BMC Bioinformatics

SOFTWARE Open Access

LAITOR - Literature Assistant for Identification of Terms co-Occurrences and Relationships

Adriano Barbosa-Silva[1,2,3], Theodoros G Soldatos[3,4], Ivan LF Magalhães[2], Georgios A Pavlopoulos[3], Jean-Fred Fontaine[1], Miguel A Andrade-Navarro[1], Reinhard Schneider[3], J Miguel Ortega[2*]

Abstract

Background: Biological knowledge is represented in scientific literature that often describes the function of genes/proteins (bioentities) in terms of their interactions (biointeractions). Such bioentities are often related to biological concepts of interest that are specific of a determined research field. Therefore, the study of the current literature about a selected topic deposited in public databases, facilitates the generation of novel hypotheses associating a set of bioentities to a common context.

Results: We created a text mining system (LAITOR: ***L****iterature* ***A****ssistant for* ***I****dentification of* ***T****erms co-****O****ccurrences and* ***R****elationships*) that analyses co-occurrences of bioentities, biointeractions, and other biological terms in MEDLINE abstracts. The method accounts for the position of the co-occurring terms within sentences or abstracts. The system detected abstracts mentioning protein-protein interactions in a standard test (BioCreative II IAS test data) with a precision of 0.82-0.89 and a recall of 0.48-0.70. We illustrate the application of LAITOR to the detection of plant response genes in a dataset of 1000 abstracts relevant to the topic.

Conclusions: Text mining tools combining the extraction of interacting bioentities and biological concepts with network displays can be helpful in developing reasonable hypotheses in different scientific backgrounds.

Background

The richness of information generated by different research groups is sometimes focused on issues that lack explicit connection with those generated by colleagues from other groups. However, currently, there are available literature mining techniques that permit to connect the knowledge generated by distinct groups and improve the understanding of some key points of their research [1]. Text mining machines have been created to mine the biological information in a trial to establish new biological concepts from previous knowledge [2-4]. These machines were proven to be reliable in extracting biological facts either analyzing full text [5,6] or just condensed information present in the abstracts of scientific papers [7,8] as stored in the MEDLINE database.

Text mining techniques for information-retrieval comprise some basic steps: to find relevant articles in the research field of interest; to identify the biological entities cited in the text, as well as to disambiguate confuse bioentity names (i.e. genes and proteins) within and among distinct species; to infer putative relationships between bioentities based on co-occurrence of biological terms in the same article, abstract, sentence or phrase [2]. Recently, AliBaba has been developed to graphically visualize information on associations between biological entities extracted from PubMed using pattern matching and co-occurrence filtering (http://alibaba.informatik.hu-berlin.de/, [9]). Later, a system called NetSynthesis [10] has been developed to permit the controlled building of biomolecular networks by users, where the searching criteria on PubMed are customized by using parse tree query language [11]. However, these systems do not permit the integration of customized dictionaries on their algorithm.

We present here a system called LAITOR (***L****iterature* ***A****ssistant for* ***I****dentification of* ***T****erms co-****O****ccurrences and* ***R****elationships*). This software was developed to normalize the bioentities names tagged in the abstracts to a user defined protein dictionary; as well as to extract their co-occurrence, along with other protein or important biotic/abiotic stimuli terms, the later implemented

* Correspondence: miguel@icb.ufmg.br
[2]Laboratório de Biodados, Dpto. de Bioquímica e Imunologia, ICB - UFMG. 31270-901, Belo Horizonte - MG, Brazil

as a customized concept dictionary. Such co-occurrences are extracted taking into consideration the presence of terms in the same sentence of scientific abstracts and adopting a set of rules to filter bioentity pairs that occur in several sentence structures (see details in Implementation). The software performed as a greatly precise method. Here, it has been used to mine protein co-occurrences related to green plant-pathogen interactions.

Implementation

Abstracts retrieval

In order to retrieve scientific abstracts related to green plants that would be related to defense mechanisms, we have used the system MedlineRanker [12]. Two MeSH http://www.nlm.nih.gov/mesh/ terms (Host-Pathogen Interactions AND Plants) have been used as "training dataset" to rank 10,000 recently-published abstract from the whole MEDLINE database. After the MedlineRanker analysis we retrieved the top 1,000 PubMed IDs from the generated rank to be loaded as "application dataset" for the next steps of our analysis [Additional file 1].

Protein tagging

LAITOR is optimized to work by analyzing tagged scientific abstracts. For this purpose, we adopted the NLPROT [13] program as LAITOR's protein tagger. The plain text format (-f txt) must be chosen for the NLPROT output file, where bioentity names present in the text are tagged between "<n>" and "</n>" tags. The tagged protein names are filtered according to a user-defined bioentity dictionary, in our case study: a plant protein name and synonym dictionary.

Protein Dictionaries

Two protein dictionaries have been generated for the development of LAITOR. The first (named human proteins dictionary) created for the evaluation of LAITOR performance (explained below) in the BioCreative II Interaction Article Subtask (IAS) [14]. The second (named plant protein dictionary) has been used in the identification of co-occurring of green-plant protein pairs retrieved for abstracts related to host-pathogen interactions.

The human protein dictionary has been created by using all the protein records deposited for *Homo sapiens* [NCBI Taxonomy id: 9606] in the UniProt-SwissProt-TrEMBL (UP-SP-TR) database. In this dictionary, the definition(s) and synonym(s) for all human UP-SP-TR proteins are included. Furthermore, for each record, the corresponding NCBI Gene symbol and synonyms were used to enrich the representative terms of said protein. At the end, the human proteins dictionary is composed by 87,537 records (IDs), comprising a total of 112,686 distinct protein terms, which have been completed by the addition of 40,234 supplementary terms from the NCBI Gene database.

