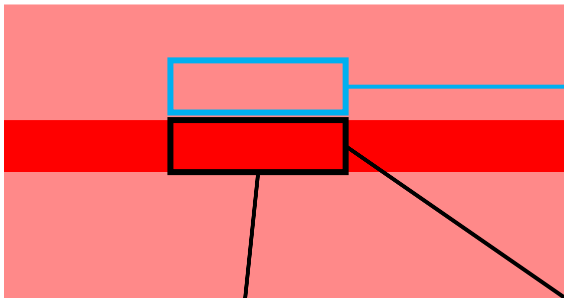
Previous Versions:



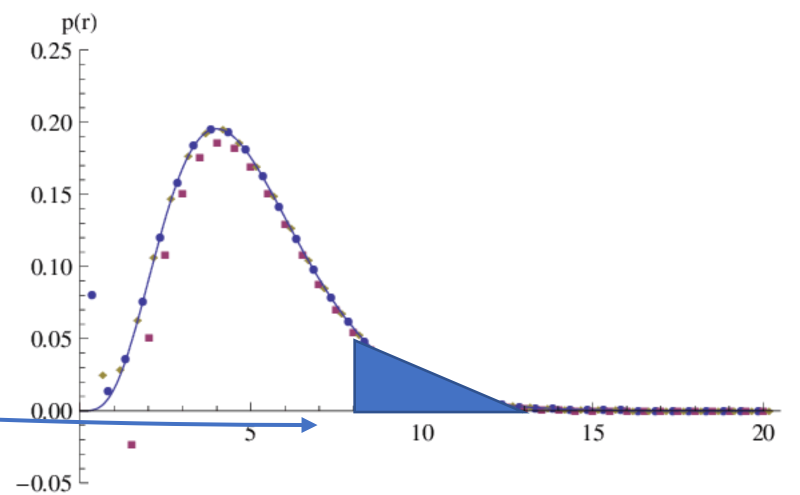
The mean value of the background regions: **a**

If we don't know anything about the central region, we would estimate it is similar with the background, which follows a Poisson distribution with expected value = **a**.

But we observed a larger value **b**.
The prob. of observing a value larger than **b** from Poisson(**a**) is:

$$p = 1 - \mathrm{cdf}(b)$$

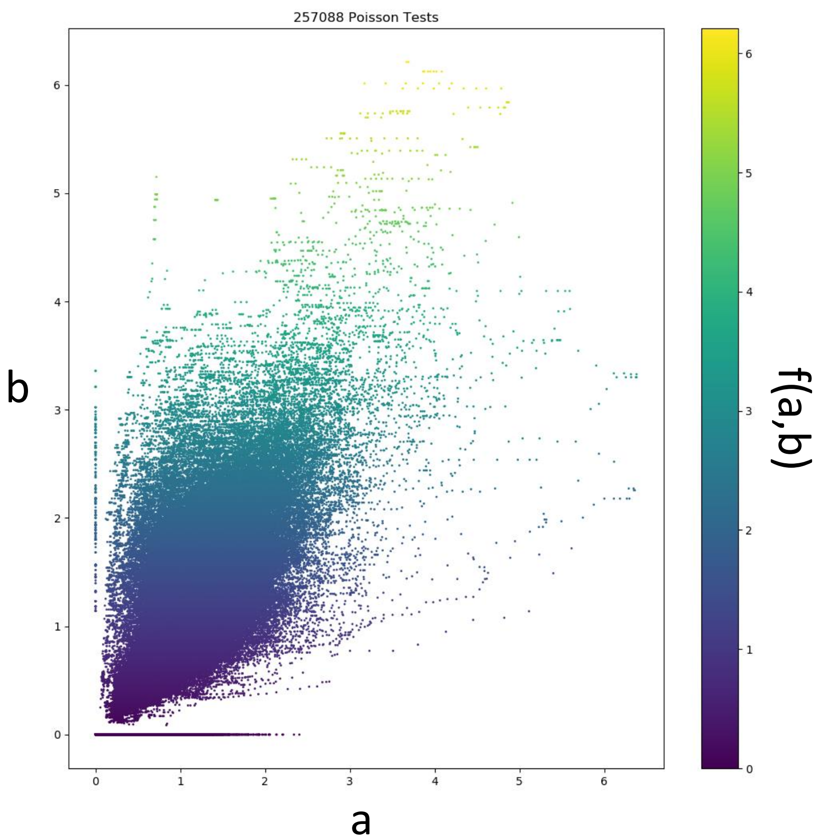
(which is a common method in loop callers.)
Calculating this is slow!

