W160–W165 Nucleic Acids Research, 2009, Vol. 37, Web Server issue Published online 11 June 2009
doi:10.1093/nar/gkp484

PLAN2L: a web tool for integrated text mining and literature-based bioentity relation extraction

Martin Krallinger[1,*], Carlos Rodriguez-Penagos[2], Ashish Tendulkar[1] and Alfonso Valencia[1]

[1]Structural Biology and Biocomputing programme, Spanish National Cancer Center (CNIO), Melchor Fernandez Almagro 3, Madrid, 28029, Spain and [2]Barcelona Media - Centre d'Innovacio, Av. Diagonal 177, 08018 Barcelona, Spain

Received March 16, 2009; Revised May 13, 2009; Accepted May 19, 2009

ABSTRACT

There is an increasing interest in using literature mining techniques to complement information extracted from annotation databases or generated by bioinformatics applications. Here we present PLAN2L, a web-based online search system that integrates text mining and information extraction techniques to access systematically information useful for analyzing genetic, cellular and molecular aspects of the plant model organism *Arabidopsis thaliana*. Our system facilitates a more efficient retrieval of information relevant to heterogeneous biological topics, from implications in biological relationships at the level of protein interactions and gene regulation, to sub-cellular locations of gene products and associations to cellular and developmental processes, i.e. cell cycle, flowering, root, leaf and seed development. Beyond single entities, also predefined pairs of entities can be provided as queries for which literature-derived relations together with textual evidences are returned. PLAN2L does not require registration and is freely accessible at http://zope.bioinfo.cnio.es/plan2l.

INTRODUCTION

Gene regulatory mechanisms and protein interactions are studied in detail to understand how complex developmental processes are controlled. Biological annotation databases provide functional descriptions of gene products, the basic components of such biological processes, through manual literature inspection, resulting generally in associations of biological entities to a set of controlled vocabulary terms contained in structured database records (1). Despite the obvious strength of controlled vocabularies for annotation consistency, information exchange and data analysis, functional annotations of proteins do not provide a straightforward way to trace back the biological evidence supporting each annotation, making it sometimes cumbersome for human domain experts to directly interpret under which biological context and experimental conditions a given annotation applies. Considering the growing amount of published articles, to manually annotate newly described functional gene product characterizations as well as maintenance and update of already annotated entities is a challenging task. This motivated recent attempts to enable a more systematic access to relevant information hidden in large literature repositories using text mining and information extraction (IE) technologies, with the aim of not only supporting the literature curation process, but especially for providing suitable information retrieval systems useful for life sciences (2). Results generated by text mining systems have the general advantage to be directly interpretable by the end users, the human domain experts, in case direct links to the textual evidences are supported. Currently existing online literature mining systems mostly focus on very particular biological aspects, such as the extraction of protein–protein interactions (3), protein–keyword co-mentions or gene regulation events without directly integrating all these heterogeneous relation types into a single application. Relevance, not only of individual bio-entities but also of their interactions and regulation events for developmental processes studied in model organisms such as the plant *Arabidopsis thaliana* have not been addressed previously using text mining approaches. *Arabidopsis thaliana*, the first plant to be completely sequenced, is being used not only for experimental research, but also increasingly by systems biology and bioinformatics approaches to understand and model central biological processes (4). Databases like TAIR [The Arabidopsis Information Resource, (5)] or UniProt (6) are providing plant biologists with valuable infrastructures of manually curated information, but only few attempts have been made to implement literature mining applications for this model organism. The Dragon Plant Biology Explorer (DPBE) was an online text mining application for plant biology based on integration and combination of collections of manually curated vocabularies compiled for

[*]To whom correspondence should be addressed. Tel: +34 91 224 6900; Fax: +34 91 224 6980; Email: mkrallinger@cnio.es