Additionally, specific genes names and synonyms for every organism deposited in the NCBI Taxonomy database that have gene records in the NCBI Gene database have been used to create LAITOR readable dictionaries. To use these dictionaries, users must inform the taxonomy identification number (Taxonomy ID) for the preferred organism followed by the extension ".dictionary" (e.g. "9606.dictionary" for "Homo sapiens" genes) during set up, as explained at LAITOR's documentation file.

For the plant dictionary, the complete Gene tab-delimited database from Entrez website has been downloaded (5,317,958 records), which comprises 505,403 different organisms (Taxonomy IDs - TAXIDs). To filter only those records related to green-plant proteins, we used the NCBI Taxonomy database to select from the Gene table only those records with a TAXID corresponding to Viridiplantae organisms, which included 99,488 different records. At the end, the plant protein dictionary contained 148 plants organisms (0.02% of total organisms) and a total of 237,077 Gene records (4.45%), which included 217,224 distinct protein symbols and 62,521 synonyms (see one example for the Gene PR1 of *Arabidopsis thaliana* [GenBank: 815949] in Additional file 2).

The resulting table displays two columns: one for the bioentity names, and the second with their respective synonyms so that it can exist as lines (records) as synonyms for each bioentity name (Additional file 2).

Name ambiguity

Another aspect explored by LAITOR, is how to handle gene name ambiguity. The strategy of using the Taxonomy database to limit the number of used entries reduced the possibility of inclusion of names of other organisms which would cause ambiguity among terms. However there are terms that commonly occur for more than one organism, or different proteins from the same organism that share the same name or synonym. To cope with this, LAITOR creates a tag file in which the ambiguous terms identified in the analysis are normalized to the same name in the protein dictionary. Such terms that match multiple protein names or that are synonyms of multiple protein names are marked in the LAITOR output. This warns users about the possibility of misinterpretation for such a term.

Concepts Dictionary

In order to check the co-occurrence and likely involvement of plant proteins names along with biotic and abiotic stimuli names, a list of previously known stimuli and their synonyms has been provided as Concept Dictionary (for example: Jasmonic Acid, Jasmonate and JA were included as the same concept). Both, Protein and

Concept Dictionaries are available as additional material [Additional files 3 and 4].

Additionally, in order to attend different contexts, we have populated all the sub-headings of NCBI's Medical Sub Headings (MeSH) Trees (available at http://www.nlm.nih.gov/mesh/trees.html) as LAITOR's concepts dictionaries, as explained at LAITOR's documentation.

Biointeractions Dictionary

A list representing the different types of interactions or relationships between proteins was generated based on previously published list [4,15]. It is composed by 76 terms, which have been included together with a total of 886 synonyms as seen in Additional file 5, Table S2. Considering all terms, the biointeraction dictionary in its entirety is composed of 963 different words.

Co-occurrence analysis

Once the abstracts to be analyzed had been retrieved and tagged for protein and gene names, biointeractions and concepts, LAITOR was used to perform a co-occurrence analysis [see Additional file 6].

At the sentence level, each line of the tagged abstracts was divided at every full stop (".") punctuation sign. We paid special attention to the presence of these full stop marks in alternative positions that did not indicate the end of the period, as in the case of species names (for example: *A. thaliana*) or protein names (for example: PDF1.2 protein).

Initially the whole abstract is screened to store the occurrence of all bioentity names. After storage of all names, each protein name is checked for its occurrence in each of the separated sentences. If a bioentity term is found, let us name this term as "Pair 1", the script checks the occurrence of a second bioentity name, "Pair 2", different from Pair 1 in the same sentence. To avoid redundancy, the script checks on-the-fly if Pair 2 is a synonym of the previously identified Pair 1 and discards such cases.

It has been previously published that 90% of the biointeractions among proteins documented in the literature adopts the pattern "Protein-Biointeraction-Protein" [16], this pattern being chosen by approaches like iHOP [15] and HomoMINT [17]. Nevertheless, we adjusted LAITOR to identify other patterns of Protein-Protein or Protein-Concept co-occurrence, as explained below.

The co-occurrences identified by LAITOR are classified into four types. From the most to the least stringent, these types are:

Type 1: Both co-occurring protein names/synonyms must not refer to the same protein (common for all types of co-occurrences), they must be present in the same sentence of the abstract and, additionally, it is required that a term from the Biointeractions Dictionary occurs in between the considered terms. An extra optional step is the identification of a biological stimuli (represented as a term from the Concepts Dictionary) term anywhere in the sentence, which is then associated to the interacting pair;

Type 2: Same as Type 1, except that the biointeraction may occur anywhere in the sentence;

Type 3: Same as Type 1, except that the occurrence of a biological term in the sentence is not required;

Type 4: All the pairs of co-occurring protein names/synonyms mentioned in the abstract are considered, whether they are in the same sentence or not.

Thus, when LAITOR performs under type 4, the other co-occurrence types are included.

Multiple co-occurrences of type 1, 2 and 3, might happen in a given sentence. To cope with this, our system was adapted to perform an overlapped search. This means that in cases where two proteins (A and B) occur along with the same biointeraction, like in the sentence "A and B regulate C", the pairs "A-regulate-C" and "B-regulate-C" are identified as type 1 co-occurrences. Note that the co-occurring pair "A-B" will be assigned type 2. Moreover, in more complex sentences such as "A is regulated by B and activates C", the system will retrieve as co-occurrences of type 1 "A-regulated-B", "A-regulated, activates-C', and "B-activates-C" (together with type 2 "A-regulated, activates-B" and type 2 A-regulated, activates-C) thus over predicting the number of different bio-interactions between the A, B and C proteins. However such complex sentences may not be very frequent. In order to determine if they are a serious problem, we performed a series of manual evaluations of the results of LAITOR's analysis on several abstract datasets.

