Deliverable 3.1b:
Peer review scientific publication on cross-context knowledge discovery in plant immune signaling

Jožef Stefan Institute and National Institute of Biology Version 1.0 FINAL

Abstract: This deliverable consists of two journal papers related to the discovery of new knowledge related to the plant immune signalling.
The first one holds a title “Discovering dependencies between domains of redox potential and plant defence through triplet extraction and copulas”. It presents a new approach to discovering dependencies between different biological domains based on copula analysis of literature mining results. More specifically, we have explored dependencies between literature from the domains of plant defence response and redox potential. Copula analysis of triplets, which are extracted by Bio3graph tool, shows that dependencies exist between these two domains indicating a potential for cross-domain literature exploration.
The second journal paper is named “Discovery of Relevant Response in Infected Potato Plants from Time Series of Gene Expression Data”. This paper presents a methodology for analyzing time series of gene expression data collected from the leaves of potato virus Y (PVY) infected and non-infected potato plants, with the aim to identify significant differences between the two sets of potato plants’ characteristic for various time points. We aim at identifying differentially-expressed genes whose expression values are statistically significantly different in the set of PVY infected potato plants compared to non-infected plants, and which demonstrate also statistically significant changes of expression values of genes of PVY infected potato plants in time. The novelty of the approach includes stratified data randomization used in estimating the statistical properties of gene expression of the samples in the control set of non-infected potato plants. A novel estimate that computes the relative minimal distance between the samples has been defined that enables reliable identification of the differences between the target and control datasets when these sets are small. The relevance of the outcomes is demonstrated by visualizing the relative minimal distance of gene expression changes in time for three different types of potato leaves for the genes that have been identified as relevant by the proposed methodology.

Document administrative information	
Project acronym:	HinLife
Project number:	J7-7303
Deliverable number:	D3.1b
Deliverable full title:	Peer review scientific publication on cross-context knowledge discovery in plant immune signaling
Document identifier:	HinLife -del-D3.1b –Cross-context knowledge discovery in plant immune signaling
Lead partner short name:	JSI and NIB
Report version:	final
Report preparation date:	31/12/2018
Lead author:	Dragana Miljkovic and Dragan Gamberger
Co-authors:	Tjaša Stare, Kristina Gruden, Nada Lavrač, Biljana Mileva-Boshkoska, Marko Bohanec
Status:	Final

Discovering dependencies between domains of redox potential and plant defence through triplet extraction and copulas

Dragana Miljkovic

Department of Knowledge Technologies,
Jožef Stefan Institute,
Ljubljana, 1000, Slovenia
Email: dragana.miljkovic@ijs.si

Nada Lavrač

Department of Knowledge Technologies,
Jožef Stefan Institute,
Ljubljana, 1000, Slovenia
and
University of Nova Gorica,
Nova Gorica, 5000, Slovenia
Email: nada.lavrac@ijs.si

Marko Bohanec

Department of Knowledge Technologies,
Jožef Stefan Institute,
Ljubljana, 1000, Slovenia
and
University of Nova Gorica,
Nova Gorica, 5000, Slovenia
Email: marko.bohanec@ijs.si

Biljana Mileva Boshkoska*

Department of Knowledge Technologies,
Jožef Stefan Institute,
Ljubljana, 1000, Slovenia
and
Faculty of Information Studies,
Novo mesto, 8000, Slovenia
Email: biljana.mileva@ijs.si
*Corresponding author

Abstract: Knowledge discovery, especially in the field of literature mining, is often involved in searching for some interconnecting concepts between two different literature domains, which might bring new understanding of both domains. This paper presents a new approach to discovering

dependencies between different biological domains based on copula analysis of literature mining results. More specifically, we have explored dependencies between literature from the domains of plant defence response and redox potential. Copula analysis of triplets, which are extracted by Bio3graph tool, shows that dependencies exist between these two domains indicating a potential for cross-domain literature exploration. Bio3graph is a rule-based natural language processing tool which extracts relations in the form (subject, predicate, object) triplets. It is publicly available at http://ropot.ijs.si/bio3graph/software/. Copula analysis was performed by using Clayton and Frank fully nested copulas and the software is publicly available at: http://source.ijs.si/bmileva/copulasfordexapps.git.

Keywords: triplets; relation extraction; modelling the domain dependence; redox potential; plant defence; knowledge discovery; literature mining; fully nested copulas.

Reference to this paper should be made as follows: Miljkovic, D., Lavrač, N., Bohanec, M. and Boshkoska, B.M. (2018) 'Discovering dependencies between domains of redox potential and plant defence through triplet extraction and copulas', *Int. J. Intelligent Engineering Informatics*, Vol. 6, Nos. 1/2, pp.61–77.

Biographical notes: Dragana Miljkovic is a Postdoctoral Researcher at the Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia. She obtained her PhD in Bioinformatics. Her research interests include natural language processing, data mining and modelling. Currently, she is the Coordinator of PD_manager project, which deals with management of Parkinson's disease and is an EU funded H2020 project.

Nada Lavrač is the Head of Department of Knowledge Technologies at Jožef Stefan Institute, Ljubljana, Slovenia. Her main research interests are in machine learning, relational data mining, knowledge management, and applications of data mining in medicine and bioinformatics. She was the Scientific Coordinator of EU projects ILPNET and SolEuNet. She is author and editor of numerous books and conference proceedings, including *Foundations of Rule Learning* (Springer 2012).

Marko Bohanec is a Senior Researcher of Department of Knowledge Technologies at Jožef Stefan Institute, Ljubljana, Slovenia. His main research interests are in decision support systems and machine learning. He was a member of many national and EU projects. He is author and editor of numerous books and conference proceedings.

Biljana Mileva Boshkoska is Postdoctoral Researcher at the Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia, and Associate Professor at the Faculty of Information Studies (FIS), Slovenia. She obtained her PhD in Computer Studies. Currently, she is the Head of the HPC at FIS.

This paper is a revised and expanded version of a paper entitled 'Detection of dependencies between literature domains through relation extraction and copulas' presented at International Conference on Control, Decision and Information Technologies, CoDIt 2016, St. Paul's Bay, Malta, 6–8 April 2016.

1 Introduction

In nature plants sense various harmful conditions, against which they have developed a certain immune mechanism. This mechanism, named plant response to stress, exhibits some differences depending on the type of the stressful stimulus. We distinguish generally between abiotic and biotic types of stress, which both impact plant survival. Abiotic stress is defined as a negative influence of non-living factors, such as extreme temperatures, winds, draught, floods, etc. on the plant. Biotic stress refers, on the other hand, to the damage that different living organisms, such as fungi, insects, weeds and various pathogens make to the plant. The result of the pathogen attack is the production of several phytohormones, among which the most crucial for the plant survival are salicylic acid (SA), jasmonic acid (JA) and ethylene (ET) (Reymond and Farmer, 1998).

Mou et al. (2003) have showed that connection exists between accumulation of SA in the cell, challenged by pathogens, and changes in redox (or reduction) potential. Redox potential is defined as a tendency of a certain molecule to acquire electrons which reduces consequentially its oxidative status. Many biological reactions, including the plant immune reactions, are of oxidation/reduction reaction type where one reacting component gets oxidised (releases electrons) and the other one gets reduced (gains electrons). Oxidation reactions often release various free radicals which can trigger chain reactions. These chain reactions, known as 'oxidative stress', might harm or even destroy the cell. Redox components, which carry fundamental information on cellular redox state, terminate these chain reactions by removing free radical intermediates, and inhibit other oxidation reactions. Redox potential is defined as a tendency of a certain molecule to acquire electrons. Fundamental information on cellular redox state is carried by redox components, which terminate particular chain reactions known as 'oxidative stress' that might harm or even destroy the cell.

There are evidences that the key redox components in the cell, such as NAD^+, NADP, glutathione, ascorbate, etc. influence gene expression triggered by biotic and abiotic stress responses (Noctor, 2006). Foyer and Noctor (2005) proposed a model for redox homeostasis where interaction of reactive oxygen species (ROS) plays a role of an interface between the signals coming from the metabolism and the ones triggered by the environment stimuli. SA mediates PATHOGEN-RELATED (PR) gene expression by altering the cellular redox potential, thereby activating transcription via the transcriptional coregulator NPR1 (Caarls et al., 2015). Tada et al. (2008) suggested that redox signals are expressed via SNO and cytosolic thioredoxins (TRXs), which are direct catalysers of NPR1 oligomer-monomer transformation, where changes in NPR1 activity are influenced by SA. Moreover, study by Fobert and Després (2005) confirms that glutathione increase, in response to pathogen attack, causes reduction and activation of NPR1.

A better understanding of the dependencies between domains of redox potential and plant defence is needed, having in mind that the influence of redox potential is still underestimated in agronomic practice (Husson, 2013). To address this challenging task we propose a new procedure, motivated by cross-domain literature mining research, introduced below. Knowledge discovery process (KDP), especially by using the approach of literature mining, often searches for some interconnecting concepts between the two different domains. For example, the KDP between domain A and domain C might bring new understanding of the two domains. Swanson (1986) has defined the ABC approach, which investigates whether agent A is connected with phenomenon C

by discovering complementary structures through interconnecting phenomenon B. If the domains A and C are known in advance, this process is named the 'closed discovery process' (Swanson, 1986). In this paper, we explore dependencies in published scientific literature of two biological domains: the domain of plant defence response to pathogen attack (domain A) and the domain of redox potential (domain C). We define literature common to both domains as bridging domain B. Next we provide two copula-based models that describe the domain dependences. The first model describes the dependences that exist between domains A and C, and the second model describes the dependences that exist among domains A, B and C. The results show that both models are supplementary. The contributions of this paper are twofold. First, linear methods have been widely used to model nonlinearity in small datasets. Here we model the dependencies by applying copula functions (Nelsen, 2006) which determines also nonlinear dependencies between variables. Second, we search for the dependencies between the biological domains, which have not been previously approached in such a way.

The proposed procedure to cross-domain literature mining follows a two-stage approach. We firstly identify important biological components and their interactions, extracted in the form of triplets (subject, predicate, object) by natural language processing (NLP) method. Secondly we use copula functions on the extracted triplets to describe dependences between the domain of plant defence and the domain of redox potential. In continuation we provide the background methodologies regarding NLP for relation extraction in the form of triplets, and different copula functions.

