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.

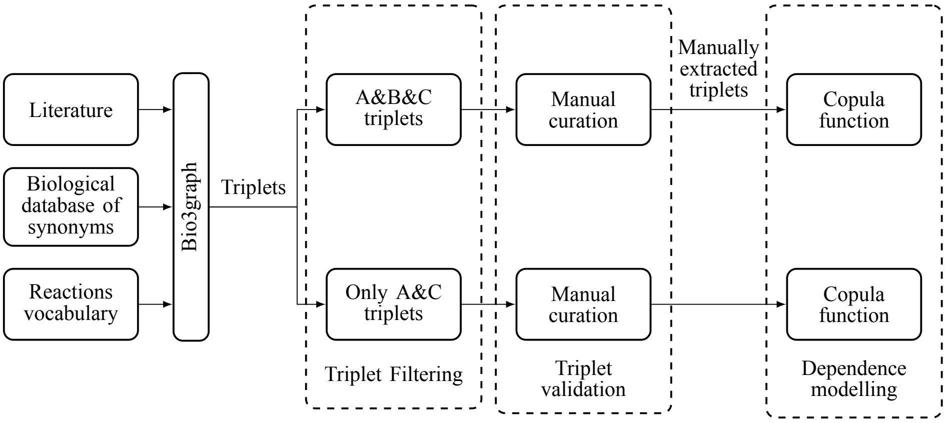
2.2 Copula functions

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):

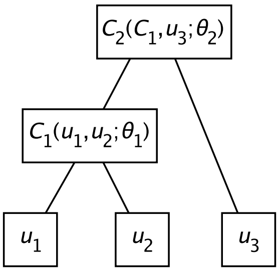
$$C_2(u_3, C_1(u_1, u_2)) \tag{3}$$

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:

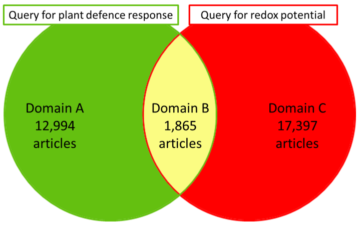
```
"arabidopsis thaliana"[All Fields] AND
( "defence"[All Fields] OR
"defense"[All Fields] OR
"ethylene"[All Fields] OR
"jasmonate"[All Fields] OR
"jasmonic acid"[All Fields] OR
"salicylate"[All Fields] OR
"salicylic acid"[All Fields] OR
"pathogen"[All Fields] OR
"virus"[All Fields])
```

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 $\Omega = \{1, 2, 3, \dots, N\}$. 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)

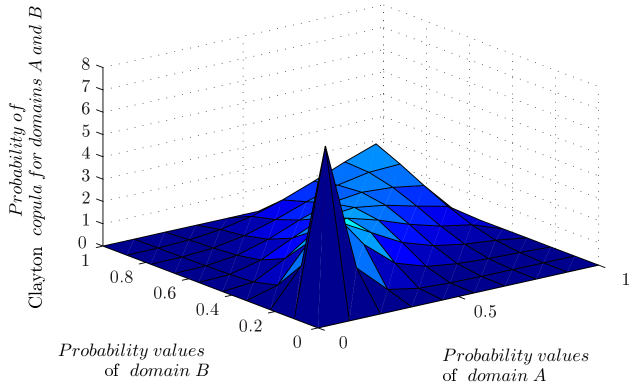


Figure 5 PDF for the Clayton copula for θ_2 (see online version for colours)

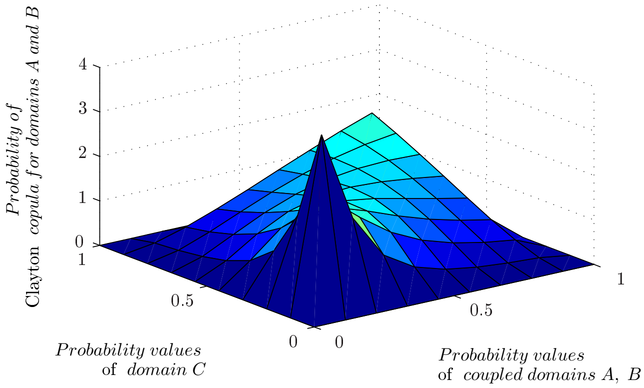


Figure 6 Probability density function for Clayton copula built on domains A and C as given in Table 3 (see online version for colours)

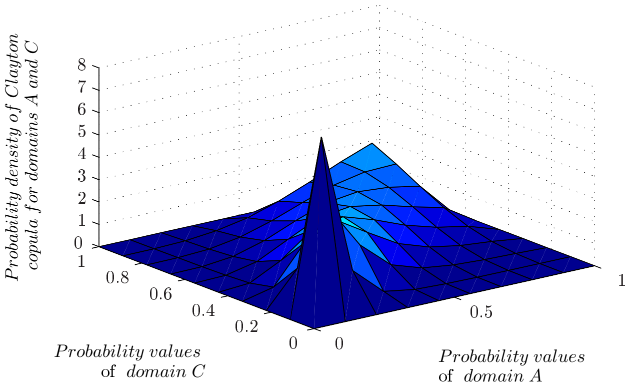
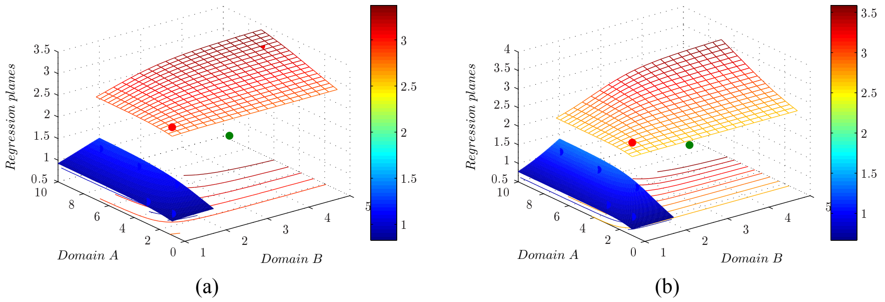


Figure 7 Predicting the values in domain C, (a) Clayton copula (b) Frank copula (see online version for colours)



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.