Project Report on

Computing the Statistical Significance of Overlap between Genome Annotations

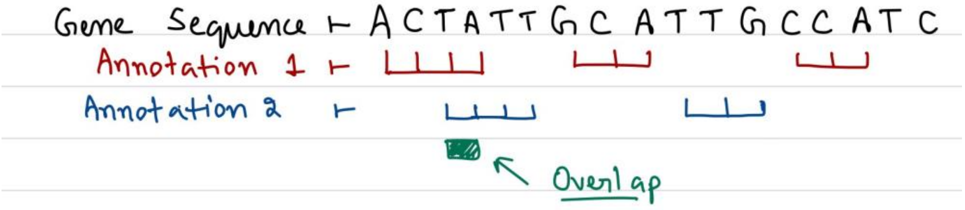
by Mandeep Singh (19534)

1. Genome Annotations

Genome annotation is the process of deriving the structural and functional information of a protein or gene from a raw data set using different analysis, comparison, estimation, precision, and other mining techniques.

It is basically annotating a genome. After sequencing a genome, we need to find the coding regions and the non coding regions of genome. In the coding regions, genome annotation tells what protein does this particular gene segment code.

Overlap between two genome annotations is explained below.



Example: TCGA-CNV Enrichment in Extra-chromosomal DNA.

2. Statistical Significance

A set of measurements or observations in a statistical study is said to be statistically significant if it is unlikely to have occurred by chance.

In scientific world, we can never prove something to be certainly true. For example, we can take a simple statement like if I am holding a ball, and if I release the ball, then it will always fall down, we can prove it to be true in only some instances, but we can not exhaust all the possibilities for indefinite amount of time. It could be possible that 100,200 years from now, it might not fall down, it will probably fall down but we can't say it with certainty. We consider the statement that it stays still and does not fall down. But when we do the experiment, it does fall down, and hence we provide the evidence against the statement that we proposed. The more and more experiments we do, the more evidence we collect against the proposal. After a certain point when we have enough evidence, we reject the idea that this ball stays in place when we release it. We did not prove that the ball always falls down when we release it, instead we disproved that the ball always stays in place when we release it.

When we say that something is statistically significant, we are comparing the result of some kind of analysis with null hypothesis. We can consider an example of jury deciding whether the defendant is guilty. Defendant is consider innocent till proven guilty. Now the innocence can be considered null hypothesis. With enough evidence in form of DNA evidence and witness testimony, jury rejects the idea of innocence.

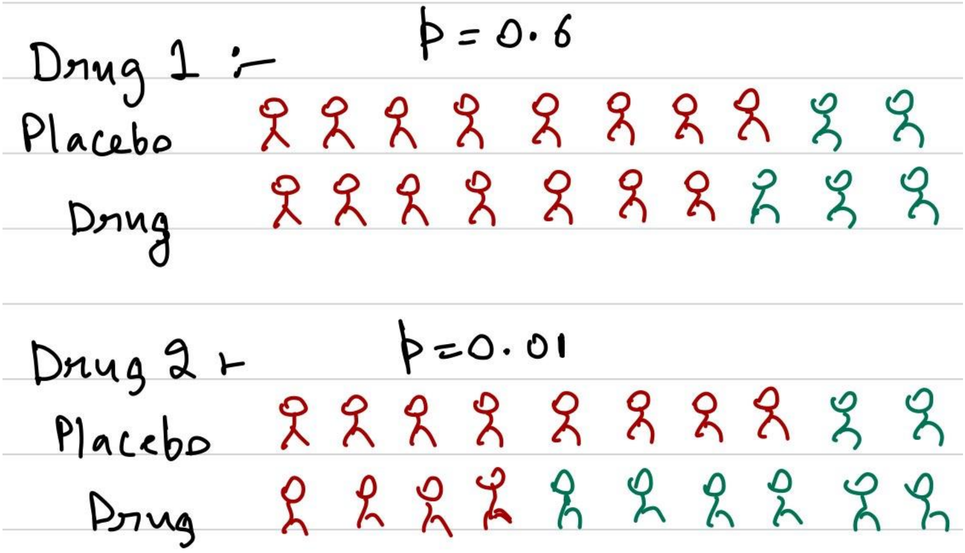
Examples of Null Hypothesis :

- Defendant is innocent
- Ball always falls down when released
- A drug is not effective for heart disease

Statistical Event basically means that we have enough evidence to reject the null hypothesis.

Generally p < 0.5 means Statistically Significant Result. Smaller the p value, greater the confidence we have in rejecting whatever the null hypothesis was.