2 Background methodologies

2.1 NLP methods

Biological information related to the plant defence and redox potential in plants is vastly stored in scientific literature, which can be either explored manually, which is a time-consuming process, or by applying automated NLP methods. In the domain of biology, many NLP tools have been developed that enable automatic extraction of relations between biological components (check bioNLP community[1] for the arising list of NLP tools in the biology field). A wide range of machine learning techniques [including the naive Bayes classifier (Craven and Kumlien, 1999), support vector machines (Donaldson et al., 2003), clustering (Hasegawa et al., 2004), etc.], rule-based systems [GeneWays (Rzhetsky et al., 2004), Chilibot (Chen and Sharp, 2004), PLAN2L (Krallinger et al., 2009), Bio3graph (Miljkovic et al., 2012)], and co-occurrence approaches have been used for relations extraction in systems biology. The closest to our Bio3graph triplet extraction approach is the GeneWays system (Rzhetsky et al., 2004), which enables the extraction, analysis, visualisation and integration of molecular pathway data, but the system is not publicly available. On the other hand, Bio3graph (Miljkovic et al., 2012) is publicly available and supports the extraction, construction and visualisation of the network topology based on the predefined component and reaction vocabularies.

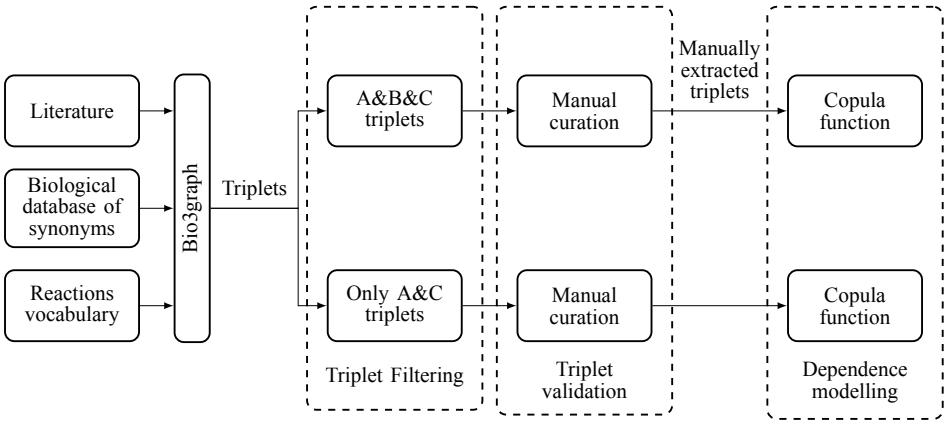
2.2 Copula functions

In probability theory, a copula is defined as a multivariate probability distribution function that is used to describe the dependences between random variables (Joe, 1997; Nelsen, 2006). Copula functions have been successfully used in various fields such as biology (Kim et al., 2008), industry (Mileva Boshkoska et al., 2015), decision making (Mileva-Boshkoska and Bohanec, 2012), etc. To use copulas, we firstly represent the domains as random variables, whose values are the triplet occurrences, and then we model their interdependence. Triplets can be considered as occurrences of events of a random variable, where the random variable is one literature domain. We are interested in the following problem. Given the number of triplet occurrences in domain A and domain C, can we say something about their interdependence expressed via domain B? Hence, we are only interested in those triplets that occur in all three domains. Occurrences of triplets in one domain and their absence in another domain at the same time, lead us to the conclusion that there is no dependence between the domains regarding the given triplet.

3 Materials and methods

The literature for domains of plant defence, redox potential and their intersection was retrieved in the form of full-text articles from the PubMed Central (PMC)[2] database. Then, for the relation extraction was used Bio3graph tool, which is implemented as a reusable workflow of NLP components for information extraction from biological literature in a format compatible with systems biology formalisms, and workflow components for graph construction and visualisation. Next, the obtained triplets are filtered regarding the domains of interest and are manually validated by expert to obtain only true positive triplets. The second step of our approach is the use of different copula functions to explore the dependencies between the two biological domains. The Figure 1 presents the overview of the proposed methodology.

Figure 1 Schematic representation of the methodology



3.1 Literature retrieval

In this study we have used full-text scientific papers stored at PMC Open Access Subset (OA). It is a constantly growing collection of publications which are accessible under a Creative Commons or similar license. The OA scientific publications are available for data mining, text mining, and information extraction using automated processing pipelines. To facilitate computer processing, the Open Archives Initiative service and the FTP service allow downloading full-text XML as well as images, PDF, and supplementary data files for all articles in the OA subset.

3.2 Triplet extraction with Bio3graph

Bio3graph is a rule-based NLP system which extracts relations in the form of triplets (subject, predicate, object) (Miljkovic et al., 2012). In biological texts, this triplet structure refers to the form (component 1, reaction, component 2). The Bio3graph includes text mining, information extraction, graph construction and graph visualisation steps, providing reusability and repeatability. An integral part of this tool is a domain specific vocabulary that is composed of two parts: a list of components and a list of reactions together with their synonyms. The components vocabulary consists of all genes, their short names and synonyms for the model plant *Arabidopsis thaliana* obtained from TAIR database (Swarbreck et al., 2008). *Arabidopsis thaliana* is a model plant, which is the most used for studies in the field of plant physiology and therefore has the most completed genomics data. Furthermore, the vocabulary for the reaction types contains synonyms for the three reaction types: activation, inhibition and binding. Separate files for each reaction type in both the passive and the active verb form are available in supporting information S4 (Miljkovic et al., 2012). Given the list of components, Bio3graph detects subject and object as component 1 and component 2, while the predicate represents the relation between the components as defined in the vocabulary of reaction types. For example, an activation reaction type is presented as: (MPK3, activates, EIN3). These triplets are more informative for systems biologists than, for example, the information obtained from co-occurrence approaches. The later obtain only the information whether component 1 and component 2 are related, but they do not extract the relation type. For this reason, we have selected triplets as a first step in our cross-domain literature mining methodology.

3.3 Copulas

In probability theory, the dependence between random variables is completely defined by their joint distribution function. The joint distribution function $H(x, y)$ for two random variables (r.v.) X and Y, specified on the same probability space, defines the probability of a random event in terms of both X and Y. It is given by:

$$H(x, y) = P[0 \leq X \leq x, 0 \leq Y \leq y] \tag{1}$$

where P is a probability function. To find the joint distribution function in analytical form, we use the Sklar's theorem (Sklar, 1959) which proves that the joint distribution function of two r.v. is equal to the copula of their uniform distributions on the unit interval [0, 1].

Theorem 1 (Sklar's theorem): Let H be a bivariate distribution function with marginal distribution functions $u_1 = F(x)$ and $u_2 = G(y)$. Then copula C exists such that for all $x, y \in \mathbb{R}$:

$$H(x, y) = C(F(x), G(y)) = C(u_1, u_2) \tag{2}$$

If $F(x)$ and $G(y)$ are continuous, then C is unique; otherwise C is uniquely determined on $Range(F) \times Range(G)$. Conversely, if C is a copula and $F(x)$ and $G(y)$ are distribution functions, then the function H defined by equations (1) and (2) is a joint distribution function.

Copulas are functions that manage to formulate the multivariate distribution in such a way that various general types of dependences including the nonlinear one may be captured. We focus on two families of bivariate Archimedean copulas: Clayton and Frank, which we extend to multivariate ones.

3.3.1 Archimedean bivariate copulas

A class of well-known copulas are the Archimedean bivariate copulas. They are constructed using functions called generator functions. The usage is mainly motivated by their convenient properties, such as symmetry and associativity.

Here we focus on Clayton and Frank Archimedean copulas. Their mathematical forms are presented in Table 1. In Table 1, the notation $\varphi_\theta(t)$ represents a so called generator function that is responsible for constructing the copula function.

Table 1 Different Archimedean copulas, their generator functions φ, borders of θ parameter

Copula type	$C_\theta(u, v)$	$\varphi_\theta(t)$	θ
Clayton	$\left[\max\left(u^{-\theta} + v^{-\theta} - 1, 0\right)\right]^{-1/\theta}$	$\frac{1}{\theta}\left(t^{-\theta} - 1\right)$	$[-1, \infty) \setminus \{0\}$
Frank	$-\frac{1}{\theta}\ln\left(1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1}\right)$	$-\ln\frac{e^{-\theta t} - 1}{e^{-\theta} - 1}$	$(-\infty, \infty) \setminus \{0\}$

3.3.2 Multivariate copulas

Table 1 presents only bivariate copulas. However, there are several approaches that describe procedures for constructing multivariate copulas (MVCs) (Fischer et al., 2009). We adopt the one described by Berg and Aas (2009) which uses nesting technique applied on bivariate Archimedean copulas to obtain a multivariate one. When nesting is

performed so that in each level the former copula is coupled with a new input variable, we obtain a copula known as fully nested Archimedean constructions (FNACs), such as the one presented in Figure 2. The basic construction element in the FNAC represents the bivariate copula. As shown in Figure 2, firstly the two nodes u_1 and u_2 are coupled forming a bivariate copula $C_1(u_1, u_2)$ with parameter θ_1. In the next step C_1 is coupled with u_3 into $C_2(u_3, C_1)$ with parameter θ_2 (Savu and Trede, 2006):

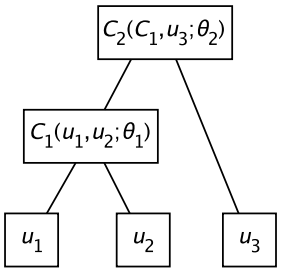
$$C_2(u_3, C_1(u_1, u_2)) \tag{3}$$

The only condition so that equation (3) represents a valid copula expression is:

$$\theta_1 \geq \cdots \geq \theta_n \tag{4}$$

The condition given in equation (4) means that the most nested copula (see copula C_2 in Figure 2 must have the highest value of the dependence parameter θ. The higher values of θ mean higher dependence between the variables.

Figure 2 Fully nested Archimedean copula



4 Results and discussion

The keywords for obtaining literature from PMC database were defined by biology experts resulting in over 30.000 full text articles. This literature was clustered into domains A, C and the bridging domain B, as explained in Section 4.1. Next, relations in the form of triplets were extracted by the Bio3graph tool, where we considered for further analysis only the triplets which appear in all three domains. In the last step of our approach copula functions revealed several dependency connections between the domains.

4.1 Retrieved literature

In order to obtain relevant literature from PMC database two queries were constructed. The queries present combination of MeSH terms and keywords that the domain experts considered important. The first query related to the domain of plant defence response, contains the following set of keywords:

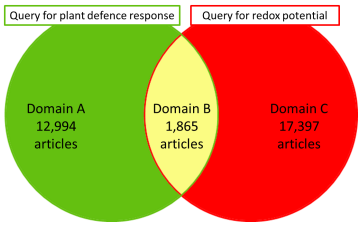
```
"arabidopsis thaliana"[All Fields] AND
( "defence"[All Fields] OR
"defense"[All Fields] OR
"ethylene"[All Fields] OR
"jasmonate"[All Fields] OR
"jasmonic acid"[All Fields] OR
"salicylate"[All Fields] OR
"salicylic acid"[All Fields] OR
"pathogen"[All Fields] OR
"virus"[All Fields])
```

and resulted in 14,859 scientific papers. Using the following second set of keywords from the domain of redox potential:

```
("redox"[All Fields] OR
"reduction"[All Fields] OR
"oxidation"[All Fields]) AND
("potential"[All Fields] OR
"state"[All Fields])
```

19,262 PMC articles were retrieved. From the two queries we formed three domains of biological papers (see Figure 3). Domain A includes papers identified exclusively by the first query. Domain C includes papers identified only by the second query. The domain B, to which we also refer as a bridging domain, contains 1,865 articles that were retrieved by both queries.

