Extending GREAT with Beta distribution p-values and spatial autocorrelation statistics

Charles Celerier (cceleri), Yifei Men (ymen),
Ahmed Bou-Rabee (bourabee), Andrew Stiles (aostiles),
Steven Lee (slee2010), and Nicholas Damien McGee (ndmcgee)

March 4, 2013

Abstract

This paper presents two extensions to GREAT. The first extension is caculating weights for arrows based on their proximity to the center of the given regulatory domains for each term. This weighting allows us to calculate a p-value from the regularized incomplete beta function $I_x(\alpha, \beta)$ where α = sum of the weights assigned to all arrows and β = (the maximum possible sum of weights) $- \alpha$. The second extension is adding spatial autocorrelation statistics. We include three of these statistics in our project: (1) the Getis-Ord General G global statistic, (2) Moran's I global statistic, and (3) Getis-Ord Gi* local statistic.

1 Weight function

Each term defines a set of transcription start sites on a genome (hg18 or mm9). Both of our extensions to GREAT require weights to be assigned to each arrow-TSS pair. Let's define the absolute distance function $d : \mathbb{N} \times \mathbb{N} \to \mathbb{Z}$ as:

$$d(A_p, TSS_p) = TSS_p - A_p.$$

We assign weights to each arrow-TSS pair based on a normal distribution with mean μ and standard deviation σ within 1Mb of the TSS and 0 otherwise. More formally, we can define a weight function $W : \mathbb{N} \times \mathbb{N} \to [0, 1]$ as:

$$W(A_p, TSS_p) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(d(A_p, TSS_p)-\mu)^2}{2\sigma^2}} & \text{if} \quad d(A_p, TSS_p) \leq 10^6 \\ 0 & \text{otherwise} \end{cases}$$

where

$$\mu = 0 \qquad \text{and} \qquad \sigma = \frac{10^6}{3}.$$

These weights are assigned by using the python script `findDist.py`.

2 Beta distribution

Given an assignment of weights for each arrow-TSS pair, we can calculate a p-value from the regularized incomplete beta function $I_x(\alpha, \beta)$. We decided to define α and β as follows:

$$\alpha = \sum_{A_p \in \text{arrows}} \sum_{TSS_p \in \text{TSSs}} W(A_p, TSS_p)$$

$$\beta = \#\text{arrows} \cdot \max\{W(A_p, TSS_p) : A_p \in \text{arrows and } TSS_p \in \text{TSSs}\} - \alpha$$

The value for x is less clear. We will try a few ideas:

$$x = \frac{\text{sum of all weights for the queried term}}{\text{sum of all weights for all terms}}$$

$$x = \frac{\text{size of the regulatory domain}}{\text{size of the genome}}$$

$$x = \frac{\text{size of regulatory domain with weight at least the mean weight}}{\text{size of the genome}}$$

The regularized incomplete beta function was already implemented in the GREAT source code. We will use that code for this statistic.

3 Spatial autocorrelation statistics

We would like to implement a type of hot spot analysis for which regulatory domains had the largest impact on the p-value (binomial or beta) for the markers on each term. The three spatial autocorrelation statistis we found can be found here.

Here are the null hypotheses for the global statistics:

1. Getis-Ord General G: "there is not spatial clustering of the arrow"
2. Moran's I: "there is not spatial clustering of the arrow associated with the regulatory domains"

Getis-Ord Gi* is a local statistic that will indicate clustering of arrows with low and high weights. This could be a useful mechanism for spotting patterns in the arrows submitted to GREAT.