RESEARCHER'S DIGEST: USING TEXT-MININIG TECHNIQUES FOR CLASSIFYING LISTS OF GENES

Juan Carlos Triviño, Juan Carlos Sánchez-Ferrero and Juan Carlos Oliveros

(jctrivino@cnb.csic.es, jcsanchez@cnb.csic.es, oliveros@cnb.csic.es)

Service of Bioinformatics for Genomics and Proteomics (BioinfoGP). Centro Nacional de Biotecnología (CNB-CSIC). Darwin 3, Campus de Cantoblanco, 28049, Madrid, Spain

Introduction

Researcher's Digest is an on-line tool to classify gene lists obtained from microarray experiments into functional groups, by using text-mining techniques applied to their annotations (free-text sentences). The results are presented in a web page where every functional group is shown together with the expression patterns (ratios) of the genes.

The method

The system first identify the relevant terms present in each sentence: words with biological prefixes or suffixes (a); gene families (b); database accession codes (c); and words with capital characters (d). The remaining words are classified as generic "monograms" (e). Additionally, the system extracts from the sentences all pairs of words with no punctuation signs in-between ("bigrams") (f).

Two genes are clustered in the same functional group if they share several relevant terms in their annotations (see figure 1).

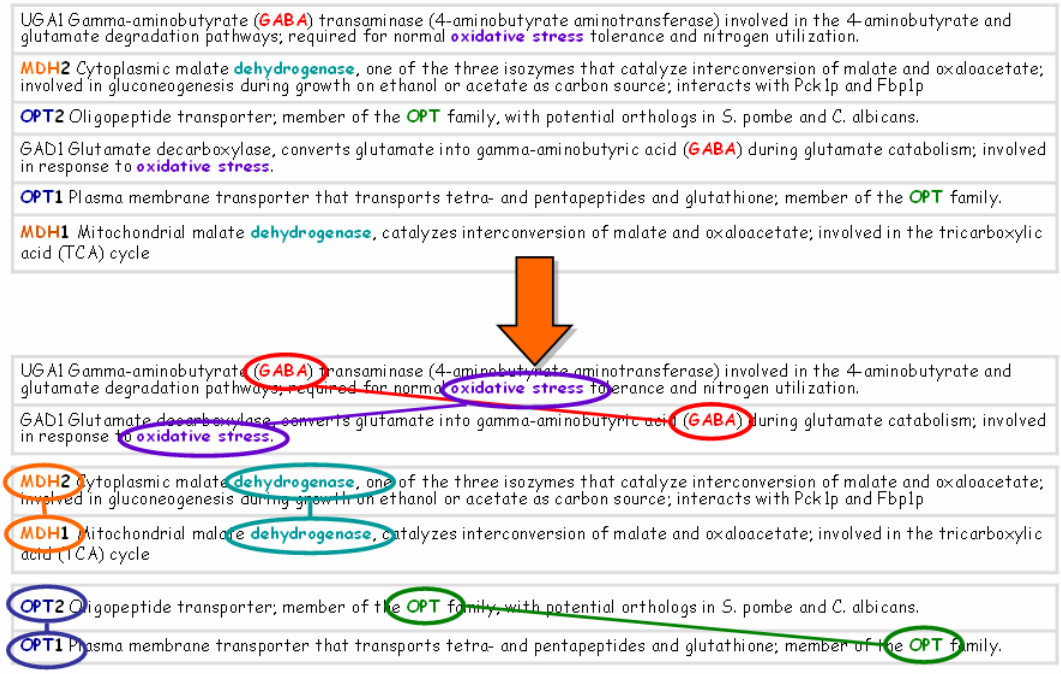


Figure 1. Upper: Relevant terms are highlighted in six annotations. Lower: Sentences sharing several highlighted terms are grouped.

The similarity (S) is calculated by the formula:

$$S = \frac{Na \times Wa + Nb \times Wb + Nc \times Wc + Nd \times Wd + Ne \times We + Nf \times Wf}{Ca \times Wa + Cb \times Wb + Cc \times Wc + Cd \times Wd + Ce \times We + Cf \times Wf}$$