Figure 3 Diagram of the domains defined in this study (see online version for colours)



Notes: The middle domain is bridging domain B, containing 1,865 papers which belong to the intersection two queries: for the plant defence response and for the redox potential. Left domain is domain A counting 12,994 scientific articles and belongs strictly to the domain of plant defence response. Domain C, the right one, contains 17,397 biological papers which belong solely to the domain of redox potential.

4.2 Extracted triplets

The result of using the Bio3graph triplet extraction algorithm is a set of 7,733 unique triplets, identified from the total of 11,492 extracted triplets. Since the objective of the study is to explore the connections between domains, only a group of 20 triplets appeared in all three domains and we have filtered them out to proceed with their

validation. The evaluation of triplets was manual and resulted in 8 triplets which were true positive[3] (see Table 2). The rest of 12 triplets were false positive[4]. False positive triplets obtained by Bio3graph were of obvious type, therefore it was not needed to introduce the validation procedure with several annotators and explore the degree of inter-annotator agreement. For example, from sentence "Light induces **CCA1** and LHY expression and **represses TOC1**." the triplet {CCA1, inhibits, TOC1} was extracted, where actually the subject in the sentence is light, and not CCA1.

All relations found by the triplet extraction algorithm are of the 'activation' type. Table 2 gives a summary of the automatically extracted relations between the biological components, providing the numbers of occurrences where each triplet was evaluated as true positive. Moreover, we have selected true positive triplets which exist only in domains A and C, where they do not appear in the domain B. These triplets are of particular interest for the cross-domain knowledge discovery since they appear in two totally separated domains. A summary of these triplets is provided in Table 3, where second and third column show number of occurrences in the domains A and C respectively. Tables 2 and 3 are used for the dependency analysis with copulas.

4.3 Detected domain dependencies through copulas

Here we explore first the dependencies between A, B and C domains and then the dependencies A and C domains excluding the bridging B domain.

4.3.1 Dependencies between A, B and C domains

To use copulas we firstly sorted triplets, according to the number of their occurrences in the domain of interest, which is the domain C (redox potential). In domain C, the number of occurrences of selected triplets is ones, twice or three times, as shown in the last column of Table 2. Based on this information, all triplets in Table 2 are grouped in three groups, as shown in the first column. The triplets IDs are given in the second column of Table 2, while the number of triplets occurrences in domains A and B are shown in the fourth and fifth columns of Table 2 consecutively. Observing Table 2, we may conclude the following. There is a positive correlation between domains A and B in groups 1 and 3. However, it is unclear what their mutual dependency with the domain of interest (domain C) is. Also a clear pattern of occurrences of triplets in different groups cannot be determined.

To provide an initial description of the mutual dependence we apply the copula functions. The question that we have to answer in order to use copula functions is how to rank the triplets meaningfully, so that we can apply copulas? Since we are interested in those triplets that occur in domain C, we have ranked them according to their number of occurrences in the domain of interest. We expect that those that occur more frequently in the domain of interest, i.e., domain C, can be found also more frequently at least in one of the domains A and C and hence would be good candidates for representing a dependence structure between the domains. From mathematical point of view, the values of domains A, B and C may get any discrete value from the space $\Omega = \{1, 2, 3, \ldots, N\}$. Consequently, domains may be considered as discrete random variables and therefore are suitable for the application of copulas. Using this approach, we have performed MVC simulations, and we provide the obtained results in Table 4.

Table 2 True positive triplets, which are extracted with Bio3graph from all three domains and are sorted and grouped according to their number of occurrences in the domain C (redox potential)

	Triplet ID	*Extracted triplet*	*Triplet occurrences in domain A*	*PMCID*	*Triplet occurrences in domain B*	*PMCID*	*Triplet occurrences in domain C*	*PMCID*
Group 1	6	arakin, activates, arabidopsis thaliana mitogen-activated protein kinase kinase 4	1	3402898	1	3325911	1	3350994
	3	agamous-like 20, activates, leafy	3	3039610; 3276106; 3777159	1	3675103	1	3669742
	2	atost1, activates, carbon dioxide insensitive 3	4	3172217; 3585266; 3548404	2	3564773	1	2978106
	4	agamous-like 25, inhibits, agamous-like 20	7	2254019; 3571917;3753250; 3278046; 2806528; 1868597	2	3572515	1	3669742
	5	agamous-like 25, inhibits, flowering locus t	10	2254019; 3540089; 3571917; 2875011; 3753250; 2777508; 2561057; 1868597	2	3572515	1	3669742
Group 2	7	agamous-like 22, inhibits, flowering locus t	3	287501	3	2605480	2	3669742
Group 3	8	ppdk, activates, pep	1	3353924	1	2173943	3	3240839
	1	phospholipase d alpha 1, activates, 7red	7	3355621; 3676348; 2814106; 3733633	5	3098243; 3645664	3	3112519; 3641713

Table 3 True positive triplets, which are extracted with Bio3graph from domains A and C and in the same time these triplets do not appear in the domain B

Extracted triplet	*Triplet occurrences in domain A*	*Triplet occurrences in domain C*
flowering locus t, activates, leafy	1	1
cyanase, activates, b-box domain protein 1	1	1
flowering locus t, activates, agamous-like 8	2	1
arabidopsis thaliana ataxia-telangiectasia mutated, activates, atnbs1	1	2
arabidopsis thaliana protein-serine kinase 1, activates, ribosomal protein s6	1	2
aprr9, inhibits, atcca1	5	1
aprr9, inhibits, late elongated hypocotyl	6	1
aprr7, inhibits, atcca1	6	1
aprr7, inhibits, late elongated hypocotyl	6	1
arabidopsis thaliana general control 0n-repressible 2, activates, arabidopsis thaliana eukaryotic translation initiation factor 3 subunit f	4	1
arabidopsis thaliana eukaryotic translation initation factor 4e1, binds, cucumovirus multiplication 2	2	1
agd10, activates, atrad51	1	1
aterf3, activates, aterf1	2	1
arabidopsis thaliana ataxia-telangiectasia mutated, activates, arabidopsis thaliana breast cancer susceptibility1	1	4
atrad50, activates, arabidopsis thaliana ataxia-telangiectasia mutated	1	2
arabidopsis meiotic recombination 11, activates, arabidopsis thaliana ataxia-telangiectasia mutated	1	2
atnbs1, activates, arabidopsis thaliana ataxia-telangiectasia mutated	1	2
arabidopsis thaliana fk506-binding protein 12, binds, target of rapamycin	1	1
enhancer of ag-4 2, activates, ag	1	1
arabidopsis thaliana sulfotransferase 1, binds, pp2a	3	1
atvps34, activates, atpip2	2	5
atvps34, activates, pip3	2	8
3'-phosphoi0sitide-dependent protein kinase 1, activates, akt1	1	1
hac1, activates, atbzip	1	1
aba insensitive 3, activates, microrna 159	3	1
arabidopsis thaliana constitutive photomorphogenic 1, activates, elongated hypocotyl 5	1	1
aha1, activates, matrix metalloproteinase	1	1
maturation of rbcl 1, binds, rbcl	1	1

Table 4 Results from FNACs

No.	*Copula type*	*Coupling order of domains in MVC*	θ_1	θ_2
1	Clayton FNAC	1-3-2 (B-C-A)	2.4226	2.1971
2	Frank FNAC	1-3-2 (B-C-A)	5.9512	3.6572
3	Frank FNAC	1-2-3 (B-A-C)	5.5649	3.8307
4	Clayton FNAC	1-2-3 (B-A-C)	3.1204	1.6065
5	Frank FNAC	2-3-1 (A-C-B)	condition (4) unfulfilled	
6	Clayton FNAC	2-3-1 (A-C-B)	condition (4) unfulfilled	

The first column in Table 4 represents the type of copula function that we have applied. The next column gives the order of coupling the domains in bivariate copulas. Using the Frank FNAC we model the dependences between intersection domain B and domain A vs. domain C, represented as (1-2-3) in Table 4; the dependencies between bridging domain B and domain C on one side and domain A on the other, represented as (1-3-2) in Table 4; and dependencies of domain A and C versus B represented as (2-3-1). The last two columns represent the values of θ_1 and θ_2, for cases where $\theta_1 \geq \theta_2$.

In Table 4, values $\theta_1 = 2.4226$ and $\theta_1 = 5.9512$ obtained with Clayton and Franc copulas, respectively, show a strong dependency between domains B and C. This observation is in line with the observed positive correlation from Table 2.

The values of $\theta_1 = 5.5649$ vs. $\theta_1 = 5.9512$, which are obtained for coupling domain B-A, and domains B-C, respectively, show that the dependence between domains B and C is stronger than between domains B and A when using Frank FNACs. On the other hand, value $\theta_2 = 3.8307$ which is higher than $\theta_2 = 3.6572$ uncovers that the overall dependency is higher, when we first couple domains B-A and then add domain C. Such values show that dependences that exist among the three domains can be better observed when looking at the domain C on one hand and A-B domains on the other, compared to the case when we look at domain A versus B-C domains.

Unlike the Frank copula, which best models values around the mode, Clayton copula models the left tails, or small values of the distributions. The values of $\theta_1 = 2.4226$ and $\theta_1 = 3.1204$ which are obtained for coupling domain B-C, and domains B-A, respectively, show that the left tail dependence between domains B-A is stronger than between domains B-C. The values $\theta_2 = 2.1971$ which is higher than $\theta_2 = 1.6065$ uncovers that the overall left tail dependency is higher, when we first couple domains B-C and then add domain A. This is of interest as we are looking exactly for triplets that occur rarely, however have a biological significance in other domains.

The last two rows in Table 4 give information about copula types and coupling order of domains for which a valid copula cannot be constructed due to unfulfilled condition (4). In particular, we refer to modelling dependencies using Clayton FNAC for the coupling order of domain 2-3-1 (A-C-B) and with Frank FNAC for the coupling order of domain 2-3-1 (A-C-B). These information reveal that modelling the domains A and C with domain B, using the data from Table 2 is not possible with Clayton and Frank copulas.

The PDFs of the Clayton copulas for θ_1 and θ_2 are given in Figures 4 and 5, respectively. Such functions could be used for predicting the occurrences of triplets in different domains, as presented in Figure 7(b).

