Incremental construction of biological networks by relation extraction from literature

Dragana Miljkovic[1,2], Vid Podpečan[1,2], Tjaša Stare[3], Igor Mozetič[1], Kristina Gruden[3], Nada Lavrač[1,2,4]

[1] Jožef Stefan Institute, Ljubljana, Slovenia
[2] Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
[3] National Institute of Biology, Ljubljana, Slovenia
[4] University of Nova Gorica, Nova Gorica, Slovenia

Abstract. This work focuses on automated incremental development of biological networks. The Bio3graph approach to information extraction from biological literature is extended with new features which allow for periodical updates of network structures using newly published scientific literature. The incremental approach is demonstrated on two use cases. First, a simple plant defence network with 37 components and 49 relations created manually by merging three existing structural models is extended in two incremental steps, yielding the final model with 183 relations. Second, a complex published network of defence response in *Arabidopsis thaliana*, containing 175 nodes and 524 relations, is incrementally updated with information extracted from recently published articles resulting in an enhanced network with 628 links. The results show that using the demonstrated incremental approach it is possible to automatically recognise new knowledge about the selected biological relations published in recent literature. The newly implemented Bio3graph extension offers an effective way of merging and visually representing the initial networks and the networks generated from texts thus enabling fast discovery of new relations which can potentially enhance the existing models.

Keywords: biological literature, biological networks, information extraction, plant defence

1 Introduction

At any level of detail, biological interactions can be modelled as networks [1]. For example, nodes can represent very different biological units ranging from atoms to individual organisms and the relations may describe atomic interactions in protein structure, molecular interactions or even species interactions. Network structures enable formal analysis of the modelled systems, mechanisms, and relations by using algorithms and methods from graph theory and other branches of discrete mathematics. The information obtained from such networks can be used in different ways to increase the understanding of biological systems. Several approaches have been recognised by Alm et al. [1]. For example, network topology can be used to propose hypotheses how the modelled systems are organised. Existing hypotheses can be tested and confirmed or rejected on the basis of the network data. Finally, existing open questions can be reformulated from a network perspective, for example, the role of the network structure in the evolutionary process and the role of evolution in shaping the network structure [1]. Studies in systems biology and graph theory have revealed that widely studied complex networks such as social networks, scientific co-authorships and the internet in fact share many features with certain biological networks, for example, the power-law node degree distribution, hierarchical modularity and small-world properties [1]. The architecture and physical properties of biological networks and networks in general are discussed in length by Wuchty et al. [2], Alm et al. [1] and Zhu et al. [3].

The structure of a biological network can be developed manually by the expert using *a priori* knowledge about entities and the relations between them. Up to date, several biological networks have been developed manually, such as the macrophage activation model developed by [4, 5], or terpenoid biosynthesis pathway [6]. On the other hand, biological networks can also be constructed automatically using computer methods to extract information from databases or textual sources. As the majority of curated human biological knowledge is produced in the form of scientific text, information extraction from the literature is an efficient way to automated construction and enhancement of biological networks. The construction of biological networks from the literature is recognised as an important task in the text-mining community and several systems for the extraction of network structures from scientific texts have been developed. Li et al. [7] and Skusa et al. [8] provide state-of-the-art reviews of available systems for biological network extraction from scientific literature and discuss aspects, phases and challenges of the topic. For example, Chilibot [9] is a web-based system which enables the search for relations by querying a certain number of entities. GeneWays [10] allows for the extraction, analysis, visualisation and integration of molecular pathway data but the system is not publicly available.

Biological networks which are the topic of this work belong to the area of plant defence modelling. In the research of plant response to stress stimuli obtaining data is particularly hard due to very long duration of experiments. Consequently, the developed models represent mostly subsets of the whole plant defence response mechanism. There were several attempts of modelling the defence mechanism of the model plant *Arabidopsis thaliana.* In the study of Su et al. [11] five gene logic networks for Arabidopsis under the normal condition and four external stimuli were constructed (short-day, long-day, bacterium and salt). One of the first attempts to model the subset of the plant defence by constructing the Boolean network was performed by Genoud et al. [12]. Devoto et al. [13] presented a similar approach to modelling one pathway of the plant defence using Boolean formalism. In the study of Miljkovic et al. [14] a complex network structure of defence response in *Arabidopsis thaliana* was developed.

The goal of this work is to extend the publicly available biological information extraction and network construction tool Bio3graph [14] to support incremental development of biological networks. The Bio3graph method is extended with functions which enable incremental development by network merging, removal of redundant relations, colour coding and network visualisation to present the newly extracted knowledge. We apply text processing components to two plant defence networks to show the potential of incremental knowledge upgrading for the mechanisms where kinetics data are sparse. The first network, which we refer to as the "simple network" was constructed from three small structural models published in the literature [15–17]. The second one, which we refer to as the "complex network" is a recently published complex plant defence network [14]. Throughout the paper both networks play the role of the "Initial network" which is extended with a "Triplet network" that is extracted automatically from the literature (see Fig. 1 where the scheme of the methodology is presented).

The rest of the paper is structured as follows. Section 2 outlines the procedures for incremental development of biological networks (literature retrieval, relation extraction in the form of *(component1, reaction, component2)* triplets and network operations) and describes the implementation. In Section 3 the results of incremental revisions of both plant defence networks are presented. The simple network was enhanced in two incremental steps whereas the second (which was only recently published) was updated once with the latest available publications. The updates of all networks are presented and discussed by means of graphical representations. The paper concludes by summarising the results of the experiments and suggesting improvements and directions for further work.

2 Materials and methods

This section presents the methods for literature retrieval and pre-processing, extraction of triplets from the texts, construction of the network structure from triplets and incremental updating of the network structure by merging, colour coding, and removal of redundant transitive relations. The implementation of the presented approach is also described.

2.1 Extraction of triplets and incremental revision of networks

The presented work on incremental revision and development of biological models is based on the existing Bio3graph [14] approach which allows for automated extraction of biological relations in the form of triplets from the literature. In the following we summarise the most important Bio3graph concepts and the proposed extensions which allow for incremental development of biological networks (see Fig. 1 for schematic overview of the methodology).