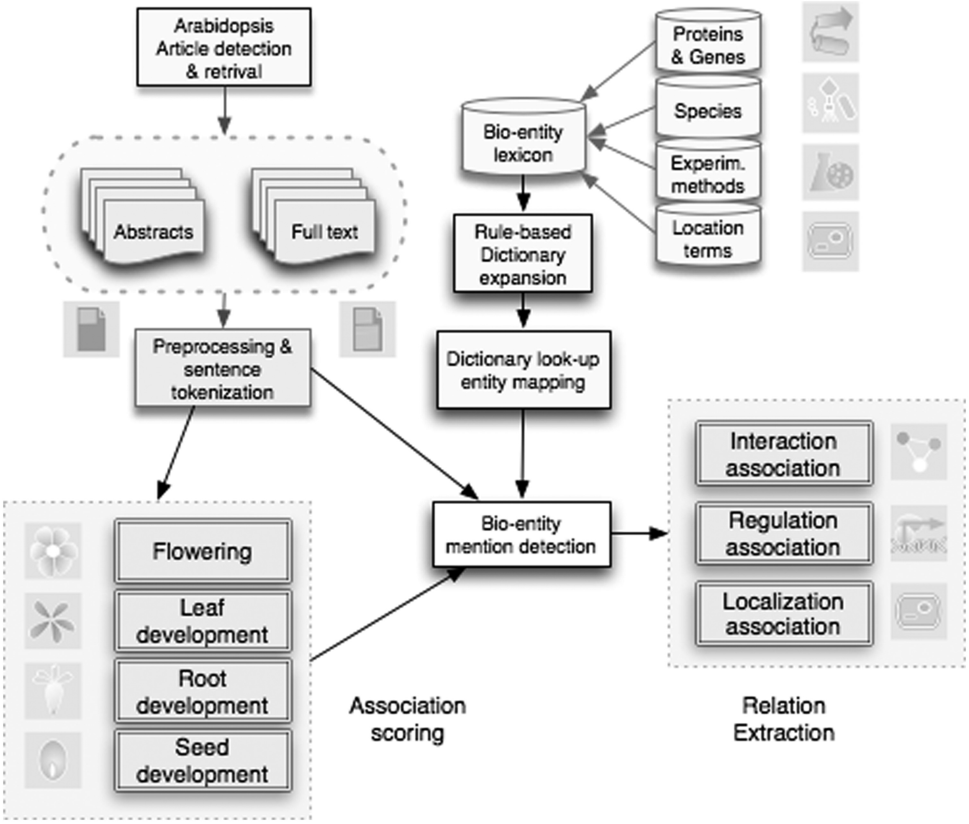


Figure 1. PLAN2L flowchart. This figure shows the main steps behind the flow chart followed by the PLAN2L system.

several topics to facilitate a more targeted literature search (7). PubSearch constitutes another system for semi-automated retrieval of literature that can be used to curate articles to extract manually annotations for gene products with Gene Ontology terms (8). It is primarily based on simple term matching and does not explore machine learning techniques to provide more sophisticated retrieval capabilities. Another system primarily used for literature curation by model organism databases is Textpresso (9), now integrated at TAIR to facilitate a better access to information relevant for *Arabidopsis*.

SYSTEM DESCRIPTION

PLAN2L is an online text mining application dedicated to improve retrieval of knowledge by integrating and scoring information extracted from textual sources for various biological topics related with the description interactome associations in *Arabidopsis*. Figure 1 shows a general flow chart of the PLAN2L system 1 (a more complete description of the technical aspects can be found in the documentation of the web system).