Figure 4 PDF for the Clayton copula for θ_1 (see online version for colours)

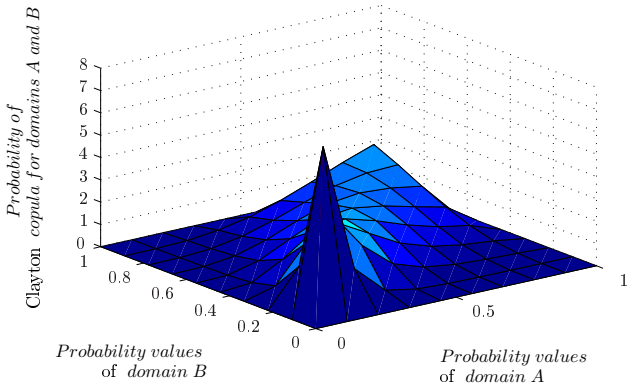


Figure 5 PDF for the Clayton copula for θ_2 (see online version for colours)

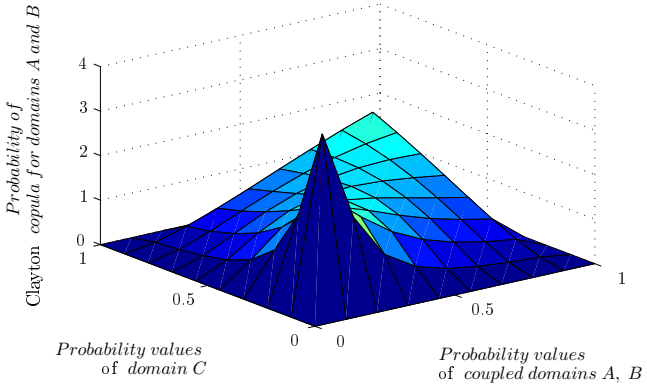


Figure 6 Probability density function for Clayton copula built on domains A and C as given in Table 3 (see online version for colours)

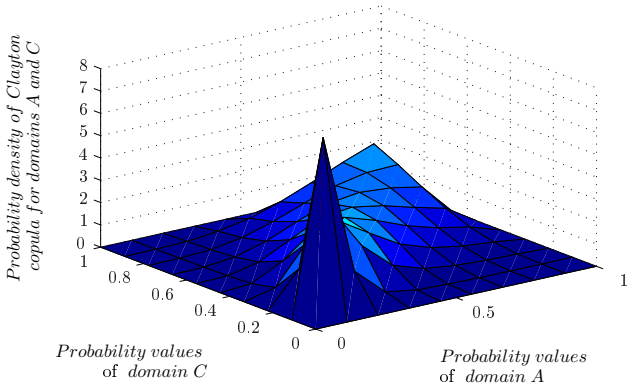
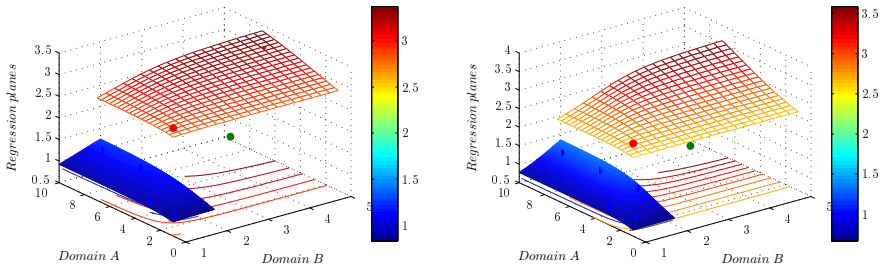


Figure 7 Predicting the values in domain C, (a) Clayton copula (b) Frank copula (see online version for colours)



(a) (b)

Notes: Z-axis are regression values obtained as a function of values in domains A and B. The models are obtained using the Clayton copula (left) and Frank copula (right).

4.3.2 Dependencies between A and C domains

Another possibility is to observe data only from domains A and C excluding the bridging domain B. Such data are provided in Table 3. To check on the linear correlation between the two datasets in Table 3, we calculated the Pearson coefficient which is -0.1392. The low negative value of the Pearson coefficient depicts very weak negative linear correlation. Thus, we propose to use copulas to depict the nonlinear dependency between the two domains. For that purpose, we built Clayton copula with $\theta = 3.3050$ and Frank copula with $\theta = -7.4437$ on these data as given in Table 5.

Table 5 Results from bivariate copulas on data in Table 2

No.	*Copula type*	*Coupling order of domains in MVC*	θ
1	Clayton	A-C	3.3050
2	Frank	A-C	−7.4437

The negative value of the θ parameter of the Franks copula depicts the negative dependency between these two domains. The probability density function for Clayton copula built on domains A and C as given in Table 3 is presented in Figure 6. It is used to describe the left tail dependences. Unlike Frank copula, Clayton copula does not depict the negative dependence, which means that it does not assign probability to joint opposite behaviour in the tails of the variable distributions. The value of $\theta = 3.3050$ models the positive dependence in the left tails of the two variables.

5 Conclusions

This paper presents an approach to discovering dependencies between different biological domains based on the copula analysis of the results obtained from relation extraction. In the illustrative example on the domains of plant defence response and

redox potential we show that dependencies exist between these two domains indicating a potential for further exploration. In future work, we plan to broaden our analysis by using also some other text mining approaches, for example co-occurrence, which might provide more triplets than the currently used Bio3graph. The presented approach can be extended to any other biomedical domain.

Acknowledgements

This work was financed by Slovenian Research Agency grants L2-7663, P1-0383, J1-7151, J7-7303 and P2-0103.

References

Berg, D. and Aas, K. (2009) 'Models for construction of multivariate dependance: a comparison study', *European Journal of Finance*, Vol. 15, Nos. 7–8, pp.639–659.

Caarls, L., Pieterse, C.M.J. and van Wees, S.C.M. (2015) 'How salicylic acid takes transcriptional control over jasmonic acid signaling', *Frontiers in Plant Science*, Vol. 6, No. 170, p.1–11.

Chen, H. and Sharp, B.M. (2004) 'Content-rich biological network constructed by mining pubmed abstracts', *BMC Bioinformatics*, Vol. 5, No. 147, pp.1–13.

Craven, M. and Kumlien, J. (1999) 'Constructing biological knowledge bases by extracting information from text sources', *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pp.77–86, AAAI Press.

Donaldson, I., Martin, J.L., de Bruijn, B., Wolting, C., Lay, V., Tuekam, B., Zhang, S., Baskin, B., Bader, G., Michalickova, K., Pawson, T. and Hogue, C. (2003) 'Prebind and textomy – mining the biomedical literature for protein-protein interactions using a support vector machine', *BMC Bioinformatics*, Vol. 4, No. 11, p.1–13.

Fischer, M., Köck, C., Schlüter, S. and Weigert, F. (2009) 'An empirical analysis of multivariate copula models', *Quantitative Finance*, Vol. 9, No. 7, pp.839–854.

Fobert, P.R. and Després, C. (2005) 'Redox control of systemic acquired resistance', *Current Opinion in Plant Biology*, Vol. 8, No. 4, pp.378–382.

Foyer, C.H. and Noctor, G. (2005) 'Redox homeostasis and antioxidant signaling: a metabolic interface between stress perception and physiological responses', *The Plant Cell*, Vol. 17, No. 7, pp.1866–1875.

Hasegawa, T., Sekine, S. and Grishman, R. (2004) 'Discovering relations among named entities from large corpora', *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, ACL '04*, Association for Computational Linguistics, Stroudsburg, PA, USA.

Husson, O. (2013) 'Redox potential (eh) and ph as drivers of soil/plant/microorganism systems: a transdisciplinary overview pointing to integrative opportunities for agronomy', *Plant and Soil*, Vol. 362, Nos. 1–2, pp.389–417.

Joe, H. (1997) *Multivariate Models and Dependence Concepts*, Chapman and Hall/CRC.

Kim, J.M., Jung, Y.S., Sungur, E., Han, K.H., Park, C. and Sohn, I. (2008) 'A copula method for modeling directional dependence of genes', *BMC Bioinformatics*, Vol. 9, No. 1, p.225.

Krallinger, M., Rodriguez-Penagos, C., Tendulkar, A. and Valencia, A. (2009) 'Plan2l: a web tool for integrated text mining and literature-based bioentity relation extraction', *Nucleic Acids Research*, Vol. 37, pp.W160–W165, Web Server issue.

Mileva-Boshkoska, B. and Bohanec, M. (2012) 'A method for ranking non-linear qualitative decision preferences using copulas', *International Journal of Decision Support System Technology*, Vol. 4, No. 4, pp.1–17.

Mileva Boshkoska, B., Boškoski, P., Debenjak, A. and Juričić, Đ. (2015) 'Dependence among complex random variables as a fuel cell condition indicator', *Journal of Power Sources*, Vol. 284, pp.566–573.

Miljkovic, D., Stare, T., Mozetič, I., Podpečan, V., Petek, M., Witek, K., Dermastia, M., Lavrač, N. and Gruden, K. (2012) 'Signalling network construction for modelling plant defence response', *PLOS ONE*, Vol. 7, No. 12, pp.e51822-1–e51822-18.

Mou, Z., Fan, W. and Dong, X. (2003) 'Inducers of plant systemic acquired resistance regulate {NPR1} function through redox changes', *Cell*, Vol. 113, No. 7, pp.935–944.

Nelsen, R.B. (2006) *An Introduction to Copulas*, 2nd ed., Springer, New York.

Noctor, G. (2006) 'Metabolic signalling in defence and stress: the central roles of soluble redox couples', *Plant Cell Environ.*, Vol. 29, No. 3, pp.409–425.

Reymond, P. and Farmer, E.E. (1998) 'Jasmonate and salicylate as global signals for defense gene expression', *Current Opinion in Plant Biology*, Vol. 1, No. 5, pp.404–411.

Rzhetsky, A., Iossifov, I., Koike, T., Krauthammer, M., Kra, P., Morris, M., Yu, H., Duboué, P.A., Weng, W.B., Wilbur, W.J., Hatzivassiloglou, V. and Friedman, C. (2004) 'Geneways: a system for extracting, analyzing, visualizing, and integrating molecular pathway data', *Journal of Biomedical Informatics*, Vol. 37, No. 1, pp.43–53.

Savu, C. and Trede, M. (2006) 'Hierarhical Archimedean copulas', *International Conference on High Frequency Finance*, Konstanz, Germany.

Sklar, A. (1959) 'Fonctions de répartition à n dimensions et leurs marges', *Publ. Inst. Statist. Univ. Paris*, Vol. 8, pp.229–231.

Swanson, D.R. (1986) 'Undiscovered public knowledge', *The Library Quarterly: Information, Community, Policy*, Vol. 56, No. 2, pp.103–118.

