The minimizer Jaccard estimator is biased and inconsistent

Paul Medvedev

joint work with
Mahdi Belbasi, Antonio Blanca, Robert S. Harris, and David Koslicki
Pennsylvania State University

Published in ISMB proceedings (here)
2022

These slides and talk are available from my homepage medvedevgroup.com

Sketching in bioinformatics

- A sketch is a much smaller representation of a dataset.
- Allows downstream algorithms
    - to be faster
    - to use less memory
    - to tackle bigger datasets
- Potential to create accuracy issues

Sketching techniques in bioinformatics	
the hyperloglog sketch	[Flajolet et al., 2007]
the modulo sketch	[Broder, 1997, Schleimer et al., 2003]
minhash with optimal densification	[Shrivastava, 2017, Zhao, 2019]
order minhash	[Marçais et al., 2019]
count minsketch	[Cormode and Muthukrishnan, 2004]

Used for sequence comparison	
to compute genomic distances	[Baker and Langmead, 2019]
to search sequence databases	[Pierce et al., 2019]
to estimate edit distance	[Marçais et al., 2019]
for k-mer counting	[Crusoe et al., 2015]

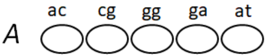
The minimizer sketch of a sequence

[Schleimer et al., 2003, Roberts et al., 2004]

Definition

- Given a sequence A, a $k \geq 2$, and a window size $w \geq 2$.
- A hash function chosen uniformly at random.
- For every w consecutive k-mers, the *minimizer* is the k-mer with the smallest hash value.
- The *minimizer sketch* of A is the set of all minimizers.

Example: A =acggat, $k = 2$, $w = 2$



Used in

- sequence alignment
  - minimap
- sequence mapping
  - mashmap
- seed generation in assembly
  - peregrine
- transcriptome clustering and error correction
  - isONclust, isONcorrect
- etc. . .

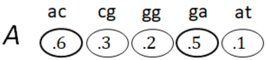
The minimizer sketch of a sequence

[Schleimer et al., 2003, Roberts et al., 2004]

Definition

- Given a sequence A, a $k \geq 2$, and a window size $w \geq 2$.
- A hash function chosen uniformly at random.
- For every w consecutive k-mers, the *minimizer* is the k-mer with the smallest hash value.
- The *minimizer sketch* of A is the set of all minimizers.

Example: $A =$acggat, $k = 2$, $w = 2$



Used in

- sequence alignment
  - minimap
- sequence mapping
  - mashmap
- seed generation in assembly
  - peregrine
- transcriptome clustering and error correction
  - isONclust, isONcorrect
- etc. . .

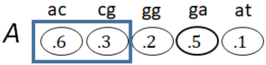
The minimizer sketch of a sequence

[Schleimer et al., 2003, Roberts et al., 2004]

Definition

- Given a sequence A, a $k \geq 2$, and a window size $w \geq 2$.
- A hash function chosen uniformly at random.
- For every w consecutive k-mers, the *minimizer* is the k-mer with the smallest hash value.
- The *minimizer sketch* of A is the set of all minimizers.

Example: $A =$acggat, $k = 2$, $w = 2$



Used in

- sequence alignment
  - minimap
- sequence mapping
  - mashmap
- seed generation in assembly
  - peregrine
- transcriptome clustering and error correction
  - isONclust, isONcorrect
- etc. . .

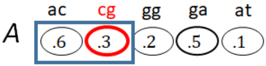
The minimizer sketch of a sequence

[Schleimer et al., 2003, Roberts et al., 2004]

Definition

- Given a sequence A, a $k \geq 2$, and a window size $w \geq 2$.
- A hash function chosen uniformly at random.
- For every w consecutive k-mers, the *minimizer* is the k-mer with the smallest hash value.
- The *minimizer sketch* of A is the set of all minimizers.

Example: $A =$acggat, $k = 2$, $w = 2$



Used in

- sequence alignment
  - minimap
- sequence mapping
  - mashmap
- seed generation in assembly
  - peregrine
- transcriptome clustering and error correction
  - isONclust, isONcorrect
- etc. . .

