A method to extract peaks in chromosome conformation capture data

Ofir Shukron

December 6, 2014

Motivation

1. Large amount of biological data can be produced in each experiment.
2. Multiple hypotheses are being conducted on it.
3. The is a growing need for methods to control the error rate of multiple hypotheses tests.
4. One would like to avoid type I error while performing multiple tests.
5. Traditional methods which reduces the probability of at least one type I error in multiple tests, were shown to be too restrictive.
6. New method of controlling the error called positive false detection rate (pFDR) was developed.
7. We want to apply this method to find significant looping event in the chromosomes using chromosome capture (CC) data.

Background

1. CC experiment capture millions of encounter events between different parts of the chromosome.
2.

Mathematical derivation

Conducting m hypothesis tests, using p-values, P, as our test statistics.
We fix a rejection region γ, and we reject the null hypothesis H is $P \leq \gamma$. $(\gamma > 0)$
Let V be the number of type I errors, out of R total rejections.
The pFDR is defined as:

$$pFDR = E\left(\frac{V}{R}|R > 0\right)$$

We assume that the null hypothesis H is true $(H = 0)$ with an a priori probability π_0 and false $(H = 1)$ with probability π_1. We write (Storey 2001, Theorem 1)

$$pFDR = \frac{\pi_0 Pr(P \leq \gamma|H = 0)}{Pr(P \leq \gamma)}$$

By the Bayes rule

$$pFDR = Pr(H = 0|P \leq \gamma)$$

Under the null hypothesis, the p-values are uniformly distributed.

$$pFDR = \frac{\pi_0 \gamma}{Pr(P \leq \gamma)}$$

Mathematical derivation

We now need an estimate of π_0 and $Pr(P \leq \gamma)$.
Let R be the total rejected hypotheses, and W the total accepted hypotheses.

$$\hat{\pi_0} == \frac{\#(P_i > \lambda)}{(1-\lambda)m} = \frac{W(\lambda)}{(1-\lambda)m} \qquad 0 \leq \lambda < 1$$

We treat λ as fixed in the following.

$$\hat{Pr}(P \leq \gamma) = \frac{R(\gamma)}{m}$$

Plugging in these estimates and remembering that the pFDR is a conditional probability measure, we have

$$pF\hat{D}R_\lambda(\gamma) = \frac{W(\lambda)\gamma}{(1-\lambda)R(\gamma)(1-(1-\gamma)^m)}$$

The equivalent of the p-values for this statistics is called the q-value.
q-values are the minimum pFDR that can occur when rejecting a statistics with t value.