Statistical Significance of Overlap between Genome Annotations

DS 202: Algorithmic Foundations of Big Data Biology

Mandeep Singh
M.Tech. CDS

Genome Annotations

Genome Annotation

- Sequencing of genome not enough
- Identifying coding and non coding regions of the genome
- Giving a tag to a particular region in genome

Statistical Significance

Statistical Significance

- In Scientific world, we can never prove something to be true
- Simple Statement : If I release the ball that I am holding, it always falls down
- No certainty that the ball falls down even when I do the experiment after 100-200 years.
- Consider the negation of the statement
- We can build evidence against this new proposed statement

Null Hypothesis

- We propose a null hypothesis whenever we talk about statistical significance
- Examples of Null Hypothesis :
  - Defendant is innocent
  - Ball always falls down when released
  - A drug is not effective for heart disease
- Statistical Significant Event basically means we have enough evidence to reject a null hypothesis

p-value

- It basically quantifies how statisitically significant an event is
- Lower the p-value, more statistically significant the event
- If p = 0.1 , it means that there is a 1 % likelihood that this observation happened by chance

Problem Formulation

Problem Formulation

- We need to find out whether the overlap between two annotations is due to some underlying biological phenomenon or is it happening just by chance
- I_f = Reference collection of m intervals
- I = Query collection of n intervals
- g denotes the length of the genomic region of interest
- Each interval is denoted by (u1,u2) where 0 <= u1 < u2 < g
- i is used to index I intervals and j is used to index I_f intervals
- Length of ith interval is li and length of jth interval is xj

Problem Formulation

- Observation : There are k intervals in I_f which overlap with I
- We use a random set of intervals I_r to measure the statistical significance of this observation
- I_r has following properties :
  - I_r has exactly n elements.
  - Intervals in I_r have same lengths as intervals in I
  - Location of intervals in I_r are such that all possible random sets are equally likely

Problem Formulation

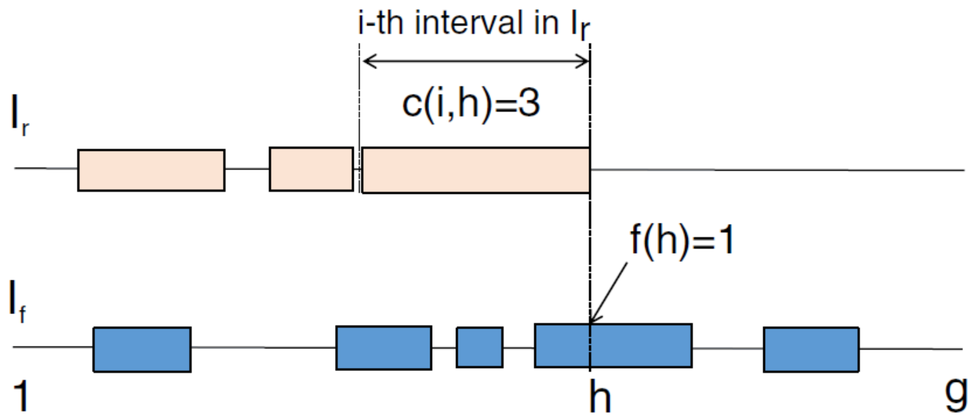
- p – value(k) = Pr(number of intervals in I_f which overlap with I_r >= k).
- Number of possible random sets is very large, and as such the problem would be computationally intractable.
- We make an assumption that intervals in I_r must have same order as I.
- Then we will use a DP algorithm to compute the p-value for all k.

DP Algorithm

- For interval i in I_r, genomic location h, (1<=h<=g), 0<=k<=m, a belonging to {0,1}, let N(i, h, k, a) denote the number of arrangements of the first i intervals in I_r such that:
  - the i-th interval ends exactly at location h.
  - k intervals in I_f are hit by the first i intervals in I_r.
  - a = 0 if the interval from I_f that spans h (if any) has not been counted earlier; a = 1 otherwise.
- We also define N1(i, h, k, a) identically to N(i, h, k, a) with the exception that the i-th interval ends at or before location h.

DP Algorithm

- Let c(i, h) denote the number of intervals in I_f which intersect with (h-li, h) in I_r.
- We also define function f where f(h) = 1 if some interval in I_f spans h and f(h) = 0 otherwise.
- Our overlap definition will change the definition of function c and f. c and f above are calculated if even single nucleotide match is considered overlap



DP Algorithm

$$N_1(i,h,k,a) = \begin{cases} N(i,h,k,a) & h = 1 \\ N(i,h,k,a) + N_1(i,h-1,k,\min\{a, f(h-1)\}) & \text{Otherwise} \end{cases}$$

$$N(i,h,k,a) = \begin{cases} 0 & h < \sum_{x=1}^{i} l_x \text{ or } k < c(i,h) - a \\ 1 & i = 1 \text{ and } k = c(i,h) - a \\ N_1(i-1, h-l_i, k-c(i,h)+a, f(h-l_i)) & \text{Otherwise} \end{cases}$$

$$1 \le i \le n, \quad 1 \le h \le g, \quad 0 \le k \le m, \quad a \in \{0,1\}.$$

DP Algorithm

$$P - \text{value}(k) = \frac{\sum_{\kappa=k}^{m} N_1(n, g, \kappa, 0)}{\sum_{\kappa=0}^{m} N_1(n, g, \kappa, 0)} \ .$$

- Denominator term is really large, so we do our calculations in log scale.
- Time Complexity and Space Complexity is O(ngm)
- We can reduce it to O(ngmv) by using a scaling factor v such that 0 < v <= 1

Poisson Binomial Approximation

- We remove the non-overlapping assumption on I_r.
- Let E_ij be the event that j-th interval in I_f intersect with I-th interval in I_r

$$p_{ij} := \Pr(E_{ij}) = \frac{l_i + x_j - 1}{g}$$

Poisson Binomial Approximation

$$P_j := \Pr(E_j) = \Pr\left(\cup_{i=1}^{n} E_{ij}\right) = 1 - \Pr\left(\cap_{i=1}^{n} \overline{E}_{ij}\right) = 1 - \prod_{i=1}^{n} \Pr\left(\overline{E}_{ij}\right) = 1 - \prod_{i=1}^{n} (1 - \Pr(E_{ij})).$$

- Let X_j be indicator variable such that X_j = 1 iff event E_j occurs. We need to compute $\Pr\left(\sum_j X_j = k\right)$
- Sum of m independent Bernaulli trials with different success probabilities is Poisson Binomial Distribution.

$$\Pr\left(\sum_{j=1}^{m} X_j = k\right) = \sum_{A \in F_k} \prod_{u \in A} P_u \prod_{v \in A^c} (1 - P_v)$$

$$P - \text{value}(k) = \Pr\left(\sum_{j=1}^{m} X_j \geq k\right)$$

Poisson Binomial Approximation

$$\pi_{k,j} = \Pr\left(\sum_{u=1}^{j} X_u = k\right)$$

$$\pi_{-1,j} = \pi_{j+1,j} = 0, j = 0, 1, \ldots, m \text{ and } \pi_{0,0} = 1.$$

$$\pi_{k,j} = P_j \pi_{k-1,j-1} + (1 - P_j)\pi_{k,j-1}, \quad 0 \le k \le m,\ 0 \le j \le m$$

- π_k,j denotes probability of getting k hits in first j intervals in I_f
- Time Complexity = O(m2)
- If intervals in I_f are clumped, then we underestimate p value and when they are far apart, we overestimate p value.

Thank You