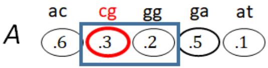
The minimizer sketch of a sequence

[Schleimer et al., 2003, Roberts et al., 2004]

Definition

- Given a sequence A, a $k \geq 2$, and a window size $w \geq 2$.
- A hash function chosen uniformly at random.
- For every w consecutive k-mers, the *minimizer* is the k-mer with the smallest hash value.
- The *minimizer sketch* of A is the set of all minimizers.

Example: $A =$acggat, $k = 2$, $w = 2$



Used in

- sequence alignment
  - minimap
- sequence mapping
  - mashmap
- seed generation in assembly
  - peregrine
- transcriptome clustering and error correction
  - isONclust, isONcorrect
- etc. . .

The minimizer sketch of a sequence

[Schleimer et al., 2003, Roberts et al., 2004]

Definition

- Given a sequence A, a $k \geq 2$, and a window size $w \geq 2$.
- A hash function chosen uniformly at random.
- For every w consecutive k-mers, the *minimizer* is the k-mer with the smallest hash value.
- The *minimizer sketch* of A is the set of all minimizers.

Example: $A =$acggat, $k = 2$, $w = 2$

Used in

- sequence alignment
  - minimap
- sequence mapping
  - mashmap
- seed generation in assembly
  - peregrine
- transcriptome clustering and error correction
  - isONclust, isONcorrect
- etc. . .

The minimizer sketch of a sequence

[Schleimer et al., 2003, Roberts et al., 2004]

Definition

- Given a sequence A, a $k \geq 2$, and a window size $w \geq 2$.
- A hash function chosen uniformly at random.
- For every w consecutive k-mers, the *minimizer* is the k-mer with the smallest hash value.
- The *minimizer sketch* of A is the set of all minimizers.

Example: $A =$acggat, $k = 2$, $w = 2$

Used in

- sequence alignment
  - minimap
- sequence mapping
  - mashmap
- seed generation in assembly
  - peregrine
- transcriptome clustering and error correction
  - isONclust, isONcorrect
- etc. . .

The minimizer sketch of a sequence

[Schleimer et al., 2003, Roberts et al., 2004]

Definition

- Given a sequence A, a $k \geq 2$, and a window size $w \geq 2$.
- A hash function chosen uniformly at random.
- For every w consecutive k-mers, the *minimizer* is the k-mer with the smallest hash value.
- The *minimizer sketch* of A is the set of all minimizers.

Example: $A =$acggat, $k = 2$, $w = 2$

Used in

- sequence alignment
  - minimap
- sequence mapping
  - mashmap
- seed generation in assembly
  - peregrine
- transcriptome clustering and error correction
  - isONclust, isONcorrect
- etc. . .

The minimizer sketch of a sequence

[Schleimer et al., 2003, Roberts et al., 2004]

Definition

- Given a sequence A, a $k \geq 2$, and a window size $w \geq 2$.
- A hash function chosen uniformly at random.
- For every w consecutive k-mers, the *minimizer* is the k-mer with the smallest hash value.
- The *minimizer sketch* of A is the set of all minimizers.

Example: $A =$acggat, $k = 2$, $w = 2$

Used in

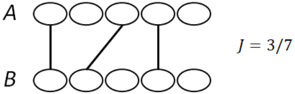
- sequence alignment
  - minimap
- sequence mapping
  - mashmap
- seed generation in assembly
  - peregrine
- transcriptome clustering and error correction
  - isONclust, isONcorrect
- etc. . .