TECHNICAL DESCRIPTION OF THE TEXT MINING PIPELINE

A document retrieval pipeline that takes into account several sources of evidence for the determining whether a given article is associated to *A. thaliana* was implemented exploiting: (i) external references derived from multiple databases providing annotations for *Arabidopsis* proteins. (ii) Organism and taxonomic name tagging using dictionary lookup based on a species lexicon derived from the NCBI Taxonomy that was automatically extended using a rule-based approach to account for typographical variants and abbreviations of species names. (iii) Keyword-based retrieval from PubMed and PubMed Central. Additionally, a full text collection of *Arabidopsis*-related articles was constructed from a local repository of open access full-text articles as well as using customized article collection tools. Plain text conversion was carried out through a combination of systems including pdftotext and PDFlib. The detection of links between the literature and protein or genes of PLAN2L is based on the construction and lookup of a gene lexicon. This gene dictionary integrated *A. thaliana* gene names and symbols derived from multiple databases, including TAIR, SwissProt and from a collection of gene and protein names identified by a machine learning-named entity recognition program (ABNER) as well as based a rule-based approach considering morphological cues and name length to identify potential *Arabidopsis* gene symbols. Lexicon expansion using manually crafted rules was carried out. To detect gene regulatory relations, we adapted an IE architecture relying on a pipeline of semantic/syntactic rules. We applied part-of-speech tagging of each word using a GENIA-trained version of Treetager (10). Some of the POS tags were automatically substituted with more semantically oriented labels (e.g. organisms, protein/gene names and activation verbs). The text with mixed syntactic and semantic tags was fed into a SCOL parser (11), which generated a tree-like structure by applying a modified CASS grammar. These rules constitute cascades of finite-state automata, and use patterns that combine both grammatical- and biological-meaning features in the linguistic structure. We implemented extensions of the rules to handle frequent phrase coordination and prepositional anaphora. The extraction of protein interaction evidence associations was addressed using a machine learning sentence classifier approach relying on manually selected interaction evidence sentences (12). The used sentence classifier relies on a Support Vector Machines algorithm trained on set of manually classified interaction evidence passages derived from a collection used at the second BioCreative challenge (12), and obtained a performance of 89.75 for precision and 92.62 for recall using a radial basis kernel function on a balanced test set. For retrieving protein localization descriptions, we explored both the use of semantic–syntactic frames for extracting a fine-grained association between proteins and subcellular location mentions together with a machine learning sentence classifier for retrieving protein localization description sentences in general. The initial step followed, consisted in the construction of a sub-cellular location dictionary that integrates location keywords and synonyms derived from SwissProt together with cellular component terms from Gene Ontology. Location term mentioning sentences were manually revised to derive hand-crafted location frames. Additionally, a location sentence classifier was constructed using a collection of 2264 protein location description sentences. A central component of PLAN2L is the scoring of each evidence sentence according to its relevance for complex temporal biological events (topics), at the cellular level (cell cycle) as well as at the level of developmental processes. We therefore implemented a classifier for scoring cell-cycle relevant abstracts and document passages. The SVM text classifier was trained on a collection of cell-cycle relevant abstracts and nonrelevant abstracts and then applied to a literature collection of abstracts and full-text articles mentioning *A. thaliana* genes. Additionally, four specific sentence classifiers for the most relevant developmental processes in higher plants, namely (i) flowering, (ii) leaf development, (iii) root development and (iv) seed development/germination have been developed. The tool provides a comprehensive approach to assist in the selection and ranking of genes, proteins, documents and terms relevant to a specific biological process for this model organism. Additional details on the different modules, their characteristics and assessment are provided at the PLAN2L web.

FUNCTIONALITY AND USAGE

The PLAN2L interface handles user-provided plain text keywords, protein/gene names or symbols. Some of the components allow additionally searching with gene or protein identifiers. The currently supported identifiers include TAIR gene identifiers and UniProt primary accession numbers. Based on a user survey that we carried out to get feedback from biologists on the PLAN2L system, aspects that were positively rated included its easy to use and intuitive query interface and the direct retrieval of evidence sentences for multiple topics. Aspects that were improved according to the user comments covered additional documentation on the system and the sentence scoring mechanism. PLAN2L supports six types of search strategies, each with its own query page, to avoid introducing unnecessary complexity through advanced search interfaces with complicated menu options. We will briefly describe each of these six search types and provide a case study to illustrate the type of results generated by PLAN2L.

(i) The basic search type allows retrieval of multiple biological topic associations for a given user query, consisting in ranked lists of evidence sentences. The default ranking is based on the association strength of the article for the organism source (i.e. *A. thaliana*), sorted in descending order. The structured result table page provides the document identifier (PMID) from which the evidence sentence has been derived, the extracted sentence text showing the color highlighted user query term as well as the relevance of the sentence for multiple biological topics: cell cycle, gene regulation, protein interaction, cellular location, flowering, leaf development, root development and seed development. Additional external links to other resources are also included, namely the BioCreative metaserver, iHOP, WikiGene and TAIR. In the results page, color-coded scores are given for each category represented visually by a unique informative glyph. Scores above classifier cutoff are displayed in green, otherwise in red. The evidence sentences can be dynamically reranked for each of the categories, and to facilitate rapid visual cueing of the association between the scores and their class, a background-coloring schema was used.

(ii) PLAN2L also facilitates retrieval of physical interaction associations for a given query protein or protein identifier. Each of the co-mentioned proteins is returned together with its corresponding identifier, interaction evidence sentence and experimental interaction detection method keywords. Ranking of the interaction association evidence sentences is based on the corresponding interaction sentence classifier score.