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T.Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A., Singh, S., Swing, V., Tissier, C., Zhang, P. and Huala, E. (2008) 'The arabidopsis information resource (TAIR): gene structure and function annotation', *Nucleic Acids Research*, Vol. 36, pp.D1009–D1014, Database issue.

Tada, Y., Spoel, S.H., Pajerowska-Mukhtar, K., Mou, Z., Song, J., Wang, C., Zuo, J. and Dong, X. (2008) 'Plant immunity requires conformational changes [corrected] of NPR1 via S-nitrosylation and thioredoxins', *Science*, Vol. 321, No. 5891, pp.952–956.

Notes

1 http://www.bionlp.org/.

2 PubMed Central is a database of full-text biomedical scientific papers that are accessed free of charge.

3 True positive triplets are triplets correctly extracted by the triplet extraction algorithm.

2 False positive triplets are ones extracted by the triplet extraction algorithm, but which do not correspond to the form {subject, predicate, object} in the sentence.

machine learning & knowledge extraction

MDPI

Article

Discovery of Relevant Response in Infected Potato Plants from Time Series of Gene Expression Data

Dragan Gamberger [1,*], Tjaša Stare [2], Dragana Miljkovic [3], Kristina Gruden [4] and Nada Lavrač [5]

1 Rudjer Bošković Institute, Bijenička cesta 54, 10000 Zagreb, Croatia
2 National Institute of Biology, 1000 Ljubljana, Slovenia; tjasa.stare@nib.si
3 Jožef Stefan Institute, 1000 Ljubljana, Slovenia; dragana.miljkovic@ijs.si
4 National Institute of Biology, 1000 Ljubljana, Slovenia; Kristina.Gruden@nib.si
5 Jožef Stefan Institute, Ljubljana, Slovenia, University of Nova Gorica, 5000 Nova Gorica, Slovenia; Nada.Lavrac@ijs.si
* Correspondence: dragan.gamberger@irb.hr; Tel.: +385-1-456-1142

Received: 2 November 2018; Accepted: 8 January 2019; Published: 16 January 2019

Abstract: The paper presents a methodology for analyzing time series of gene expression data collected from the leaves of potato virus Y (PVY) infected and non-infected potato plants, with the aim to identify significant differences between the two sets of potato plants' characteristic for various time points. We aim at identifying differentially-expressed genes whose expression values are statistically significantly different in the set of PVY infected potato plants compared to non-infected plants, and which demonstrate also statistically significant changes of expression values of genes of PVY infected potato plants in time. The novelty of the approach includes stratified data randomization used in estimating the statistical properties of gene expression of the samples in the control set of non-infected potato plants. A novel estimate that computes the relative minimal distance between the samples has been defined that enables reliable identification of the differences between the target and control datasets when these sets are small. The relevance of the outcomes is demonstrated by visualizing the relative minimal distance of gene expression changes in time for three different types of potato leaves for the genes that have been identified as relevant by the proposed methodology.

Keywords: gene expression time series; potato virus infections; variance-stabilized data; randomization test; stratified randomization; relative minimal distance of samples

1. Introduction

Potato (*Solanum tuberosum* L.) is the most widely grown tuber crop in the world, and the fourth largest food crop in terms of fresh produce, after rice, wheat, and tomato. Potato virus Y (PVY) is a member of the Potyviridae family and, economically, it is one of the most important potato pathogens, with PVY^{NTN} being, worldwide, an aggressive isolate that induces severe symptoms in sensitive potato cultivars [1–3]. The interaction between a plant and its pathogen initiates a complex signaling network, resulting in massive changes of the gene activity and extensive reprogramming of the cell metabolism [4,5].

Salicylic acid (SA) has shown to mediate resistance in many compatible plant-virus interactions and its deficiency leads to impairment of the defense responses and increased susceptibility to pathogen attacks [6,7]. Compatible interaction is a term broadly used in plant pathology that refers to the interaction between the pathogen and the plant that leads to successful infection, while incompatible interaction stands for successful plant resistance: i.e., the host's ability to limit pathogen multiplication [8]. Recently, we performed a time series analysis of defense responses in compatible potato-PVY^{NTN} interaction using the tolerant cultivar Désirée [4]. Although the plant's fitness was

almost unaffected, the virus multiplied in the inoculated leaves from five days post inoculation (dpi) on and the spread of viral RNA to upper leaves was detected at 7 dpi [9]. To determine the role of SA in this interaction, the NahG-Désirée transgenic line that expresses salicylate hydroxylase, which catalyzes the conversion of SA to catechol [7,10,11], was also analyzed. In contrast to the non-transgenic plants of cv. Désirée, the SA-deficient transgenic NahG-Désirée showed a greater susceptibility to PVY^{NTN}. Symptoms in the terms of pronounced yellowing and necrotic lesions started to appear on the site of infection from 4 dpi, and became more pronounced in later days. The appearance of the symptoms in NahG-Désirée corresponded to the first detection of viral multiplication at 4 dpi [4]. The dynamics of whole transcriptome changes of cultivar Désirée and NahG-Désirée was analysed in inoculated and systemically infected leaves following 0, 1, 3, 4, 5, and 7 dpi.

This paper proposes a methodology aimed at systematic identification of genes that have statistically significant differences of gene expression values between the PVY infected samples and the non-infected (mock) samples at various time points of the recorded time series data. The identified genes present the input for expert analysis and reasoning, aiming to uncover why different potato cultivars differ in terms of resistance to PVY, with the ultimate goal to provide novel insights into the relevant biological processes.

The difficulty of the problem is due to a small number of samples (typically only three samples per a time point per a given potato type) and more than 37,000 candidate genes that have to be tested for their significance. Therefore, we cannot use the standard statistical approaches like the Student's *t*-test or the Mann-Whitney U test for this dataset. A potentially interesting approach that has been specifically developed for gene expression data performs a differential comparison of sets of genes that are constructed based on their biological functions [12]. A problem of this methodology for our application in the domain of potato plant time series analysis is the often missing information about functions of genes and their functional groups. Standard approaches to longitudinal gene expression analysis are based on spline-based methods for short time sequences [13] and on the approximation of noisy time sequences with simple and smooth functions [14]. These techniques are appropriate for discriminating among complete sequences and are less effective for detecting of differences in specific time points. Finally, techniques for analysis of longitudinal data in medical applications that use within-subject correlation to increase the power of statistical tests [15] are not applicable because potato leaves must be removed from the plants for the analysis of the transcriptome and, therefore, our samples are from physically different leaves.

The novel approach proposed in this work is based on the randomization test concepts [16]. The applicability and usefulness of randomization in gene expression statistical analysis has been previously demonstrated [17,18]. In our approach, we construct large stratified randomized gene sets on which we compute the statistical properties of genes, without taking into account the differences between the infected and non-infected plants; we then use this distribution to estimate which of the actual genes have statistically different expression values distinguishing between PVY infected and mock potato samples.

The next section presents the data and the methodology used for the identification of relevant genes from a time series of small data samples. Section 3 presents the results that illustrate the type and quality of the outcomes of this methodology. Finally, Section 4 provides a summary and discusses the limitations of the proposed methodology.

2. Materials and Methods

The data analysed in this paper are deposited in the NCBI Gene Expression Omnibus, and are accessible through GSE58593 [19].

2.1. Data

Plant material has been grown and manipulated as follows [5]. Potato (*Solanum tuberosum* L.) cv. Désirée and transgenic potato plants of the same cultivar deficient in SA signaling (NahG-Désirée)

were propagated in tissue culture. Two weeks after node segmentation, they were transferred to soil in a growth chamber, and kept at 21 ± 2 °C in the light and 18 ± 1 °C in the dark, at a relative humidity of 75% ± 2%, with 70–90 µmol/m^2/s^2 radiation (L36W/77 lamp, Osram, Germany) and a 16-h photoperiod. After four weeks of growth in soil, the potato plants were inoculated with PVY^{NTN} (isolate NIB-NTN, AJ585342) using sap prepared from homogenized leaves of tissue culture-grown infected potato plantlets of cv. Pentland squire. For the mock-inoculated plants, the same procedure was performed with sap from healthy potato plants.

On the day of inoculation, three leaves from three non-treated plants for each genotype (Désirée, NahG-Désirée) were collected, which were designated as controls, i.e., at the time point zero days post inoculation (dpi). PVY^{NTN}-and mock-inoculated leaf samples were collected on 1, 3, 4, 5, and 7 dpi. Three plants for each treatment were used.

Total RNA from the inoculated leaves was extracted, DNase treated, purified, and quality controlled as described previously by [5]. A one-colour based hybridization protocol was performed on the custom 60-mer oligo microarrays (4 × 44K; AMADID 015425) designed by the Potato Oligo Chip Initiative [20]. For each sample at least 1 µg total RNA was sent for analysis at IMGM Laboratories GmbH, Planegg, Germany. The raw data were analysed in the R Project for Statistical Computing program (R Development Core Team, 2011; version 2.13.2), using the packages Agi4x44PreProcess [21] and Limma [22].

The microarray features were filtered according to the Agilent quality control flags: if the feature was determined to be well above background, if the noise did not exceed a threshold, and if it was not saturated (IsNOTWellAboveBG, IsSaturated, and IsFeatNonUnifOL [20]) in at least 10% of the total microarray count, then that particular microarray feature was retained for further analysis. The raw data of the remaining 37,865 (from a total of 42,034) features was robust spline normalized ('rsn'; see [23]). The empirical Bayes method [22] was used to detect differentially-expressed genes between PVY^{NTN}-and mock-inoculated plants at each time point and for each genotype with corrected $p \leq 0.05$ [24]. Functional analysis of differentially-expressed genes was performed using the MapMan software tool [25] using the ontology adapted for potato [26].

We analysed separately the data from upper non-inoculated and bottom inoculated leaves for the Désirée potatoes, while for NahG potatoes we analysed only the bottom inoculated leaves. In the rest of the paper, Désirée upper leaves are referred to as NT upper, Désirée bottom leaves as NT bottom, while NahG-Désirée potatoes are referred to as NAHG. Table 1 provides an illustration of normalized gene expression data for three out of 37,865 genes. The presented data are for days 1 and 3 for NT upper potato leaves. They illustrate the variability inherent to gene expression measurements.

Table 1. Normalized gene expression data for three genes for NT upper potato leaves. For days 1 and 3, there are three samples for PVY infected plants and three samples for mock plants. The second row presents the values of the three samples before the start of the experiment.

