Improving Gene Functional Analysis in Ethylene-induced Leaf Abscission using GO and ProteInOn

Sara Domingos[1], Cátia Pesquita[2], Francisco M. Couto[2], Luis F. Goulao[3], Cristina Oliveira[1]

[1] Instituto Superior de Agronomia, Universidade Técnica de Lisboa, Tapada da Ajuda, 1349-017 Lisboa, Portugal
saradomingos@gmail.com, crismoniz@isa.utl.pt
[2] Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749–016 Lisboa, Portugal
cpesquita@xldb.di.fc.ul.pt, fcouto@di.fc.ul.pt
[3] Eco-Bio, Instituto de Investigação Científica Tropical IP, Av. da República, Quinta do Marquês,2784-505 Oeiras, Portugal
goulao@iict.pt

Abstract. In this study we apply a new strategy to select candidate genes involved in the abscission of citrus leaves based on the identification of the most meaningful functional aspects of differentially expressed genes using ProteInOn, a Gene Ontology (GO) based web tool. We submitted a previously analyzed dataset obtained from a microarray experiment to the tool, to obtain the most meaningful GO terms annotated to the given gene products, highlighting the relevant functional processes involved with a greater detail than the original analysis and supporting identification of the genes more likely to be involved in the studied phenomenon. This strategy proved to be effective by complementing the previous analysis with a wider knowledge about the biological processes and gene products more relevant to ethylene-induced leaf abscission.

Keywords: ethylene-induced leaf abscission, *Citrus clementina*, functional analysis, meaningful terms, semantic similarity.

1 Introduction

Abscission is a cell separation process by which plant organs are detached, by reduction of cell adhesion and cell wall disassembly in specialized cells, named the abscission zone, and represents an important agricultural phenomenon which affects final yield [1]. The accumulation of data generated by high-throughput techniques of gene expression analysis requires bioinformatics tools and defined vocabularies to describe genes functions that allow its integration and computation.

This study aims to improve functional analysis of a list of differentially expressed genes, using the Gene Ontology (GO) and the ProteInOn web tool. To perform it, we

used microarray experiment results obtained in [1] where a functional categorization from MIPS (Munich Information Center for Protein Sequences) was applied. MIPS functional catalogue database (FunCat) provides hierarchical information derived from a few species [2] assigning a *p*-value for each category. This requires translation into *Arabidopsis thaliana* orthologues. Nowadays, functional categorization of transcripts is mostly based on GO which provides a controlled vocabulary to describe protein function for all eukaryotes, based on 3 orthogonal ontologies: biological processes, cellular components and molecular functions [3]. To analyze the gene set, we chose ProteInOn, a GO-based web tool focused on identifying the most meaningful functional aspects of related gene products [4]. It uses the hierarchical structure of GO to find representative terms, while typical enrichment strategies use direct annotations and usually fail to find general but meaningful annotations. ProteInOn also measures the specificity of each GO term based on information content (IC) and calculates semantic similarity (SSM) which is a function that, given two sets of terms annotating two entities, returns a value reflecting the closeness in meaning between them. There are more than other 68 enrichment analysis tools available *(e.g.* Onto-Express, FatiGO), based on the premise that co-functioning genes should have a higher potential to be selected, if a biological process is irregular in a study [6]. Our main goal was to apply the functional analysis strategy detailed in [7] and to compare these results with the original ones, while comparing different gene products functional classifications, GO and FunCat. By using updated databases and bioinformatics tools to analyze previously results, we aim to provide insight into the transcriptome abscission-related to support further experiments.

2 Methods

The data set used for functional analysis was obtained from [1]. The authors aimed to identify and classify genes involved in the ethylene-induced leaf abscission process in *Citrus clementina*, and used a cDNA microarray to analyze transcriptome of laminar abscission zone (LAZ) and petiole of leaf explants (Pet).

The functional analysis strategy applied here was as follows:

i) A list of genes was organized according to moment and organ where genes were over-represented [1] and translated into UniProtKB accession numbers.

ii) GO annotations were analyzed through ProteInOn, using UniProtKB accession numbers as input. For genes absent in UniProtKB database, the code of a putative *A.th* orthologue was used.

iii) The ten most meaningful GO terms were selected. Meaningfulness was given by the term representativity measure (e-value) and calculated using the global

frequency of each GO term annotation in the database as an estimator of its probability of occurrence. For each GO term, a protein taken randomly from the dataset is a random event with two outcomes: success if annotated to t, or failure. The probability of observing at least k successes in a random set of n gene products is given by the probability of success P(t) (1). IC values were calculated using (2), where f(t) is the frequency of annotation of the term t [7].

