# Extending GREAT with probabilistic assignment of regions to genes

The recently published Genomic Regions Enrichment of Annotations Tool (McLean 2010, "GREAT improves functional interpretation of cis-regulatory regions") performs functional analysis of genomic regions by associating the regions with nearby genes, applying the functional annotations of the genes to the regions, and then performing statistical enrichment tests. GREAT assigns each gene a regulatory domain, and for a given gene, a base pair in the genome either falls inside the gene's regulatory domain or does not. In other words, the assignment of genomic locations to a gene is binary (either assigned or not assigned). While this simple model works well in practice, it does not account for uncertainty in assigning regions to genes. Consider the following locus:


Currently region A is associated with the green gene, and region B is associated with the red and blue genes. This has two drawbacks: 1) region A is not too far from the red gene and may regulate that gene but the presence of the green gene "blocks" the extension of the red regulatory domain; 2) region B is assigned equally to the red and blue genes, but it if more likely to regulate the blue gene based on proximity. To address both challenges, you will extend GREAT to use "soft" probabilistic assignment of regions to genes. For example, region B might be 90% assigned to the blue gene and 10% assigned to the red gene. Additionally, the regulatory domain of the red gene would extend to include region A rather than abruptly ending at the green gene. To accomplish this:

- Define a "gravity" function that specifies how strongly a gene "pulls" at a genomic region at a given distance from its transcription start site. Have the gravity function drop to 0 at a reasonable distance (perhaps 1Mb) to simplify calculation.
- For a given basepair in the genome, assign it to all genes that "pull" at it, weighted by how strongly each gene pulls at the location.
- Define an extension to the GREAT region-based binomial test appropriate to the probabilistic assignment of regions to genes.
- Implement your "probabilistic" GREAT.
    - We will provide you with a list of all human genes with their transcription start sites.
    - We will provide you with mappings of ontology terms (functions) to human genes.
- Apply your "probabilistic" GREAT to the SRF ChIP-seq example set and the limb p300 example set from the GREAT paper.

- Compare the results of your GREAT to the published tool:
    - Do you highlight any additional relevant terms?
    - Do the p-values of enriched terms become stronger or weaker?
    - Which regions do you associate with relevant terms that were not associated by the published tool? Looking at these regions in the genome browser, does the assignment of region to gene seem appropriate?

## Suggested Checkpoint Goals

- Implement functionality that determines what fraction of each basepair is assigned to each gene (independent of ontology terms at first) for a generic gravity function.
- Choose a parameterized gravity function (perhaps choose a few).
- Define an extension to the region-based binomial test.

## Implementation Notes

The directory `/afs/ir/class/cs173/finalProjects/GREATRegDoms/` contains the following useful data.

- A file of transcription start sites for all genes in hg18 and mm9
- A command-line version of the GREAT executable and its ontology data for hg18 and mm9 (to compare to the published tool)
- The SRF (hg18) and p300 limb (mm9) data sets
- A README.txt file describing the various file formats and how to run the command line GREAT