The minimizer Jaccard estimator

[Jain et al., 2017, Jain et al., 2018a, Jain et al., 2018b]

Jaccard similarity

- percent of k-mers that are shared
- Let S be the number of k-mers that appear in both sequences.
- Let U be the number of k-mers that appear in at least one of the sequences.
- $J(A, B) = \frac{S}{U}$



The *minimizer Jaccard estimator* $\widehat{J}$

- $\widehat{J}(A, B) = J(A_{\text{sketch}}, B_{\text{sketch}})$
- The Jaccard of the minimizer sketches

Mapper application (mashmap)

- Used to reject putative mappings

Main Question

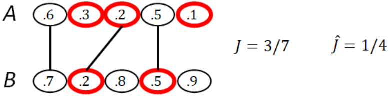
- Is $\widehat{J}$ an accurate estimator of Jaccard?

The minimizer Jaccard estimator

[Jain et al., 2017, Jain et al., 2018a, Jain et al., 2018b]

Jaccard similarity

- percent of k-mers that are shared
- Let S be the number of k-mers that appear in both sequences.
- Let U be the number of k-mers that appear in at least one of the sequences.
- $J(A,B) = \frac{S}{U}$



The *minimizer Jaccard estimator* $\widehat{J}$

- $\widehat{J}(A,B) = J(A_{\text{sketch}}, B_{\text{sketch}})$
- The Jaccard of the minimizer sketches

Mapper application (mashmap)

- Used to reject putative mappings

Main Question

- Is $\widehat{J}$ an accurate estimator of Jaccard?

Measuring accuracy

Bias and consistency

Bias

- An estimator $\widehat{J}$ is *unbiased* if $\mathrm{E}[\widehat{J}] = J$

Consistency

- An estimator $\widehat{J}$ is *consistent* if $\lim \mathrm{E}[\widehat{J}] = J$

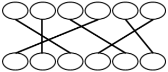
For example, the minhash Jaccard estimator is unbiased [Broder, 1997]

Evidence of bias in mashmap

	divergence (%)		
true	10.0	5.0	1.0
mashmap	11.1	5.9	1.4

Why would the minimizer Jaccard estimator be biased?

- Jaccard does not depend on the order of k-mers but the estimator does.

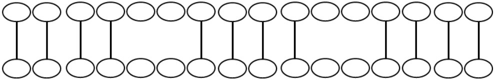


Example: one of the sequences is a k-mer permutation of the other.

- Jaccard is one
- Jaccard estimator depends on randomness so has a distribution
- Jaccard estimator can never be ≥ 1, so expectation is < 1

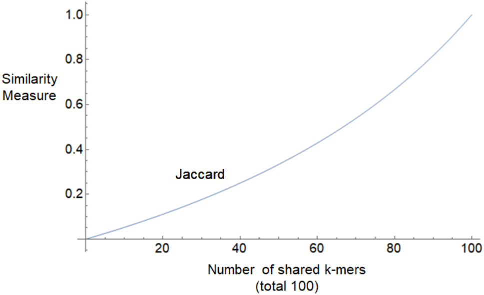
More intuition behind bias

What about cases where the order does not change?



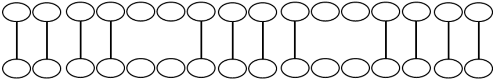
Question: Why should $\widehat{J}$ be UNbiased?

- Higher $J \iff$ More shared k-mers $\iff$ more shared minimizers $\iff$ Higher $\widehat{J}$
- But then lots of functions exist that have this property
- On the other hand, $\widehat{J}$ and J have the same structure (a ratio)
- But $\mathrm{E}[\widehat{S}]$ is not proportional to S.



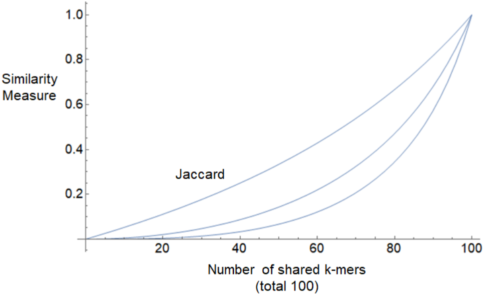
More intuition behind bias

What about cases where the order does not change?



Question: Why should $\widehat{J}$ be UNbiased?