Performance evaluation

Protein term co-occurrences at sentence level of scientific abstracts might be potentially useful for the prediction of literature-based protein-protein interactions. Therefore, we have tested the performance of LAITOR to find protein-protein interaction data in abstracts. For this purpose, we have used the BioCreative II test dataset for the Interaction Article Subtask (IAS) as gold standard [14]. This "performance evaluation dataset" is composed of relevant (3,529) and irrelevant (1,957) abstracts for the curation of protein-protein interactions present in the MINT and IntAct databases [18]. Once LAITOR identifies a co-occurring protein pair in an abstract, this is considered to be positively (relevant) classified. After the classification of all gold standard abstracts the precision and recall are calculated for each of the four co-occurrence types (1-4), and the performance compared to methods participating in the BioCreative II challenge. A receiver operating curve (ROC) was created by using the package ROCR [19]. Positive and negative performance

evaluation datasets are provided as additional material [Additional file 7].

Network representation

A protein and stimuli co-occurrence analysis created by LAITOR from PubMed abstracts is parsed from a general output file into a tab-delimited text file (extension .co) that is used as input by most network visualization software. As default, LAITOR generate inputs for two of these programs: EMBL Medusa [20] and EMBL Arena3D [21], which provide networks in one- and multi-dimensional charts, respectively, enabling the complex output generated by LAITOR to be efficiently handled.

Results and Discussion

LAITOR's developmental pipeline

LAITOR has been developed by combining a flexible rule-based method together with a pre-defined vocabulary match approach. Figure 1 illustrates the pipeline for LAITOR's development, which is explained in detail in the following sections.

LAITOR uses as input a set of scientific abstracts as stored in the records of the MEDLINE database. Abstracts are analyzed individually for co-occurrences, which are extracted and classified into four types according to the rules described in Implementation section. Additional file 8, Figure S1 exemplifies a tagged sentence extracted from the PubMed article identified by PMID 19061405. The co-occurrence analysis starts by (i) the creation of a list with the occurring bioentities (proteins or genes, [see Additional file 2, Table S2]) and stimuli names present in precompiled dictionaries (see Implementation), for the whole abstract. In the example the names detected were: HSP90, RAR1 and SGT1. (ii) Further, each sentence is queried for the co-occurrences of different bioentity names establishing pairs. In this