$$P\,(x1 \geq k) = \sum_{i=k}^{n} \binom{n}{i} P(t)^{I} (1-p(t))^{n-I} \quad . \tag{1}$$

$$IC(t) = -\log_2 f(t) \quad . \tag{2}$$

iv. The SSM between selected gene products was calculated with *simGIC* measure [5], which has been shown to outperform other measures. SimGIC is given by the sum of the IC of each shared term by two gene products, A and B, divided by the sum of the IC of all their terms (3):

$$simGIC\,(A,B) = \frac{\sum_{te\{GO(A) \cap GO(B)\}} IC(t)}{\sum_{te\{GO(A) \cup GO(B)\}} IC(t)} \quad . \tag{3}$$

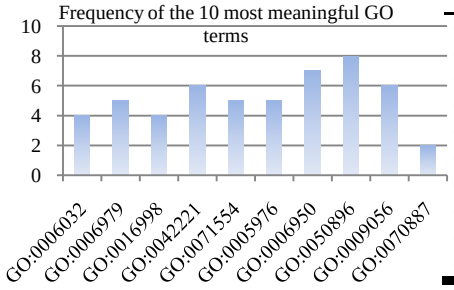
v. Gene products with SSM > 80% were clustered and the remaining ones were analyzed.
vi. GO annotations for both the clusters and relevant individual gene products were analyzed and compared with their previous MIPS classification.

3 Results

The list of differentially expressed genes in LAZ and Pet was divided by the two main moments after ethylene treatment (Supplementary Data 1). In the first moment (6h and 12h after treatment) ten differently expressed genes were found in the LAZ. In the second moment (24h and 36 h after treatment) 23 over-represented genes were found in the LAZ and 31 in the Pet, three of them in common with the first phase. The initial set was composed by 61 gene products of which 56 had UniProtKB accession number, 52 had annotations on GO, of which 84% were inferred from electronic annotations. The authors of the experimental assay associated each over-represented transcript to the matching gene, first identified in a single species, which we used as input in UniProtKB, instead of finding the *A.th* orthologue to use MIPS [1].

From the term representativity analysis we obtained a list of relevant GO terms ranked by e-value. Figure 1 and Supplementary Data 2 show the ten most meaningful *biological process* terms of each set, since it is the most interesting GO type for the identification of the relevant genes in a biological phenomenon. For the LAZ, in the

first moment, the most meaningful GO terms had a low IC, but in the second moment we found annotations with a higher specificity, such as lipid transport and cell wall organization (Supplementary Data 2). Figure 1 shows the ten most meaningful terms of over-represented gene products in the Pet, which have a high IC and relevant biologic meaning for the abscission process, as chitin and other cell wall molecular catabolic processes, cell wall biosynthesis, polysaccharide metabolism, response to stress and oxidative stress and response to stimulus and chemical stimulus. We calculated the SSM for the subset of gene products that are annotated with these meaningful GO terms and built clusters of gene products with SSM > 80% (Table 1). Relevant GO terms associated to remaining gene products were also selected (Supplementary Data 3). Then the annotations of the selected gene products were analyzed (Table 2). In the LAZ, in the first phase, we found gene products involved in photosynthetic processes, response to biotic and abiotic stimulus, protein metabolism, lipid metabolism and floral organ abscission, while in the second phase gene products were related to carbohydrate and protein metabolism, molecular transport and cell wall organization (Tables 1, 2 and Supplementary Data 3). In the PET, there was more processes involved (Supplementary Data 4): DNA transcription, cell wall organization and biosynthesis, response to abiotic stress, oxidation-reduction, hormonal regulation, protein metabolism, molecular catabolism and transport.



Term Id	Term Name	IC
GO:0006032	chitin catabolic process	0.45
GO:0006979	response to oxidative stress	0.37
GO:0016998	cell wall macromolecule catabolic p.	0.38
GO:0042221	response to chemical stimulus	0.25
GO:0071554	cell wall organization or biogenesis	0.27
GO:0005976	polysaccharide metabolic process	0.24
GO:0006950	response to stress	0.19
GO:0050896	response to stimulus	0.16
GO:0009056	catabolic process	0.20
GO:0070887	cellular response to chemical stimulus	0.39

Fig. 1. Ten most meaningful GO terms and corresponding IC, of *biological process* type, annotated to gene products over-represented in Pet, in the second moment (Source: ProteInOn).

Table 1. Clusters of gene products over-represented in LAZ e Pet.

	Clusters of gene products (Uniprot ac. number)	Term name
LAZ/ second moment	Q6EV47, Q8L5S8	lipid transport
	Q81746, Q94C86	cellular cell wall organization
	A0MKC8, Q8L4F2, Q9SF40	translation
Pet/ second moment	O82547, Q43752, Q8H985, Q8H986	chitin catabolic process cell wall macromolecule catabolic proc.
	A9PHA0, A9UFX7, Q0ZA67	response to oxidative stress oxidation-reduction process

Table 2. Functional characterization of isolated gene products over-represented in LAZ, in the first moment after ethylene treatment, of *biological process* GO terms type.

