in bionformatics workshop (DSB)
Feb 11-12, 2021

This paper will appear in RECOMB 2021 and is available on bioRxiv.

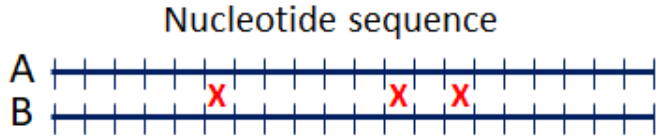
Talk outline

1. Introduce model
2. Motivating applications
3. Number of mutated k-mers
   - 3.1 expectation
   - 3.2 variance
   - 3.3 hypothesis test
   - 3.4 confidence interval
4. Other random variables
5. Experimental results

Simple model

Generative model

- Start with a genome A
- Mutate every nucleotide with probability r_1
- Get a new genome B
- Assume that all k-mers are unique.



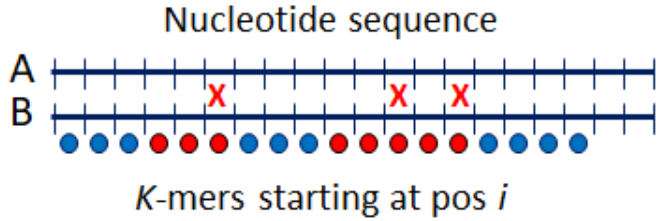
Simple model

Generative model

- Start with a genome A
- Mutate every nucleotide with probability r_1
- Get a new genome B
- Assume that all k-mers are unique.

What do we observe?

- not the nucleotide sequences
- N_{mut}
  - Number of mutated k-mers



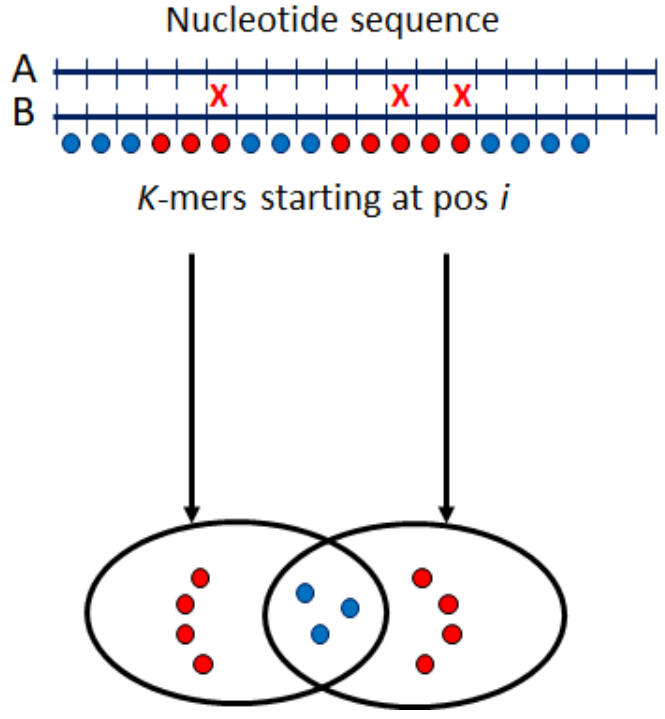
Simple model

Generative model

- Start with a genome A
- Mutate every nucleotide with probability r_1
- Get a new genome B
- Assume that all k-mers are unique.

What do we observe?

- not the nucleotide sequences
- N_{mut}
  - Number of mutated k-mers
- Jaccard
  - $J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{L - N_{mut}}{L + N_{mut}}$



