www.elsevier.com/locate/cbm

Improving Information Retrieval in Functional Analysis

Juan C Rodriguez, Germán A González, Cristóbal Fresno, Andrea S Llera, Elmer A Fernández

PII: S0010-4825(16)30244-X
DOI: http://dx.doi.org/10.1016/j.compbiomed.2016.09.017
Reference: CBM2504

To appear in: *Computers in Biology and Medicine*

Received date: 20 April 2016
Revised date: 21 September 2016
Accepted date: 22 September 2016

Cite this article as: Juan C Rodriguez, Germán A González, Cristóbal Fresno Andrea S Llera and Elmer A Fernández, Improving Information Retrieval i Functional Analysis, *Computers in Biology and Medicine* http://dx.doi.org/10.1016/j.compbiomed.2016.09.017

Improving Information Retrieval in Functional Analysis

Juan C Rodriguez[a,b], Germán A González[a,c], Cristóbal Fresno[a], Andrea S Llera[d], Elmer A Fernández[a,e,*]

[a] UA AREA CS. AGR. ING. BIO. Y S, Universidad Católica de Córdoba, CONICET, Córdoba, Argentina
[b] Facultad de Matemática, Astronomía y Física, Universidad Nacional de Córdoba, Córdoba, Argentina
[c] Instituto Nacional de Cáncer, MinSal, Córdoba, Agentina
[d] IIBBA, Fund. Instituto Leloir, CONICET, Buenos Aires, Argentina
[e] Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de Córdoba, Córdoba, Argentina

Abstract

Transcriptome analysis is essential to understand the mechanisms regulating key biological processes and functions. The first step usually consists of identifying candidate genes; to find out which pathways are affected by those genes, however, functional analysis (FA) is mandatory. The most frequently used strategies for this purpose are Gene Set and Singular Enrichment Analysis (GSEA and SEA) over Gene Ontology. Several statistical methods have been developed and compared in terms of computational efficiency and/or statistical appropriateness. However, whether their results are similar or complementary, the sensitivity to parameter settings, or possible bias in the analyzed terms has not been addressed so far. Here, two GSEA and four SEA methods and their parameter combinations were evaluated in six datasets by comparing two breast cancer subtypes with well-known differences in genetic background and patient outcomes. We show that GSEA and SEA lead to different results depending on the chosen statistic, model and/or parameters. Both approaches provide complementary results from a biological perspective. Hence, an Integrative Functional Analysis (IFA) tool is proposed to improve information retrieval in FA. It provides a common gene expression analytic framework that grants a comprehensive and coherent analysis. Only a minimal user parameter setting is required, since the best SEA/GSEA alternatives are integrated. IFA utility was demonstrated by evaluating four prostate cancer and the TCGA breast cancer microarray datasets, which showed its biological generalization capabilities.

Keywords: Big Omics Data, Gene Set Enrichment Analysis, Functional Class Scoring, Over Representation Analysis, Singular Enrichment Analysis, Biological Insight, Knowledge Discovery, Breast Cancer, R Framework.

1. Introduction

Cancer is so heterogeneous that single the analysis of differentially expressed (DE) genes is not enough to gain biological insight of this complex disease [1]. On the contrary, it is the starting point for an interpretation process in which biologists search for patterns using different information sources [2]. The process to uncover those functionalities is known as Functional Analysis (FA), which is based on the assessment not of individual genes but of genes grouped due to their association with a biological mechanism (gene sets), under the assumption that their coordinated action impacts the same biological process [2, 3]. There are two main approaches to perform this task: Over Representation Analysis and Functional Class Scoring [4, 5]. According to Huang et al., the most commonly used methods in those categories are Singular and Gene Set Enrichment Analysis (SEA and GSEA), respectively [6]. The former uses an interest gene list as input, which is usually the DE gene list. Then, given a statistical test based on a contingency table, each term is evaluated and considered enriched if the observed proportion of DE genes in the term differs from the expected distribution when compared against a background reference (BR). One of the main criticisms towards SEA is that it requires a user-defined DE gene list (usually by setting a threshold) [2, 4, 5, 7, 8, 9]. GSEA methods have overcome this limitation by using all gene expression levels available in the experiment. These genes are sorted according to some metric related to the analyzed phenotype.

Several SEA and GSEA algorithms have been proposed [9] with their own assumptions and input parameters, which could potentially lead to different results. Indeed, some gene sets such as the ones provided by the Gene Ontology (GO) Consortium [10] are organized in some particular structure that yields additional penalization strategies to consider. Therefore, selecting the appropriate algorithm and its parameter settings is not trivial decision to make for researchers that face a biological problem and has not been comprehensively addressed. In addition, what each method returns from an information retrieval point of view is not clear; moreover, whether these results are independent of the method and parameters, complementary or are equally useful is also unclear. Manoli et al. [4] compared

*Corresponding author at: UA AREA CS. AGR. ING. BIO. Y S, Universidad Católica de Córdoba, CONICET, Córdoba, Argentina
E-mail address: efernandez@bdmg.com.ar