(iii) In case of regulatory associations, the typical user-specified query consists in a protein/gene name or in which the system returns a tabular regulation evidence summary. To facilitate the interpretation of the directionality of a given regulatory event, a color-coding strategy was used to highlight the regulator and regulated gene, together with a key sentence display. To qualify the type of regulation, PLAN2L labels each regulation evidence as activation, repression or undefined (for unclear cases).

(iv) The sub-cellular location evidence sentences for a given protein can also be retrieved by PLAN2L. Complementing the location sentence scoring of the basic search option, more accurate location sentence retrieval, integrating also location term co-occurrence and location association words is available through PLAN2L.

(v) As *A. thaliana* is one of the most important model organisms for studying the plant cell-cycle, PLAN2L integrates a cell-cycle relevance ranking of PubMed abstracts using a document classification tool. The *Arabidopsis* literature can be searched using gene names or keywords and the resulting abstracts are ranked according to their cell-cycle relevance (cell-cycle score from the document classifier). Also cell-cycle relevant terms are tagged. The species ambiguity score of abstracts can be used to determine how specific a document is for *Arabidopsis*, something useful when several species are mentioned in the same article.

(vi) In order to enable more effective searches of association evidences existing between two biological entities of interest, PLAN2L provide the association retrieval interface, where for two entities entered by the user, the application returns a set of sentence where both entities are co-mentioned. Each evidence sentence is scored according to the biological topics supported for the basic search query. The default ranking is based on the interaction sentence classifier scores. The typical time taken for text processing is 2–3 min, but in case of heavy server load and the actual query type, some jobs might take between 4 and 5 min, upon completion of the search task PLAN2L displays its results together with hyperlinks to the resources providing links to the original article.