	Term Name	O80397	Q8L6Y9	Q9ARO7
LAZ/ first moment	regulation of stomatal complex patterning	√		
	regulation of stomatal complex development	√		
	floral organ abscission	√		
	jasmonic acid metabolic process			√
	plant-type hypersensitive response	√		
	defense response, incompatible interaction	√		
	oxylipin biosynthetic process			√
	protein phosphorylation	√	√	
	response to wounding			√
	response to stress	√		√

4 Discussion and Conclusion

Using the current versions of UniProtKB and GO annotations, we obtained an improvement on the coverage of the initial gene list with only four terms without GO annotations associated comparing with MIPS categorization [1]. Over-represented gene products with manual annotations were 15.38% of the transcripts, which probably means that the studied phenomenon has been gaining interest.

By selecting the clusters of gene products with high SSM and the relevant isolated gene products, we found gene products with an interesting set of annotations for ethylene-induced leaf abscission. Analyzing this restricted list, for example in the Pet, in the second moment after ethylene treatment (Supplementary Data 4) we found that they are associated with other important processes such as hormonal regulation, molecular transport, DNA transcription, protein metabolism, aromatic compounds metabolism and organ formation. In the last phase of abscission, a protective layer is generated in the organ that remains attached to the plant [1], which can be related with the gene product annotated with *organ formation* term. MIPS categorization indicated that transcriptional activation during abscission was involved in defense, transport mechanisms, signal transduction and protein biosynthesis, comparing it with our strategy we obtained a more comprehensive characterization, including cell wall biosynthesis, hormonal regulation and biosynthesis, protein metabolism, catabolic process, molecular transport and response to biotic/abiotic stresses, and other detailed information. The adopted strategy using ProteInOn, with the empirical choice of the ten most meaningful terms and clustering based on SSM > 80%, proved to be effective in providing a wider understanding of the process under study. A gene product that seems to have a key role is ATMKK4 (O80397) which is annotated with very relevant terms: *floral organ abscission*, *stomatal complex patterning and*

development regulation, *plant-type hypersensitive response*, *defense*, *response to stress and protein phosphorylation.* For this specific gene product, and to account for the bias in basing our analysis in a GO 2011 version and comparing it to results obtained with a MIPS 2008 version, we investigated its functional description in MIPS 2011 and GO 2008. In MIPS 2011 it is related to defense, response to biotic stimulus, protein phosporylation and protein kinase function, while in GO 2008 it is annotated with sensitive response, protein phosphorylation and defense. Hence, GO did gain relevant information about this gene since 2008, while MIPS did not.

By reanalyzing previously obtained results with more up to date bioinformatics tools and databases, we were able to improve the functional analysis of differentially expressed genes and the identification of gene products relevant to the studied phenomenon. This strategy can be an important tool to guide future experiments, saving time and materials by avoiding experiment replication.

Supplementary Data. Supplementary data are available at http://xldb.fc.ul.pt/xldb/images/e/e2/Supplementary_Data.pdf .

Acknowldegements. This work was supported by the Portuguese Fundação para a Ciência e Tecnologia through the Multiannual Funding Programme for LaSIGE, and the PhD grants SFRH/BD/42481/2007 and SFRH/BD/69076/2010.

5 References

1. Agusti, J., Merelo, P., Cercós, M., Tadeo, F.R., Talón, M.: Ethylene-induced Differential Gene Expression during Abscission of Citrus Leaves. J. Exp. Bot. 59, 2717-33 (2008)
2. Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Guldener, U., Mannhaupt, G., Munsterkotter, M., Mewes, H.W.: The FunCat, a Functional Annotation Scheme for Systematic Classification of Proteins from Whole Genomes. Nucleic Acids Res. 32, 5539-45 (2004)
3. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nat. Genet. 25(1):25-29 (2000)
4. Faria, D., Pesquita, C., Couto, F., Falcao, A.,: ProteInOn: A Web Tool for Protein Semantic Similarity. DI/FCUL TR 07-6, Department of Informatics, University of Lisbon (2007)
5. Pesquita C, Faria D, Bastos H, Falcão AO, Couto F.: Evaluating GO-based Semantic Similarity Measures. ISMB/ECCB 2007 SIG Meeting Program Materials, International Society for Computational Biology (2007)
6. Huang, daW., Sherman, B.T., Lempicki, R.A.: Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37,1-13 (2009)
7. Bastos, H., Tavares, B., Pesquita, C., Faria, D., Couto F.: Application of Gene Ontology to Gene Identification. In: Yu, B., Hinchcliffe, M., (eds) In Silico Tools for Gene Discovery. Springer Science (in press)