Incremental revision of biological networks from texts

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

[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 automatic extraction of relations between biological components from the literature to support incremental development of biological models. The incremental approach is illustrated by automatic expansion of a partial plant defence response network with 36 components and 50 relations, which was created manually by merging three existing structural models. Two incremental steps of automated extraction with the Bio3graph tool yielded the final model with 36 components and 237 relations. The results show that the existing biological networks can be considerably extended by mining publicly available biomedical articles, and that numerous valid relations can be extracted automatically.

Keywords: information extraction, biological literature, biological networks

1 Introduction

Systems biology approach has demonstrated its success in modelling complex biological processes [1]. However, one needs to understand first the network structure before analysing its dynamics. There are various ways to represent a network topology, including the directed graphs formalism as used in the Systems Biology Graphical Notation by Le Novère et al. [2] or the modified EPN (mEPN) scheme proposed by Raza et al. [3]. Depending on how thoroughly the biological mechanism was thoroughly studied, different information sources can be used to construct its network. This includes pathway databases such as the KEGG Pathway [4], Reactome [5] and BioCyc [6], integrated knowledge sources such as ONDEX [7] and Biomine [8], and the scientific literature itself. Taking into account that the majority of human biological knowledge is deposited in the form of scientific articles, retrieving information from the literature can considerably enhance the structure of biological networks.

Scientific literature can be inspected manually or processed by using automated tools for literature mining. Numerous biological models were manually

constructed based on an in-depth literature survey, such as the macrophage activation model developed by Raza et al.[3, 9]. On the other hand, state-of-the-art technologies enable information extraction from scientific texts in an automated way by means of text processing techniques such as co-occurence discovery, natural language processing (NLP) and text mining (see e.g., the research advances of the emerging bioNLP[5] community).

Similar to our work which involves the extraction of a set of (component1, reaction, component2) triplets from biology texts, several existing NLP tools enable the extraction of interactions between the components (e.g., see the review by Ananiadou et al. [10]). Examples of rule-based NLP approaches (for the NLP approaches see the review by Cohen et al. [11]) include GeneWays [12], Chilibot [13] and the approach proposed by Ono et al. [14]. Combined methods, including co-occurrence-based approaches, such as the one developed by Suiseki et al. [15] and upgraded in the BioRAT system by Corney et al. [16], are less appropriate for systems biology as the information retrieved is partial and can therefore not be directly transformed into a graph structure used for signalling network modelling. In most systems, the information is retrieved only from abstracts; an exception is BioRAT which can process full texts, albeit using a quite general vocabulary [16].

The goal of our work is to extend our publicly available biological information extraction tool Bio3graph[6] [17] to support incremental development of biological networks by means of semi-automatic extraction of triplets. The results of Bio3graph and incremental updates of the biological networks are presented by using different colour schemes in the network visualisation software to emphasise the newly extracted knowledge. Three already published models related to the plant defence response [20–22] are selected. These models are transformed and merged into a single edge-labelled directed network to which we refer as the initial manual network in the rest of the paper. In order to analyse the incremental development of the model a time point is introduced which corresponds to the earliest publication date of the three publications. This time point is used to mark the knowledge available prior to the first publication and the knowledge available from the time point onwards.

The rest of the paper is structured as follows. Section 2 discusses data acquisition and briefly outlines the triplet extraction procedure. Performance of triplet extraction is also presented. In Section 3 the transformation of three pathway schemata into a directed network is presented first. Two incremental updates of the network using automated triplet extraction from the literature are discussed next using also graphical representations of the networks. The paper concludes by summarising the results and pointing out possible improvements and directions for further work.

[5] See `http://www.bionlp.org` for BioNLP tools and research advances.
[6] `http://ropot.ijs.si/bio3graph`

2 Materials and methods

This section presents the methods for data acquisition, extraction of triplets from literature and construction of the network structure. The triplet extraction method Bio3graph, which forms the basis of biological network extraction from text, is briefly outlined (we refer the reader to [17] for a detailed description).

2.1 Data acquisition

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). PubMed Central provides E-utilities which enable programmatic access to the Open Access Subset of PubMed Central. E-utilies are accessible via the HTTP protocol using GET and POST commands, and return the response in a structured XML document. We have implemented data acquisition in the Python programming language using the ESearch and EFetch functions provided by E-utilities. First, we formulated a general database query which should cover as much literature as possible regarding signalling pathways for defence response in Arabidopsis thaliana. The query is as follows:

```
arabidopsis thaliana AND (defence OR defence OR ethylene
OR jasmonate OR jasmonic acid OR pathogen OR salicylate OR
salicylic acid)
```

Listing 1: PubMed query.