- Higher $J \iff$ More shared k-mers $\iff$ more shared minimizers $\iff$ Higher $\widehat{J}$
- But then lots of functions exist that have this property
- On the other hand, $\widehat{J}$ and J have the same structure (a ratio)
- But $\mathrm{E}[\widehat{S}]$ is not proportional to S.



Theoretical results

General sequence pairs

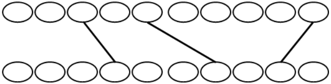
- It is enough to consider the simpler case when A and B are duplicate-free.
- Theorem 1
  - For all pairs of duplicate-free sequences,

    $$\mathcal{B} - \varepsilon \leq \mathrm{E}[\widehat{J}] - J \leq \mathcal{B} + \varepsilon.$$

    where
    - $\mathcal{B}$ is a function of the matching graph and
    - $0 \leq \varepsilon \leq \frac{100w^2}{\sqrt[3]{L}}$
- Theorem 2
  - Assume the two sequences are padded.
  - If $0 < J < 1$, then $\mathcal{B}$ is strictly negative.
  - If $J = 0$, then $\mathcal{B} = 0$.
- In other words:
  - Minimizer Jaccard estimator is biased.
  - The bias can be derived from the matching graph
    - i.e. layout of the shared k-mers

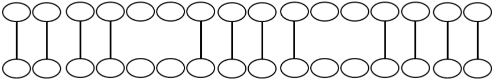
Theoretical results

Examples of concrete families of sequence pairs



Theorem 3: sparsely matched sequences

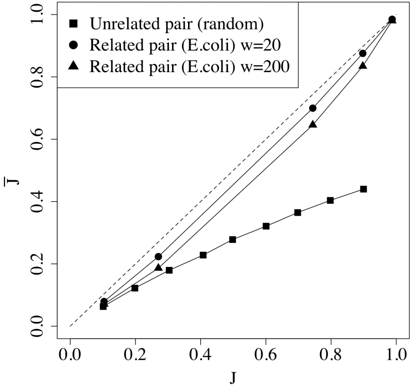
- Assume:
    - distance between any two shared k-mers is $\geq 2w - 1$.
    - two sequences are padded.
- Then, $\mathcal{B}(A, B; w) \leq -J(A, B)\frac{3w^2 - 3w}{8w^2 - 2}$.
- In other words, $E[\widehat{J}]$ is at most 63% of J.



Theorem 4: densely matched sequences

- Assume B is generated from A by mutating every j^{th} nucleotide, for $j > w + 2k$.
- $\mathcal{B}(A, B; w) = \frac{2\ell(\ell j + k)h(w)}{(\ell(j+k) + 2k - \ell h(w))(\ell(j+k) + 2k)}$,
    - where $\ell = \frac{L-k}{j}$, $h(w) = \frac{(w+1)(1-2(H_{2w} - H_w))}{2}$ and $H_n = \sum_{i=1}^{n} \frac{1}{i}$.
- For $k = 15$, $w = 10$, $L \approxeq 10,000$, $j = 43$,
    - $E[\widehat{J}]$ is about 10% smaller than J.

Extent of empirical bias on real sequences



- Related sequence pairs
  - Take arbitrary substring of *E.coli* with $L = 10,000$.
  - Mutate every position with probability $r_1 \in \{.001, .005, .01, .05, .1\}$.
  - Set $k = 16$ and $w \in \{20, 200\}$.
- Unrelated sequence pairs
  - Given a target Jaccard j and fixed k,
  - . . . generate uniformly random string
  - . . . long enough to have the expected Jaccard of j.
  - $w = 20$

Effect of bias on mashmap divergence esimates

	divergence (%)		
true	10.0	5.0	1.0
mashmap	11.1	5.9	1.4
mashmap “Poisson-adjusted”	10.5	5.7	1.4
mashmap adjusted + unbiased	10.0	5.0	1.0

Experiment

- Simulate a read from *E.coli* with substitutions to achieve the given divergence
- Map reads to *E.coli* with mashmap
- mashmap estimates sequence divergence from Jaccard estimator
    - Mash formula

Table

- median divergence reported by mashmap, over 100 trials.
- “Poisson-adjusted”
    - Correction to Mash formula
- “unbiased”
    - Replace $\widehat{J}$ with J

Conclusion

Summary

- The minimizer Jaccard estimator suffers from bias and inconsistency
- On unrelated sequences, the bias can be drastic
- On pairs of sequences related via mutations the bias is less but is substantial
- The bias cannot be removed by changing the window size
- Bias causes inaccurate estimates of divergence during mapping

Where does this leave the minimizer Jaccard estimator?

