# An equivalence test to detect functional similarity between feature lists based on the joint enrichment of gene ontology terms.

Pablo Flores Muñoz[1,2]

[1]Universitat Politècnica de Catalunya/Faculty of Mathematics and Statistics/Department of Statistics and Operational Research
[2]Escuela Superior Politécnica de Chimborazo (ESPOCH)/Faculty of Sciences

In the current era, omics technologies such as high-throughput experiments have significantly transformed the fields of biology and medicine. These advances enable the generation of large volumes of biological data, such as gene lists, proteins, and other biological features, under different experimental conditions. Although getting this large amount of information represents a breakthrough, it is crucial to develop appropriate statistical methods to analyze and extract knowledge from these data.

In this context, the present study proposes a statistical method based on an equivalence hypothesis test to evaluate biological similarity between feature lists. The central idea is that two or more feature lists can be considered biologically similar if they share a significant proportion of enriched GO terms.

First, the choice of the Sorensen index is justified as an appropriate metric for assessing the dissimilarity of joint enrichment between the lists under comparison. Next, the sampling distribution of this measure is studied both theoretically and through approximation using the Bootstrap method, which proves to be particularly effective when the enrichment level is low. Based on these distributions, an equivalence hypothesis test is developed, along with its corresponding irrelevance threshold, which is less arbitrary than the thresholds commonly used in equivalence approaches.

Furthermore, the R package goSorensen has been developed, published, and is available on the Bioconductor platform. This informatics tool allows for the efficient application of the proposed methodology.

Additionally, a dissimilarity matrix is constructed based on the irrelevance threshold, which defines when two lists are significantly equivalent. This matrix provides an inferential measure of how close or distant the compared lists are from each other. The graphical representation and interpretation of this matrix, such as in an MDS-Biplot, is useful for identifying the GO terms associated with the formation of equivalence between lists.

Finally, it is important to note that the proposed methodology has been rigorously evaluated and applied to real gene lists, with an exhaustive comparison of the results obtained against other similar comparison methods.

**Keywords**: Equivalence test, feature lists, functional similarity, Gene Ontology, High-throughput experiments.