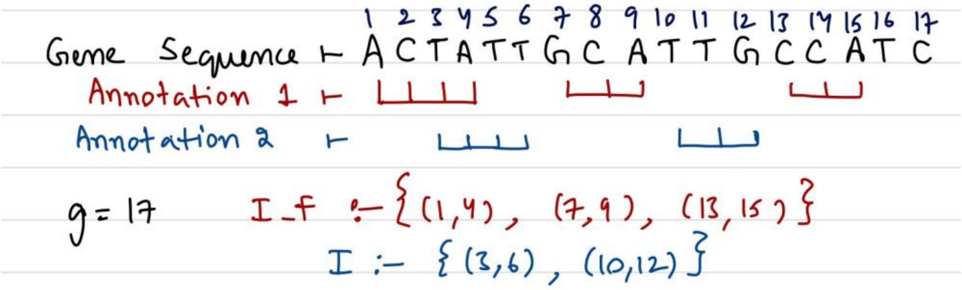
We can take an example of drug testing. Whenever a new drug is discovered for treatment of some disease then it needs to undergo trials. A random sample of people are given a placebo and other sample is given the actual drug. In example below we have 2 cases. Red color indicates the person still has the disease, green color indicates he/she no longer has the disease. For drug 1, the effect is not much different from placebo. p value of 0.6 means that there is a 60% likelihood that this observation might have happened by chance. In second case, drug is being effective in more people and p value of 0.01 means that there is only 1% likelihood we can get this observation by chance. This means that if the null hypothesis were true that is the drug is not effective for disease, the chances of us getting the result that we did is only 1%. Given the observation, the much more likely case is that null hypothesis is actually false. It will be really unusual for us to have this observation if the drug really didn't work.



3. 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.

Let I_f denote a ‘‘reference’’ collection of m intervals and I denote a ‘‘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. For the toy example considered above, g, I_f and I would be as shown below.



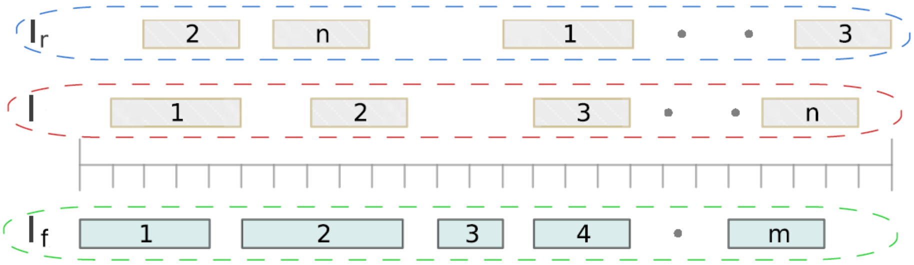
Suppose there are k intervals in I_f which overlap with any interval in I. We basically need to find out that these number of overlaps are due to some biological phenomenon or k or more number of intervals can overlap even by chance with a high probability.

Null Hypothesis : The k intervals are overlapping just by chance and not due to any biological phenomenon.

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.

I_r, I and I_f schematic is given below:



p-value(k) = Probability(number of intervals in I_f which overlap with any interval in 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.

4. DP Algorithm

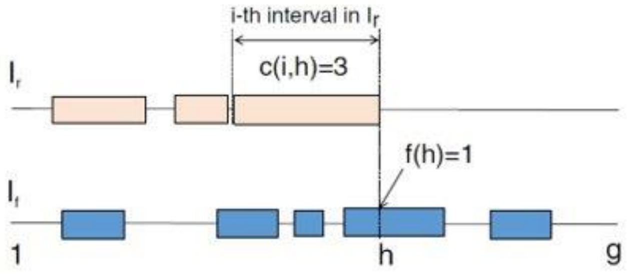
For interval i in I_r, genomic location h, (1<=h<=g), 0<=k<=m, a belonging to {0,1}, let N(i, h, k, a) denote the number of arrangements of the first i intervals in I_r such that:

- 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.

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. f and c calculation for a particular example is shown below.



DP Algorithm :

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

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

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

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

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

5. 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.

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

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

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.

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

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

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

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

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

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.

6. References

1. Josep F. Abril, Sergi Castellano, Encyclopedia of Bioinformatics and Computational Biology, 2019

2. Module 6.1 STAT:1010, The University of Iowa

3. Statistical Significance and p-Values Explained Intuitively - YouTube

4. Sarmashghi S, Bafna V. Computing the Statistical Significance of Overlap between Genome Annotations with iStat. Cell Syst. 2019 Jun 26;8(6):523-529.e4. doi: 10.1016/j.cels.2019.05.006. Epub 2019 Jun 12. PMID: 31202632; PMCID: PMC7200088.