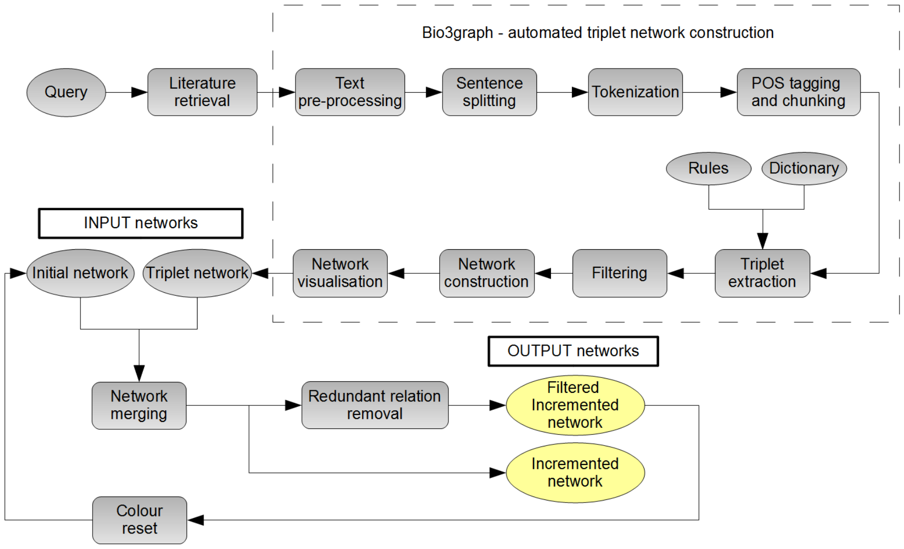

Fig. 1: Scheme of the methodology for incremental construction of biological networks using information extraction from literature.

The Bio3graph approach is essentially a workflow of processing components which extract triplets of the form *(component1, reaction, component2)* using natural language processing tools. The workflow consists of the following steps: (1) literature retrieval, (2) text pre-processing, (3) sentence splitting, (4) tokenization, part-of-speech (POS) tagging and chunking, (5) triplet extraction and filtering, and (6) network construction and visualisation. In addition, the incremental extension of Bio3graph implements (7) network merging, (8) redundant relation removal and (9) colour reset. We define the inputs to the incremental extension as follows (see Fig. 1). The existing network which is the subject of incremental enhancement is called the “Initial network” and the result of Bio3graph is called the “Triplet network”. The incremental extension of Bio3graph produces two outputs: “Incremented network”, a result of merging the Initial and the Triplet network, and “Filtered incremented network”, a result of redundant transitive relation removal from the “Incremented network”. In the following we discuss all steps of the approach in more details.

Literature retrieval. The collection of relevant scientific publications about various aspects of the selected case study topic (*Arabidopsis thaliana* defence response) was obtained from PubMed Central (PMC), a freely accessible online archive of biomedical and life sciences literature, which is developed and managed by National Library of Medicine's National Center for Biotechnology Information (NCBI). As of May 2013, PMC database hosts more than 2.7 million articles for which full text is available, either as HTML/XML or PDF or both. NCBI also provides the Entrez Programming Utilities (E-utilities), which enable programmatic access to the Entrez query and database system covering a variety of biomedical data, including nucleotide and protein sequences, gene records, three-dimensional molecular structures, and the biomedical literature [18]. The E-utilities are accessible via the HTTP protocol using GET and POST commands, and return the response in a structured XML document.

PMC also provides the PMC Open Access Subset (OA), a growing collection of publications which are made available under a Creative Commons or similar license. The OA subset is a valuable source of reviewed scientific publications which are readily 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 to download full-text XML as well as images, PDF, and supplementary data files for all articles in the OA subset.

To obtain sets of documents to increment networks in both use cases we have used the PMC Advanced Search Builder to construct the query which should cover as much literature as possible regarding the defence response signalling pathways in *Arabidopsis thaliana*. The query is as follows:

```
"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]
)
```

The query was used for both use cases only with the following differences. For the first use case with the simple model all publications regardless of the publication date were collected (the query was performed in May 2012). On the other hand, to increment the complex model in the second use case the earliest publication date was set to the latest date of any publication used by the authors of the model [14] (April 5th, 2011). Also, in the simple use case the keyword "virus" was excluded from the query and the source document set was not limited to the PMC OA subset in order to collect as much knowledge as possible (the most important non-OA publications were considered and extracted manually as PMC does not allow automated downloading of any publications outside of the OA subset).

For the simple network the query yielded 10,299 documents out of which some were available only as PDF and were left out. In order to time-stamp them we have collected pub-date tags and extracted the earliest available date (which in most cases corresponds to the classic publication date or the electronic publication). The final corpus, containing 10,207 documents, was divided in two datasets which were used in two incremental steps of the triplet extraction by Bio3graph.

In the case of the complex network, the query resulted in 2,988 full-text publications which were also subject to automated triplet extraction leading to an incremental enhancement of a complex, recently published network.

Text pre-processing. In this step basic text pre-processing is performed. For example, the improper formatting of chemical formulae (which results from the conversion of structured XML into unstructured plain text) is corrected, e.g., “H 2 O 2” is replaced by “H2O2”. Also, replacements such as “SA-treatment” into “SA treatment” are performed. Finally, the citation artifacts such as “et al.” or “et al;” are converted into “ETAL.” in order to avoid mismatching with the abbreviation for ethylene (et).

Sentence splitting. Sentence splitting is the process of breaking the homogenous text into sentences. For this task Bio3graph uses Natural Language Toolkit (NLTK) [19], a well known natural language processing library for Python. Sentence splitting is performed using NLTK’s recommended sentence tokenizer, currently the Punkt sentence splitter [20] which uses an unsupervised algorithm to build a model for abbreviation words, collocations, and words that start the sentences.

Tokenization, POS tagging and chunking. Tokenization is the process of breaking the input text into words, symbols and other meaningful elements called tokens. Tokenization is followed by POS tagging which assigns POS tags to words, i.e., it labels words as nouns, verbs, adjectives, etc. Finally, chunking is the process of segmenting and labelling multi-token sequences such as noun phrases (NP) or verb phrases (VP).

In Bio3graph all three functions are performed by the GENIA tagger [21] which offers POS tagging, chunking and named entity recognition from English texts. In our setting the GENIA tagger is preferred over its counterpart provided by the NLTK toolkit. The reason is that general purpose POS taggers typically do not perform well on biomedical text while the GENIA tagger is trained on the GENIA corpus and the PennBioIE corpus [22], and reportedly works well on various types of biomedical documents [21]. In Bio3graph the GENIA tagger is available through the developed interface which performs tokenization, POS tagging and chunking on a given sentence, and returns a list of tuples of the form

$$< word,\ base\,form,\ POS\,tag,\ chunk\,tag,\ named\,entity\,tag >$$

where chunk tags are in the widely used IOB format[5]. For example, a tree representation of the IOB chunk structure of a simple statement is shown in Fig. 2.

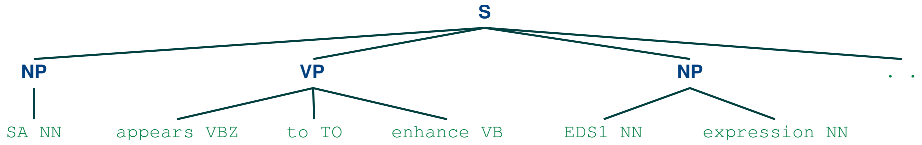

Fig. 2: A tree representation of the results of POS tagging and chunking of the sentence "SA appears to enhance EDS1 expression." using the GENIA tagger.