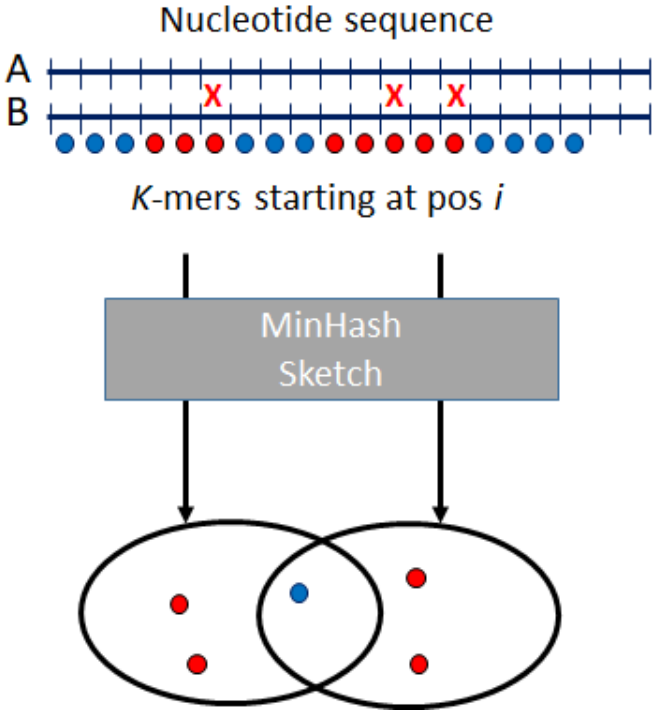
Simple model

Generative model

- Start with a genome A
- Mutate every nucleotide with probability r_1
- Get a new genome B
- Assume that all k-mers are unique.

What do we observe?

- not the nucleotide sequences
- N_{mut}
  - Number of mutated k-mers
- Jaccard
  - $J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{L - N_{mut}}{L + N_{mut}}$
- Minhash Jaccard
  - $A_{sk} \triangleq$ minhash sketch of A
  - $B_{sk} \triangleq$ minhash sketch of B
  - $\hat{J} = J(A_{sk}, B_{sk})$



Motivating applications

Mash distance [Ondov et al., 2016]

- Take two evolutionary related sequences
- Observe $\hat{J}$ from two genomes
- Assume that genomes evolved under the simple model
- Estimate r_1 from $\hat{J}$.
- What about a confidence interval for r_1?
  - Given that the two sequences evolved under this simple model, and we observe N_{mut}, what is an interval that will contain r_1 with 95% probability?

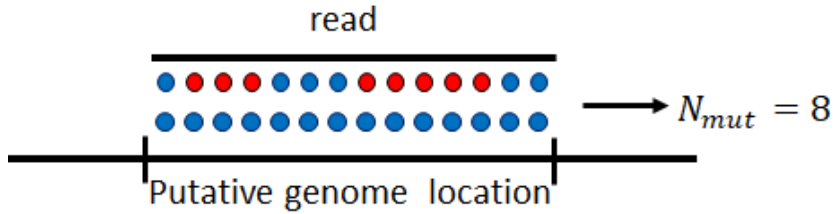
Motivating applications

Mash distance [Ondov et al., 2016]

- Take two evolutionary related sequences
- Observe $\hat{J}$ from two genomes
- Assume that genomes evolved under the simple model
- Estimate r_1 from $\hat{J}$.
- What about a confidence interval for r_1?
  - Given that the two sequences evolved under this simple model, and we observe N_{mut}, what is an interval that will contain r_1 with 95% probability?

Alignments of reads to de Bruijn graph (minimap2, jabba, lorma)

- A read is generated from a genome location
  - sequencing error rate r_1.
- Is a putative genome location the one that generated the read?
  - We observe N_{mut}
  - Want to accept/reject this alignment, with 95% chance of being correct.
- A hypothesis test with significance level 95% for N_{mut}
  - Given r_1 what is the range into which N_{mut} would fall with 95% probability?

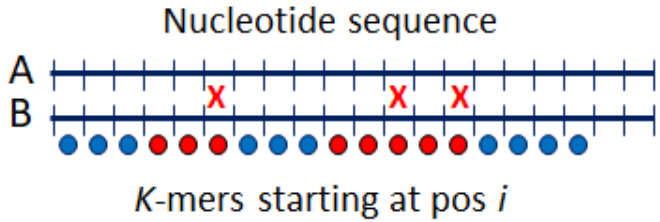


Distribution of N_{mut}

Expectation

Expectation is easy.

- Let X_i be the indicator r.v. if k-mer starting at position i is mutated.
- Let $\mathrm{E}[X_i] \triangleq r_k = (1 - (1 - r_1)^k)$ be the probability that a k-mer is mutated.
- $N_{mut} = \sum X_i$
- $\mathrm{E}[N_{mut}] = \mathrm{E}[\sum X_i] = L\mathrm{E}[X_i] = Lr_k$.