- Makes sure you empirically test the bias for your application before using it.
- Make sure that the bias is acceptable in your application
  - Read mapping?
    - probably fine
  - Fine grained phylogeny reconstructions?
    - probably not
- Or, use a consistent sketch
  - FracMinHash [Hera et al., 2022, Pierce et al., 2019]
  - mod hash [Broder, 1997]
  - universe minimizer [Ekim et al., 2021]

References I

Baker, D. N. and Langmead, B. (2019).
Dashing: fast and accurate genomic distances with hyperloglog.
Genome biology, 20(1):1–12.

Broder, A. Z. (1997).
On the resemblance and containment of documents.
In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pages 21–29. IEEE.

Cormode, G. and Muthukrishnan, S. (2004).
An improved data stream summary: The count-min sketch and its applications.
In *Latin American Symposium on Theoretical Informatics*, pages 29–38. Springer.

Crusoe, M. R., Alameldin, H. F., Awad, S., Boucher, E., Caldwell, A., Cartwright, R., Charbonneau, A., Constantinides, B., Edvenson, G., Fay, S., et al. (2015).
The khmer software package: enabling efficient nucleotide sequence analysis.
F1000Research, 4.

Ekim, B., Berger, B., and Chikhi, R. (2021).
Minimizer-space de bruijn graphs.
bioRxiv.

Flajolet, P., Fusy, É., Gandouet, O., and Meunier, F. (2007).
Hyperloglog: the analysis of a near-optimal cardinality estimation algorithm.
In *Discrete Mathematics and Theoretical Computer Science*, pages 137–156.

Hera, M. R., Pierce-Ward, N. T., and Koslicki, D. (2022).
Debiasing FracMinHash and deriving confidence intervals for mutation rates across a wide range of evolutionary distances.
Technical report.

References II

Jain, C., Dilthey, A., Koren, S., Aluru, S., and Phillippy, A. M. (2017).
A fast approximate algorithm for mapping long reads to large reference databases.
In *International Conference on Research in Computational Molecular Biology*, pages 66–81. Springer.

Jain, C., Koren, S., Dilthey, A., Phillippy, A. M., and Aluru, S. (2018a).
A fast adaptive algorithm for computing whole-genome homology maps.
Bioinformatics, 34(17):i748–i756.

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., and Aluru, S. (2018b).
High throughput ani analysis of 90k prokaryotic genomes reveals clear species boundaries.
Nature communications, 9(1):1–8.

Marçais, G., DeBlasio, D., Pandey, P., and Kingsford, C. (2019).
Locality-sensitive hashing for the edit distance.
Bioinformatics, 35(14):i127–i135.

Marçais, G., Pellow, D., Bork, D., Orenstein, Y., Shamir, R., and Kingsford, C. (2017).
Improving the performance of minimizers and winnowing schemes.
Bioinformatics, 33(14):i110–i117.

Pierce, N. T., Irber, L., Reiter, T., Brooks, P., and Brown, C. T. (2019).
Large-scale sequence comparisons with sourmash.
F1000Research, 8.

Roberts, M., Hayes, W., Hunt, B. R., Mount, S. M., and Yorke, J. A. (2004).
Reducing storage requirements for biological sequence comparison.
Bioinformatics, 20(18):3363–3369.