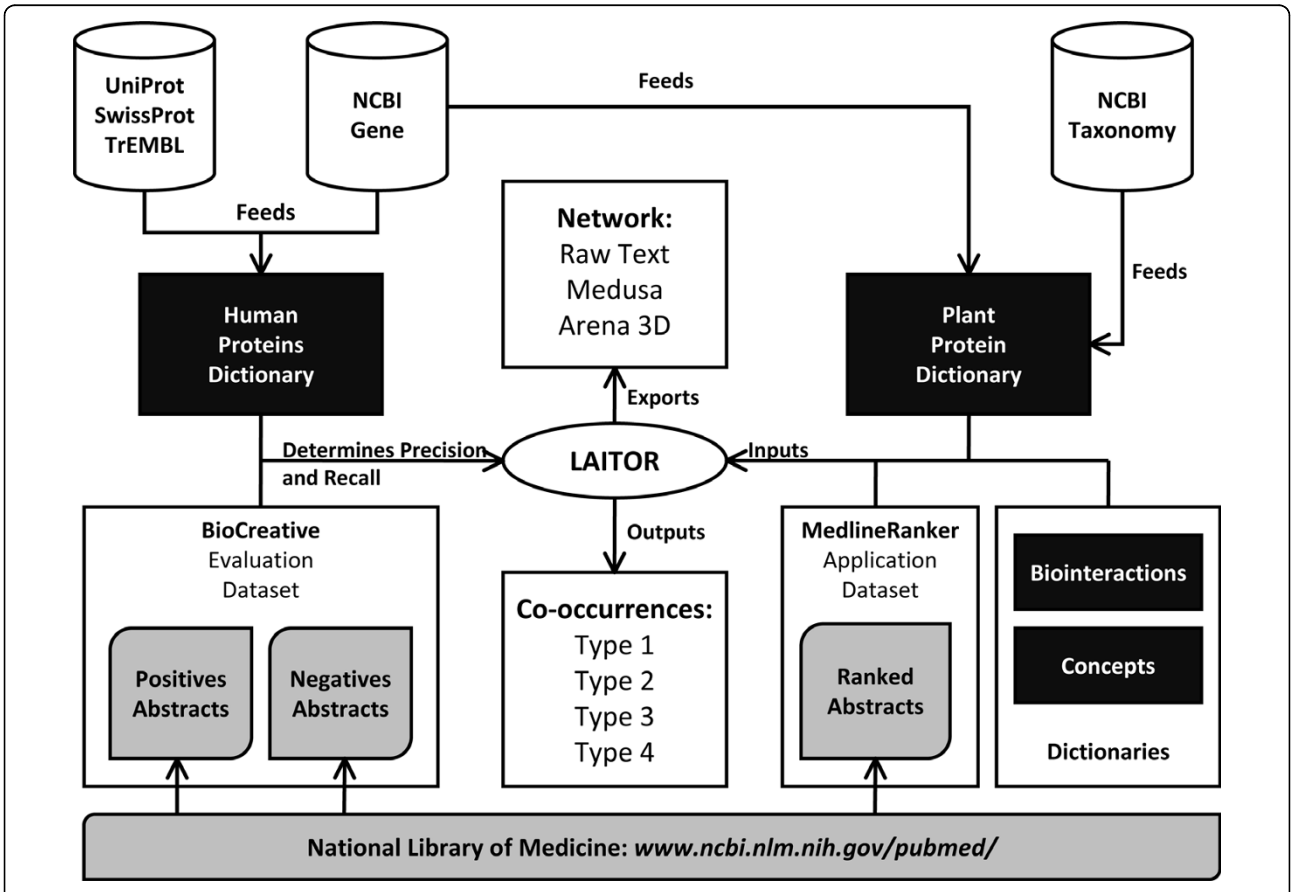

Figure 1 Pipeline for LAITOR's development. LAITOR's has been evaluated for correct classification of abstracts relevant to curation of protein-protein interactions from BioCreative II challenge (evaluation dataset). Co-occurrences of terms from the human proteins dictionary in these abstracts have been used as an indicator of relevance. Precision and recall have been measured as 0.89 and 0.48 respectively for the type 1 of co-occurrences. Afterwards, for abstracts ranked to be related to host-pathogen interactions in plants (application dataset), LAITOR has generated a list of co-occurrences and a network representation of the terms from the plant protein dictionary which could be found in this dataset. **Symbol key**: dark rectangles: dictionaries; grey shapes: abstract datasets; cylinders: public databases; ellipse: LAITOR script; white rectangles: LAITOR's outputs;

example the co-occurrences of the types 1, 2 and 3 are defined as follows.

Type 1: the pairs HSP90 and RAR1, as well as, HSP90 and SGT1 were both extracted with the interleaved biointeraction term "interact" associating the members of each pair (see Additional file 5, Table S2 for example of a Biointeraction term representation in the Biointeraction Dictionary).

Type 2: the pair RAR1 and SGT1 was extracted, with the occurrence of the biointeraction "interact" in the same sentence, however not interleaved.

Type 3: Other co-occurrences of the protein terms (HSP90, RAR1 and SGT1) found in the same sentence were considered as co-occurrences of type 3.

Furthermore, the combinations of all the bioentity names identified in the abstract, except synonyms, are considered as co-occurrences of type 4 (see Implementation for explanation).

Evaluation against BioCreative II

LAITOR was compared to the Interaction Article Subtask (IAS) of BioCreative II text mining challenge [14]. Table 1 shows that LAITOR could predict abstracts considered relevant for the curation of protein-protein interaction (evaluation dataset) with a maximum precision of 0.89 and a corresponding recall of 0.48 considering type 1 co-occurrences (bioentities co-occur within the same sentence, and they are interleaved by some biointeraction term; see Implementation for a detailed description). Among the 19 evaluated methods for the IAS task, LAITOR's predictions (considered to be a non SVM-based prediction) demonstrated to be the second most precise method keeping a reasonable sensitivity (recall) index. In predictions using the co-occurrence types 2-4, which do not require the presence of a biointeraction term, LAITOR produced results with a precision ranging from 0.82 to 0.85, a recall ranging from 0.61 to 0.70 and a F-score ranging from 0.60 to 0.72 (See Table 1 for values for each type). This implies that LAITOR's detection of protein co-occurrences with biointeraction terms improves precision that the expense of a small reduction of recall and therefore increases the likelihood of filtered protein pairs from such abstracts will indeed display biologically relevant fact.

Table 1 LAITOR evaluation against BioCreative II IAS subtask.

Type	Precision	Recall	F-score	Accuracy
1	0,89	0,48	0,63	0,63
2	0,85	0,61	0,71	0,68
3	0,83	0,62	0,72	0,68
4	0,81	0,70	0,60	0,60

Manual examination of some false-positive abstracts showed that although the biointeraction was not correctly identified, the selected sentences described a relevant biological interaction. For example, this sentence: "*Taken together, these results suggest that loss of RPA1 activates the Chk2 signaling pathway in an ATM-dependent manner*" (PMID: 15620706), was interpreted as RPA1 activates Chk2 because the term "activates" was found between the protein names RPA [Entrez Gene id: 6117] and Chk2 [Entrez Gene id: 11200]. The sentence actually indicates a different relation but it is informative in terms defining a functional relation between these two proteins.

In further comparison of LAITOR's performance with other methods from the BioCreative II challenge in order to correctly classify the IAS gold standard abstracts, we scored LAITOR's prediction of these abstracts with a score S = 5-T where T, that is the type of co-occurrence, ranges from 1 to 4, according to the presence of at least one sentence displaying a co-occurrence of types 1 to 4 (adopting S = 0 when no co-occurrence is detected in the abstract). Then, we calculated the area under the receiver operating curve (AROC), corresponding to 0.74 (Figure 2).

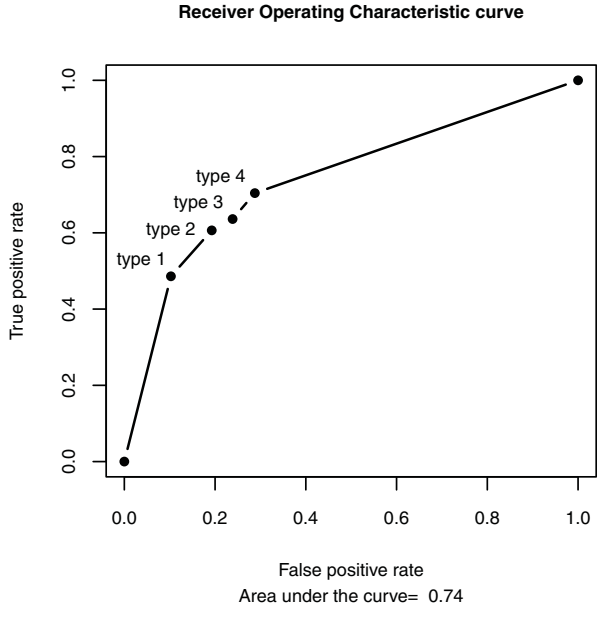

Figure 2 Receiver Operating Characteristic (ROC) curve for LAITOR predictions. The corresponding area under the curve (AROC) is 0.74, calculated using the four types of interactions found in such an abstract as a measurement of its overall predictive power. Note that higher types include the lowest ones.

Case-study: co-occurrence analysis of terms related to a plant-pathogen interaction dataset

We performed a case study by applying LAITOR to generate a list of green plant's protein co-occurrences related to host-pathogen interactions. Plants respond to diverse environmental stimuli, biotic and abiotic, by mobilizing specific protein networks used to identify its source and to activate the cellular mechanisms to surpass changes caused by stressful conditions. Commonly, the adaptative responses found in plants are flexible and the same subset of proteins/genes can be activated by different types of stimuli, including defense against pathogens or tolerance under severe environmental conditions [22]. Therefore, a system like LAITOR used in this context should be expected to be useful in suggesting novel roles for known protein interactions.

Moreover, this topic is important for plant biotechnological and physiological studies, since (i) diverse economically important crops are attacked by several phytopathogens in the field, which is prejudicial for agricultural practices along the world [23], and (ii) cultivated lands are often affected, for instance, by severe abiotic conditions such as high salinity [24], drought [25], over-flooding [26] or extreme cold [27]. As a result of this interest, during the last few decades several efforts have been dedicated to characterize these mechanisms, which resulted in a fair amount of related publications deposited in MEDLINE. These data comprises proteins or entire protein networks that are used by plants, as well as chemicals identified to have a key role in the signaling pathways that establish the plant adaptative responses. Jasmonic acid (JA) [28], ethylene (ET) [29] and salicylic acid (SA) [30,31] are examples of phytohormones employed by plants that act as signaling molecules in diverse defense response networks [32]. This wealth of data facilitates a text mining procedure such as LAITOR.

A total of 1,000 abstracts on the topic of green plant's host-pathogen interactions were retrieved with MedlineRanker [26] (application dataset) and analysed with LAITOR, of which 79 displayed at least one filtered co-occurrence. From the total 9,823 parsed sentences (including titles), 116 provided co-occurrences of the different types and pairs of bioentities (Table 2). A total of 263 pairs were retrieved from the application dataset.

In this dataset, a total of 68 different biointeraction terms could be identified among the co-occurring pairs, considering that the co-occurrences of type 3 do not restrict the filtering of biointeraction terms in the sentences. The top 10 most-common biointeraction terms and their frequencies within the application dataset are shown in Additional file 9, Table S3.

Table 2 Survey of sentences and pairs extraction using the LAITOR algorithm on application dataset.

Type	Sentences	Pairs
1	25	52
2	35	66
3	24	27
4	N. A.*	21
Total	116	263

*N.A.: not applicable, as LAITOR does not consider sentences to extract co-occurrences of type 4.

Network visualization

LAITOR generates a network file relating the co-occurrences extracted. The nodes represent bioentities and the edges their co-occurrences in the set of abstracts used as input. Each edge is annotated by the type of co-occurrence from strictest (type 1) to least strict (type 4).

As an example we generated a network for a total of 51 nodes and 143 edges found in the application dataset only representing the co-occurrences of type 1, in order to reduce the complexity of the network [Additional file 10, Figure S2]. We illustrate the relevance for the analysis of using the dictionary of concepts in Additional file 11, Figure S3. It can be noticed that the displayed subnetwork with 9 proteins (Additional file 11, Figure S3A; this is one of the subnetworks of the network represented in Additional file 10, Figure S2) gained two more members (catalase and SOD) when the concepts "oxidative stress" and "jasmonic acid" were also considered [see Additional file 11, Figure S3B]. The top 10 most-common terms present in the concept dictionary and their observed frequencies within the application dataset are shown in Additional file 12, Table S4.

Hypothesis generation example

One of the most interesting applications of a co-occurrence based text mining analysis is the support given to new hypothesis generation [33,34]. Here we explore this functionality in LAITOR by examining the involvement of a common member of the photosystem response and disease signaling in *Arabidopsis* [see Additional file 13, Figure S4].

Accessing the abstracts analyzed by LAITOR and listed in Additional file 13, Figure S4B we observe that the *Arabidopsis thaliana* gene *RPS4* (RESISTANT TO *P. SYRINGAE* 4 [Entrez GeneID: 834561]) confers resistance to the bacterial pathogen *Pseudomonas syringae* carrying the avirulence gene *avrRps4* [Entrez GeneID: 3555344, PMID: 8589423]. We can use LAITOR to find genes that could be hypothetically involved in resistance mechanisms regulated by *RPS4*. LAITOR associates this gene to several other genes. In

the topic of resistance against pathogens *EDS1* stands out: we can see that RPS4 requires the gene *EDS1* (ENHANCED DISEASE SUSCEPTIBILITY1 [Entrez GeneID: 823964]) to confer *avrRps4*-independent resistance in tomato plants transiently expressing *RPS4* [PMID: 15447648]. Using LAITOR we can see that there is another pathogen resistance gene that, similarly to *RPS4*, also requires EDS1, although in a different context [see Additional file 13, Figure S4A]. This is *PAD4* (PHYTOALEXIN DEFICIENT4 [Entrez GeneID: 824408]), which confers resistance against the phloem-feeding green peach aphid (GPA) infesting *Arabidopsis*, and also requires its signaling and stabilizing partner *EDS1* [PMID: 17725549].