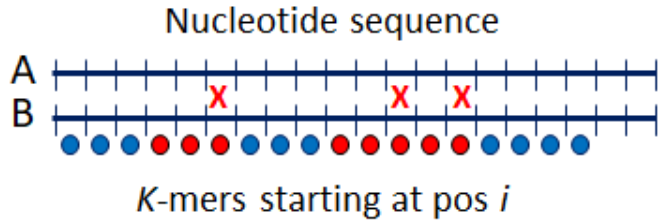


Distribution of N_{mut}

Expectation

Expectation is easy.

- Let X_i be the indicator r.v. if k-mer starting at position i is mutated.
- Let $\mathrm{E}[X_i] \triangleq r_k = (1 - (1 - r_1)^k)$ be the probability that a k-mer is mutated.
- $N_{mut} = \sum X_i$
- $\mathrm{E}[N_{mut}] = \mathrm{E}[\sum X_i] = L\mathrm{E}[X_i] = Lr_k$.



Is N_{mut} a binomial?

- Binomial is sum of *independent* Bernoulli trials
- But nearby X_is are dependent.

Dependency lemma and variance

Lemma

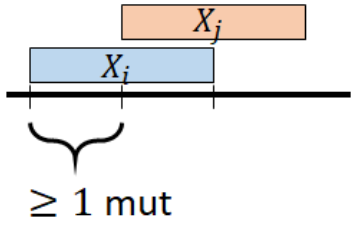
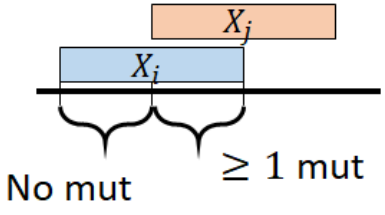
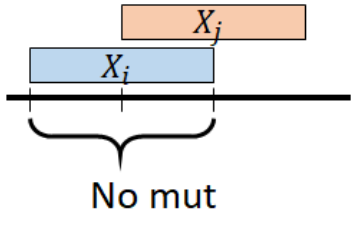
- If $j - i \geq k$, then X_i and X_j are independent
- If $j - i < k$, $\Pr[X_i = 1, X_j = 1] = 2r_k - 1 + (1 - r_1)^{k+j-i}$

Dependency lemma and variance

Lemma

- If $j - i \geq k$, then X_i and X_j are independent
- If $j - i < k$, $\Pr[X_i = 1, X_j = 1] = 2r_k - 1 + (1 - r_1)^{k+j-i}$

Proof

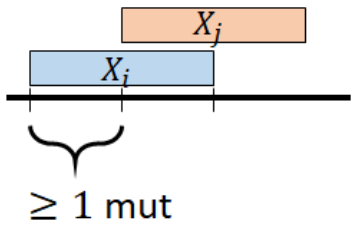
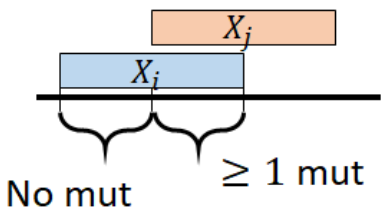
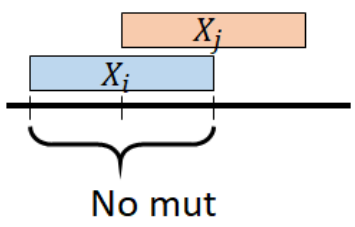
Case	Pr[Case]	$\Pr[X_i = 1, X_j = 1 \mid \text{Case}]$
	$1 - (1 - r_1)^{j-i}$	$\Pr[X_j = 1] = r_k$
	$(1 - r_1)^{j-i}(1 - (1 - r_1)^{k-j+i})$	1
	N/A	0

Dependency lemma and variance

Lemma

- If $j - i \geq k$, then X_i and X_j are independent
- If $j - i < k$, $\Pr[X_i = 1, X_j = 1] = 2r_k - 1 + (1 - r_1)^{k+j-i}$

Proof

Case	Pr[Case]	$\Pr[X_i = 1, X_j = 1 \mid \text{Case}]$
	$1 - (1 - r_1)^{j-i}$	$\Pr[X_j = 1] = r_k$
	$(1 - r_1)^{j-i}(1 - (1 - r_1)^{k-j+i})$	1
	N/A	0

Lemma

- $\mathrm{Var}[N_{mut}] = L(1 - r_k)(r_k(2k + \frac{2}{r_1} - 1) - 2k) + o(L)$

M-dependent variables and Main Technique Theorem

A sequence of L random variables $X_0, \ldots, X_{L-1}$ is said to be **m-dependent** if there exists a bounded m such that if $j - i > m$, then the two sets $\{X_0, \ldots, X_i\}$ and $\{X_j, \ldots, X_{L-1}\}$ are independent [Hoeffding et al., 1948].

