Peak detection using positive False Discovery Rate Applications in chromosome capture data

Ofir Shukron

December 9, 2014

Motivation

1. Multiple hypotheses are being tested on large amount of experimental data.
2. Many time it is required to find outlying observations (peaks).
3. When finding many peaks, a criteria to control the error rate is needed.
4. One would like to reduce type I errors.
5. Restrictive traditional methods controlled the probability of at least one type I error.
6. New method of controlling the error called positive False Discovery Rate (pFDR) was developed[1].
7. We want to apply this method to find frequent specific looping events in the chromosomes using chromosome capture (CC) data.
8. Under the assumption of a polymer model, the peaks will be treated individually in the reconstruction of polymer structure from encounter data.

[1] Storey JD. A direct approach to false discovery rates. J. R. Statist. Soc. B (2002)64, Part 3, pp. 479498

A simple model

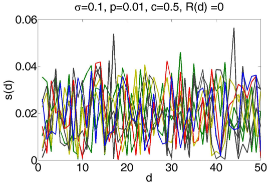
Assuming n realizations of a process $R(d) \in \mathcal{C}^0, \quad d \in \mathbb{R}$ with noise term $F(d) = \{f_1(d), f_2(d), ..., f_n(d)\}, \quad f_i \sim \mathcal{N}(0,1) \quad \forall i$, such that

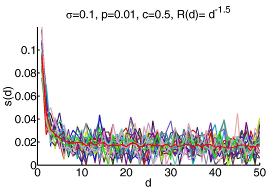
$$s_i(d) = R(d) + \sigma f_i(d), \qquad i = 1..n$$

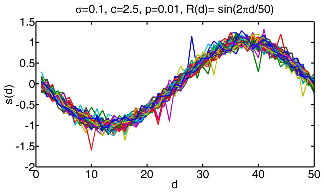
Assume $\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_n\}$ are n realizations of a random pulse process, e.g characterized by $\lambda_i(d) \sim Bin(1, p \ll 1)$ such that,

$$s_i(d) = R(d) + \sigma f_i(d)(1 + c\lambda_i(d))$$

with $c = const$

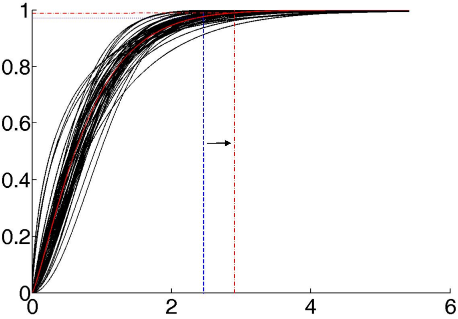






Approach

1. Estimate the expected signal and signal density (parametric or empirical) and set rejection region (value) Γ.
2. Calculate signals' densities for each d and p-values according to the rejection region.
3. Reminder: $p - value(t) = \min_{\{\Gamma; t \in \Gamma\}} \{Pr(T \in \Gamma | H = 0)\}$
4. Shrink the rejection region to reduce type I errors/balance Type II errors.



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.

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

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

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

The optimal λ is determined by minimizing the MSE of the bootsrap version of the pFDR[2].

[2] Storey JD. A direct approach to false discovery rates. J. R. Statist. Soc. B (2002)64, Part 3, pp. 479498

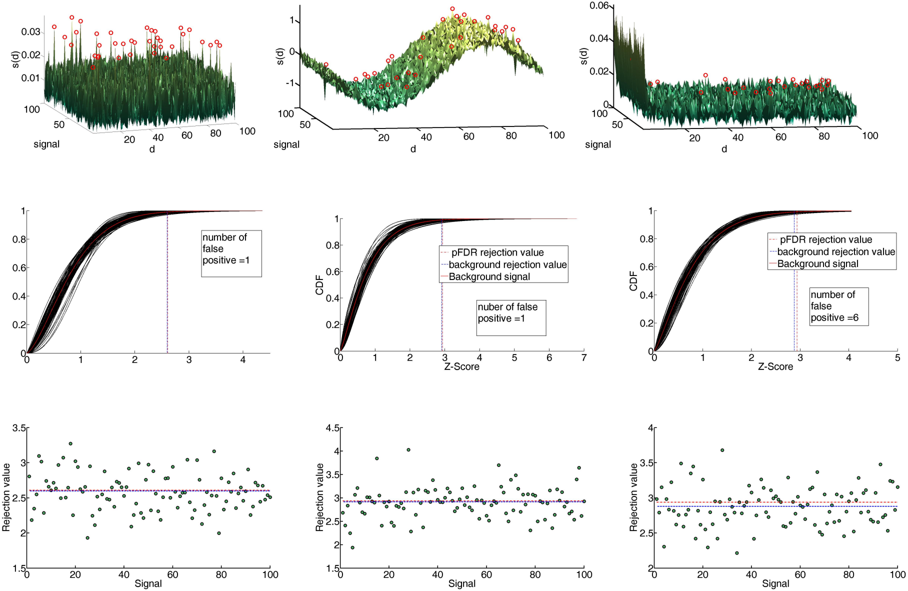
How do we do it in practice

1. For N signals, $s_i(d)$, $i = 1...N$.
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 $F_B^{-1}(\max\{P_d|q(P_d) < \alpha\})$

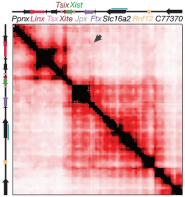
In the following examples we use $\alpha = 0.01$.