Triplet extraction and filtering. In this step of the Bio3graph workflow the aim is to extract triplets of the form *(component1, reaction, component2)*. We assume that the grammatical structure of the triplets is such that the *component1* and *component2* are part of the NPs while the reaction is a part of the VP. The triplet extraction procedure performs matching of (NP, VP, NP) patterns to a manually crafted vocabulary of components and reactions while also satisfying a number of rules.

We have composed two different vocabularies for the two use cases. Essentially, both share the reactions vocabulary but use different component vocabularies. The reactions vocabulary specifies various words (verbs) and phrases which can represent activation, binding and inhibition but also contains passive forms. Altogether, more than 150 different reaction terms are recognised without counting their numerous forms and synonyms (see Supporting Information 4 in [14]). Component vocabularies of both networks contain also numerous synonyms and short names (e.g., SA, Salicylate, 2-Hydroxybenzoic acid and o-Hydroxybenzoic acid are synonyms for salicylic acid).

A number of simple rules limit the number of spurious triplet patterns. First, the rules do not allow for patterns where the NPs are separated by more than one VP (on the other hand, they allow soft matching of multi-word reaction terms such that the VP and reaction phrase must overlap in at least one word). Second, hypothetical triplets are filtered out. This is accomplished by searching for words such as "possibly", "to determine", etc. in the sentence, and auxiliary verbs like "may", "can" and "would" in the VP. Third, mutant-related triplets are also discarded by recognising terms such as "mutant" and "line" in the NP. Fourth, triplets which are too general and refer to the whole pathway instead of some specific component are also not allowed. Finally, triplets where the first and the second component are the same, and triplets with a negation in the VPs are filtered out.

[5] According to the IOB tagging scheme each token is tagged with one of three special chunk tags, I (inside), O (outside), or B (begin). I and B tags are suffixed with the chunk type.

Network construction and visualisation. In the final step of the Bio3graph workflow, the extracted set of triplets is transformed into a network structure (a directed multigraph). Each triplet yields a set of nodes and an arc that points from the first component to the second component of the triplet. Additional information can be also assigned to nodes and arc, such as the sentence from which the triplet was extracted, the id of the source document, and the time related to the source document (e.g., publication date). The time attribute adds a temporal quality to the network structure, and allows for the analysis of development of the network structure through time.

Network merging. In order to allow for incremental updates of an existing model using Bio3graph (or any other biological network construction method) the existing model and newly extracted network have to be merged. The merging process produces a union of the networks and applies colour coding to relations in order to distinguish between known, new, and duplicate relations.

All biological networks discussed in this work (and biological networks in general) are directed edge-labelled graphs with several types of relations. Therefore, the data structure used for merging must support the most general type of graphs which is called a multigraph. A multigraph supports duplicate relations, relations of different types and cycles.

The merging procedure merges the input networks into a single network using the following colour coding: existing relations originating from the old network are coloured in black, newly discovered relations originating from the new network are coloured in red while the re-discovered existing relations originating from the new network are coloured in green. Other existing information about nodes and relations is also preserved during merging.

Redundant relation removal. Automated extraction of biological relations with Bio3graph can yield relations which may not appear in the existing model (the subject of incremental revision) but do not contain new biological knowledge. Such relations, which are known as transitive relations in graph theory, represent only a shortcut of a chain of biological relations. For example, the new relation *A activates C* does not represent new biological knowledge given the chain *A activates B activates C*.

In general, transitive relations can be removed by computing the transitive reduction of the directed network. Transitive reduction yields a new network on the same set of nodes with as few edges as possible to maintain the same reachability relation. For a finite, directed acyclic network the transitive reduction is a unique subnetwork which is also the minimum equivalent network. However, the transitive reduction of directed networks with cycles is not unique and is not necessarily a subnetwork. This means that the transitive reduction of general biological networks – which typically contain cycles – is not applicable as it may produce several equivalent networks and also introduce new relations.

For this reason, we have developed a procedure which does not exhibit the mentioned limitations. Given an existing network and a new network, the procedure evaluates all relations in the new network. For each relation in the new network the procedure tries to find a path in the existing network. If such a path exists, the new relation is transitive and thus redundant. If no such path exists, the new relation is not redundant as it represents a new piece of knowledge. It should be noted that we do not make any assumptions about the existing network and that each type of relation is considered separately, i.e., the path must contain only relations of the same type.

Fig. 3 shows an example of a transitive relation in a simple graph. The redundant transitive relation *v Activates x* is shown in grey. On the other hand, the relation *v Activates z* is not transitive as no alternative path consisting only of relations of the same type exists between *v* and *z*.

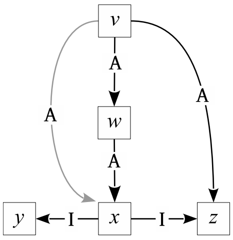

Fig. 3: An example of a redundant transitive relation in a simple graph. The relation *v Activates x*, shown in grey, is transitive. It does not contain new biological knowledge and is thus redundant.

Colour reset. The incremental revision of the Initial network with a Triplet network extracted from the literature can be used again in the next iteration (see Fig. 1). The only requirement is that the colours of relations are reset to the default colour (black) so that merging and colour coding can be performed correctly using a next Triplet network obtained by Bio3graph from a new set of documents.

2.2 Implementation

Our implementation of incremental development of biological networks is built as an extension of Bio3graph. Therefore, we only discuss the implementation of new components and the complete integrated solution as a scientific workflow as the implementation of Bio3graph is presented in length in [14].

Literature retrieval. We have implemented literature retrieval in the Python programming language using the ESearch and EFetch functions provided by PMC E-utilities. The implementation accepts the query constructed manually or by using the Advanced Search Builder and invokes ESearch to obtain the identifiers of the corresponding articles. The identifiers are then matched against the downloaded archives of the PMC OA subset and full-text XML files are extracted. Our XML parser, which is used to transform the XML files into plain text data, is set to ignore the following XML tags which do not contain relevant textual data and may contain unwanted special characters or words with excessive length (they can cause problems in some language processing components): *xref*, *table*, *graphic*, *ext-link*, *media*, and *inline-formula*.

Network merging. The network merging component was implemented using the NetworkX[6] Python library which can be natively integrated into the Bio3graph workflow in Orange4WS [23]. To maintain the compatibility with the Bio3graph network representation in Biomine's graph format we have also implemented a bidirectional transformation between the Biomine's [24] network format and NetworkX data structures which preserves all existing information concerning nodes and relations. For example, if the positions of the nodes in the visualisation canvas are available they will be preserved during merge which is essential for the efficient visual comparison of the networks.

Redundant relation removal. The discovery of transitive relations also relies on the NetworkX library. It is implemented as a separate component which accepts the existing and the new network and returns a list of redundant relations. In this way, the relations can be reported to the user, removed from the merged network or even marked with a different colour in a merged network to aid the visual evaluation of the network.