- N_{mut} **is sum of m-dependent variables**, with $m = k - 1$.
- Sum of m-dependent variables is asymptotically normal [Hoeffding et al., 1948].
- Stein's method also gives us the rate of convergence [Ross, 2011].
- We can derive hypothesis test using same strategy as with Binomial
- **Main Technique Theorem**
  - Let X be a sum of m-dependent Bernoulli random variables.
  - Then, $X \in \mathrm{E}[X] \pm z_\alpha \sqrt{\mathrm{Var}(X)}$ with limiting* probability α,
    - z_α is value of inverse Normal CDF at $(1 - \alpha)/2$

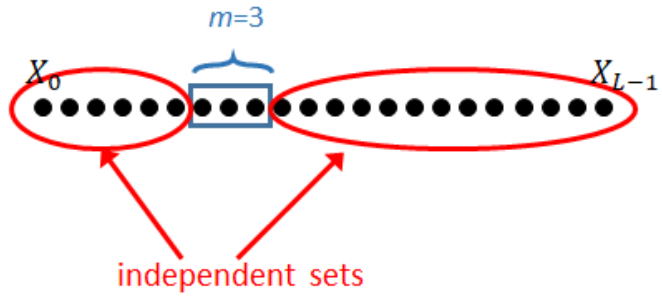
M-dependent variables and Main Technique Theorem

A sequence of L random variables $X_0, \ldots, X_{L-1}$ is said to be **m-dependent** if there exists a bounded m such that if $j - i > m$, then the two sets $\{X_0, \ldots, X_i\}$ and $\{X_j, \ldots, X_{L-1}\}$ are independent [Hoeffding et al., 1948].

- N_{mut} **is sum of m-dependent variables**, with $m = k - 1$.
- Sum of m-dependent variables is asymptotically normal [Hoeffding et al., 1948].
- Stein's method also gives us the rate of convergence [Ross, 2011].
- We can derive hypothesis test using same strategy as with Binomial
- **Main Technique Theorem**
  - Let X be a sum of m-dependent Bernoulli random variables.
  - Then, $X \in \mathrm{E}[X] \pm z_\alpha \sqrt{\mathrm{Var}(X)}$ with limiting* probability α,
    - z_α is value of inverse Normal CDF at $(1 - \alpha)/2$

M-dependent variables and Main Technique Theorem

A sequence of L random variables $X_0, \ldots, X_{L-1}$ is said to be **m-dependent** if there exists a bounded m such that if $j - i > m$, then the two sets $\{X_0, \ldots, X_i\}$ and $\{X_j, \ldots, X_{L-1}\}$ are independent [Hoeffding et al., 1948].



- N_{mut} **is sum of m-dependent variables**, with $m = k - 1$.
- Sum of m-dependent variables is asymptotically normal [Hoeffding et al., 1948].
- Stein's method also gives us the rate of convergence [Ross, 2011].
- We can derive hypothesis test using same strategy as with Binomial
- **Main Technique Theorem**
  - Let X be a sum of m-dependent Bernoulli random variables.
  - Then, $X \in \mathrm{E}[X] \pm z_\alpha \sqrt{\mathrm{Var}(X)}$ with limiting* probability α,
    - z_α is value of inverse Normal CDF at $(1 - \alpha)/2$

N_{mut} and Jaccard

Hypothesis tests and confidence intervals

Corollary of Main Technique Theorem

- $N_{mut} \in Lr_k \pm z_\alpha \sqrt{\mathrm{Var}(N_{mut})}$ with limiting* probability α,

  *assuming r_1 and k are independent of L

To compute CI for r_1,

- Numerically find the range of r_1 for which N_{mut} is in the test range.