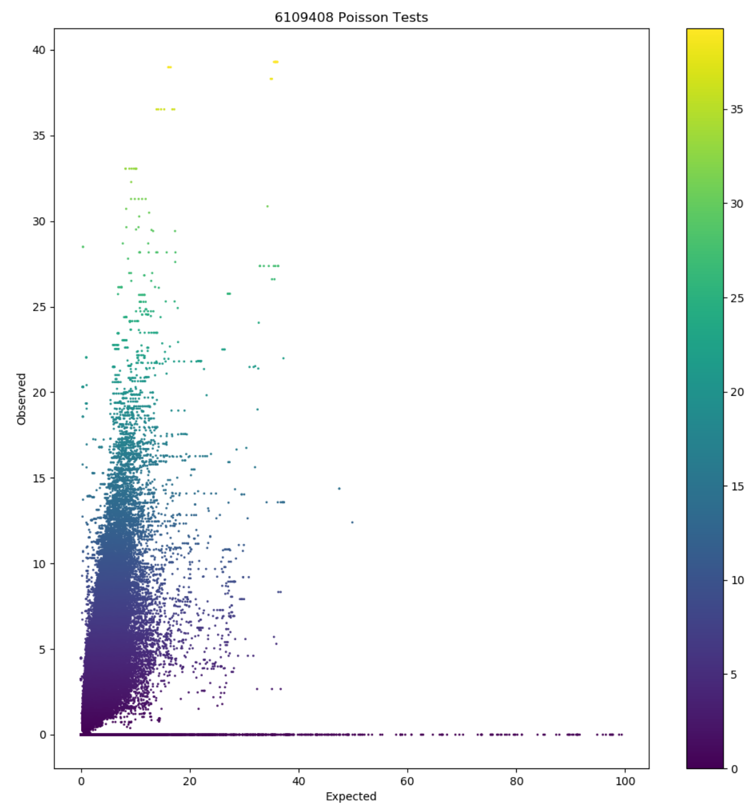


* Let $x \sim$ Poission(a) and $f(a,b) = -\log_{10}[P(x > b)] = -\log_{10}[1 - \mathrm{cdf}(b)]$.

GM12878 chr1 25Kb resolution
- 257,088 calculations of f(a, b)
- Time: 219s (the entire program)
- 8 stripes

HFF chr1 1Kb resolution
- 6,109,408 calculations
- 3960s
- 207 stripes





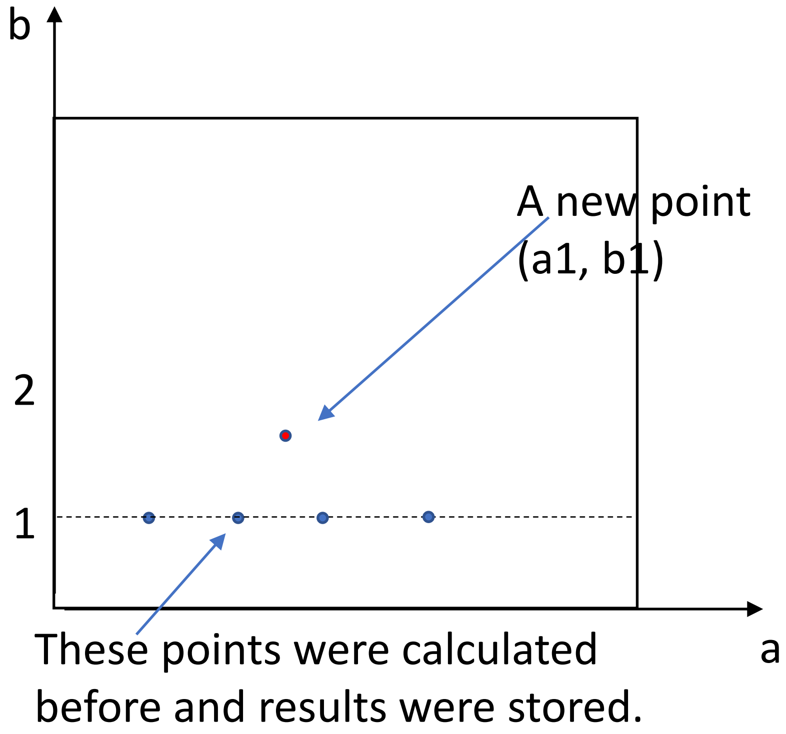
Strategy:

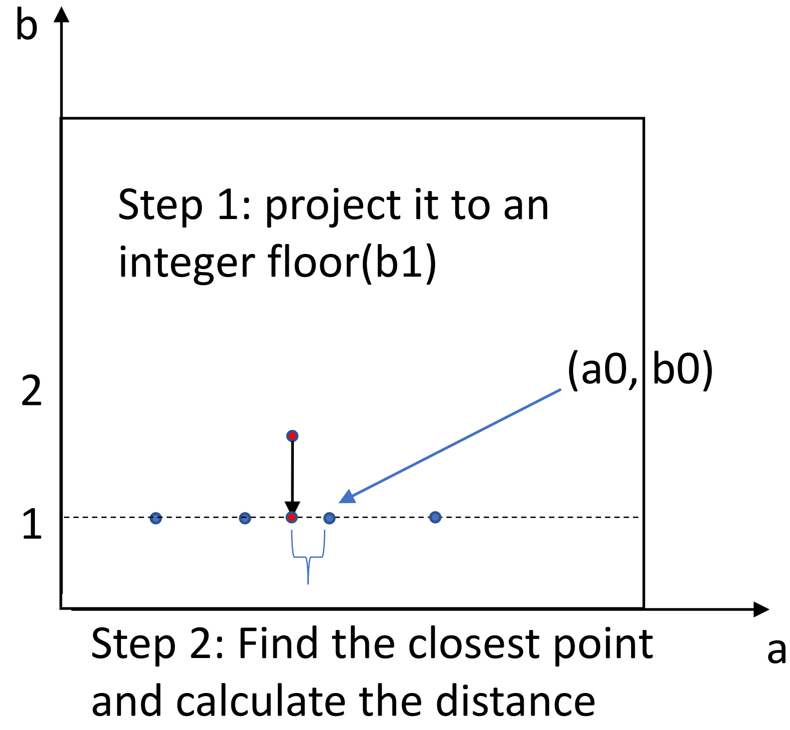
- Record the “expected-observed” pair and their p value results for each calculation
- If a new pair (e.g., 4.99-9.05) is similar to a calculated pair (e.g., 5.00-9.00), we direct use the p value from the previous pair
- There should be a “tolerance level”

Let x~Poission(a) and $f(a,b) = -\log_{10}[P(x > b)] = -\log_{10}[1 - cdf(b)]$.
Based on the definition of Poisson dist.