Now, LAITOR shows that *PAD4* is related to three genes: *LSD1* [Entrez GeneID: 827786], *SIZ1* [Entrez GeneID: 836163], and *WIN3* [Entrez GeneID: 831173]. In more detail, a win3-T *Arabidopsis* (*WIN3*) mutant shows greatly reduced resistance to the bacterial pathogen *Pseudomonas syringae* carrying the avirulence gene *avrRpt2* and expression of this gene at an infection site partially requires *PAD4* [PMID:17918621]. The small Ubiquitin-like Modifier E3 Ligase (encoded by the gene *SIZ1*) interacts epistatically with PAD4 to regulate pathogenesis related gene expression and disease resistance [PMID: 17163880]. Finally, the disease resistance signaling components EDS1 and PAD4 are essential regulators of the cell death pathway controlled by LSD1 in *Arabidopsis* [PMID: 11595797].

Given the fact that both RPS4 and PAD4 require EDS1, one could explore weather or not these three known targets of PAD4 (SIZ1, WIN3, LSD1) could also be targets of RPS4, a fact not represented in the literature as evidenced by the absence of matches for the PubMed query "RPS4 AND (SIZ1 OR WIN3 OR LSD1)" [see Additional file 13, Figure S4C]. This example highlights the potential of LAITOR to unearth undiscovered public knowledge [35] using the condensed information of abstracts [36]. Thus, the system is able to extract precise information from the sentences in abstracts that can be used to generate new hypotheses.

Current limitations of LAITOR

The main limitations of the system can be classified as those producing false positives and those producing false negatives co-occurring pairs. False negatives are mainly due to terms not recognized to be gene/protein names, and to failure to recognize a biointeraction. The first problem can be solved by improving the tagging mechanism and the underlying dictionaries. We approach the second by manually adding to the dictionary of biointeractions those that we find to be common. Some false positives co-occurrences are caused due to misrecognition of gene/protein names and/or biointeractions. The current tagging is conservative and therefore does not increase false identification of gene/protein names (see Material and Methods); it actually constitutes the slower step of the method. This ensures that the identified biointeractions actually point to relevant sentences. Most falsely identified biointeractions were originating from sentences with large numbers of genes. We are considering adding an option to dismiss sentences with more than two gene/proteins as a choice for users requiring greater accuracy.

Comparison to other similar systems specialized in co-occurrence extraction

LAITOR is, as far as we know, the only method of co-occurrence detection along with customized that has been designed as standalone software to be included as part of other systems. However, LAITOR has some methodological particularities that merit comparison to recently developed systems that apply biological term co-occurrence as part of their functionalities.

STRING [37] is a web resource focused on a pre-compiled list of protein-protein interactions extracted by different methods. STRING uses Natural Language Processing [38] to search for statistically relevant co-occurrences of gene names, and also extract a subset of semantically specified interactions. Similarly, iHOP [15] is focused on the navigation of the scientific literature using biological term co-occurrence networks as a natural way of accessing PubMed abstracts. iHOP's text mining approach retrieves and ranks all the sentences for a given gene according to significance, impact factor of published journal, publication date or syntax structures where the gene occurs (i.e. gene-biointeraction-gene pattern). Furthermore, iHOP uses MeSH terms as source for information about gene function, what could be comparable to LAITOR's concepts search. Similarly to iHOP, co-occurrence methods have been developed for plant-directed literature analysis using *Arabidopsis thaliana* as a model [39]. This system, called PLAN2L, also classifies the extracted terms and co-occurrences as being related to physical and regulatory events for developmental processes, as well as with sub-cellular context, for that PLAN2L uses from co-occurrence to syntactic/semantic rule-based algorithms and supervised machine learning methods.

Although being designed for different purposes, we compared the features among LAITOR, STRING and iHOP (Table 3), once that these systems use biological term co-occurrences as part of their text mining strategies.

The main novelty of LAITOR in comparison to previous published software, besides the implementation of the concepts search, is the possibility to customize the dictionaries to be considered in the co-occurrence analysis (bioentities and biointeractions).

Reflecting this flexibility, we have included in the current LAITOR's distribution package a set of genes

Table 3 Comparison of features between LAITOR, STRING and iHOP.

Features	LAITOR	STRING	iHOP
Software type	Command-line script	Website application	Website application
Information sources	Any type of text loaded by the user (e.g. PubMed, OMIM, Wikipedia)	PubMed, SGD, OMIM, The Interactive Fly.	PubMed
Text limit	Any type of tagged text	Only abstracts	Only abstracts
Protein name tagging	Depends of external software (NLPROT), confers against loaded dictionary	YES, filtered by selected organism	YES, filtered by selected organism
List of used synonyms	Flexible user-based dictionary input	Variety of pre-compiled dictionaries	Entrez Gene, FlyBase, UniProt and HUGO Nomenclature Committee
Explores biological concepts	YES, finds user loaded concepts linked to a co-occurring pair at sentence level.	NOT	YES, searches species names, MeSH and compound terms
Extracts co-occurrences among proteins	YES, considering whole text and isolated sentences	YES, limited to the whole abstract	YES, at sentence level only
Extracts interactions among proteins	YES, considering a biointeractions dictionary defined by the user	NOT	YES, considering a pre-compiled biointeractions dictionary
Terms co-occurrences	YES, extracts terms mentioned in the full text or in isolated sentences at different structures which are scored differently	YES, extract terms mentioned together in abstracts, more often than what would be expected by chance based on their overall occurrence	YES, extracts terms mentioned in isolated sentences
Semantic understanding	YES, extracts the biointeractions and concepts linked to an extracted pair at sentence level in different co-occurrence types	NOT, only checks co-occurrences of terms	YES, extracts the biointeractions and concepts linked to an extracted pair at sentence level
Co-occurrence frequency report	YES, displays the frequency that a pair co-occurred in general sentences, and for each found biointeraction	YES, only the number of times that a pair co-occurred in each abstract	NOT
Outputs network	YES, in tabular format and in pre-compiled formats for third-part applications (ARENA3D, MEDUSA)	YES, displays the network in the browser from selected abstracts	YES, users can build a network by adding a set of nodes per time by selecting desired abstracts