The query yielded 10,299 documents[7] out of which 1,100 were available only as pdf and were left out of the constructed document corpus. Our XML parser, which was used to transform the EFetch XML responses into plain text data, was set to ignore the following XML tags which do not contain relevant textual data: *xref*, *table*, *graphic*, *ext-link*, *media*, and *inline-formula*. In order to time-stamp the documents we have collected pub-date tags and extracted the earliest available date (which in most cases corresponds to the classic publication date or electronic publication). The distribution of all retrieved document according to their earliest known date is shown in Fig. 1. The final corpus contained 9,157 documents which were then used in the automated extraction of triplets and construction of network structures.

2.2 Extraction of triplets and network construction

We have implemented a pipeline for automated construction of network structures representing signalling pathways and interactions between components,

[7] As only a certain part of PubMed document database is accessible through E-utilities due to copyright restrictions (bulk download and crawlers are not allowed) we have manually saved the corresponding documents as HTML and transformed them into plain text.

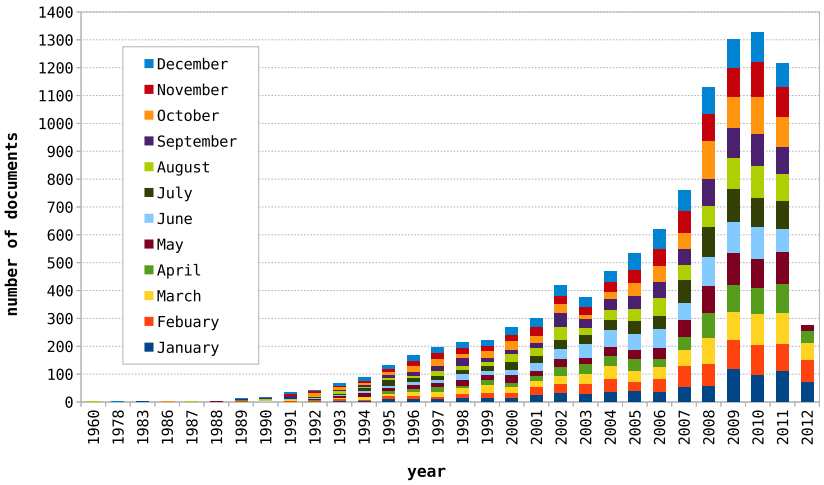

Fig. 1: The distribution of documents obtained by querying PubMed Central with the query from Listing 1 according to the earliest mentioned publication date. In total 10,207 documents are shown. Note that the column for 2012 is not complete as the query was performed in May 2012.

which is based on extraction of triplets of the form (component1, reaction, component2) using natural language processing tools. The pipeline consists of the following steps: (1) data retrieval, (2) text filtering, (3) sentence splitting, (4) tokenization, POS tagging and chunking, (5) triplet extraction and filtering, and (7) network construction. For a detailed description of the procedure, applied triplet extraction rules and implementation details we refer the reader to [17].

2.3 Performance of the triplet extraction algorithm

We have evaluated the presented triplet extraction method on 50 manually annotated full-text articles and obtained the overall (all three reaction types) precision and recall of 42.6% and 62.3% [17], which is comparable to other systems [16] for extracting relations on full-text documents. The annotated corpus and evaluation details are available as Supplement Information S6 in [17].

3 Results and discussion

To illustrate the incremental development of a biological network we have manually constructed an initial network representing subsets of plant defence mechanism. For this purpose, three schemata describing the salicylic acid (SA), jasmonic acid (JA) and ethylene (ET) pathways were selected from the scientific publications [20–22] and transformed into a directed network (see Fig. 2).

Too general components such as lipid, lesion, pathogen, etc. were not implemented in the directed network. On the other hand, to prevent the loss of

connections between components we have added several reaction products as nodes. A result of this transformation is an edge-labelled directed network, visualised with Biomine visualisation engine [8], that contains 36 nodes (biological components) and 50 vertices representing the relations between them (see Fig. 2). Among these nodes, the SA, JA and ET are the most crucial for the plant defence response. The types of relations are activation (abbreviated as A) and inhibition (abbreviated as I). The nature of the interactions from the schemata was easily recognisable, and the transformation was accomplished with respect to these types.