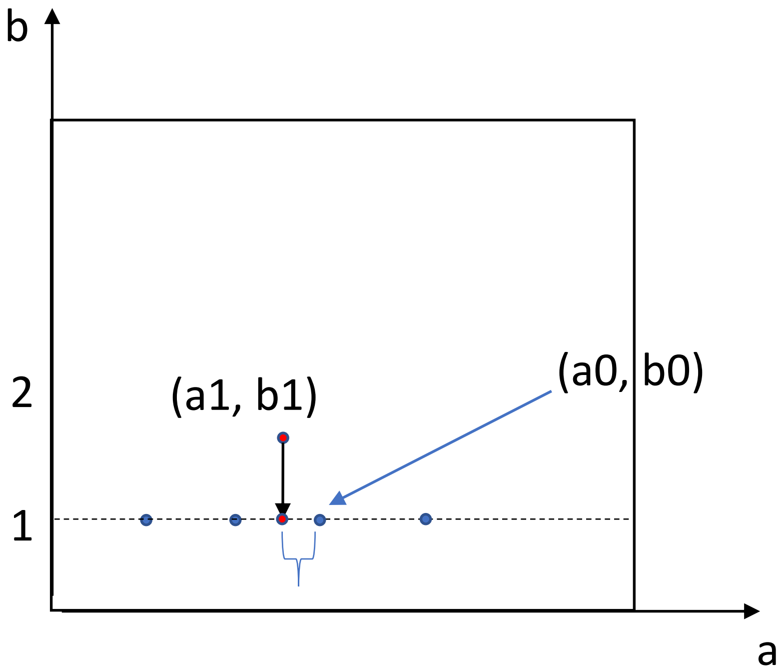
- $f(a, b) = f(a, floor(b))$. E.g., $f(5, 10) = f(5, 10.5) = f(5, 10.99) = 1.863$
- $f(a, b) \approx f(a+\Delta a, b)$ when Δa is small. E.g., $f(5, 10) = 1.863$, $f(5.05, 10) = 1.835$, $f(5.1, 10) = 1.807$

Strategy:

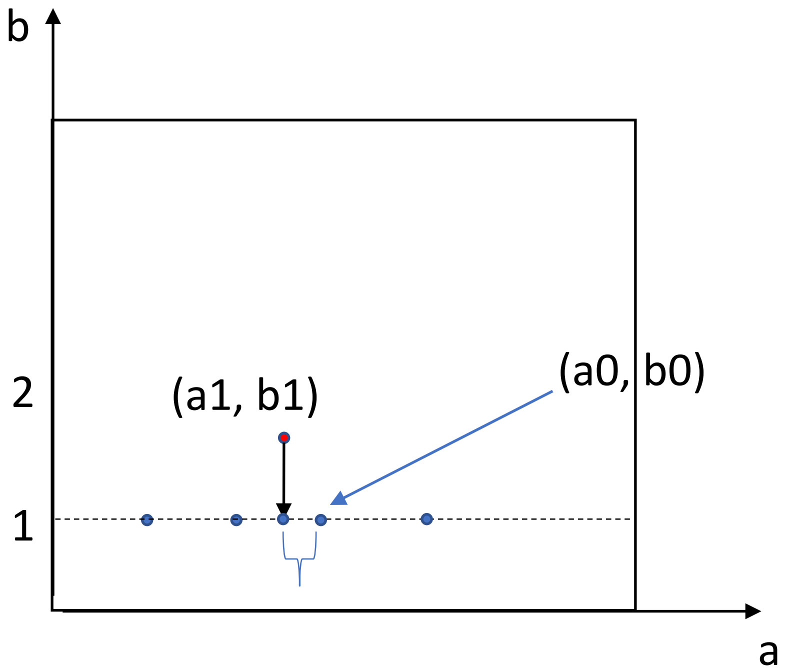

b
a
2
1
A new point
(a1, b1)
These points were calculated
before and results were stored.


b
a
2
1
Step 1: project it to an
integer floor(b1)
(a0, b0)
Step 2: Find the closest point
and calculate the distance

Strategy:



If the distance is shorter than (tolerance * a1), we directly use f(a0, b0) to estimate f(a1, b1). Do not save f(a1, b1).