	Gene 10557			**Gene 21013**			**Gene 29447**		
Untouched (day zero)	7.73	7.63	7.62	5.53	5.45	5.48	7.70	7.38	7.50
mock day 1	7.40	7.53	7.39	5.41	5.40	5.55	7.77	7.86	7.88
PVY day 1	7.35	7.07	7.25	5.64	5.52	5.55	7.52	8.01	8.11
mock day 3	7.36	7.57	7.54	5.63	5.58	5.50	7.56	7.66	7.70
PVY day 3	7.35	7.19	7.34	5.48	5.53	5.48	7.87	7.84	7.71

2.2. Methodology

For domain expert analysis the most interesting are the genes that significantly change their expression value for infected plants at a specific point in time. Such genes are characterized by two properties: there is statistically significant difference between PVY values at time points X and X–1, while

the values at these time points do not change in mock samples, and there is a statistically significant difference between PVY and mock values at time point X. Figure 1 illustrates these two conditions.

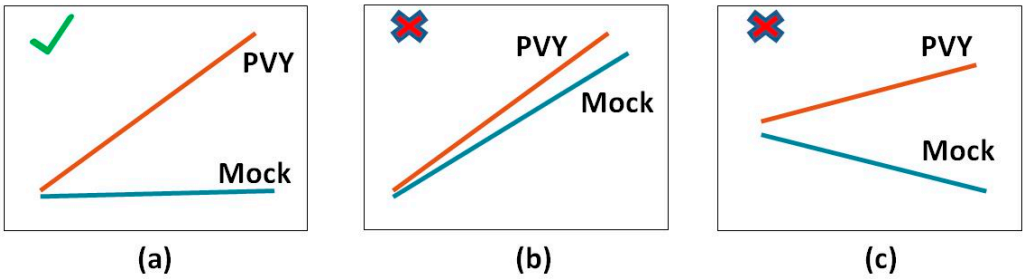


Figure 1. Illustration of gene expression changes that are interesting for domain expert evaluation in subfigure (**a**), in contrast to those that are not relevant as shown in subfigures (**b**,**c**).

This means that we have to solve two separate tasks: (1) In the first task, for each time point, we have to identify the genes whose PVY values have changed significantly from the previous time point; and (2) the second task concerns the identification of genes that at the given time points have substantially different values for the PVY and mock samples. Only the genes that satisfy both conditions are interesting for expert analysis. Systematic identification of complete lists of genes that satisfy both conditions is the aim of the proposed methodology.

While the two tasks are performed on different data, the goal of both of them is to identify sets of genes that are substantially differently expressed between the target and the control samples. In the first task the PVY samples at time point X are the target set, while the PVY samples at time point X–1 are the control set. On the other hand, in the second task, the PVY and mock samples at time point X are the target and the control set, respectively.

Our basic task is the identification of genes that have statistically significant differences in their expression between the target and the control set. When the target and the control sets are small, in our case consisting of only three samples per each set, the standard statistical tests are not applicable. An alternative approach is possible due to the fact that variance has been stabilized in preprocessing of gene expression data by a model based transformation [23]. The approach consists of two steps: definition of an appropriate measure for computing the difference between the target set and the control set, and by the identification of reliable ranges when the actual values of this measure for some concrete gene can be accepted as statistically significantly different from the no-difference assumption.

A natural selection for the measure of difference between two sets of samples is the relative difference between the average values for samples in different sets, referred as RDA. It is defined as the difference between the average value of the target set and the average value of the control set, divided by the average value for the control set. An alternative measure is relative minimal difference (RMD). In this work we use the latter because of its property that a single measurement error, regardless how large it is, cannot substantially increase its absolute value. This property is important for preventing false positive discoveries. Additionally, for a large fraction of randomly generated data, their RMD values are either equal to zero or they have very small values. This property is, therefore, beneficial for the estimation of distributional characteristics of this measure on random data and the recognition of genes whose measured expression values are significantly different.

2.2.1. Relative Minimal Distance

Relative minimal distance (RMD) is defined as follows: its value is positive if all target samples have larger values than the control samples and its value is negative, if all target samples have lower values than the control samples. Furthermore, if there is at least one target sample larger than some control sample and at least one target sample with a value lower than some control sample, then the RMD value is, by definition equal to zero, regardless of the actual values of the samples. RMD also

always has a value of zero when a pair of target and control samples has identical values. The concept is illustrated in Figure 2.

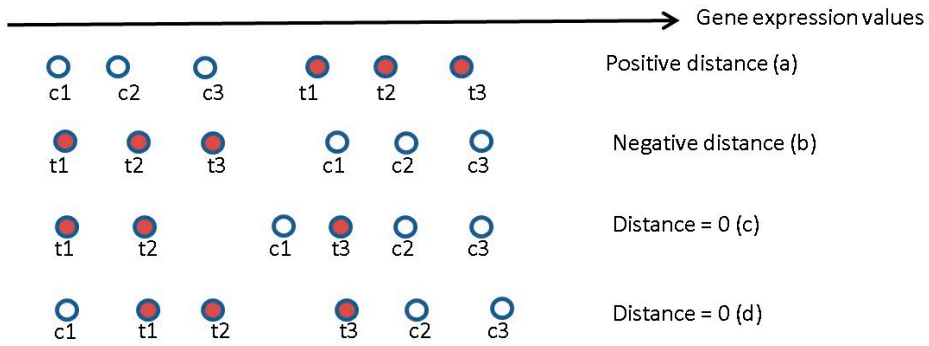


Figure 2. Illustration of the concept of Relative Minimal Distance between samples t1–t3 in the target set (red) and samples c1–c3 in the control set (white).

When all input target and control gene expression values are positive (tx, cx > 0), non-zero RMD values for positive cases are defined by the relation:

$$\mathrm{RMD} = (\mathrm{tmin} - \mathrm{cmax})/\mathrm{cmax} \tag{1}$$

while for negative cases it is:

$$\mathrm{RMD} = (\mathrm{tmax} - \mathrm{cmin})/\mathrm{tmax} \tag{2}$$

where tmin, tmax, cmin, and cmax are the minimal and maximal values for the samples in the target and control sets, respectively.

An important property of RMD is that for random differences between the target and control sets many RMD values will be equal to zero or their value will tend to be small. For example, in the potato plant gene analysis domain with typically three samples both in the target and the control set we can expect that about 90% of genes that behave randomly will have RMD values equal to zero. In the case where there were five samples per set available for the analysis, less than 1% of RMD values for random variables will be different from zero. This property does not depend on the actual intra-set variance. A negative aspect of the RMD measure is that if the intra-set variance is high then it can happen that, even for really significantly different gene sets, the RMD value can also be equal to zero.

Figure 3 presents the distribution of real gene expression data and the corresponding RMD values for one out of 37,865 genes for NAHG potatoes. The RMD values are computed for the differences of gene expression values between the PVY and mock samples. Gene 08407 has a large negative RMD value for day 1 and a small positive RMD vale for day 4. The RMD value for day 1 is statistically significant (see the next section).

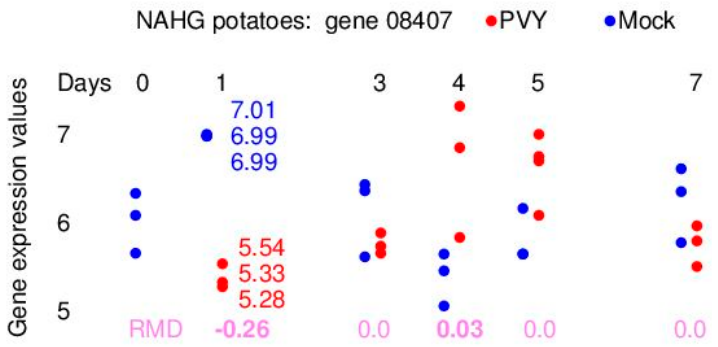


Figure 3. Illustration of changes in time for real gene expression data and the corresponding RMD values.

2.2.2. Critical Regions for Accepting RMD Values as Statistically Significant

Let us assume that we have the target and the control sets with a small number of samples and that there is a large set of genes that are candidates for being uncovered in terms of significantly different values for these two sets. Additionally, let us assume the input data have stabilized variance and there is some measure that can be used to quantify the difference in gene expression values. In the rest of the paper we use the RMD measure defined in the previous section, but note that the methodology can also be used for some other measure of the user's choice. The goal is to identify critical regions of RMD values or, in other words, to compute how large some RMD values must be in order that we can claim that the gene has significantly different expression values for the two sets.

The underlying idea is to construct randomized sets of gene expression values and to compute statistics of RMD values on this data. These statistics will determine the critical regions for RMD values that are acceptable as statistically significantly different from the random data. Since, for randomized data, there are a large number of non-zero RMD values and the probability of positive and negative values is equal, we can conclude that non-zero RMD values computed for randomized data will be normally distributed and that their average value will be equal to zero. This fact is illustrated in Figure 4 for NT-upper potato leaves for day 1. The figure presents distributions of RMD values for real PVY versus mock data and for the randomized data when the PVY and mock values have been shuffled.

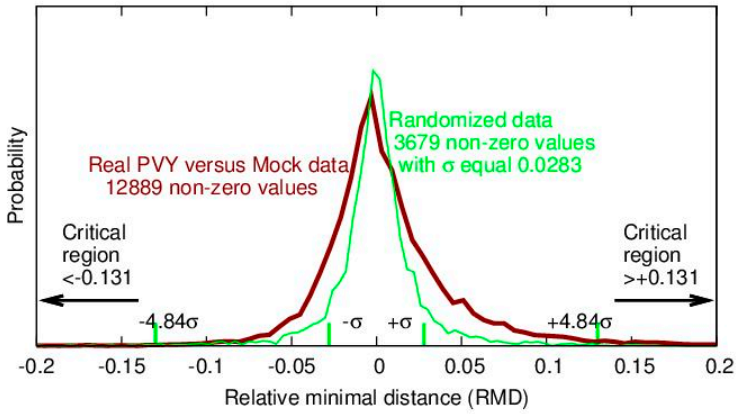


Figure 4. Distribution of real and randomized non-zero RMD values presenting the PVY versus mock data for NT upper potato leaves for day 1.

The critical regions of real RMD values that can be accepted as statistically significant are computed from the standard deviation of the non-zero RMD values for randomized data. The Bonferroni correction has to be used because we have to test the statistical significance of a large number of genes. Theoretically this means that for the two-tailed significance level of 5% and 37,865 genes that have to be tested, some real RMD value can be accepted as statistically significant if its absolute value is at least 4.84 times greater than the standard deviation computed for non-zero RMD values for randomized data (see Figure 4). Factor 4.84 is determined by using function pnorm from the R language, as follows:

$$1\text{-pnorm}\ (4.84 = 6.5 \times 10^{-7} \sim (0.05/2)/37865. \quad (3)$$

Practically the factor can be slightly smaller because for all the genes with real RMD values equal to zero we know that they are not significant. The significance test should be done only for non-zero real RMD values and their number is known for each specific case. Tables 2 and 3 present the data used for computing the critical regions for the first and the second task, respectively.

Table 2. Computation of regions for statistically significant RMD values for the first task.