The procedure, described in Section 2, is implemented using the path discovery procedures available in the NetworkX library. The search for an existing path in the existing network is performed by generic function *has_path(G, source, target)* which is essentially instantiated to the bidirectional shortest path search which executes a breadth-first search from both the source and the target and returns a list of nodes in the path or an empty list if such path does not exist.

Colour reset. Reset of the colours of relations works by modifying the attributes of the relations which are stored in the NetworkX MultiDiGraph data structure. The implemented bidirectional transformation from this data structure to the Biomine's format can be used to export the structure and properties of the reset network into a portable text file.

The workflow. The proposed extension of Bio3graph was implemented as a scientific workflow in the same service-oriented data mining environment Orange4WS [23] where Bio3graph was developed and implemented. By utilising Orange4WS the following benefits were achieved. First, incremental revision and development is natively integrated with Bio3graph. Second, workflow-based implementation ensures repeatability of experiments and makes the modifications

[6] `http://networkx.github.io`

and extensions of the developed workflow easy. Finally, the workflow-based solution is shareable and can be used anywhere where Orange4WS is available.

The implementation of the incremental network development approach in the Orange4WS environment is shown in Fig. 4. The first part of the workflow implements Bio3graph (loading of documents, preprocessing and parsing, loading of vocabularies, triplet extraction, and network construction) while the second part implements incremental development (network merging, colour coding, removal of redundant transitive relations, and visualisation of incrementally constructed networks). It should be noted, however, that only one incremental step is composed in the workflow. Additional steps can be performed by repeating the two parts of the workflow: triplet extraction with Bio3graph followed by incremental revision.

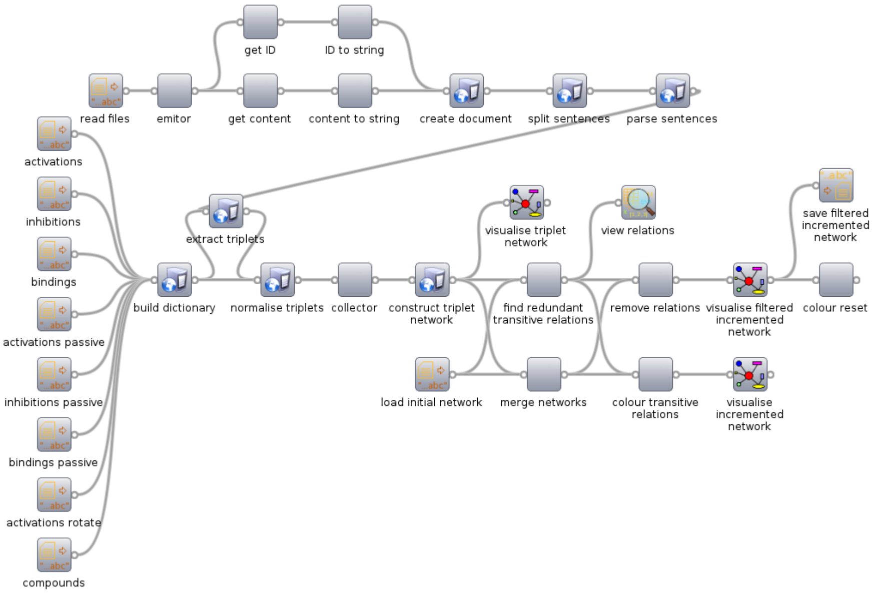

Fig. 4: A screenshot of the workflow implementing the proposed incremental revision of biological networks. The first part of the workflow implements the Bio3graph approach for automated triplet network extraction from biological literature while the second part implements incremental extension of biological networks.

The workflow works as follows. First, the dictionary has to be constructed as it is needed for the triplet extraction algorithm. This is accomplished by loading the dictionary files which are passed to the web service which constructs the dictionary structure. The parallel branch of the workflow is used to prepare the data. A collection of text files is sent to the *Emitor* component which simulates the for-loop by outputting the elements of the input list one by one. Each emitted document is passed to the web service which creates the document data structure. Each instance of this data structure is then forwarded to the sentence splitting component which is followed by POS tagging with the GENIA tagger. The document, tokenised and parsed, is sent to the triplet extraction web service which requires also the dictionary. The extracted triplets (if any) are subjected to the normalisation process where the names of the involved components and the reaction are replaced by the corresponding base names (for example, "influence accumulation" is replaced by "activate" and "SA" is substituted for "o-Hydroxybenzoic acid"). The extracted triplets from all documents are collected by the *Collector* component which closes the emulated for-loop. The Bio3graph part of the workflow concludes with the construction of a network from triplets, to which we refer as a Triplet network, and its visualisation.

The second part of the workflow, which performs incremental revision of network starts by loading an Initial network from a file which will be the subject of incremental enhancement. This model and the Triplet network are sent to the component which discovers and reports redundant transitive relations. In parallel, the networks are merged into an Incremented network which is colour coded marking differently the relations that belong solely to the Initial network, the ones in the networks' intersection and the new ones. The discovered redundant relations are then removed from the Incremented network and finally the Filtered Incremented network is visualised and saved. Alternatively, the redundant relations are not removed but coloured differently in the Incremented network, which is useful for a visual comparison. In the very last step of the workflow the colours of relations in the Filtered Incremented network are reset to black which makes the network ready for the next incremental revision which can be performed by providing a new set of documents and repeating the execution of the entire workflow.

3 Results and discussion

This section presents the results of two experiments in which two different biological networks available in the literature were incrementally extended. The first experiment is performed on a simple model which is a subset of the plant defence mechanism while the second experiment extends a recently published complex plant defence model structure.

3.1 Simple plant defence network

The Initial network in this experiment was constructed manually from the published figures (structural models) in scientific publications [15–17]. It was expanded in two incremental steps using Bio3graph and its incremental extension on a time-labelled collection of documents.

The Initial network. We have manually constructed the Initial network from structural models published in the scientific literature. Three schemata describing the salicylic acid (SA), jasmonic acid (JA) and ethylene (ET) pathways [15–

17] were selected and transformed into a directed networks with multiple relations (see Figs. 5, 6 and 7). To obtain the Initial network all three were merged into a single network which contains 37 nodes (biological components) and 49 links. The merged network is shown in Fig. 8.A.

Among all the represented components, SA, JA and ET are the most crucial for plant defence. The types of relations between the nodes are *activation* (abbreviated as A) and *inhibition* (abbreviated as I). The nature of interactions from the schemata was easily recognisable, and the transformation was accomplished with respect to these types. Too general components such as lipid, lesion, pathogen, etc. were not implemented in the Initial network. On the other hand, to prevent the loss of connections between components we have added several reaction products as nodes.

The Triplet network. Triplet extraction with Bio3graph requires a predefined vocabulary of components and reactions. We have developed the component vocabulary from the list of the Initial network nodes that represent biological components. Small compounds and proteins were considered. In addition, we have acquired the list of component synonyms from TAIR [25] and iHOP [26] sources. The vocabulary of reactions with reaction synonyms was used from Supporting Information S4 in [14]). Besides the activation and inhibition reaction types that exist in the Initial network, we have also taken into account the additional *binding* (abbreviated as B) reaction type.