Suppose we observe $T = f(N_{mut})$

- $f(x)$ is a monotone function
- e.g. Jaccard $= \frac{L - N_{mut}}{L + N_{mut}}$

Corollaries

- With limiting* probability α,
  - $f(N_{mut}) \in f(Lr_k \pm z_\alpha \sqrt{\mathrm{Var}(N_{mut})})$
  - $J \in \left(\frac{L - Lr_k - z_\alpha \sqrt{\mathrm{Var}(N_{mut})}}{L + Lr_k + z_\alpha \sqrt{\mathrm{Var}(N_{mut})}}, \frac{L - Lr_k + z_\alpha \sqrt{\mathrm{Var}(N_{mut})}}{L + Lr_k - z_\alpha \sqrt{\mathrm{Var}(N_{mut})}} \right)$

Minhash Jaccard estimator

a.k.a. Mash distance

Two layers of randomness

- Mutation process
  - We can apply our Main Technique
- Sketching process
  - Our Main Technique does not apply
  - . . . because sketch uses global information
  - We use a different approach

Theorem

- With limiting* probability α, $j_{low} \leq \hat{J} \leq j_{high}$

Islands and oceans

Island definition

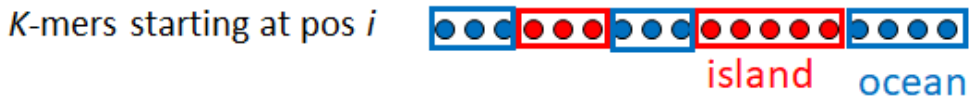
- An *island* is a maximal interval of mutated k-mers.
- Sequence can be partitioned into alternated islands and oceans.

- Number of islands is $\sum_i B_i$.
  - $B_i = 1$ iff the k-mer at pos i is mutated and at at $i + 1$ is not.
  - $B_{L-1} = 1$ is special end case.

Islands and oceans

Island definition

- An *island* is a maximal interval of mutated k-mers.
- Sequence can be partitioned into alternated islands and oceans.

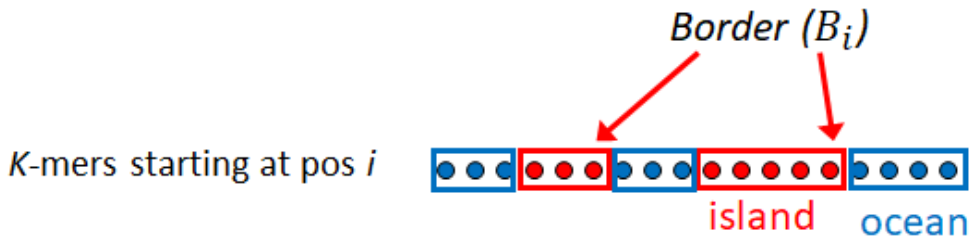


- Number of islands is $\sum_i B_i$.
  - $B_i = 1$ iff the k-mer at pos i is mutated and at at $i + 1$ is not.
  - $B_{L-1} = 1$ is special end case.

Islands and oceans

Island definition

- An *island* is a maximal interval of mutated k-mers.
- Sequence can be partitioned into alternated islands and oceans.

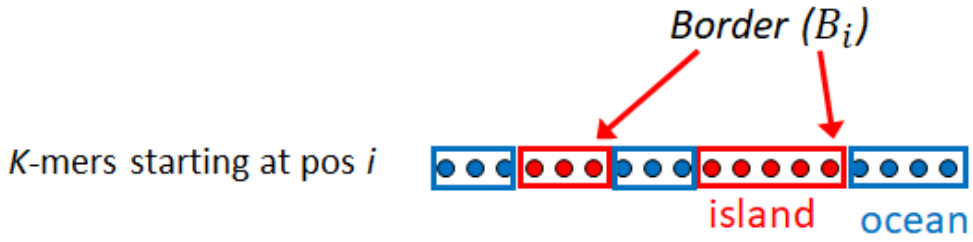


- Number of islands is $\sum_i B_i$.
  - $B_i = 1$ iff the k-mer at pos i is mutated and at at $i + 1$ is not.
  - $B_{L-1} = 1$ is special end case.

Islands and oceans

Island definition

- An *island* is a maximal interval of mutated k-mers.
- Sequence can be partitioned into alternated islands and oceans.



- Number of islands is $\sum_i B_i$.
  - $B_i = 1$ iff the k-mer at pos i is mutated and at at $i+1$ is not.
  - $B_{L-1} = 1$ is special end case.