symbols/synonyms dictionaries pre-compiled from GeneDB records and divided by all the organisms deposited NCBI's Taxonomy Database http://www.ncbi.nlm.nih.gov/Taxonomy, in addition to the green plants dictionary used in the test case described above, making it possible to use LAITOR virtually for any species with gene data. Furthermore, in order to provide users with a wide set of relevant dictionaries for the concepts search, we compiled LAITOR's concepts dictionaries for each of the NCBI's Medical Subject Headlines (MeSH) main tree structures http://www.nlm.nih.gov/mesh/trees2008.html. The information about how to use these dictionaries is available in the documentation file of LAITOR.

Conclusions

We presented here a new text mining software component called LAITOR, which performs co-occurrence analysis of scientific abstracts where biological entities are filtered from the tagged text using a user defined bioentity dictionary as support. Subsequently, a rule based system is used to detect the co-occurrence of such names along with biointeraction and, optionally, other biological terms provided by the Concepts Dictionary (such as stimuli), in scientific abstracts. We provide here an example of knowledge discovery by applying LAITOR to a subset of abstracts published about defense mechanisms in *Arabidopsis*. In this example, genes from different contexts (light and pathogen responses) have been placed together. Additionally, we have explored a new feature in biological text mining, which is the application of a user pre-defined concept dictionary in order to mine the literature and gather facts previously not reported together. Here, we have evidenced that the inclusion of the concept "oxidative stress" in the analysis conducted for *Arabidopsis* abstracts has brought two new members to a predicted gene network thought to be related to "jasmonic acid" signaling pathway.

Taken together, our results suggest that LAITOR is very precise in identifying abstracts of scientific literature mentioning interactions between genes and proteins. LAITOR is able to extract very variable types of

protein co-occurrences, no matter how they have been cited in the abstract. In our future work, we intend to adapt LAITOR components to an on-line tool, in which users, as well as computers (using the web services technology) will be able to load their desired literature and perform a LAITOR-based co-occurrence analysis that, integrated with other databases (for example, KEGG [40]), will provide a flexible framework for literature mining-based knowledge discovery.

Availability and requirements

LAITOR is distributed under the General Public License (GPL). Access http://laitor.sourceforge.net to obtain LAITOR's repository and its documentation from SourceForge.net.

LAITOR requires Linux as operating system, PHP version 5.3.2 or superior, MySQL version 5.0.45 or superior to run. Additional information is found on-line in the LAITOR documentation file.

Additional file 1: Application dataset.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-70-S1.TXT]

Additional file 2: Table S1: Example of a protein term and its synonyms representation in the Protein Dictionary.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-70-S2.DOC]

Additional file 3: Plant protein dictionary.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-70-S3.TXT]

Additional file 4: Concepts dictionary.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-70-S4.TXT]

Additional file 5: Table S2: Example of a biointeraction term represented in the Biointeraction Dictionary.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-70-S5.DOC]

Additional file 6: LAITOR co-occurrence pipeline.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-70-S6.PPT]

Additional file 7: Performance evaluation dataset.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-70-S7.ZIP]

Additional file 8: Figure S1: Example of a tagged phrase output.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-70-S8.TIFF]

Additional file 9: Table S3: Top-10 biointeraction terms most cited in the green plants application analysis.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-70-S9.DOC]

Additional file 10: Figure S2: Full network created by LAITOR from application dataset.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-70-S10.TIFF]

Additional file 11: Figure S3: Co-occurrence sub-networks generated by LAITOR.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-70-S11.TIFF]

Additional file 12: Table S4: Top-10 concepts terms mostly cited in the co-occurrence analysis.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-70-S12.DOC]

Additional file 13: Figure S4: Hypothesis generation supported by LAITOR output.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-70-S13.TIFF]

Abbreviations

PMID: PubMed Identifier; IAS: Interaction Article Subtask.

Acknowledgements

We are grateful to Venkata Satagopam, Evangelos Pafilis and RS for the training given to ABS at EMBL-Heidelberg during his external Ph.D training in Germany. This work has been developed as part of ABS Ph.D thesis which has been sponsored by Foundation for Research Support of Minas Gerais State (FAPEMIG), Brazilian Ministry of Education (CAPES/ME) and Brazilian Ministry of Science and Technology (CNPq/MCT). This work was supported by grants from Germany's National Genome Research Network (Bundesministerium für Bildung und Forschung) and from The Helmholtz Alliance on Systems Biology (Helmholtz-Gemeinschaft Deutscher Forschungszentren).

Author details

[1]Computational Biology and Data Mining Group, Max-Delbrück Center for Molecular Medicine. Robert-Rössle-Strasse. 10, D-13125, Berlin, Germany. [2]Laboratório de Biodados, Dpto. de Bioquímica e Imunologia, ICB - UFMG. 31270-901, Belo Horizonte - MG, Brazil. [3]European Molecular Biology Laboratory, EMBL-Heidelberg, Meyerhofstrasse 1, 69117, Heidelberg, Germany. [4]LIFE Biosystems GmbH. Poststrasse 34, D-69115, Heidelberg, Germany.

Authors' contributions

ABS created the main idea of the article. ILFM and TGS helped in the development and initial discussion of LAITOR algorithm. ABS and ILFM developed the prototype scripts. ABS developed the final scripts. TGS and RS provided the biointeraction dictionaries. GAP idealized the graph outputs. ABS performed the evaluation and application experiments. JFF and MAAN idealized the concept search and helped in the evaluation experiment. ABS wrote the article. JMO, MANN and RS corrected the article. JMO and RS supervised the initial development of LAITOR. JMO and MAAN supervised the final development of LAITOR. All authors read and approved the final version of the article.