	Average RMD Value for Randomized Data	Standard Deviation of RMD Values for Randomized Data	Number of Non-Zero RMD Values for Randomized Data	Number of Non-Zero RMD Values for Real Data	Stand. dev. in the Non-Critical Region (with Bonferroni Correction)	Absolute Critical Value for RMD Acceptance as stat. Significant
Day				**NT upper**		
0–1	0.0000	0.0339	3603	16,936	+/−4.68	0.159
1–3	0.0008	0.0305	3509	10,378	+/−4.58	0.140
3–5	−0.0018	0.0314	3690	13,147	+/−4.63	0.145
5–7	−0.0007	0.0321	3677	5,733	+/−4.45	0.143
				NT bottom		
0–1	−0.0001	0.0277	3850	19,444	+/−4.71	0.130
1–3	0.0001	0.0285	3857	8918	+/−4.55	0.130
3–5	0.0001	0.0283	3833	7791	+/−4.52	0.128
5–7	−0.0002	0.0282	3797	5450	+/−4.44	0.125
				NAHG		
0–1	−0.0002	0.0343	3658	14,028	+/−4.64	0.159
1–3	−0.0009	0.0346	3730	16,563	+/−4.67	0.162
3–5	0.0001	0.0287	2129	9119	+/−4.55	0.131
5–7	0.0003	0.0259	2168	4196	+/−4.38	0.113

Table 3. Computation of regions for statistically significant RMD values for the second task.

	Average RMD Value for Randomized Data	Standard Deviation of RMD Values for Randomized Data	Number of Non-Zero RMD Values for Randomized Data	Number of Non-Zero RMD Values for Real Data	Stand. dev. in the Non-Critical Region (with Bonferroni Correction)	Absolute Critical Value for RMD Acceptance as stat. Significant
Day				**NT upper**		
1	0.0001	0.0283	3679	12,889	+/−4.62	0.131
3	0.0001	0.0314	3587	14,459	+/−4.65	0.146
5	0.0004	0.0247	3637	8853	+/−4.54	0.112
7	−0.0004	0.0274	3591	13,546	+/−4.63	0.127
				NT bottom		
1	0.0003	0.0254	3831	8992	+/−4.55	0.116
3	0.0004	0.0283	3831	10,454	+/−4.58	0.130
5	0.0005	0.0373	3857	10,705	+/−4.58	0.171
7	0.0003	0.0397	3855	8377	+/−4.53	0.180
				NAHG		
1	0.0005	0.0370	3884	12,543	+/−4.62	0.171
3	0.0001	0.0222	3790	6509	+/−4.48	0.097
5	−0.0001	0.0220	2107	3896	+/−4.37	0.096
7	0.0001	0.0361	3786	4949	+/−4.42	0.160

The second, third, and fourth columns of the two tables contain the values computed for randomized data sets among which the standard deviation of non-zero RMD values is the most relevant information. The following two columns contain the number of non-zero RMD values on real data and the corresponding factor for the computation of the critical region with the Bonferroni correction. The last column contains the critical values for each day and the type of potato plants. If the absolute value of RMD for some gene is greater than this value, then the gene can be accepted as having statistically significant differences between the target and the control sets.

An important practical issue is how to generate randomized gene expression data. A simple and effective method is by shuffling the real data:

- The shuffling could be done on the complete gene expression data set irrespective of the meaning of the data. It this case the data are mixed irrespective of the potato types, days, and infection status. This approach can result in incorrect estimation of the standard deviations if the differences between the gene expression data for different potato types and/or PVY versus mock data are large.
- Since we do not have expert knowledge on whether and/or how these differences are relevant, we implemented a better approach based on stratified data shuffling.
- For the first task we shuffle only the PVY values for the same potato type within different days. In this way we randomize time related information and the standard deviation is computed from the real PVY data for the specific potato type as if there were no changes of PVY in time.
- For the second task, we shuffle only PVY and mock data for the specific day and the same potato type. In this way we randomize only the differences between the infected and non-infected plants, while the potential time-related differences and the differences between the potato types remain present also in the randomized datasets.

2.2.3. Combination of the Two Tasks

The first step in the identification of significant genes is the computation of critical regions. This is performed by the approach based on the construction of randomized data sets described in the previous section. After we have computed the critical regions, the process of identification of significant genes is very simple: all genes whose absolute value of RMD is above the critical values presented in the last columns of Tables 3 and 4 are accepted as statistically significant. The genes in the resulting lists are ordered according to the descending absolute RMD value.

Table 4. Number of significant genes per task.

	Number of Genes with Significant PVY Changes between Two Time Points	Number of Genes with Significant PVY/mock Differences at the Final Time Point	Number of Genes Satisfying Both Conditions (Final Solution)
Day		**NT-upper**	
0–1	407	240	180
1–3	191	144	41
3–5	258	99	7
5–7	40	205	30
		NT-bottom	
0–1	943	230	173
1–3	96	78	1
3–5	89	243	-
5–7	30	195	5
		NAHG	
0–1	258	38	1
1–3	324	109	35
3–5	254	81	35
5–7	90	125	4

For each time interval the process has to be repeated two times. First we construct an ordered list of genes that significantly changed their PVY values in the given time interval, followed by computing the list of genes that at the end of the time interval have significantly different values in PVY and mock samples. The final list that presents the result of this methodology consists of the genes that are present in both previous lists and their RMD values in these lists have the same sign.

The result of this process is illustrated in Table 4. The table presents in its second column the number of genes identified as relevant by the first task, in the third column are numbers of relevant genes identified by the second task, while the last column presents the number of genes satisfying both conditions.

The main conclusion from Table 4 is that the final solutions include much less genes than the lists generated by the first and by the second task independently. This means that it makes sense to search for the agreement of both conditions and that the genes in the final solution deserve biological evaluation by the domain expert.

The largest numbers of genes detected as relevant are for the changes in the days 0–1 interval for the NT potatoes, which are large both for the upper and the bottom leaves. The result confirms the expert's knowledge that Désirée reaction to infection is stronger and that the reaction is the strongest immediately after PVY infection.

3. Results

The results of the application of the presented methodology are lists of relevant genes that are computed for every time interval. These lists serve as input for expert evaluation. The first step in this evaluation is the analysis of functions of relevant genes. For NT upper leaves functions of three genes whose gene expression values changed most significantly for the first day after infection are: *Argnine/serine-rich splicing factor*, *Thioredoxin*, and *DNA binding protein*. For these three genes their gene expression values of PVY infected samples are statistically significantly higher than the corresponding values of mock samples. For NT bottom leaves, the three genes whose expression values changed most significantly in the same time period are: *Chlorophyll a-b binding protein 3C chloroplastic*, *cell wall protein*, and *YTH domain family 2.* For all these three genes their expression values for PVY infected samples have also increased. In contrast, for NAHG potatoes there is only one gene (*Maleylacetoacetate isomerase glutathione S-transferase)* whose gene expression values have statistically significantly changed and its expression values in PVY infected samples have changed in the opposite direction.

3.1. Biological Evaluation of Selected Genes

In accordance with our previous analysis, photosynthesis-related genes (e.i. gene encoding for *chlorophyll a-b binding protein*) are differentially expressed in NT bottom leaves at first day post inoculation (dpi). The same day the gene encoding for YTH domain family protein, involved in calcium signalling as well as transcripts for cell wall protein are differentially expressed. SA-deficiency alters fast transcriptional response resulting in *maleylacetoacetate isomerase* being identified as a relevant transcript. Interestingly, the changes in gene expression in the upper leaves are detected already in the first time period, with transcripts of *arginine/serine-rich splicing factor*, *thioredoxin*, and *DNA binding protein* being differentially expressed, suggesting a fast systemic plant response.

The differences in gene expression of wild type potato plants in the bottom leaves are detected mostly in the last time interval; from 5–7 day transcriptional regulator (*MYB transcription factor*), gene involved in calcium signalling (*calmodulin-like protein*) and sugar metabolism (*hexulose-6-phosphate isomerase*) are regulated upon virus infection. In NahG genotype (NAHG), different transcripts were identified as important regulators showing the importance of SA hormone in the regulatory process. In the upper leaves, reprograming of gene expression is also noted, showing that the plant response is not limited only to the site of virus entry and identifying genes that have a role in systemic plant response (NT upper).

3.2. Visualization of RMD Values

Results visualization is beneficial for understanding of the meaning and for inspecting the relevance of the results. A standard approach is to present the average values and their changes in time. In Figure 5 we present the data for the gene *chlorophyll a-b binding protein 3C* which has been identified as the most relevant for NT bottom leaves for interval day 0–1. From the figure it can be concluded that NT upper and NT bottom have higher expression values of PVY infected samples than the values of the mock samples in the period day 1–3. The problem with Figure 5 is that it presents six curves and it is rather difficult to capture all potentially relevant relations.

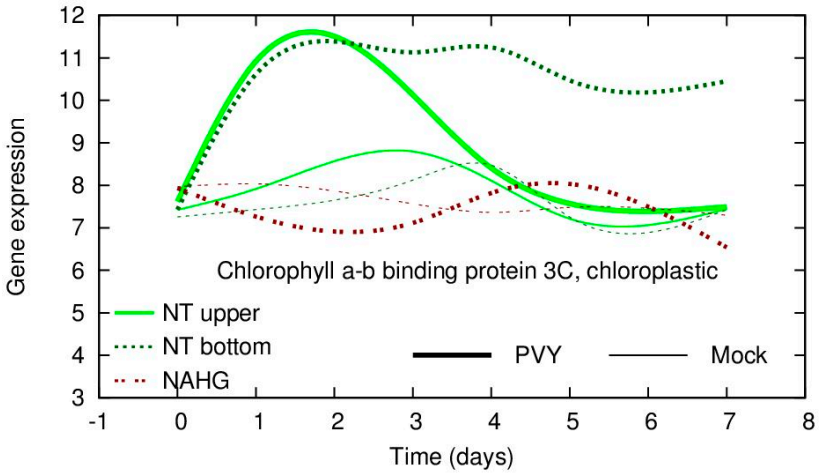


Figure 5. Average values for PVY infected samples (thick lines) and the mock samples (thin lines) for three types of potato leaves for the gene which is detected as relevant for NT bottom leaves in the interval day 0–1.

A figure that is much easier for interpretation can be obtained by the visualization of RMD values as illustrated in Figure 6.