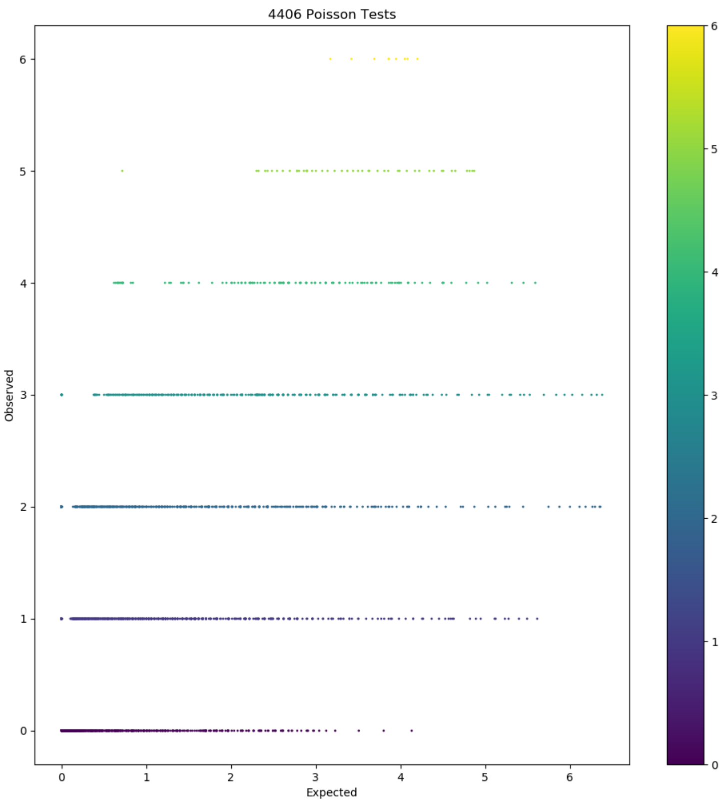
If the distance is longer than (tolerance * a1), we calculate the accurate value of f(a1, b1), i.e., f(a1, b0), and save this result for future use.

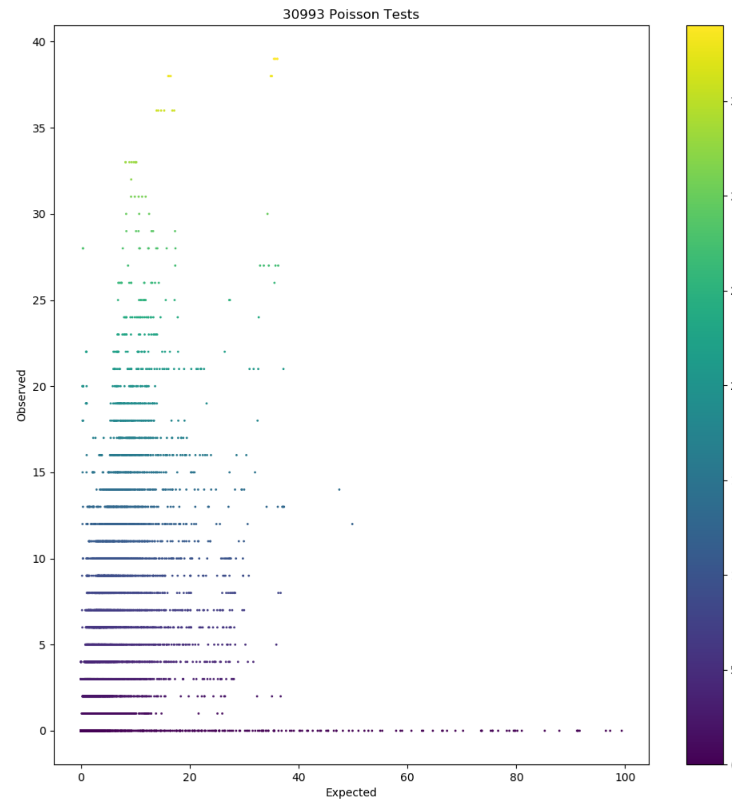
* The searching-inserting operation can be further accelerated by balanced binary tree.

Result (tolerance = 0.02):

GM12878 chr1 25Kb resolution
- 4,406 calculations
- 87s
- 9 stripes

HFF chr1 1Kb resolution
- 30,993 calculations
- 843s
- 201 stripes





The stripes also match quite well.