Case study: AGAMOUS and LEUNIG

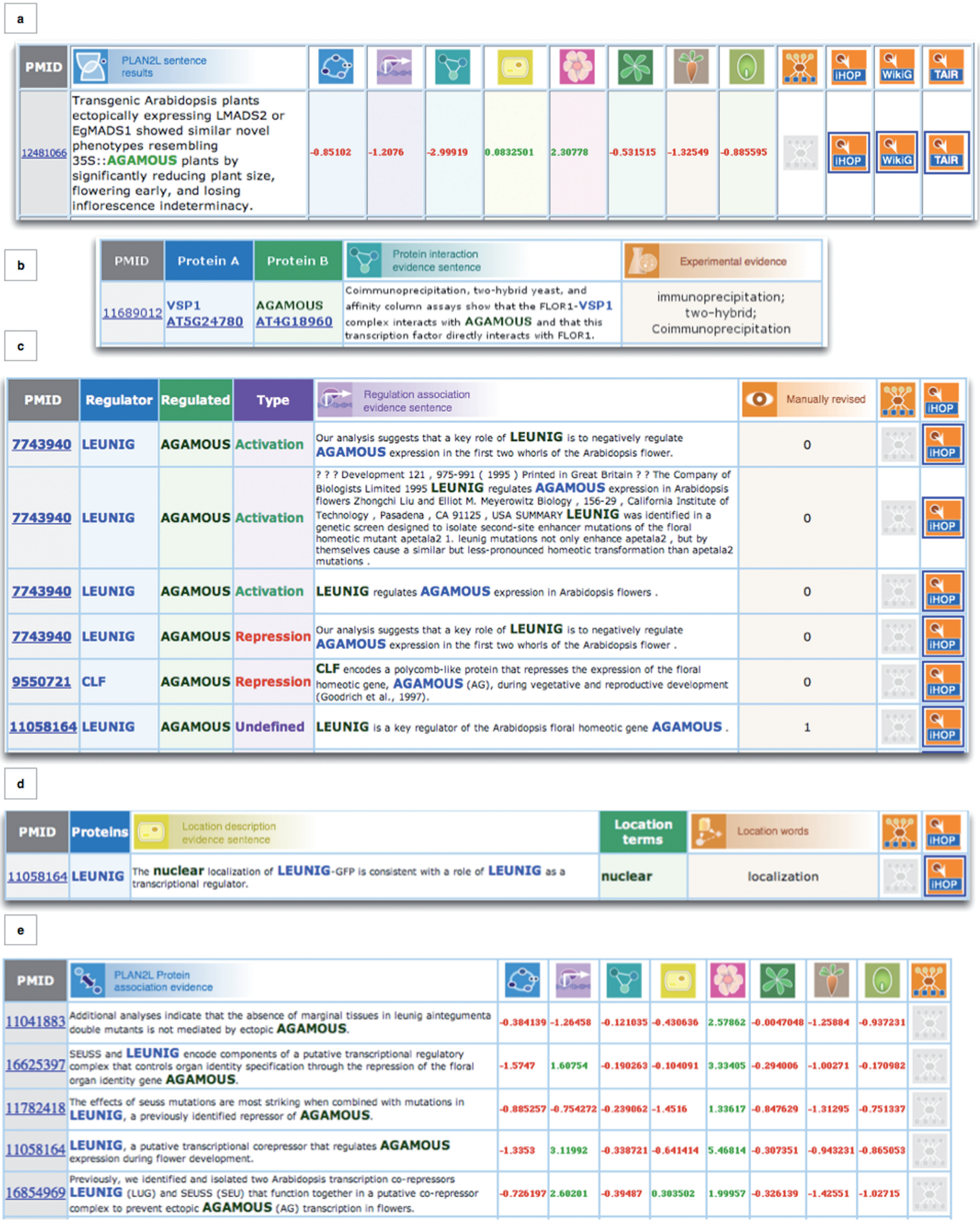
As an example, to illustrate the kind of output generated by PLAN2L, we searched the system using the AGAMOUS (TAIR locus AT4G18960). The basic search for this bio-entity returns a range of descriptive sentences, Figure 2a shows the first of them. Each of these evidence sentences has a range of associated topic scores. By looking at the different sentences and their scores, it becomes clear that this gene is important for the flowering process. This is consistent with information obtained through cross-checking with existing annotations reported for this gene, being described in the TAIR database as a floral homeotic gene that specifies floral meristem and carpel and stamen identity. When trying to determine potential interaction partners of AGAMOUS, by the interaction search facility of PLAN2L, retrieved hits include the FLOR1-VSP1 complex (Figure 2b). Regulatory associations of AGAMOUS on the other side show several hits where this gene is regulated by LEUNIG and CLF (Figure 2c). Examining the results more carefully shows that there are cases where the regulation type was correctly labeled as 'Repression', but in other cases PLAN2L could not determine correctly the actual regulation type, although the directionality (regulator versus regulated gene) was detected correctly. Based on this regulatory association, LEUNIG, the negative regulator of AGAMOUS should be a transcription factor localized in the nucleus. This could be confirmed when querying the location extraction module with this protein, obtaining as first hit the following evidence sentence: 'The nuclear localization of LEUNIG-GFP is consistent with a role of LEUNIG as a transcriptional regulator' (Figure 2d). In case additional evidence sentences for the association of AGAMOUS and LEUNIG in terms of their biological context topic is interesting, for instance, their association might have an implication in the flowering process, this can be easily found out using the association search type. Providing these two entities as queries results in several association evidence sentences (Figure 2e), and when examining their corresponding sentence scores for various topics, it becomes clear that they are in a regulatory association and that they might have an implication in the flowering process.

Implementation, user testing and availability

PLAN2L is mainly written in Python and uses the Zope web application server (www.zope.org) to display the online results. Some of the protein normalization modules are implemented in C, and a collection of additional preprocessing and NLP components are written in Perl. The sentence classifier component relied on the SVMLight package. The relational database server is PostgreSQL, hosts part of the underlying text corpora and lexical resources. The Zope server runs on a HP 360 G5 Intel(R) Xeon(TM) CPU 3.40 GHz machine with 3 GB de RAM. The initial version of this system (PLAN2L-aratreg) has been online since 10 January 2007, and has been improved based on the demands of the EU funded DIAMONDS consortium. Since the first of January 2009, the system had been tested by over 890 visitors. Individual user feedback and requests are being collected to improve the practical usefulness of PLAN2L for the plant science community. The PLAN2L system is available at: http://zope.bioinfo.cnio.es/plan2l. PLAN2L has been tested on the most common browsers (Firefox, Safari and Internet Explorer), accounting for 95.88% of the PLAN2L users. Online help, including documentation and a tutorial using prerun example cases as well as additional details on the system evaluation, and user feedback (FAQ) are provided together with the online application.

