

A method to extract peaks in chromosome conformation capture data

Ofir Shukron

December 7, 2014

Motivation

- 1 Multiple hypotheses are being tested on large amount of experimental data.
- 2 A criteria to control the error rate is needed.
- 3 To control the error, one would like to reduce type I error.
- 4 Traditional methods controlled the probability of at least one type I error, but were too restrictive.
- 5 New method of controlling the error called positive False Discovery Rate (pFDR) was developed¹.
- 6 We want to apply this method to find frequent specific looping events in the chromosomes using chromosome capture (CC) data.

¹Storey JD. A direct approach to false discovery rates. J. R. Statist. Soc. B (2002)64, Part 3, pp. 479-498

Background

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

Mathematical derivation

Conducting m hypothesis tests, using p-values, P , as our test statistics.

We fix a rejection region $\gamma = [0, \gamma]$, and we reject the null hypothesis H is $P \leq \gamma$. ($\gamma > 0$)

Let V be the number of type I errors and R the total number of 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 null hypotheses, and W the total accepted hypotheses.

$$\hat{\pi}_0 = \frac{\#(P_i > \lambda)}{(1 - \lambda)m} = \frac{W(\lambda)}{(1 - \lambda)m}, \quad 0 \leq \lambda < 1$$

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

$$p\hat{FDR}_\lambda(\gamma) = \frac{W(\lambda)\gamma}{(1 - \lambda)R(\gamma)(1 - (1 - \gamma)^m)}$$

The equivalent of the p-values for the pFDR is called the q-value.

q-values are the minimum pFDR that can occur when rejecting a statistics with a value t .

$$q = \inf_{\{\gamma; t \in \Gamma\}} pFDR(\gamma)$$

How do we do it in practice

- 1 For the 5C data of N encounter signals, $s_i(d)$, $i = 1 \dots N$.
- 2 Calculate the background (expected) signal, $\mu(d) = \frac{1}{N_d} \sum_{i_d=1}^{N_d} (s_{i_d}(d))$, with i_d the index of available observation in position d .
- 3 Calculate the background distribution $F_B(d)$.
- 4 For each d calculate the distribution, $F_d(z)$ of the z-score, $z_d(i_d) = \frac{s_{i_d}(d) - \mu(d)}{\sigma_d}$
- 5 Remark: if we are interested in the peaks, truncate the negative values of the z-scores.
- 6 For the rejection value γ of the null distribution, calculate $P_d = F_d(F_B^{-1}(\gamma))$
- 7 Calculate the pFDR and the associated q-values, and set a threshold α .
- 8 set the new threshold at $F_B^{-1}(\max\{P_d | q(P_d) < \alpha\})$

The 5C data