Steps to derive hypothesis test for number of islands

- Derive $\Pr[B_i = 1, B_j = 1]$.
- Confirm that B_i and B_j are independent if they are far apart.
- Derive $\mathrm{E}(N_{island})$ and $\mathrm{Var}(N_{island})$
- Apply Main Technique Theorem
  - $N_{island} \in \mathrm{E}(N_{island}) \pm z_\alpha \sqrt{\mathrm{Var}(N_{island})}$ with limiting* probability α.

Summary of theoretical results

the expectation, variances, and intervals derived in the paper

Variable	Expectation	Variance	α interval
N_{mut}	Lq	$L(1-q)(q(2k+\frac{2}{r_1}-1)-2k)$	$Lq \pm z_\alpha \sqrt{\mathrm{Var}(N_{mut})}$
N_{island}	$Lr_1(1-q)$	$Lr_1(1-q)(1-r_1(1-q)(2k+1))$	$\mathrm{E}[N_{island}] \pm z_\alpha \sqrt{\mathrm{Var}(N_{island})}$
N_{ocean}	$Lr_1(1-q)$	$Lr_1(1-q)(1-r_1(1-q)(2k+1))$	$\mathrm{E}[N_{ocean}] \pm z_\alpha \sqrt{\mathrm{Var}(N_{ocean})}$
Jaccard	—	—	(see prev slide)
minhash Jaccard	—	—	(j_{low}, j_{high})
C_{ber} **	$\frac{L(1-q)(1+r_1(k-1))}{L+k-1}$	see paper	$\mathrm{E}[C_{ber}] \pm z_\alpha \sqrt{\mathrm{Var}(C_{ber})}$

** Coverage by exact regions [Miclotte et al., 2016]

*Only higher order terms are shown here, see paper for exact expressions.

Experimental results

N_{mut} confidence intervals

Simulation experiments

- Starting sequence with no dup k-mers
- 10,000 replicates for each cell.
- Report fraction of replicates for which the true r_1 falls into the predicted 95% CI.

$L = 10,000$	r_1			
	0.001	0.01	0.1	0.2
$k = 100$	0.95	0.95	—	—
51	0.95	0.95	0.96	—
21	0.95	0.94	0.95	0.95

Experimental results

N_{mut} confidence intervals

Simulation experiments

- Starting sequence with no dup k-mers
- 10,000 replicates for each cell.
- Report fraction of replicates for which the true r_1 falls into the predicted 95% CI.

$L = 10,000$	r_1			
	0.001	0.01	0.1	0.2
$k = 100$	0.95	0.95	—	—
51	0.95	0.95	0.96	—
21	0.95	0.94	0.95	0.95

$L = 1,000$	r_1			
	0.001	0.01	0.1	0.2
$k = 100$	0.95	0.96	—	—
51	0.94	0.95	0.94	—
21	0.93	0.95	0.95	0.95

Experimental results

N_{mut} confidence intervals

Simulation experiments

- Starting sequence with no dup k-mers
- 10,000 replicates for each cell.
- Report fraction of replicates for which the true r_1 falls into the predicted 95% CI.

$L = 10,000$	r_1			
	0.001	0.01	0.1	0.2
$k = 100$	0.95	0.95	—	—
51	0.95	0.95	0.96	—
21	0.95	0.94	0.95	0.95

$L = 1,000$	r_1			
	0.001	0.01	0.1	0.2
$k = 100$	0.95	0.96	—	—
51	0.94	0.95	0.94	—
21	0.93	0.95	0.95	0.95

$L = 100$	r_1			
	0.001	0.01	0.1	0.2
$k = 100$	0.91	1.00	—	—
51	0.91	1.00	1.00	—
21	0.91	0.96	1.00	1.00

Experimental results

N_{mut} confidence intervals

Simulation experiments

- Starting sequence with no dup k-mers
- 10,000 replicates for each cell.
- Report fraction of replicates for which the true r_1 falls into the predicted 95% CI.