The collection of full-text documents for triplet extraction with Bio3graph was divided into two sets according to the defined time point. We used the time point of November 2001, which is the earliest publication date of the three observed publications [15–17]. The first set of documents (published before November 2001) contains 1,714 publications while the second one contains 8,493 publications (published after November 2001). Using the two sets of documents two sets of triplets were obtained with the Bio3graph method. We refer to the first set as triplets before the time point and to the second set as the triplets after the time point.

Some of the extracted triplets appear in several sentences but we count only the number of unique triplets. We introduce the term *correct triplet* in the following way: if the triplet is a true positive (TP) in at least one sentence of the whole text corpus, it is considered to be a *correct triplet.* The extracted triplets were inspected manually and classified as correct or false positives (FP).

The summary of triplet extraction from documents before and after the time point is presented in Table 1. The Triplet network for the first incremental step is configured from the set of correct triplets before time point (Fig. 8.B) while the Triplet network for the second incremental improvement consists of the correct triplets after the time point.

First incremental step. The first incremental improvement of the Initial network is performed with the Triplet network consisting of correct triplets before time point of November 2001. The result of this enhancement is the Incremented

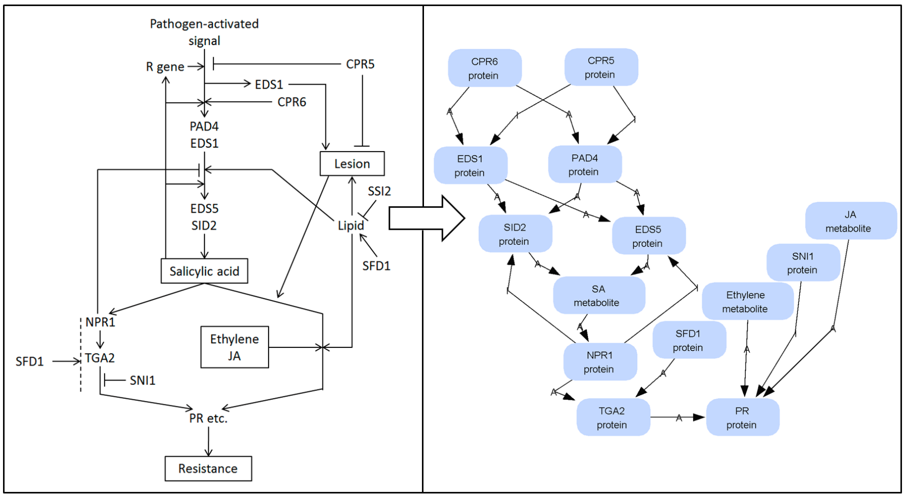

Fig. 5: Transformation of the SA model available in literature into a directed network with labelled edges. The model originates from the study of Shah [17].

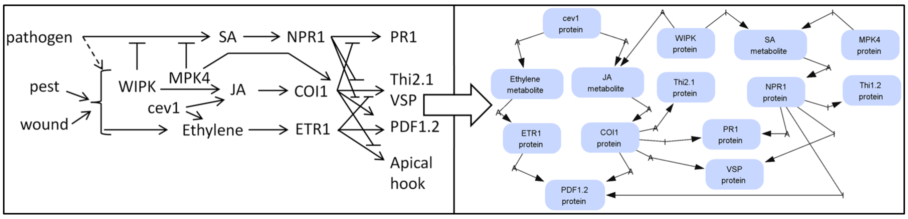

Fig. 6: Transformation of the crosstalk between SA, JA and ET pathways available in literature into a directed network with labelled edges. The model originates from the study of Turner et al. [16].

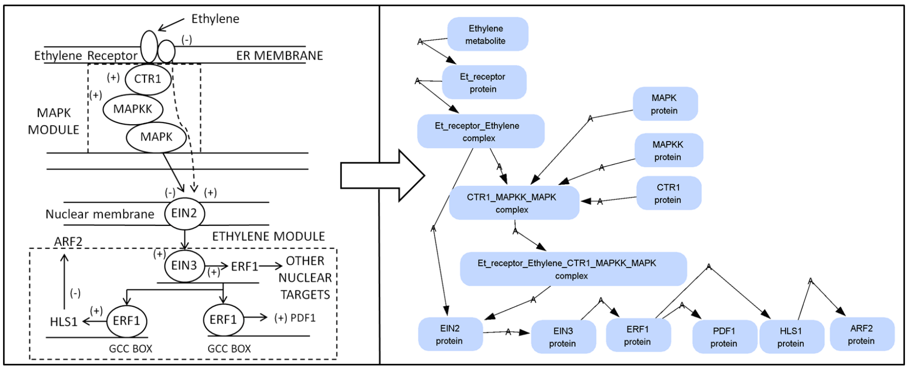

Fig. 7: Transformation of the ET model from literature into a directed network with labelled edges. The model originates from the work of Gonzalez-Garcia et al. [15].

Table 1: The summary of triplet extraction from documents before and after the time point for the simple plant defence network.

Reaction types	Triplets before time point			Triplets after time point		
	Total	Correct	FP	Total	Correct	FP
Activation	52	26	26	231	92	139
Inhibition	19	7	12	157	43	114
Binding	3	2	1	30	17	13
All reactions	74	35	39	418	152	266

network with 37 nodes and 78 relations shown in Fig. 8.C. Green, red and pink arcs represent the correct triplets discovered by Bio3graph from the biomedical texts already available at the time point, while the black arcs are the relations present in the Initial network. The summary of relation types in the network is show in Table 2.

In the Incremented network in Fig. 8.C the green arcs represent the intersection between the Initial and the Triplet network. The red and pink arcs represent the newly discovered relations not present in the Initial network. However, the arcs coloured in pink are transitive and thus redundant as they do not introduce new knowledge into the underlying biological model. The Initial network, however, can contain transitive relations but they do not interfere with our transitive relation discovery procedure as described in Section 2.1 as such relations are only searched for in the new Triplet network. The incremental extension of Bio3graph supports the removal of such relations. The result of this operation is shown in Fig. 8.D and represents the final Filtered Incremented network. The knowledge in this network which is most interesting for a domain expert is represented by red arcs (newly discovered biological relations from the literature).

Table 2: The summary of relations of the Incremented network shown in Fig. 8.C. The initial links originate only from the Initial network, while the intersection, new redundant and new links originate from the Triplet network. The intersection links are the common relations of the Initial and the Triplet network. The new redundant links are the transitive relations while the most interesting are the new links, which represent exclusively new relations discovered by the Bio3graph tool.