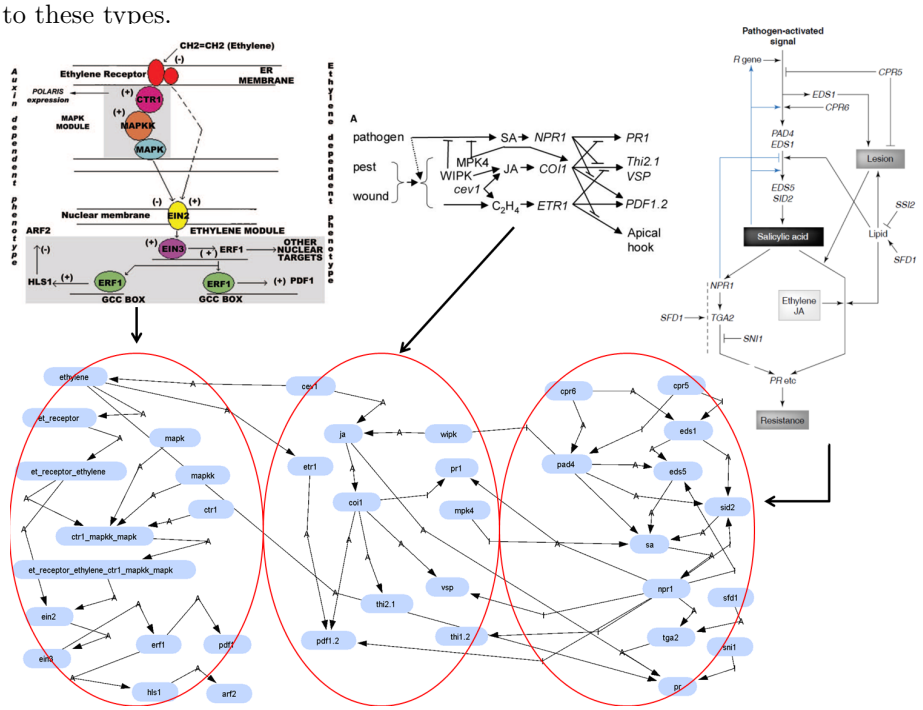

Fig. 2: Transformation of the three models (A, B, C) available in the literature into a directed network with labelled edges. Models A, B, and C originate from [20–22].

The manually constructed initial network provided the basis for the complementary automatic triplet extraction from biomedical literature. As a result, a vocabulary of components was developed from the list of the network nodes that represent biological components. For this vocabulary only single molecules, such as small compounds or proteins but not biological complexes were considered. In addition, we have acquired the list of component synonyms from TAIR [18] and iHOP [19] sources. The vocabulary of reactions was also developed using reaction synonyms (see Supporting Information S4 in [17]. Apart from the activation and inhibition reaction types that exist in the initial manual model, we have taken into account an additional binding reaction type.

Incremental development of the network structure was performed using the time point of November 2001, the earliest publication date of the three observed

Table 1: The summary of extracted triplets from the biological texts before and after the time point.

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

publications [20–22] (see Fig. 2B). Using this time point, two sets of triplets were obtained.

Since some of the extracted triplets appear in several sentences, we have defined 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 validated as correct or false positives (FP). The summary of the first set of extracted triplets before and after the time point is presented in Table 1.

After obtaining the initial manual model, the first set of triplets before the time point was inspected manually. The enhancement of the manual model topology with the correct triplets resulted in an extended network structure shown in Fig. 3. Red arcs represent the connections discovered automatically from the biomedical texts already available at the time point. This means that the underlying knowledge about the related components and reactions already existed at that time, and that these automatically discovered connections by the triplet extraction algorithm enhances and speeds up the process of understanding certain functions of plant defence response.

Following the enrichment of the network structure with the new relations from the triplet set before the time point, the next incremental step is performed. In this phase, the initial manual network together with the newly discovered relations from the first set of triplets before the time point is considered to be the initial network structure (all the arcs of the graph in Fig. 3 are now black).

The network constructed from the second set of triples, obtained from documents published after the time point, is then added. The final network structure is shown in Fig. 4. Note, however, that according to 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 initial 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.