DISCUSSION AND CONCLUSION

We described a text mining system, PLAN2L that enables exploration of literature information at different levels of granularity, from retrieval of gene description sentences derived from multiple documents, to qualified biological relations important to understand the sub-cellular context and both physical as well as regulatory interaction networks of bio-entities. PLAN2L extracts biological information from both abstracts as well as full-text articles and



PMID	PLAN2L sentence results										iHOP	WikiG	TAIR
12481066	Transgenic Arabidopsis plants ectopically expressing LMADS2 or EgMADS1 showed similar novel phenotypes resembling 35S::AGAMOUS plants by significantly reducing plant size, flowering early, and losing inflorescence indeterminacy.	-0.85102	-1.2076	-2.99919	0.0832501	2.30778	-0.531515	-1.32549	-0.885595		iHOP	WikiG	TAIR

PMID	Protein A	Protein B	Protein interaction evidence sentence	Experimental evidence

11689012	VSP1
	AT5G24780
		AGAMOUS
		AT4G18960
			Coimmunoprecipitation, two-hybrid yeast, and affinity column assays show that the FLOR1-VSP1 complex interacts with AGAMOUS and that this transcription factor directly interacts with FLOR1.	immunoprecipitation;
				two-hybrid;
				Coimmunoprecipitation

PMID	Regulator	Regulated	Type	Regulation association evidence sentence	Manually revised		iHOP
7743940	LEUNIG	AGAMOUS	Activation	Our analysis suggests that a key role of LEUNIG is to negatively regulate AGAMOUS expression in the first two whorls of the Arabidopsis flower.	0		iHOP
7743940	LEUNIG	AGAMOUS	Activation	? ? ? Development 121 , 975-991 (1995) Printed in Great Britain ? ? The Company of Biologists Limited 1995 LEUNIG regulates AGAMOUS expression in Arabidopsis flowers Zhongchi Liu and Elliot M. Meyerowitz Biology , 156-29 , California Institute of Technology , Pasadena , CA 91125 , USA SUMMARY LEUNIG was identified in a genetic screen designed to isolate second-site enhancer mutations of the floral homeotic mutant apetala2 1. leunig mutations not only enhance apetala2 , but by themselves cause a similar but less-pronounced homeotic transformation than apetala2 mutations .	0		iHOP
7743940	LEUNIG	AGAMOUS	Activation	LEUNIG regulates AGAMOUS expression in Arabidopsis flowers .	0		iHOP
7743940	LEUNIG	AGAMOUS	Repression	Our analysis suggests that a key role of LEUNIG is to negatively regulate AGAMOUS expression in the first two whorls of the Arabidopsis flower .	0		iHOP
9550721	CLF	AGAMOUS	Repression	CLF encodes a polycomb-like protein that represses the expression of the floral homeotic gene, AGAMOUS (AG), during vegetative and reproductive development (Goodrich et al., 1997).	0		iHOP
11058164	LEUNIG	AGAMOUS	Undefined	LEUNIG is a key regulator of the Arabidopsis floral homeotic gene AGAMOUS .	1		iHOP

PMID	Proteins	Location description evidence sentence	Location terms	Location words		iHOP
11058164	LEUNIG	The nuclear localization of LEUNIG-GFP is consistent with a role of LEUNIG as a transcriptional regulator.	nuclear	localization		iHOP

PMID	PLAN2L Protein association evidence									
11041883	Additional analyses indicate that the absence of marginal tissues in leunig aintegumenta double mutants is not mediated by ectopic AGAMOUS.	-0.384139	-1.26458	-0.121035	-0.430636	2.57862	-0.0047048	-1.25884	-0.937231	
16625397	SEUSS and LEUNIG encode components of a putative transcriptional regulatory complex that controls organ identity specification through the repression of the floral organ identity gene AGAMOUS.	-1.5747	1.60754	-0.190263	-0.104091	3.33405	-0.294006	-1.00271	-0.170982	
11782418	The effects of seuss mutations are most striking when combined with mutations in LEUNIG, a previously identified repressor of AGAMOUS.	-0.885257	-0.754272	-0.239062	-1.4516	1.33617	-0.847629	-1.31295	-0.751337	
11058164	LEUNIG, a putative transcriptional corepressor that regulates AGAMOUS expression during flower development.	-1.3353	3.11992	-0.338721	-0.641414	5.46814	-0.307351	-0.943231	-0.865053	
16854969	Previously, we identified and isolated two Arabidopsis transcription co-repressors LEUNIG (LUG) and SEUSS (SEU) that function together in a putative co-repressor complex to prevent ectopic AGAMOUS (AG) transcription in flowers.	-0.726197	2.60201	-0.39487	0.303502	1.99957	-0.326139	-1.42551	-1.02715	