Reaction types	Initial links	Intersection links	New redundant links	New links
Activation	32	6	10	10
Inhibition	11	0	0	7
Binding	0	0	0	2
All reactions	43	6	10	19

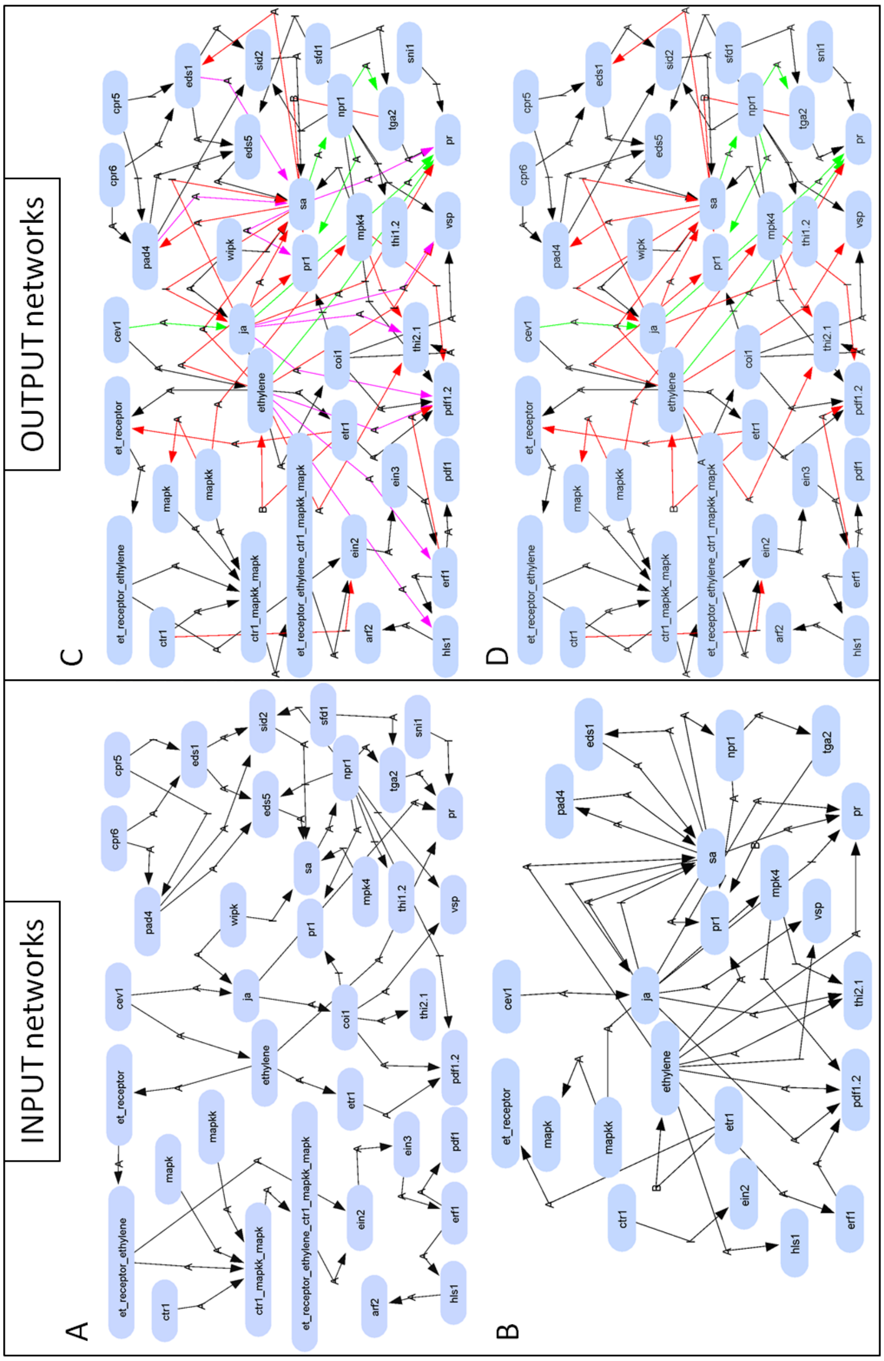

Fig. 8: The enhancement of the Initial network (A) with the correct triplets obtained from documents published before the time point (B). The left side represents the input networks for the incremental extension of Bio3graph while the right side represent the output networks. A) The Initial network created by merging the manually constructed three graphs from the literature shown in Figs. 5, 6 and 7. B) The Triplet network constructed from the correct triplets extracted with Bio3graph. C) The Incremented network obtained by merging the Initial and the Triplet network. The relations present only in the Initial network are coloured in black while the relations present also in the the Triplet are coloured in green, red or pink. Relations present in both Initial network and triplet nework are coloured in green, newly discovered relations are coloured in red while the newly discovered, redundant transitive relations are coloured in pink. D) The Filtered Incremented network obtained from the Incremented network by removing the redundant transitive relations.

Second incremental step. The second step incremental step is performed in an analogue way as the first. The input networks for the incremental extension of Bio3graph are as follows. The Initial network is the Filtered Incremented network shown in Fig. 8.D, but all of its relations are now marked as known (all arc are reset to the initial black colour). The Triplet network is constructed from the set of correct triplets after the time point of November 2001.

The result of merging of the two input networks is the Incremented network with 37 nodes and 183 relations shown in Fig. 9. The relations are summarised in Table 3. The removal of the redundant transitive relations which are shown in Fig. 9 in pink yields the final Filtered Incremented network which is also the final result of the first experiment.

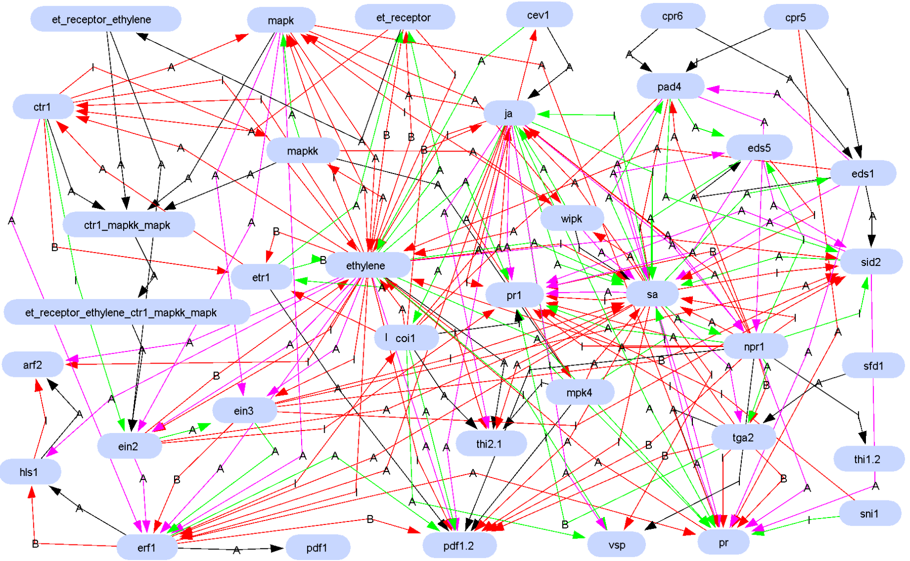

Fig. 9: The final Incremented network after two incremental steps. The new relations from the second set of correct triplets are shown in red. Pink arcs represent redundant transitive relations from the second set of correct triplets which are newly discovered.