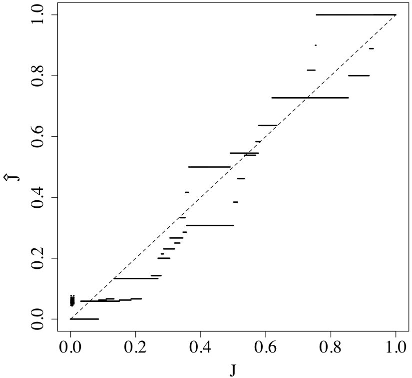
Schleimer, S., Wilkerson, D. S., and Aiken, A. (2003).
Winnowing: local algorithms for document fingerprinting.
In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pages 76–85. ACM.

References III

Shrivastava, A. (2017).
Optimal densification for fast and accurate minwise hashing.
In *International Conference on Machine Learning*, pages 3154–3163. PMLR.

Zhao, X. (2019).
BinDash, software for fast genome distance estimation on a typical personal laptop.
Bioinformatics, 35(4):671–673.

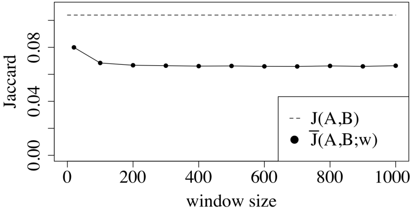
Extent of empirical bias during the mapping process



Simulate (roughly) the behavior of mashmap.

- *A* is a 1,000 long arbitrary substring of *E.coli*.
- Compare against all 1,000-long mapping locations.
- Each point represents *A* against a putative mapping location *B*.
- We use default params of mashmap ($k = 16$ and $w = 200$).

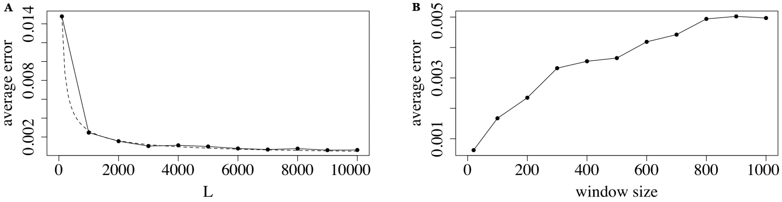
Effect of window size on bias



Pair of sequences related via simple mutation process

- Mutation rate of 0.1
- $L = 10,000$
- $k = 16$
- $w \in \{20, 100, 200, \ldots, 1000\}$
- 50 mutation replicates.

Empirical accuracy of our $\mathcal{B}$ formula



- For the previous examples,
    - the empirical error is roughly one/two orders of magnitude smaller than bias
- The error decreases with L
    - The average J, over the mutation replicates, is between .101 and .106.
    - The average empirical bias ranges between -0.023 and -0.027.
- The error increases with window size
    - The average J is .104.

How we prove bias

high level teaser only

Need to derive $\mathrm{E}[\widehat{J}]$.

- Let $\widehat{S}$ be the number of minimizers shared between the two sequences.
- Let $\widehat{U}$ be the number of minimizers in at least one of the sequences.
- $\widehat{J} = \frac{\widehat{S}}{\widehat{U}}$

Step 1: Dealing with the ratio

- i.e. $\mathrm{E}[\frac{\widehat{S}}{\widehat{U}}] \neq \frac{\mathrm{E}[\widehat{S}]}{\mathrm{E}[\widehat{U}]}$.
- Lemma: $\left|\mathrm{E}\left[\frac{\widehat{S}}{\widehat{U}}\right] - \frac{\mathrm{E}[\widehat{S}]}{\mathrm{E}[\widehat{U}]}\right| \leq \frac{13w^2}{\sqrt[3]{L}}$.

Step 2: Deriving $E[\widehat{U}]$

- Lemma: $\mathrm{E}[\widehat{U}] = 2\left(\frac{2L}{w+1}\right) - \mathrm{E}[\widehat{S}]$.
    - follows from [Schleimer et al., 2003, Roberts et al., 2004].