Experiments with *E.Coli*

- Simulation done on *E.Coli* sequence
- CI calculator only observes
  - set of k-mers before (A)
  - set of k-mers before (B)
- CI calculator defines
  - $L = (|A| + |B|)/2$.
  - $N_{mut} = L - |A \cap B|$

$L = 10,000$	r_1			
	0.001	0.01	0.1	0.2
$k = 100$	0.95	0.95	—	—
51	0.95	0.95	0.96	—
21	0.95	0.94	0.95	0.95

$L = 1,000$	r_1			
	0.001	0.01	0.1	0.2
$k = 100$	0.95	0.96	—	—
51	0.94	0.95	0.94	—
21	0.93	0.95	0.95	0.95

$L = 100$	r_1			
	0.001	0.01	0.1	0.2
$k = 100$	0.91	1.00	—	—
51	0.91	1.00	1.00	—
21	0.91	0.96	1.00	1.00

E.Coli	r_1			
$r_1 =$	0.001	0.01	0.1	0.2
$k = 100$	0.95	0.95	—	—
51	0.95	0.95	0.95	—
21	0.95	0.94	0.93	0.94

Experimental results

Mash distance (i.e. minhash Jaccard estimator)

- Table 1 in [Ondov et al., 2016] tested the point estimate on a range of values.
  - $k = 21$
  - $L = 4,500,000$
  - Varying sketch size and r_1
- We replicate their experiments, but instead predict 95% CIs
  - 1,000 replicates for each cell

	$r_1(r_k)$		
	.05(.659)	.15(.967)	.25(.998)
sketch size = 100	0.97	1.00	1.00
1,000	0.96	0.97	1.00
10,000	0.95	0.96	0.96
100,000	0.95	0.95	0.96
1,000,000	0.94	0.95	0.96

Experimental results

Mash distance (i.e. minhash Jaccard estimator)

- Table 1 in [Ondov et al., 2016] tested the point estimate on a range of values.
  - $k = 21$
  - $L = 4,500,000$
  - Varying sketch size and r_1
- We replicate their experiments, but instead predict 95% CIs
  - 1,000 replicates for each cell

	$r_1 (r_k)$		
	.05(.659)	.15(.967)	.25(.998)
sketch size = 100	0.97	1.00	1.00
1,000	0.96	0.97	1.00
10,000	0.95	0.96	0.96
100,000	0.95	0.95	0.96
1,000,000	0.94	0.95	0.96

- We also simulated with *E.coli*.

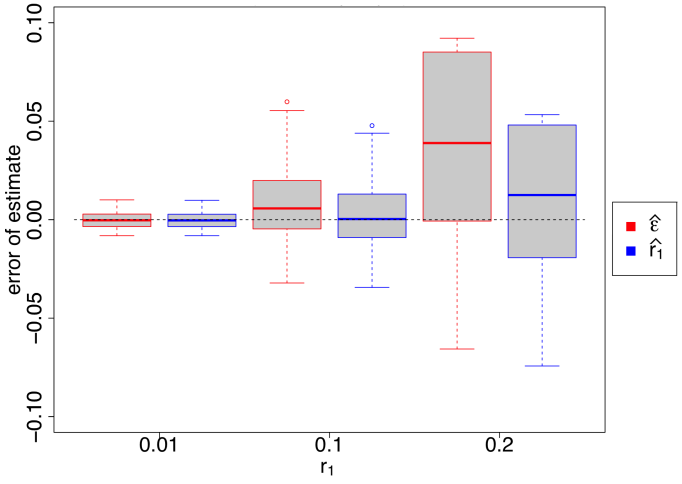
	$r_1 (r_k)$		
	.05(.659)	.15(.967)	.25(.998)
sketch size = 100	0.97	1.00	1.00
1,000	0.97	0.96	1.00
10,000	0.96	0.96	0.97
100,000	0.94	0.95	0.96

Experimental results

Minimap2 [Li, 2018] and Jabba [Miclotte et al., 2016] read filtering

Minimap2

- Filters out alignment if r_1 estimate is far from error rate
- Estimates r_1 from the number of seeds that match a location
  - $\hat{\epsilon} = \frac{1}{k} \log \frac{n}{m}$
- Using our model improves r_1 estimate.



Experimental results

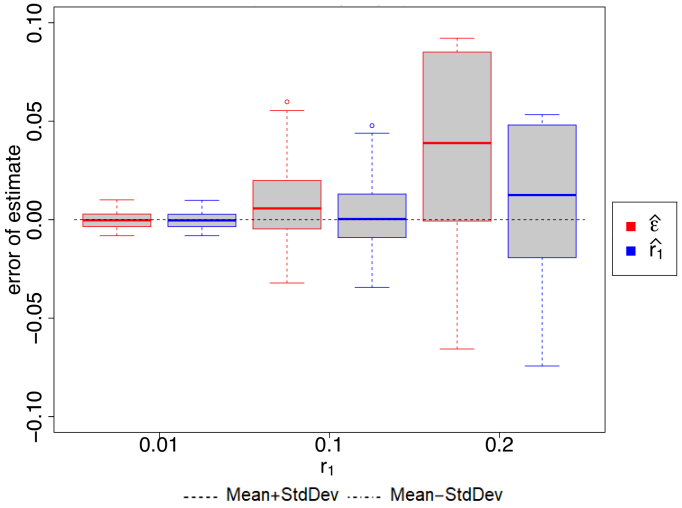
Minimap2 [Li, 2018] and Jabba [Miclotte et al., 2016] read filtering