The starting network, constructed from the three schemata describing the SA, JA and ET pathways initially contained 49 relations. Using Bio3graph in the course of two incremental steps we have confirmed 37 relations (shown as green in Fig. 9) which represent almost 75% of all relations present in the starting manually configured network. This shows that using Bio3graph as a starting point followed by incremental updates as new publications appear it is possible to confirm existing information but also propose new candidates (relations) for expert analysis. New candidates (shown as red arcs in Fig. 9) have the potential to generate new hypothesis in biological experiments where the functionality of the link is tested.

Table 3: The summary of relations of the Incremented network shown in Fig. 9. The initial links originate only from the Initial network, while the intersection, new redundant and new links originate from the Triplet network. The intersection links are the common relations of the Initial and the Triplet network. The new redundant links are the transitive relations while the most interesting are the new links, which represent exclusively new relations discovered by the Bio3graph tool.

Reaction types	Initial links	Intersection links	New redundant links	New links
Activation	22	26	32	34
Inhibition	9	9	1	33
Binding	0	2	0	15
All reactions	31	37	33	82

3.2 Complex plant defence network

To explore the potential of incrementally extending an existing, validated model using automated triplet extraction from literature we have selected a complex network structure to complement the small-scale experiment.

In the second experiment the Initial network was a complex plant defence network published in the study by Miljkovic et al. [14]. It contains in total 175 nodes and 524 relations. Since we did not introduce new components into the network, the vocabulary of components for the Bio3graph tool remained the same (Supporting Information S3 in [14]). Also, the vocabulary of reactions was the same as in the first experiment (also available as Supporting Information S4 in [14]).

The network, published as Supporting Information S10 in [14], was used as the Initial network and all arcs were reset to black colour. The Triplet network was constructed from the correct triplets extracted with Bio3graph from the set of 2,988 publications which were published after the latest publication used by the authors of [14] in the construction of the complex plant defence network. Manual validation of 399 unique triplets resulted in a set of 156 correct triplets. The Initial and the Triplet network were merged into the Incremented network (the summary of the relations is shown in Table 4). The evaluation of the newly discovered relations reveals that they mostly represent cross-talk connection between the SA, JA and ET pathways.

While exploring the new links (red arcs) in the Incremented network, we have observed an interesting pattern related to the discovery of binding relations. Most of the 13 new binding relations connect the components that are already connected either through activation or inhibition reaction type. The new links provide an additional explanation that first these components physically bind and then the activation or inhibition occurs.

Among the newly discovered links, biologically very interesting is *(ros, inhibits, npr1)*. Earlier studies reveal that ROS, more specifically H2O2, and SA together work as a self-amplifying system [27, 28]. However, after consulting the publication from where the triplet was extracted [29] we have found out that the new results in the study of Peleg et al. [30] indicate the presence of the negative regulation of NPR1 transport by H2O2.

In addition, newly discovered triplets that are biologically interesting are *(myc2, inhibits, b-chi)* and *(myc2, inhibits, pdf1.2)* which were extracted from [31]. Both links are extracted from the same sentence: "MYC2 is a negative regulator of the JA-responsive pathogen defense genes PDF1.2 and B-CHI." In the Initial network the relation between MYC2 and b-CHI components already exists: *(myc2, activates, b-chi)*. It was acquired manually by the authors of the network [14]. The discovery of the new link of a contradictory relation type indicates necessity of further exploration of the relation between MYC2 and b-CHI components. The second link *(myc2, inhibits, pdf1.2)* is also biologically interesting as it represents a cross-talk connection between JA and ET pathway where the component of the JA pathway has diminishing influence of the product of the ET pathway.

For the final evaluation of the network structure, one should keep in mind that most of the automatically extracted relations can be considered as "indirect" and that intermediate molecules participating in the network can be discovered by thorough inspection of the corresponding sentences or by performing additional wet-lab experiments.

Table 4: The summary of triplet extraction from biological texts for the complex plant defence network. The initial links originate only from the Initial network, while the intersection, new redundant and new links originate from the Triplet network. The intersection links are the common relations of the Initial and the Triplet network. The new redundant links are the transitive relations while the most interesting are the new links, which represent exclusively new relations discovered by the Bio3graph tool.

Reaction types	Initial links	Intersection links	New redundant links	New links
Activation	279	43	47	26
Inhibition	100	6	2	16
Binding	48	3	0	13
Produces	45	0	0	0
All reactions	472	52	49	55

4 Conclusion

This paper presents an approach to incremental development of biological networks by extending the existing tool Bio3graph with new components that perform literature retrieval from the PMC Open Access Subset and incremental upgrading of networks. The developed literature retrieval procedures enable easy access to the freely available articles by integrating E-utilities and parsing of XML data. The extended Bio3graph tool provides efficient tracking and visualisation of new knowledge obtained from biological literature. By applying the triplet extraction incrementally on time-labelled data one can follow the development of knowledge about certain biological phenomena and discover new relations which can potentially enhance already developed models (networks). Note also that according to the user's preferences more than one time point can be defined. For example, if the overall goal is to inspect a fine-grained development of the starting model, it is recommended to set as many time points as needed so that one batch of newly discovered relations does not contain more than a few relations. Furthermore, the incremental extension offers the removal of transitive relations which are redundant with respect to the existing network.

We have applied the extended Bio3graph method to a time-labelled collection of biomedical documents obtained from the PMC database in order to incrementally enrich two different networks. The first network has a simple structure and is configured from three published structural models. The network is enhanced throughout two phases which demonstrate the incremental approach. The second network is a recently published complex plant defence network. By extending this complex structure the experts have detected several interesting links among the newly discovered relations that might be subject to further experimental validation, e.g., the link between MYC2 and b-CHI components which contradicts previously published results.

In the future we plan to include the GENIA sentence splitter [32] which is trained on the GENIA corpus[7] [33] and employs a classification model based on maximum entropy modelling. Moreover, we plan to improve the triplet extraction by using fast deep parsing instead of chunking, and fine tune the rules for triplet extraction and filtering. The current implementation of Bio3graph discovers new relations, but does not enable automated discovery of new components as it employs a manually constructed vocabulary. To further evolve the network structure, new components could be added to the vocabulary to find additional relations. We plan to implement named entity recognition and automatic discovery of synonyms which will enable automated construction of the components vocabulary.

We expect that the extended version of the Bio3graph tool will assist the construction and enhancement of network structures that model other biological mechanisms. The results show that publicly available sources of biomedical literature, such as PMC database, offer a good starting point for computer-assisted development of plant defence models, and that approaches such as the presented incremental method can contribute to the discovery of potentially interesting relations. The obtained results show the potential of the developed method but also indicate the need for further development to improve the accuracy and utility of information extraction.