The incremental development of the network structure provides an easily accessible and verifiable source of knowledge updates of a given biological model. This approach can be particularly useful when the components of particular pathways are known while the crosstalk connections between them remain un-

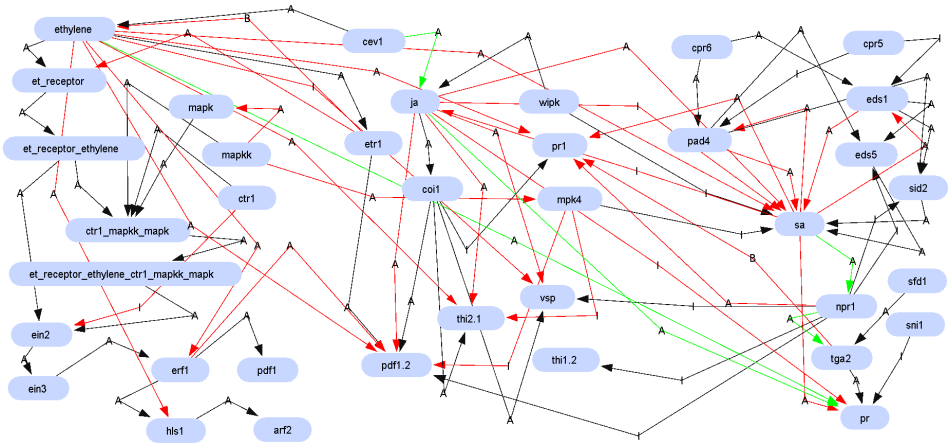

Fig. 3: The enhancement of the manual model topology with the correct triplets obtained from documents published before the time point. Black arcs originate from the manual model while red arcs represent newly discovered relations (green arcs are present in both sets).

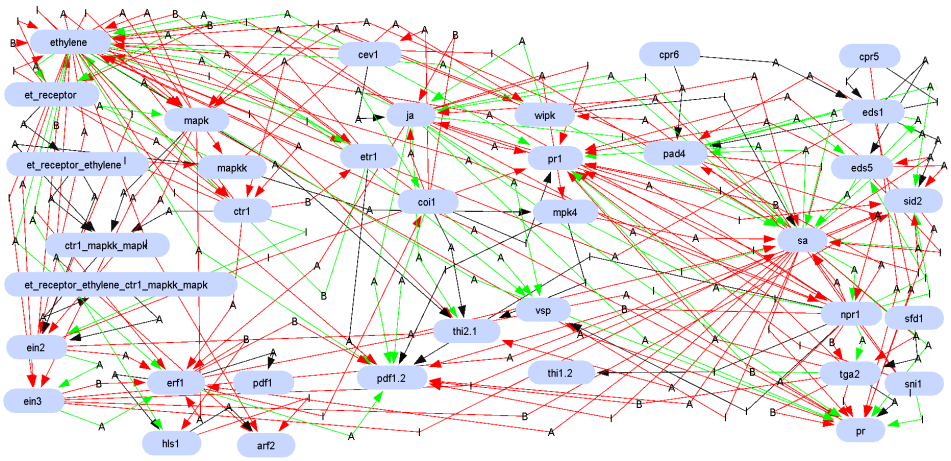

Fig. 4: The final network structure which integrates the manual model topology, triplets obtained from documents published before the time point, and triplets obtained from documents published after the time point. Note that in this figure only the last set of triplets is represented as an incremental update (red arcs) for an easier overview.

clear. For the final evaluation of the network topology, 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 in a thorough inspection of the corresponding sentences.

Most of these relations indicate a cross-talk between the sub-pathways or a feed-back regulation of the crucial components in the model. However, one should keep in mind that the procedure we propose discovers new relations and not new components. To further evolve the network topology, new components could be added to the vocabulary to find additional relations. However, it is possible to upgrade our triplet extraction tool Bio3graph with the additional feature that would automatically discover new biological components (named entity recognition). Implementation of this feature is planned in our future work.

4 Conclusion

In this work we presented an incremental approach to the extraction of relationships between biological entities from literature to assist the development of biological networks. The presented method extracts relations between components in the form of triplets from which a network structure was constructed. We have applied the developed method to a time-labelled collection of biomedical documents obtained from PubMed Central in order to incrementally enrich a network which we constructed from three models available in the literature. The results show that publicly accessible sources of biomedical literature such as PubMed Central offer a good starting point for the computer-assisted development of plant defence models, and that approaches such as our incremental method can contribute to the discovery of potentially interesting relations. By applying the triplet extraction incrementally on time-labelled data one can follow the development of knowledge about certain biological relations, and discover new relations which can potentially enhance already developed models.

In the future work 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. Also we plan to add the named entity recognition and automatic discovery of components' synonyms to automatically construct the vocabulary of components for the Bio3graph. Experts evaluation of the importance of reactions in pathways could also be meaningful.

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 ENVISION (Environmental Services Infrastructure with Ontologies) under the Grant Agreement No. 249120.