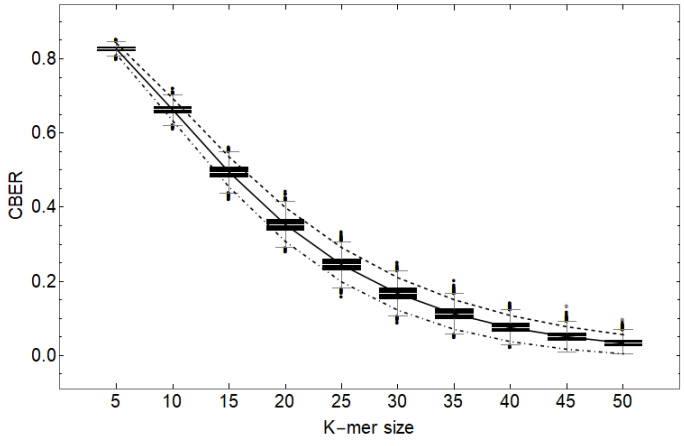
Minimap2

- Filters out alignment if r_1 estimate is far from error rate
- Estimates r_1 from the number of seeds that match a location
  - $\hat{\epsilon} = \frac{1}{k} \log \frac{n}{m}$
- Using our model improves r_1 estimate.

Jabba

- Filters out alignment if coverage by exact regions (C_{ber}) "significantly deviates" from expectation.
- What is "significantly"?
- We can use a hypothesis test for C_{ber}





Conclusion

- Simple mutation model has been widely used but never studied in depth
- We show a technique for deriving hypothesis tests and confidence intervals
  - Exploit the fact that k-mer dependecies are local
- We derive these for a few natural random variables.
- Can we predict when the approximations stop working?
  - E.g. in Binomial, this is when $np(1-p)$ is low

Variable	Expectation	Variance	α interval
N_{mut}	Lq	$L(1-q)(q(2k+\frac{2}{r_1}-1)-2k)$	$Lq \pm z_\alpha \sqrt{\mathrm{Var}(N_{mut})}$
N_{island}	$Lr_1(1-q)$	$Lr_1(1-q)(1-r_1(1-q)(2k+1))$	$\mathrm{E}[N_{island}] \pm z_\alpha \sqrt{\mathrm{Var}(N_{island})}$
N_{ocean}	$Lr_1(1-q)$	$Lr_1(1-q)(1-r_1(1-q)(2k+1))$	$\mathrm{E}[N_{ocean}] \pm z_\alpha \sqrt{\mathrm{Var}(N_{ocean})}$
Jaccard	—	—	(see prev slide)
minhash Jaccard	—	—	(j_{low}, j_{high})
C_{ber} **	$\frac{L(1-q)(1+r_1(k-1))}{L+k-1}$	see paper	$\mathrm{E}[C_{ber}] \pm z_\alpha \sqrt{\mathrm{Var}(C_{ber})}$

References I

Hoeffding, W., Robbins, H., et al. (1948).

The central limit theorem for dependent random variables.
Duke Mathematical Journal, 15(3):773–780.

Li, H. (2018).

Minimap2: pairwise alignment for nucleotide sequences.
Bioinformatics, 34(18):3094–3100.

Miclotte, G., Heydari, M., Demeester, P., Rombauts, S., Van de Peer, Y., Audenaert, P., and Fostier, J. (2016).

Jabba: hybrid error correction for long sequencing reads.
Algorithms for Molecular Biology, 11(1):1–12.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., and Phillippy, A. M. (2016).

Mash: fast genome and metagenome distance estimation using minhash.
Genome biology, 17(1):132.

Ross, N. (2011).

Fundamentals of Stein's method.
Probability Surveys, 8:210–293.