Figure 2. Example case. (**a**) Basic search using AGAMOUS as a query. (**b**) Results obtained when searching for interaction associations using the same query term. (**c**) Regulatory associations for AGAMOUS. (**d**) Localization result for the regulator of AGAMOUS, the LEUNIG transcription factor. (**e**) Association evidence sentences for AGAMOUS and LEUNIG.

integrates different language processing strategies from simple co-occurrence to syntactic/semantic rule-based algorithms and supervised machine learning methods. PLAN2L is intended to be useful for general retrieval, topic-specific retrieval as well as for finding association evidences for user-specified entities and for knowledge and hypothesis confirmation. A similar strategy as used for PLAN2L can be easily adapted to other model organisms as well as specific biological topics with minimal additional manual data preparation.

ACKNOWLEDGEMENTS

Many thanks to Florian Leitner and other users including Ernesto Ortiz, Cameron MacPherson, Nathan McCorkle, Colleen Doherty and Anais Baudot for useful feedback.

FUNDING

ENFIN (LSHG-CT-2005-518254); the BIOSAPIENS (LSG-CT-2003-503265); DIAMONDS (LSHG-CT-2004-512143) projects. Funding for open access charge: ENFIN (LSHG-CT-2005-518254).

Conflict of interest statement. None declared.

REFERENCES

1. Berardini,T.Z., Mundodi,S., Reiser,L., Huala,E., Garcia-Hernandez,M., Zhang,P., Mueller,L.A., Yoon,J., Doyle,A., Lander,G. *et al*. (2004) Functional annotation of the Arabidopsis genome using controlled vocabularies. *Plant Physiol*., **135**, 745–755.
2. Krallinger,M., Valencia,A. and Hirschman,L. (2008) Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol*., **9(Suppl. 2)**, S8–S8.
3. Hoffmann,R. and Valencia,A. (2005) Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, **21(Suppl. 2)**, ii252–ii258.
4. Bevan,M. and Walsh,S. (2005) The Arabidopsis genome: a foundation for plant research. *Genome Res*., **15**, 1632–1642.
5. Swarbreck,D., Wilks,C., Lamesch,P., Berardini,T.Z., Garcia-Hernandez,M., Foerster,H., Li,D., Meyer,T., Muller,R., Ploetz,L. *et al*. (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res*., **36**, D1009–D1014.
6. Schneider,M., Bairoch,A., Wu,C.H. and Apweiler,R. (2005) Plant protein annotation in the UniProt Knowledgebase. *Plant Physiol*., **138**, 59–66.
7. Bajic,V.B., Veronika,M., Veladandi,P.S., Meka,A., Heng,M.-W., Rajaraman,K., Pan,H. and Swarup,S. (2005) Dragon plant biology explorer. A text-mining tool for integrating associations between genetic and biochemical entities with genome annotation and biochemical terms lists. *Plant Physiol*., **138**, 1914–1925.
8. Yoo,D., Xu,I., Berardini,T.Z., Yon Rhee,S., Narayanasamy,V. and Twigger,S. (2006) PubSearch and PubFetch: a simple management system for semiautomated retrieval and annotation of biological information from the literature. *Curr. Protoc. Bioinformatics*, **Chapter 9**, Unit9.7–Unit9.7.
9. Müller,H.-M., Kenny,E.E. and Sternberg,P.W. (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*., **2**, e309–e309.
10. Schmid,H. (1994) Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK, pp. 44–49.
11. Abney,S. (1996) Partial parsing via finite-state cascades. *Nat. Lang. Engg*., **2**, 337–344.
12. Krallinger,M., Morgan,A., Smith,L., Leitner,F., Tanabe,L., Wilbur,J., Hirschman,L. and Valencia,A. (2008) Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. *Genome Biol*., **9(Suppl. 2)**, S1–S1.