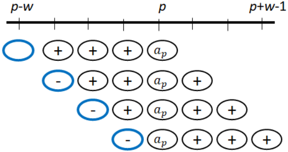
Step 3: Deriving $\mathrm{E}[\widehat{S}]$

- See following slides for intuition

Deriving $\mathrm{E}[\widehat{S}]$

Reducing to sum of charging indicator variables

- Let p be a position of a minimizer.
- Take the leftmost window in which p is a minimizer.
- Let i be the position right before that.
- We say that p *charges* i [Schleimer et al., 2003, Marçais et al., 2017].
- We know that $a_{i+1}, \ldots, a_{i+w}$, except for a_p, are more than a_p
- What about a_i?
  - If $i > p - w$, then $a_i < a_p$.
  - If $i = p - w$, then a_i can be anything.



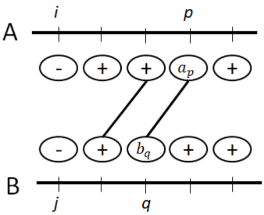
- Let $X^A_{i,p}$ indicate if p charged i in sequence A.
- Let $X^B_{j,q}$ indicate if q charged j in sequence B.

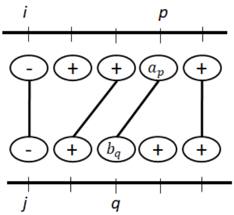
$$\mathrm{E}[\widehat{S}] = \sum_i \sum_j \sum_{\substack{p,q \\ \text{s.t.} \\ A_p = B_q}} \Pr[X^A_{i,p} = 1, X^B_{j,q} = 1]$$

$$= \sum_i \sum_j \sum_{\substack{p,q \\ \text{s.t.} \\ A_p = B_q}} \int_0^1 \Pr[X^A_{i,p} = 1, X^B_{j,q} = 1 \mid a_p = b_q = x]\, dx$$

Deriving $\mathrm{E}[\widehat{S}]$

Computing $\Pr[X^{A}_{i,p} = 1, X^{B}_{j,q} = 1 \mid a_p = b_q = x]$





- Example for intuition
  - $w = 4$
- Left example
  - Two values that need to be less than x
  - Five values that need to be more than x
  - Prob that both charging indicators are one, conditioned on x, is
    - $x^2(1-x)^5$
- Right example
  - One value that needs to be less than x
  - Four values that need to be more than x
  - Prob that both charging indicators are one, conditioned on x, is
    - $x(1-x)^4$

Deriving $\mathrm{E}[\widehat{S}]$

- $\Pr[X^A_{i,p} = 1, X^B_{j,q} = 1 \mid a_p = b_q = x] = x^\alpha (1-x)^\beta$
  - α and β depend on the layout of the shared k-mers in $A_i, \ldots, A_{i+w}$ and $B_j, \ldots, B_{j+w}$.
- What matters?
  - How many k-mers are shared between $\{A_{i+1}, \ldots, A_{i+w}\}$ and $\{B_{j+1}, \ldots, B_{j+w}\}$?
  - Does $A_i = B_j$?
  - Is $A_i \in \{B_{j+1}, \ldots, B_{j+w}\}$?
  - Is $B_j \in \{A_{i+1}, \ldots, A_{i+w}\}$?
  - Does $A_{i+w} = B_{j+w}$?
  - Is $A_{i+w} \in \{B_{j+1}, \ldots, B_{j+w}\}$?
  - Is $B_{j+w} \in \{A_{i+1}, \ldots, A_{i+w}\}$?
- The answer to these questions define a *configuration*
- For each configuration, its *number* is the number of index pairs i, j that are in that configuration.
- $\mathrm{E}[\widehat{S}]$ is a linear of combination of configuration numbers,
  - weighted by functions of the form $\int_0^1 x^\alpha (1-x)^\beta \, dx$.
- We get a closed form solution to
  - $\mathcal{B} \triangleq \mathrm{E}[\widehat{J}] - J \approx \frac{\mathrm{E}[\widehat{S}]}{\frac{4L}{w+1} - \mathrm{E}[\widehat{S}]} - J$.