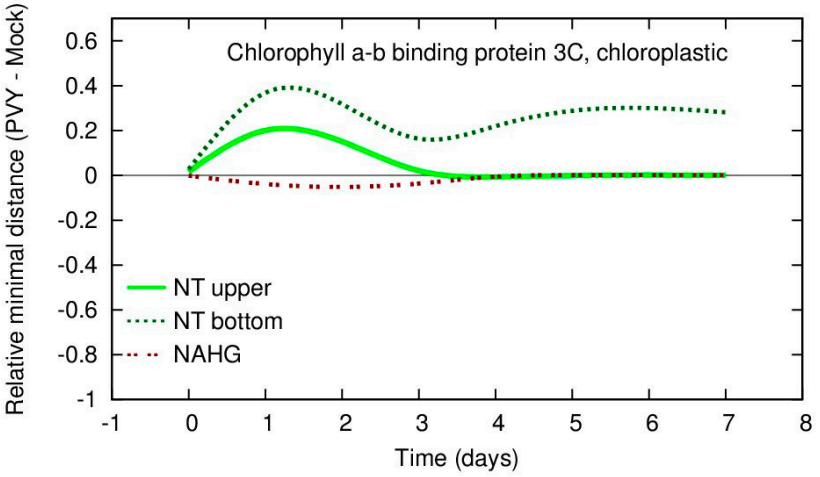


Figure 6. RMD values for three types of potato leaves for the same data as presented in Figure 5.

From Figure 6 it is very clear that the selected gene has a very large difference between the gene expression values of the PVY and mock samples for the NT upper and NT bottom potatoes leaves for days 1–3 and that this value has strongly increased from day 0. Additionally, the figure demonstrates that in the same period NAGH potato plants have slightly decreased values for this gene. This observation may be a trigger for expert evaluation of the differences between various potato types. By evaluating both Figures 5 and 6 at the same time it may be concluded that for PVY-infected samples,

the NT upper potato leaves have increased expression values of this gene in the period day 1–3 even more than the NT bottom leaves, but that this difference is not so significant in terms of RMD values. This observation suggests that the variability for PVY and mock data for this gene for the NT upper leaves is large. The result demonstrates the usefulness of visualization of both real gene expression values and their RMD measure for small sets of samples.

4. Discussion

High throughput gene expression profiling has emerged over the last decades as one of the most important and powerful approaches in life science research. Additionally, systematic characterization of temporal changes in mRNA levels under different conditions identifies genes relevant for a specific biological response. The task of detecting genes with statistically relevant properties in the setting with a very large number of genes and a small number of samples is a common experimental setting caused by limited sample/tissue access. The contribution of the work is in the definition of a novel measure for characterization of a difference between samples in two classes. When gene expression values have stabilized variance in data preprocessing then stratified data randomization can be used to estimate the statistical properties of this measure for genes that do not differ between target and control samples. Genes whose expression values are statistically significantly differently expressed when compared with control samples and that in some time point have statistically significantly changed expression values in the biological response sequence when compared with the previous time point are selected as specific for this biological response. The approach based on relative minimal distance is very simple and efficient in detecting significant genes with strong stringency. It can be applied to any number of target and control samples, but its application is particularly justified when the number of samples is very small.

The main drawback is that the proposed methodology can result in a very high false negative rate (type II error, a large number of genes that are not detected as significant but that are actually differentially expressed). Namely, a single measurement error may have a consequence that a highly significant gene has a very low minimal distance which can be even equal to zero. This means that the relevance of the uncovered genes in terms of their differential expression is statistically justified, but that we cannot be sure that the resulting set of relevant genes is complete. This fact must be taken into account when analysing the sets of genes detected as relevant.

The methodology enables that the false positive rate (type I error when a gene is detected as significant although it is not significant) can be easily controlled by changing the number of standard deviations in the non-critical region. For example, for the case with the Bonferroni correction for 37,865 genes it is enough to increase the non-critical region from 4.84 to 5.57 standard deviations in order to obtain a probability $p < 0.001$ instead of $p < 0.05$ that is used in Tables 2 and 3. Some of RMD values in the potato domain are more than ten standard deviations far from the mean values meaning that they are very statistically significant. Six out of seven genes whose functions are presented in Section 3 are statistically significant with $p < 0.001$.

Author Contributions: T.S. performed the data curation. D.G. implemented the methodology. K.G. and T.S. performed the validation. D.G. and T.S prepared the original draft. D.M., N.L, and K.G. performed the review and editing.

Funding: This work was financially supported by the Slovenian Research Agency (ARRS) grant HinLife: Analysis of Heterogeneous Information Networks for Knowledge Discovery in Life Sciences (J7-7303) and research programme Knowledge Technologies (P2-0103).

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Scholthof, K.-B.G.; Adkins, S.; Czosnek, H.; Palukaitis, P.; Jacquot, E.; Hohn, T.; Hohn, B.; Saunders, K.; Candresse, T.; Ahlquist, P.; et al. Top 10 plant viruses in molecular plant pathology. *Mol. Plant Pathol.* **2011**, *12*, 938–954. [CrossRef]
2. Kogovšek, P.; Ravnikar, M. Physiology of the Potato–Potato Virus Y Interaction. In *Progress in Botany: Vol. 74*; Lüttge, U., Beyschlag, W., Francis, D., Cushman, J., Eds.; Progress in Botany; Springer: Berlin/Heidelberg, Germany, 2013; pp. 101–133. ISBN 978-3-642-30967-0.
3. Singh, R.P.; Valkonen, J.P.T.; Gray, S.M.; Boonham, N.; Jones, R.A.C.; Kerlan, C.; Schubert, J. Discussion paper: The naming of Potato virus Y strains infecting potato. *Arch. Virol.* **2008**, *153*, 1–13. [CrossRef]
4. Baebler, Š.; Stare, K.; Kovač, M.; Blejec, A.; Prezelj, N.; Stare, T.; Kogovšek, P.; Pompe-Novak, M.; Rosahl, S.; Ravnikar, M.; et al. Dynamics of Responses in Compatible Potato—Potato virus Y Interaction Are Modulated by Salicylic Acid. *PLoS ONE* **2011**, *6*, e29009. [CrossRef]
5. Stare, T.; Ramšak, Ž.; Blejec, A.; Stare, K.; Turnšek, N.; Weckwerth, W.; Wienkoop, S.; Vodnik, D.; Gruden, K. Bimodal dynamics of primary metabolism-related responses in tolerant potato-Potato virus Y interaction. *BMC Genom.* **2015**, *16*, 716. [CrossRef]
6. Jovel, J.; Walker, M.; Sanfaçon, H. Salicylic acid-dependent restriction of Tomato ringspot virus spread in tobacco is accompanied by a hypersensitive response, local RNA silencing, and moderate systemic resistance. *Mol. Plant Microbe Interact.* **2011**, *24*, 706–718. [CrossRef]
7. Sánchez, G.; Gerhardt, N.; Siciliano, F.; Vojnov, A.; Malcuit, I.; Marano, M.R. Salicylic acid is involved in the Nb-mediated defense responses to Potato virus X in Solanum tuberosum. *Mol. Plant Microbe Interact.* **2010**, *23*, 394–405. [CrossRef]
8. Glazebrook, J. Contrasting mechanisms of defense against biotrophic and necrotrophic pathogens. *Annu. Rev. Phytopathol.* **2005**, *43*, 205–227. [CrossRef]
9. Little, T.J.; Shuker, D.M.; Colegrave, N.; Day, T.; Graham, A.L. The Coevolution of Virulence: Tolerance in Perspective. *PLoS Pathog.* **2010**, *6*, e1001006. [CrossRef]
10. Baebler, Š.; Witek, K.; Petek, M.; Stare, K.; Tušek-Žnidarič, M.; Pompe-Novak, M.; Renaut, J.; Szajko, K.; Strzelczyk-Żyta, D.; Marczewski, W.; et al. Salicylic acid is an indispensable component of the Ny-1 resistance-gene-mediated response against Potato virus Y infection in potato. *J. Exp. Bot.* **2014**, *65*, 1095–1109. [CrossRef]
11. Halim, V.A.; Vess, A.; Scheel, D.; Rosahl, S. The role of salicylic acid and jasmonic acid in pathogen defence. *Plant Biol.* **2006**, *8*, 307–313. [CrossRef]
12. Hejblum, B.P.; Skinner, J.; Thiébaut, R. Time-Course Gene Set Analysis for Longitudinal Gene Expression Data. *PLoS Comput. Biol.* **2015**, *11*, e1004310. [CrossRef]
13. Storey, J.D.; Xiao, W.; Leek, J.T.; Tompkins, R.G.; Davis, R.W. Significance analysis of time course microarray experiments. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 12837–12842. [CrossRef]
14. Berk, M.; Hemingway, C.; Levin, M.; Montana, G. Longitudinal Analysis of Gene Expression Profiles Using Functional Mixed-Effects Models. *Adv. Stat. Methods Anal. Large Data-Sets* **2013**, 57–67.
15. Guo, X.; Qi, H.; Verfaillie, C.M.; Pan, W. Statistical significance analysis of longitudinal gene expression data. *Bioinformatics* **2003**, *19*, 1628–1635. [CrossRef]
16. Hooton, J.W. Randomization tests: Statistics for experimenters. *Comput. Methods Prog. Biomed.* **1991**, *35*, 43–51. [CrossRef]
17. Kallio, A.; Vuokko, N.; Ojala, M.; Haiminen, N.; Mannila, H. Randomization techniques for assessing the significance of gene periodicity results. *BMC Bioinform.* **2011**, *12*, 330. [CrossRef]
18. Wang, X.; Tian, J. A gene selection method for cancer classification. *Comput. Math. Methods Med.* **2012**, *2012*, 586246. [CrossRef]
19. GEO Accession Viewer. Available online: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE58593 (accessed on 10 January 2019).
20. Kloosterman, B.; De Koeyer, D.; Griffiths, R.; Flinn, B.; Steuernagel, B.; Scholz, U.; Sonnewald, S.; Sonnewald, U.; Bryan, G.J.; Prat, S.; et al. Genes driving potato tuber initiation and growth: Identification based on transcriptional changes using the POCI array. *Funct. Integr. Genom.* **2008**, *8*, 329–340. [CrossRef]
21. Lopez-Romero, P. Agi4x44PreProcess. Available online: http://bioconductor.org/packages/Agi4x44PreProcess/ (accessed on 26 October 2018).

22. Smyth, G.K.; Michaud, J.; Scott, H.S. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* **2005**, *21*, 2067–2075. [CrossRef]
23. Lin, S.M.; Du, P.; Huber, W.; Kibbe, W.A. Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic Acids Res.* **2008**, *36*, e11. [CrossRef]
24. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B (Methodol.)* **1995**, *57*, 289–300. [CrossRef]
25. Thimm, O.; Bläsing, O.; Gibon, Y.; Nagel, A.; Meyer, S.; Krüger, P.; Selbig, J.; Müller, L.A.; Rhee, S.Y.; Stitt, M. MAPMAN: A user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* **2004**, *37*, 914–939. [CrossRef]
26. Rotter, A.; Usadel, B.; Baebler, S.; Stitt, M.; Gruden, K. Adaptation of the MapMan ontology to biotic stress responses: Application in solanaceous species. *Plant Methods* **2007**, *3*, 10. [CrossRef]