



1. DATOS BÁSICOS DEL TFG:

Título: Métodos ponderados para el análisis de sobrerepresentación de anotaciones en análisis bioinformáticos

Descripción general (resumen y metodología):

El análisis de sobrerepresentación de anotaciones funcionales, también llamado análisis de enriquecimiento en bioinformática, es una técnica ampliamente utilizada en este campo para interpretar la información asociada a grandes listas de variables (genes). Los métodos clásicos aplican el test exacto de Fisher, basado en la distribución hipergeométrica donde se asume que todos los genes tienen la misma probabilidad de selección. Sin embargo, en datos reales existen sesgos asociados a diferentes factores que introducen heterogeneidad en las probabilidades de selección, como es la longitud génica o al número de anotaciones asociadas a cada gen.

Este TFG propone estudiar métodos estadísticos que corrigen dichos sesgos mediante distribuciones hipergeométricas no centrales, especialmente la distribución de Wallenius. Se compararán métodos clásicos y ponderados mediante simulaciones y datos reales de bioinformática usando R.

Tipología: Estudio de casos, teóricos o prácticos, relacionados con la temática del Grado.

Objetivos planteados:

El objetivo general de este trabajo es evaluar el impacto de los sesgos de selección y anotación en los análisis de enriquecimiento funcional y estudiar métodos de inferencia basados en distribuciones hipergeométricas no centrales para corregir dichos sesgos.

Se plantean los siguientes objetivos específicos:

- Formalizar el problema de sesgo en análisis de sobrerepresentación en bioinformática-
- Estudiar la distribución hipergeométrica no central de Wallenius como modelo probabilístico para muestreo ponderado
- Comparar, mediante simulación y/o datos reales, el comportamiento de métodos clásicos y ponderados en términos de control del error tipo I y robustez frente al sesgo.

Bibliografía básica:

- Mark D. Young, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*, 2010, 11(2):R14. DOI: 10.1186/gb-2010-11-2-r14
- Jovana Maksimovic, Oshlack A, Phipson B. Gene set enrichment analysis for genome-wide DNA methylation data. *Genome Biology*, 2021, 22:173. DOI: 10.1186/s13059-021-02388-x
- Kristen Glass, Michelle Girvan. Annotation Enrichment Analysis: An Alternative Method for Evaluating the Functional Properties of Gene Sets. *Scientific Reports*, 2014, 4:4191. DOI: 10.1038/srep04191
- Agner Fog. Calculation Methods for Wallenius' Noncentral Hypergeometric Distribution. *Communications in Statistics - Simulation and Computation*, 2008, 37(2):258-273. DOI: 10.1080/03610910701790269

Recomendaciones y orientaciones para el estudiante:

Es recomendable conocimientos de programación en R avanzados

Plazas: 2

2. DATOS DEL TUTOR/A:

Nombre y apellidos: PEDRO MARÍA CARMONA SÁEZ

Ámbito de conocimiento/Departamento: ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

Correo electrónico: pcarmona@ugr.es

3. COTUTOR/A DE LA UGR (en su caso):

Nombre y apellidos: PEDRO JESÚS FEMIA MARZO

Ámbito de conocimiento/Departamento: ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

Correo electrónico: pfemia@ugr.es

4. COTUTOR/A EXTERNO/A (en su caso):

Nombre y apellidos:

Correo electrónico:

Nombre de la empresa o institución:

Dirección postal:

Puesto del tutor en la empresa o institución:

Centro de convenio Externo:

5. DATOS DEL ESTUDIANTE:

Nombre y apellidos:

Correo electrónico: