# Modelling a Biological System: Network Creation by Triplet Extraction from Biological Literature

Dragana Miljkovic[1,⋆], Vid Podpečan[1], Miha Grčar[1], Kristina Gruden[3], Tjaša Stare[3], Marko Petek[3], Igor Mozetič[1], and Nada Lavrač[1,2]

[1] Jožef Stefan Institute, Ljubljana, Slovenia
{dragana.miljkovic,vid.podpecan,miha.grcar}@ijs.si,
{igor.mozetic,nada.lavrac}@ijs.si
[2] University of Nova Gorica, Nova Gorica, Slovenia
[3] National Institute of Biology, Ljubljana, Slovenia
{kristina.gruden,tjasa.stare,marko.petek}@nib.si

**Abstract.** The chapter proposes an approach to support modelling of plant defence response to pathogen attacks. Such models are currently built manually from expert knowledge, experimental results, and literature search, which is a very time consuming process. Manual model construction can be effectively complemented by automated model extraction from biological literature. This work focuses on the construction of triplets in the form of subject-predicate-object extracted from scientific papers, which are used by the Biomine automated graph construction and visualisation engine to create the biological model. The approach was evaluated by comparing the automatically generated graph with a manually developed Petri net model of plant defence. This approach to automated model creation was explored also in a bisociative setting. The emphasis is not on creative knowledge discovery, but rather on specifying and crossing the boundaries of knowledge of individual scientists. This could be used to model the expertise of virtual scientific consortia.

## 1 Introduction

The mechanism of a plant's defence response to virus attacks is a hot topic of current biological research. Despite a vivid interest in creating a holistic model of plant defence, only partial and oversimplified models of the entire defence system are created so far.

The motivation of biologists to develop a more comprehensive model of the entire defence response is twofold. Firstly, it will provide a better understanding of the complex defence response mechanism in plants by highlighting important connections between biological molecules and understanding how the mechanism operates. Secondly, prediction of experimental results through simulation saves time and indicates further research directions to biological scientists. The development of a more comprehensive model of plant defence for simulation purposes raises three research questions:

⋆ Corresponding author.

M.R. Berthold (Ed.): Bisociative Knowledge Discovery, LNAI 7250, pp. 427–437, 2012.

- What is the most appropriate formalism for representing the plant defence model?
- How to extract the model, i.e. network structure; more precisely, how to retrieve relevant molecules and relations between them?
- How to determine network parameters such as initial molecules values, types and speeds of the reactions, threshold values, etc.?

Having studied different representation formalisms, we have decided to represent the model of the given biological network in the form of a graph. This chapter addresses the second research question, i.e. automated extraction of the graph structure through information retrieval and natural language processing techniques. We propose a methodology to support modelling of plant defence response to pathogen attacks, and present its implementation in a workflow which combines open source natural language processing tools, data from publicly available databases, and hand-crafted knowledge. The evaluation of the approach is carried out using a manually crafted Petri net model which was developed by fusing expert knowledge and the results of manual literature search.

The structure of the chapter is as follows. Section 2 presents existing approaches to modelling plant defence and discusses their advantages and shortcomings. Section 3 briefly presents our manually crafted Petri net model, followed by Section 4 which proposes a methodology used for automated model extraction from the biological literature. Section 5 explores the results of model extraction in a bisociative setting, where extracted knowledge of different scientists is combined. Section 6 concludes the chapter and proposes directions for further work.

## 2 Related Work

Due to the complexity of the plant defence mechanism, the challenge of building a general model for simulation purposes is still not fully addressed. Early attempts to accomplish simulation by means of Boolean formalism from experimental microarray data [5] have already indicated the complexity of defence response mechanisms, and highlighted many crosstalk connections. However, several of the interconnecting molecules were not considered in the model presented in that work. These intermediate molecules as well as the discovery of new connections between them are of particular interest for biological scientists.

Other existing approaches, such as the MoVisPP tool [6], attempt to automatically retrieve information from databases and transfer the pathways into the Petri net formalism. MoVisPP is an online tool which automatically produces Petri net models from KEGG and BRENDA pathways. However, not all pathways are accessible, and the signalling pathways for plant defence do not exist in the databases.

Tools for data extraction and graphical representation are also related to our work as they are used to help experts to understand the underlying biological principles. They can be roughly grouped according to their information sources: databases (Biomine [14][4], Cytoscape [15], ProteoLens [8], VisAnt [7], PATIKA [2]), databases and experimental data (ONDEX [9], BiologicalNetworks [1]), and

literature (TexFlame [10]). More general approaches, such as visualisation of arbitrary textual data through triplets [13] are also relevant. However, such general systems have to be adapted in order to produce domain-specific models.

## 3 Manually Constructed Petri Net Model of Plant Defence Response

This section presents a part of the manually crafted Petri net model using the Cell Illustrator software [12]. We briefly describe the development cycle of the model and show some simulation results.

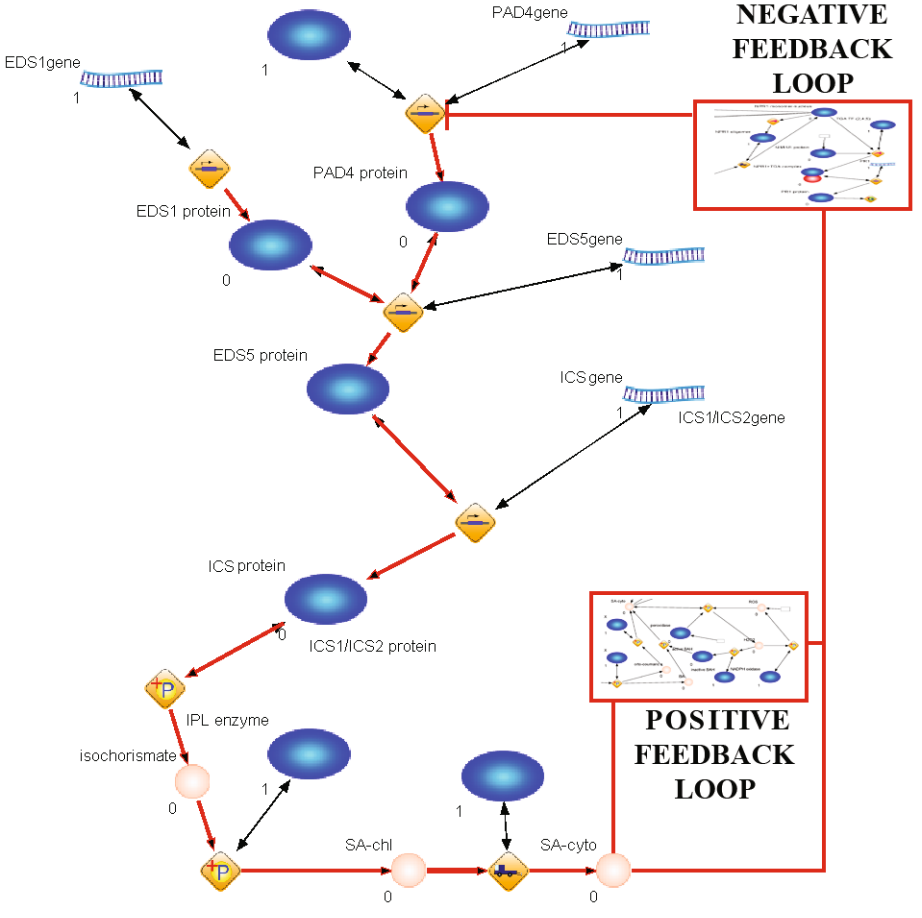


**Fig. 1.** A partial and simplified Petri net model of SA biosynthesis and signalling pathway in plants. Biological molecules SA-chl and SA-cyto represent SA located in different parts of the cell. Both SA-chl and SA-cyto of this figure correspond to node SA in the graph of Figure 4.

A Petri net is a bipartite graph with two types of nodes: places and transitions. Standard Petri net models are discrete in terms of variables and sequence of events, but their various extensions can represent both qualitative and quantitative models. The Cell Illustrator software implements an extension of Petri nets, called hybrid functional Petri net, which was used in our study. In the hybrid functional Petri net formalism, the speed of transitions depends on the amount of input components and both, discrete and continuous places, are supported.

Our manually crafted Petri net model of plant defence currently represents a complex network where molecules and reactions, according to the Petri net formalism, correspond to places and transitions, respectively. A part of the model of salicylic acid (SA) biosynthesis and signalling pathway, which is one of the key components in plant defence, is shown in Figure 1.

Early results of the simulation already show the effects of positive and negative feedback loops in the SA pathway as shown in Figure 2. The light grey line represents the level of SA-chl (SA in chloroplast) that is not part of the positive feedback loop. The dark grey line represents SA-cyto, same component in cytoplasm, that is a part of the positive feedback loop. The peak of the dark grey line depicts the effect of the positive feedback loop which rapidly increases the amount of SA-cyto. After reaching the peak, the trend of the dark grey line is negative as the effect of the negative feedback loop prevails.

The represented Petri net model consists of two types of biological pathways: metabolic and signalling. The metabolic part is a cascade of reactions with small compounds as reactants, and was obtained from KEGG database. The signalling part is not available in the databases and had to be obtained from the literature.

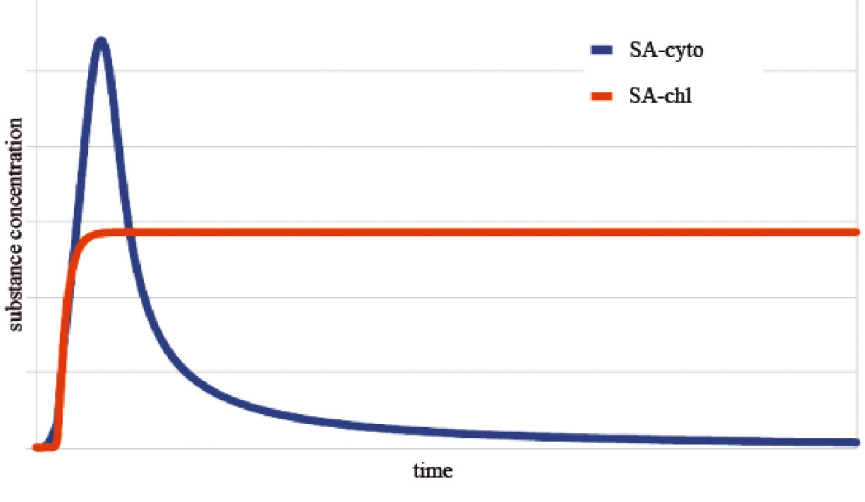


**Fig. 2.** Simulation results of the Petri net model of SA pathway. The light grey line represents the level of SA-chl, i.e. SA in chloroplast that is not part of the positive feedback loop. The dark grey line represents the same component in cytoplasm, SA-cyto, which is in the positive feedback loop.

The biological scientists have manually extracted relevant information related to this pathway within a period of approximately two months. Keeping in mind that the SA pathway is only one out of three pathways involved in plant defence response, it is clear that a purely manual approach would be very time-consuming.

## 4 Automated Extraction of Plant Defence Response Model from Biological Literature

The process of fusing expert knowledge and manually obtained information from the literature as presented in the previous section turns out to be time consuming and error-prone. Therefore, we suggest the automated extraction of information from the scientific literature relevant to the construction and curation of such models. The proposed methodology consists of a series of text mining and information retrieval steps, which offer reusability and repeatability, and can be easily extended with additional components. For natural language processing we employed functions from the NLTK library [11], which were transformed into web services and used in the proposed triplet extraction, graph construction, visualisation and exploration workflow. Additionally, the GENIA tagger [16] for biological domains was used to perform part-of-speech tagging and shallow parsing. The data was extracted from PubMed[1] and Web of Science[2] using web-service enabled access.

The methodology for information retrieval from public databases to support modelling of plant defence is shown in Figure 3. Computer-assisted creation of plant defence models from textual data, is performed by using following services:

1. PubMed web service and Web of Science search to extract the article data,
2. PDF-to-text converter service, which is based on Poppler[3], an open source PDF rendering library,
3. natural language processing web services based on NLTK: tokenizer, shallow parser (chunker), sentence splitter,
4. the GENIA tagger,
5. filtering components, e.g. negation removal, etc.

The goal of this study is to extract sets of triplet in the form:

$$(Subject, Predicate, Object)$$

from biological texts which are freely available. The defence response related information is obtained by employing the vocabulary which we have manually developed for this specific field. *Subject* and *Object* are biological molecules such as proteins or small compounds, and their names with synonyms are built

[1] PubMed is a free database that comprises biomedical journals and online books.
[2] Web of Science is an online academic citation index constructed to access multiple databases and explore cross-disciplinary studies and specialized subfields within a scientific domain.
[3] `http://poppler.freedesktop.org/`

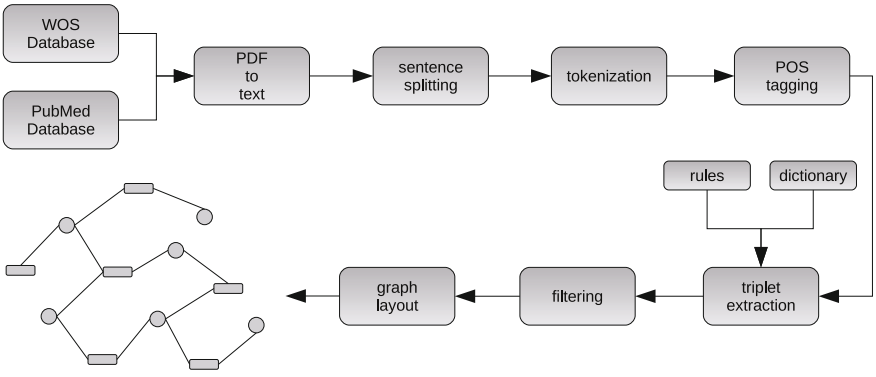


**Fig. 3.** Methodology for information retrieval from public databases to support modelling of plant defence response

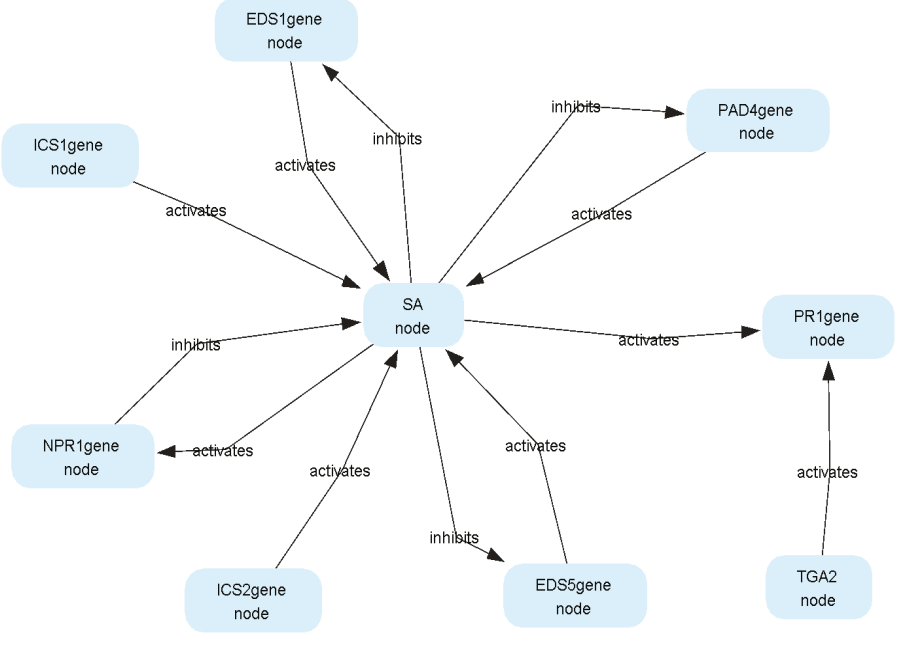


**Fig. 4.** A graph constructed from a set of triplets, extracted from ten documents, visualised using the Biomine visualisation engine

into the vocabulary. *Predicate* represents the relation or interaction between the molecules. We have defined three types of reactions, i.e. *activation*, *inhibition* and *binding*, and the synonyms for these reactions are also part of the vocabulary. An example of such a triplet is shown below:

$$(PAD4\,protein,\ activates,\ EDS5\,gene)$$

Such triplets, if automatically found in text, composed and visualised as a graph, can assist the development of the plant defence Petri net model. Triplet extraction is performed by employing simple rules to find the last noun of the first phrase as *Subject*. *Predicate* is a part of a verb phase located between the noun phrases. *Object* is then detected as a part of the first noun phrase after the verb phrase. In addition to these rules, pattern matching from the dictionary is performed to search for more complex phrases in text to enhance the information extraction. The graph is then constructed and visualised using the Biomine graph construction and visualisation engine [14]. An example of such a graph is shown in Figure 4.

While such automatically extracted knowledge currently cannot compete — in terms of details and correctness — with the manually crafted Petri net model, it can be used to assist the expert in the process of building and curation of the model. Also, it can provide novel and relevant information to the biological scientist.

## 5 Two Modelling Scenarios

### 5.1 An Illustrative Example

Consultation with biological scientists resulted in the first round of experiments performed on a set of ten most relevant articles from the field which were published after 2005. Figure 4 shows the extracted triplets, visualised using the Biomine graph visualiser.

SA appears to be the central component in the graph, which confirms the biological fact that SA is indeed one of the three main components in plant defence. The information contained in the graph of Figure 4 is similar to the initial knowledge obtained from biological scientists by manual information retrieval from the literature[4]. Such a graph, however, cannot provide the cascade network type which is closer to reality (and to the manually crafted Petri net model).

The feedback from the biologists was positive. Even though this approach cannot completely substitute human experts, biologists consider it a helpful tool in speeding up information acquisition from the literature. The presented results indicate the usefulness of the proposed approach but also the necessity to further improve the quality of information extraction.

### 5.2 Crossing the Boundaries of Individual Readers

The goal of the second experiment is to elicit differences in knowledge and interests between different scientists. We take a simplifying assumption that each

[4] It is worth noting that before the start of joint collaboration between the computer scientists and biologists, the collaborating biological scientists have tried to manually extract knowledge from scientific articles in the form of a graph, and have succeeded to build a simple graph representation of the SA biosynthesis and signalling pathway.

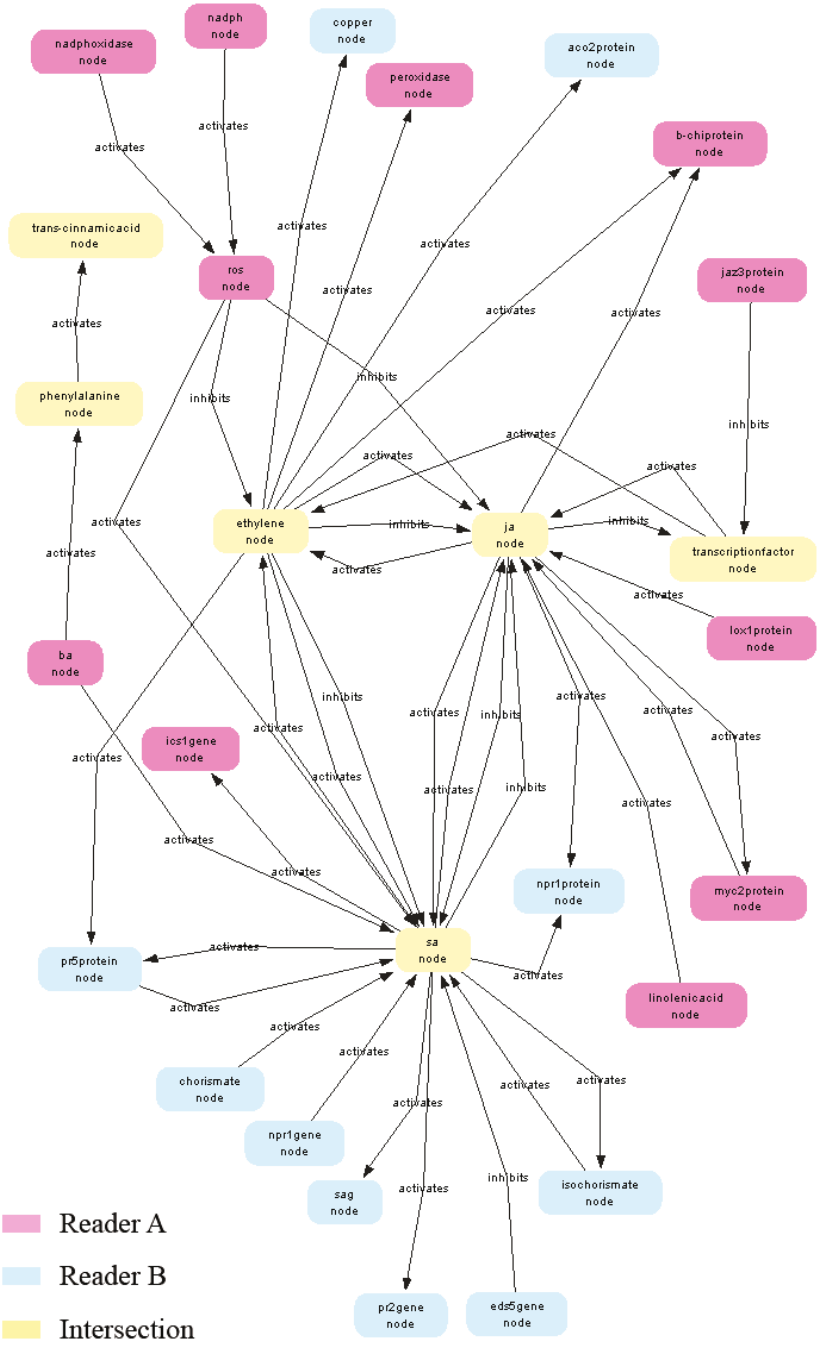


**Fig. 5.** A model constructed from a set of triplets extracted from 122 documents, read by two different readers and displayed using the Biomine graph visualisation engine

scientists' knowledge corresponds to a set of papers it read. The extracted triplets and subgraph thus model her/his subjective, habitual knowledge [3]. By combining subjective knowledge bases we obtain a join BisoNet where the intersecting subgraph represents a bridging graph pattern of bisociation. In particular, we extracted triplets from a set of 122 documents, read by two biology experts:

**Reader A:** Reader A (colored dark grey) has read 91 papers, of which 13 unique triplets were extracted automatically.

**Reader B:** Reader B (colored medium grey) has read 31 papers, of which 21 unique triplets were extracted automatically.

**Intersections:** Eight common triplets, extracted from 91 publications read by reader A and from 31 publications read by reader B, were colored in light grey colour.

Figure 5 shows the model extracted from 122 articles read by the two readers (two biological scientists). Besides supporting the automatic model construction, there are other benefits from visualising knowledge of different domain experts as illustrated in Figure 5. For instance, one can clearly see which nodes are in the intersection of interest of the two experts (coloured light grey in Figure 5).

This could indicate the areas of joint interest which the two experts might want to investigate jointly in more detail, e.g., to get answers to some yet unexplored research question in the intersection of their domains of expertise. On the other hand, this visualisation enables to see also who has some unique expertise in the field, with no intersection with other experts (coloured dark and medium grey in Figure 5). If applied to modelling the knowledge of larger consortia of readers, this type of information could be used to determine the complementarities of research groups.

The proposed approach to modelling and visualisation of knowledge extracted from the literature could be used also for modelling the know-how of large project consortia where it is hard to track the expertise of all project participants. Consequently, the proposed approach to cross-context modelling may be viewed as a step towards creating virtual laboratory knowledge models.

## 6 Conclusion

In this chapter we presented a methodology which supports the domain expert in the process of creation, curation, and evaluation of plant defence response models by combining publicly available databases, natural language processing tools, and hand-crafted knowledge. The methodology was implemented as a reusable workflow of software services, and evaluated using a hand crafted Petri net model. This Petri net model has been developed by fusing expert knowledge, experimental results and biological literature, and serves as a baseline for evaluation of automatically extracted plant defence response knowledge, but it also enables computer simulation and prediction.

This chapter presented also an approach to modelling the knowledge of different domain experts, by visualising the intersections as well as complementarities

of their expertise, with a potential of providing a global overview of the expertise of consortia members. This type of modelling can be used to analyze and monitor knowledge of larger groups of experts to establish how their knowledge grows and evolves in terms of time and research topics.

**Acknowledgments.** This work has been supported by the European Commission under the 7th Framework Programme FP7-ICT-2007-C FET-Open, contract no. BISON-211898, the European Community 7th frame-work program ICT-2007.4.4 under grant number 231519 e-Lico: An e-Laboratory for Interdisciplinary Collaborative Research in Data Mining and Data Intensive Science, AD Futura scholarship and the Slovenian Research Agency grants P2-0103 and J4-2228. We are grateful to Claudiu Mihăilă for the initial implementation of the triplet extraction engine, and Lorand Dali and Delia Rusu for constructive discussions and suggestions.



## References

1. Baitaluk, M., Sedova, M., Ray, A., Gupta, A.: BiologicalNetworks: visualization and analysis tool for systems biology. Nucl. Acids Res. 34(suppl. 2), W466–W471 (2006)
2. Demir, E., Babur, O., Dogrusoz, U., Gursoy, A., Nisanci, G., Cetin-Atalay, R., Ozturk, M.: PATIKA: An integrated visual environment for collaborative construction and analysis of cellular pathways. Bioinformatics 18(7), 996–1003 (2002)
3. Dubitzky, W., Kötter, T., Schmidt, O., Berthold, M.R.: Towards Creative Information Exploration Based on Koestler's Concept of Bisociation. In: Berthold, M.R. (ed.) Bisociative Knowledge Discovery. LNCS (LNAI), vol. 7250, pp. 11–32. Springer, Heidelberg (2012)
4. Eronen, L., Hintsanen, P., Toivonen, H.: Biomine: A Network-Structured Resource of Biological Entities for Link Prediction. In: Berthold, M.R. (ed.) Bisociative Knowledge Discovery. LNCS (LNAI), vol. 7250, pp. 364–378. Springer, Heidelberg (2012)
5. Genoud, T., Trevino Santa Cruz, M.B., Metraux, J.-P.: Numeric Simulation of Plant Signaling Networks. Plant Physiology 126(4), 1430–1437 (2001)
6. Hariharaputran, S., Hofestädt, R., Kormeier, B., Spangardt, S.: Petri net models for the semi-automatic construction of large scale biological networks. Springer Science and Business. Natural Computing (2009)
7. Hu, Z., Mellor, J., Wu, J., DeLisi, C.: VisANT: data-integrating visual framework for biological networks and modules. Nucleic Acids Research 33, W352–W357 (2005)
8. Huan, T., Sivachenko, A.Y., Harrison, S.H., Chen, J.Y.: ProteoLens: a visual analytic tool for multi-scale database-driven biological network data mining. BMC Bioinformatics 9(suppl. 9), S5 (2008)

9. Köhler, J., Baumbach, J., Taubert, J., Specht, M., Skusa, A., Röuegg, A., Rawlings, C., Verrier, P., Philippi, S.: Graph-based analysis and visualization of experimental results with Ondex. Bioinformatics 22(11), 1383–1390 (2006)
10. Le Novère, N., Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., Demir, E., Wegner, K., Aladjem, M.I., Wimalaratne, S.M., Bergman, F.T., Gauges, R., Ghazal, P., Kawaji, H., Li, L., Matsuoka, Y., Villéger, A., Boyd, S.E., Calzone, L., Courtot, M., Dogrusoz, U., Freeman, T.C., Funahashi, A., Ghosh, S., Jouraku, A., Kim, S., Kolpakov, F., Luna, A., Sahle, S., Schmidt, E., Watterson, S., Wu, G., Goryanin, I., Kell, D.B., Sander, C., Sauro, H., Snoep, J.L., Kohn, K., Kitano, H.: The Systems Biology Graphical Notation. Nature Biotechnology 27(8), 735–741 (2009)
11. Loper, E., Bird, S.: NLTK: The Natural Language Toolkit. In: Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, pp. 62–69. Association for Computational Linguistics, Philadelphia (2002)
12. Matsuno, H., Fujita, S., Doi, A., Nagasaki, M., Miyano, S.: Towards Biopathway Modeling and Simulation. In: van der Aalst, W.M.P., Best, E. (eds.) ICATPN 2003. LNCS, vol. 2679, pp. 3–22. Springer, Heidelberg (2003)
13. Rusu, D., Fortuna, B., Mladenić, D., Grobelnik, M., Sipoš, R.: Document Visualization Based on Semantic Graphs. In: Proceedings of the 13th International Conference Information Visualisation, pp. 292–297 (2009)
14. Sevon, P., Eronen, L., Hintsanen, P., Kulovesi, K., Toivonen, H.: Link Discovery in Graphs Derived from Biological Databases. In: Leser, U., Naumann, F., Eckman, B. (eds.) DILS 2006. LNCS (LNBI), vol. 4075, pp. 35–49. Springer, Heidelberg (2006)
15. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T.: Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Research 13, 2498–2504 (2003)
16. Tsuruoka, Y., Tateishi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S., Tsujii, J.: Developing a Robust Part-of-Speech Tagger for Biomedical Text. In: Bozanis, P., Houstis, E.N. (eds.) PCI 2005. LNCS, vol. 3746, pp. 382–392. Springer, Heidelberg (2005)