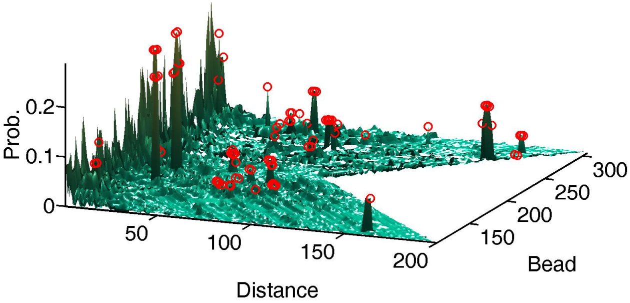
Synthetic examples

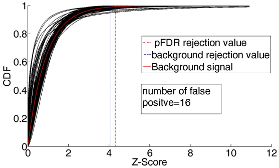
(1) $[R = 0, \sigma = c = 0.5, p = 0.01]$, (2) $[R = \sin(\frac{2\pi d}{100}), \sigma = 0.5, c = 2.5, p = 0.01]$ (3) $[R = d^{-1.5}, \sigma = c = 0.5, p = 0.01]$



Finding peaks of the 5C data

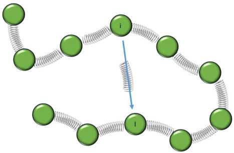


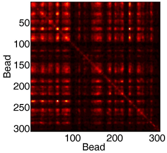




From encounter probability to chromosome structure

1. What do we do with the peaks after we've found them?
2. Assuming a Rouse model, one option is to connect with a spring any two beads corresponding to peaks.
3. If beads *i* and *l* correspond to a peak:

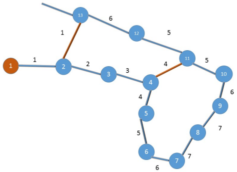




4. The encounter histogram on the right does not look like the experimental data.
5. The hight of the peak has to be taken into account.

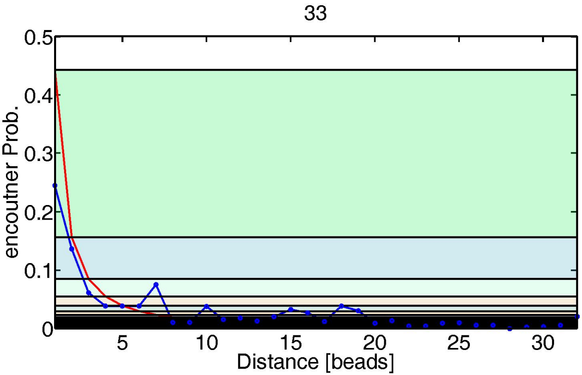
From encounter probability to chromosome structure

1. Trivially, connecting beads, the distance along the chain shortens.



2. In the figure, distance along the chain from bead 1. Added connections marked in orange
3. The encounter probability should carry information about the distances between beads.
4. Assuming a Rouse model, we know that $Pr(encounter(i,l)) \sim dist(i,l)^{-1.5}$ in 3D.

Projecting encounter Probabilities onto the encounter curve



we see that the encounter probability at distance 7 for bead 33 corresponds to distance 3 under the assumption of the Rouse model.
We have enough data to discover the distances between beads under the assumption of a polymer model (data not shown)

The spring constant corresponding to peaks

What shall we do if the encounter probability is higher than the expected probability of the nearest neighbor?

1. For a Rouse chain the spring constant is $k = \frac{3k_B T}{b^2}$
2. We need to distinguish nearest neighbors encounter probability from encounter probability stemming from different spring constants.
3. The bead distance probability in 3D is $P(r) = \left(\frac{3}{2\pi b^2}\right)^{1.5} \exp\left(-\frac{3r^2}{2b^2}\right)$
4. setting $r = b$ for nearest neighbor, we get in steady state $P(b) = \left(\frac{3}{2\pi e b^2}\right)^{1.5}$.
5. estimating nearest neighbor probability, $\hat{P}(b)$ from the data *without peaks*, and equating to $P(b)$, we get $b^2 = \left(\frac{3}{2\pi e}\right) \hat{P}(b)^{1.5}$
6. Using the relation for the spring constant $k = \frac{k_B T}{b^2}$, we get $k = \frac{2\pi e k_B T}{3\hat{P}(b)^{1.5}}$
7. since $D = \frac{k_B T}{\xi} = \frac{k_B T}{6\pi \eta_s a}$, we get $k = \frac{4\pi e D \eta_s a}{\hat{P}(b)^{1.5}}$, if we have access to these parameters, otherwise
8. assuming we observe $P_{il} > \hat{P}(b)$ in the encounter probability signal, then the peak correspond to nearest neighbor and the estimation for k is $\frac{2k_B T \pi e}{3P_{il}}$

Summary

1. I have presented the pFDR as means of controlling the error when searching for peaks in signals
2. The pFDR was applied on the CC data to eliminate false positive peaks.
3. location of the peaks will be used when identifying parameters of the chain (spring constant)
4. future work will include incorporation of different spring constant and simulations with heterogeneous polymer.