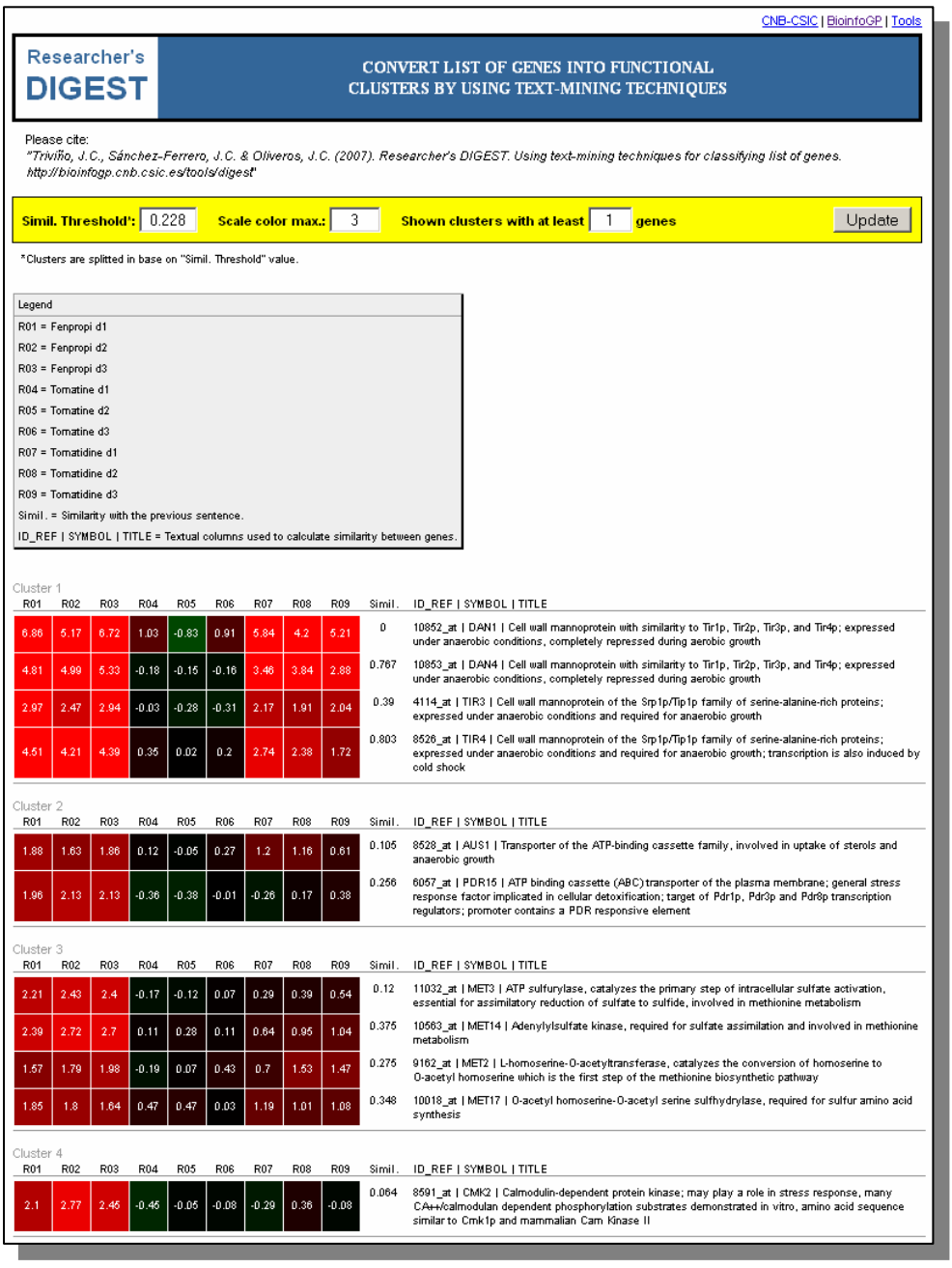
Where Nx is the number of common relevant terms of each category (a,b,c,d,e,f); Wx is a weight that determines the importance of every type of term (defined by the user) and Cx is the maximum possible value of Nx for the two sentences.

Example of results (as provided by the software):

This example is available at Researcher's Digest user form. It demonstrates the capabilities of the software by classifying a gene list obtained from a real microarray experiment (1).

When presenting the results, a similarity value (threshold) is proposed by the system to separate the genes into different groups. This value can be changed by the user to increase / decrease the final number of clusters.

Figure 2. Part of the results of classfying the 100 most variable genes from GEO dataset GDS2196 (1)



(1) Simons V, Morrissey JP, Latijnhouwers M, Csukai M et al. Dual effects of plant steroidal alkaloids on Saccharomyces cerevisiae. Antimicrob Agents Chemother 2006 Aug;50(8):2732-40.

Researcher's Digest is accessible at: http://bioinfogp.cnb.csic.es/tools/digest