

Appendix 1

Threshold selection

The threshold window size k is chosen such that the expected number of significant tests for each window size k is equal and the significance level of the total test is α . Using extreme value theory Zhang deduces the following simple relationship between the significance level and the total intensity λ , $\alpha = 1 - e^{-\lambda}$. Hence $\lambda = -\log(1 - \alpha)$. The number of non-overlapping windows of size k ($\#window_k$) is equal to number of observation ($nobs$) in the CpG-island divided by window size k . Therefore the threshold t_k of window size k is chosen such that the intensity $\lambda_k = k \cdot \lambda_1$. The total intensity $\lambda = \sum_{k \in K} \lambda_k = \sum_{k \in K} k \cdot \lambda_1$, here K is the set of window sizes to be examined. The expected number of significant windows of size k is given by

$$\mathbb{E}[\text{Significant Windows}] = \lambda_k \cdot window_k = k \cdot \lambda_1 \frac{nobs}{k} = \frac{\lambda}{\sum_{k \in K} \lambda} \cdot nobs = \frac{-\log(1 - \alpha)}{\sum_{k \in K} \lambda_k} \cdot nobs$$

and hence independent of the window size k . Hence we can derive number of significant windows allowed keeping the significance level α . Using simulation we can find the optimal threshold such that

$$\mathbb{E}[\text{Significant Windows}] = \frac{-\log(1 - \alpha)}{\sum_{k \in K} \lambda_k} \cdot nobs$$

for the various window sizes.