References

1. H.A. Kestler, C. Wawra, B. Kracher and M. Kühl, "Network modeling of signal transduction: establishing the global view", Bioessays, vol. 30, pp. 1110-1125, 2008.
2. N. Le Novère, M. Hucka, H. Mi, S. Moodie, F. Screiber, A. Sorokin, et al., "The Systems Biology Graphical Notation", Nat Biotechnol, vol. 27, pp. 735-741, 2009.

3. S. Raza, K.A. Robertson, P.A. Lacaze, D. Page, A.J. Enright, P. Ghazal, et al. "A logic-based diagram of signalling pathways central to macrophage activation", BMC Syst Biol, vol 2, 2008.
4. M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes", Nucleic Acids Res, vol. 28, pp. 27-30, 2000.
5. N. Tsesmetzis, M. Couchman, J. Higgins, A. Smith, J. H. Doonan, G. J. Seifert, et al. "Arabidopsis reactome: A foundation knowledgebase for plant systems biology", Plant Cell, vol. 20, pp. 1426-1436, 2008.
6. M. Krummenacker, S. Paley, L. Mueller, T. Yan and P. D. Karp, "Querying and computing with BioCyc databases", Bioinformatics, vol. 21, pp. 3454-3455, 2005.
7. J. Köhler, J. Baumbach, J. Taubert, M. Specht, A. Skusa, A. Ruegg, C. Rawlings, et al. "Graph-based analysis and visualization of experimental results with ONDEX," Bioinformatics, vol. 22, pp. 1383-1390, 2006.
8. L. Eronen and H. Toivonnen, "Biomine: predicting links between biological entities using network models of heterogeneous databases", BMC Bioinformatics, vol. 13, pp. 119, 2012.
9. S. Raza, N. McDerment, P. A. Lacaze, K. Robertson, S. Watterson, Y. Chen, et al. "Construction of a large scale integrated map of macrophage pathogen recognition and effector systems", BMC Syst Biol, vol. 4, pp. 63, 2010.
10. S. Ananiadou, S. Pyysalo, J. Tsujii and D.B. Kell, "Event extraction for systems biology by text mining the literature", Trends Biotechnol, vol. 28, pp. 381-390, 2010.
11. K.B. Cohen and L. Hunter, "Getting started in text mining", PLoS Comput Biol, PLoS Comput Biol 4(8): e20, 2008.
12. A. Rzhetsky, I. Iossifov, T. Koike, M. Krauthammer, P. Kra, M. Morris, et al. "GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data", J Biomed Inform, vol. 37, pp. 43-53, 2004.
13. H. Chen and B.M. Sharp, "Content-rich biological network constructed by mining PubMed abstracts", BMC Bioinformatics, vol. 5, pp. 147, 2004.
14. T. Ono, H. Hishigaki, A. Tanigami and T. Takagi, "Automated extraction of information on protein-protein interactions from the biological literature", Bioinformatics, vol. 17, pp. 155-161, 2001.
15. C. Blaschke and A. Valencia, "The frame-based module of the SUISEKI information extraction system", IEEE Intell Syst, vol. 17, pp. 14-20, 2002.
16. D.P.A. Corney, B.F. Buxton, W.B. Langdon and D.T. Jones, "BioRAT: extracting biological information from full-length papers", Bioinformatics, vol. 20, pp. 3206-3213, 2004.
17. D. Miljković, T. Stare, I. Mozetič, V. Podpečan, M. Petek, K. Witek, et al. "Signalling Network Construction for Modelling Plant Defence Response", PLoS ONE 7(12): e51822, 2012.
18. D. Swarbreck, C. Wilks, P. Lamesch, T.Z. Berardini, M. Garcia-Hernandez, H. Foerster, et al. "The Arabidopsis Information Resource (TAIR): gene structure and function annotation", Nucleic Acids Res, vol. 36, pp. D1009-D1014, 2008.
19. R. Hoffmann and A. Valencia, "A gene network for navigating the literature", Nat Genet, vol. 36, pp. 664-664, 2004.
20. J.S. González-García and J. Díaz, "Information theory and the ethylene genetic network", Plant Signal Behav. vol. 6, pp. 1483-98, 2011.
21. J. G. Turner, C. Ellis and A. Devoto, "The jasmonate signal pathway", Plant Cell, vol. 14 Suppl:S153-64, 2002.
22. J. Shah, "The salicylic acid loop in plant defense", Curr Opin Plant Biol, vol. 6, pp. 365-71, 2003.