[7] GENIA corpus is a semantically annotated corpus for bio-textmining.

Acknowledgment

This work was supported by the Slovenian Research Agency grants P4 0165, J4-2228, J4-4165, P2-0103, AD Futura scholarship and FP7 project MUSE (Machine Understanding for interactive StorytElling) under the Grant Agreement No. 296703.

References

1. Alm E, Arkin AP. Biological networks. Current opinion in structural biology. 2003 Apr;13(2):193–202.
2. Wuchty S, Ravasz E, Barabási AL. The Architecture of Biological Networks. In: Deisboeck TS, Kresh JY, editors. Complex Systems Science in Biomedicine. Topics in Biomedical Engineering International Book Series. Springer US; 2006. p. 165–181.
3. Zhu X, Gerstein M, Snyder M. Getting connected: analysis and principles of biological networks. Genes & Development. 2007 May;21(9):1010–1024.
4. Raza S, Robertson KA, Lacaze PA, Page D, Enright AJ, Ghazal P, et al. A logic-based diagram of signalling pathways central to macrophage activation. BMC Systems Biology. 2008;2.
5. Raza S, McDerment N, Lacaze PA, Robertson K, Watterson S, Chen Y, et al. Construction of a large scale integrated map of macrophage pathogen recognition and effector systems. BMC Systems Biology. 2010;4.
6. Hawari AH, Hussein ZAM. Simulation of a Petri net-based Model of the Terpenoid Biosynthetic Pathway. BMC Bioinformatics. 2010 Feb;11(1):83+.
7. Li C, Liakata M, Rebholz-Schuhmann D. Biological network extraction from scientific literature: state of the art and challenges. Briefings in Bioinformatics. 2013;.
8. Skusa A, Rüegg A, Köhler J. Extraction of biological interaction networks from scientific literature. Brief Bioinform. 2005 Sep;6(3):263–276.
9. Chen H, Sharp BM. Content-rich biological network constructed by mining PubMed abstracts. Bmc Bioinformatics. 2004;5. Chen, H Sharp, BM.
10. Rzhetsky A, Iossifov I, Koike T, Krauthammer M, Kra P, Morris M, et al. GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. Journal of Biomedical Informatics. 2004;37(1):43–53.
11. Su Y, Wang S, Li E, Song T, Yu H, Meng D. Analysis of Gene Logic Networks for Arabidopsis. Current Bioinformatics. 2013;8(2):244–252.
12. Genoud T, Santa Cruz MBT, Metraux JP. Numeric simulation of plant signaling networks. Plant Physiology. 2001;126(4):1430–1437.
13. Devoto A, Turner JG. Jasmonate-regulated Arabidopsis stress signalling network. Physiologia Plantarum. 2005;123(2):161–172.
14. Miljkovic D, Stare T, Mozetič I, Podpečan V, Petek M, Witek K, et al. Signalling Network Construction for Modelling Plant Defence Response. PLoS ONE. 2012 12;7(12):e51822.
15. Gonzalez-Garcia JS, Diaz J. Information theory and the ethylene genetic network. Plant Signal Behav. 2011 Oct;6(10):1483–1498.
16. Turner JG, Ellis C, Devoto A. The Jasmonate Signal Pathway. The Plant Cell Online. 2002;14(suppl 1):S153–S164.
17. Shah J. The salicylic acid loop in plant defense. Curr Opin Plant Biol. 2003 Aug;6(4):365–371.

18. Sayers E. A General Introduction to the E-utilities. In: Entrez Programming Utilities Help. Bethesda, Maryland, US: National Center for Biotechnology Information; 2011. .
19. Bird S, Klein E, Loper E. Natural language processing with Python. O'Reilly; 2009.
20. Kiss T, Strunk J. Unsupervised Multilingual Sentence Boundary Detection. Comput Linguist. 2006 dec;32(4):485–525.
21. Tsuruoka Y, Tateishi Y, Kim JD, Ohta T, McNaught J, Ananiadou S, et al. Developing a Robust Part-of-Speech Tagger for Biomedical Text. In: Bozanis P, Houstis EN, editors. Advances in Informatics. vol. 3746. Berlin, Heidelberg: Springer Berlin Heidelberg; 2005. p. 382–392.
22. And SK, Kulick S, Bies A, Liberman M, Mark M, Winters S, et al. Integrated Annotation for Biomedical Information Extraction. In: Proceedings of the HLT/-NAACL 2004 Workshop: Biolink 2004; 2004. .
23. Podpečan V, Zemenova M, Lavrač N. Orange4WS Environment for Service-Oriented Data Mining. Computer Journal. 2012;55(1):82–98.
24. Eronen L, Toivonen H. Biomine: predicting links between biological entities using network models of heterogeneous databases. BMC Bioinformatics. 2012;13:119.
25. Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, et al. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Research. 2008;36:D1009–D1014.
26. Hoffmann R, Valencia A. A gene network for navigating the literature. Nature Genetics. 2004;36(7):664–664.
27. Leon J, Lawton MA, Raskin I. Hydrogen Peroxide Stimulates Salicylic Acid Biosynthesis in Tobacco. Plant Physiol. 1995 Aug;108(4):1673–1678.
28. Rao MV, Paliyath G, Ormrod DP, Murr DP, Watkins CB. Influence of salicylic acid on H2O2 production, oxidative stress, and H2O2-metabolizing enzymes. Salicylic acid-mediated oxidative damage requires H2O2. Plant Physiol. 1997 Sep;115(1):137–149.
29. Petrov VD, Van Breusegem F. Hydrogen peroxide-a central hub for information flow in plant cells. AoB Plants. 2012;2012:pls014.
30. Peleg-Grossman S, Melamed-Book N, Cohen G, Levine A. Cytoplasmic H2O2 prevents translocation of NPR1 to the nucleus and inhibits the induction of PR genes in Arabidopsis. Plant Signal Behav. 2010 Nov;5(11):1401–1406.
31. Lu X, Jiang W, Zhang L, Zhang F, Zhang F, Shen Q, et al. *AaERF1* Positively Regulates the Resistance to *Botrytis cinerea* in *Artemisia annua*. PLoS ONE. 2013 02;8(2):e57657.
32. Saetre R, Yoshida K, Yakushiji A, Miyao Y, Matsubayashi Y, Ohta T. AKANE System: Protein-Protein Interaction Pairs in the BioCreAtIvE2 Challenge, PPI-IPS subtask. In: Hirschman L, Krallinger M, Valencia A, editors. Proceedings of the Second BioCreative Challenge Workshop; 2007. .
33. Kim JD, Ohta T, Tateisi Y, Tsujii J. GENIA corpus–semantically annotated corpus for bio-textmining. Bioinformatics. 2003;19 Suppl 1:i180–182.