Received: 7 August 2009
Accepted: 1 February 2010 Published: 1 February 2010

References

1. Andrade MA, Bork P: **Automated extraction of information in molecular biology.** *FEBS Lett* 2000, **476**:12-17.
2. Krallinger M, Valencia A: **Text-mining and information-retrieval services for molecular biology.** *Genome Biol* 2005, **6**:224.

3. Kostoff RN, DeMarco RA: **Extracting information from the literature by text mining.** *Anal Chem* 2001, **73**:370A-378A.
4. Blaschke C, Andrade MA, Ouzounis C, Valencia A: **Automatic extraction of biological information from scientific text: protein-protein interactions.** *Proc Int Conf Intell Syst Mol Biol* **1999**:60-67.
5. Natarajan J, Berrar D, Dubitzky W, Hack C, Zhang Y, DeSesa C, Van Brocklyn JR, Bremer EG: **Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line.** *BMC Bioinformatics* 2006, **7**:373.
6. Yu H, Hatzivassiloglou V, Friedman C, Rzhetsky A, Wilbur WJ: **Automatic extraction of gene and protein synonyms from MEDLINE and journal articles.** *Proc AMIA Symp* 2002, 919-923.
7. Schuemie MJ, Weeber M, Schijvenaars BJ, van Mulligen EM, Eijk van der CC, Jelier R, Mons B, Kors JA: **Distribution of information in biomedical abstracts and full-text publications.** *Bioinformatics* 2004, **20**:2597-2604.
8. Rodriguez-Penagos C, Salgado H, Martinez-Flores I, Collado-Vides J: **Automatic reconstruction of a bacterial regulatory network using Natural Language Processing.** *BMC Bioinformatics* 2007, **8**:293.
9. Plake C, Schiemann T, Pankalla M, Hakenberg J, Leser U: **AliBaba: PubMed as a graph.** *Bioinformatics* 2006, **22**:2444-2445.
10. Tari L, Hakenberg J, Gonzalez G, Baral C: **Querying parse tree database of Medline text to synthesize user-specific biomolecular networks.** *Pac Symp Biocomput* **2009**:87-98.
11. Thu PH, Baral C, Gonzales G: **Generalized text extraction from molecular biology text using parse tree database querying.** Technical Report TR-08-004, Arizona State University 2008.
12. Fontaine JF, Barbosa-Silva A, Schaefer M, Huska MR, Muro EM, Andrade-Navarro MA: **MedlineRanker: flexible ranking of biomedical literature.** *Nucleic Acids Res* 2009, **37**:W141-146.
13. Mika S, Rost B: **NLProt: extracting protein names and sequences from papers.** *Nucleic Acids Res* 2004, **32**:W634-637.
14. Krallinger M, Leitner F, Rodriguez-Penagos C, Valencia A: **Overview of the protein-protein interaction annotation extraction task of BioCreative II.** *Genome Biol* 2008, **9(Suppl 2)**:S4.
15. Hoffmann R, Valencia A: **Implementing the iHOP concept for navigation of biomedical literature.** *Bioinformatics* 2005, **21(Suppl 2)**:ii252-258.
16. Blaschke C, Valencia A: **The potential use of SUISEKI as a protein interaction discovery tool.** *Genome Inform* 2001, **12**:123-134.
17. Persico M, Ceol A, Gavrila C, Hoffmann R, Florio A, Cesareni G: **HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms.** *BMC Bioinformatics* 2005, **6(Suppl 4)**:S21.
18. Chatr-aryamontri A, Kerrien S, Khadake J, Orchard S, Ceol A, Licata L, Castagnoli L, Costa S, Derow C, Huntley R, Aranda B, Leroy C, Thorneycroft D, Apweiler R, Cesareni G, Hermjakob H: **MINT and IntAct contribute to the Second BioCreative challenge: serving the text-mining community with high quality molecular interaction data.** *Genome Biol* 2008, **9(Suppl 2)**:S5.
19. Sing T, Sander O, Beerenwinkel N, Lengauer T: **ROCR: visualizing classifier performance in R.** *Bioinformatics* 2005, **21**:3940-3941.
20. Hooper SD, Bork P: **Medusa: a simple tool for interaction graph analysis.** *Bioinformatics* 2005, **21**:4432-4433.
21. Pavlopoulos GA, O'Donoghue SI, Satagopam VP, Soldatos TG, Pafilis E, Schneider R: **Arena3D: visualization of biological networks in 3D.** *BMC Syst Biol* 2008, **2**:104.
22. Fujita M, Fujita Y, Noutoshi Y, Takahashi F, Narusaka Y, Yamaguchi-Shinozaki K, Shinozaki K: **Crosstalk between abiotic and biotic stress responses: a current view from the points of convergence in the stress signaling networks.** *Curr Opin Plant Biol* 2006, **9**:436-442.
23. Rommens CM, Kishore GM: **Exploiting the full potential of disease-resistance genes for agricultural use.** *Curr Opin Biotechnol* 2000, **11**:120-125.
24. Tuteja N: **Mechanisms of high salinity tolerance in plants.** *Methods Enzymol* 2007, **428**:419-438.
25. Seki M, Umezawa T, Urano K, Shinozaki K: **Regulatory metabolic networks in drought stress responses.** *Curr Opin Plant Biol* 2007, **10**:296-302.
26. Jackson MB, Colmer TD: **Response and adaptation by plants to flooding stress.** *Ann Bot (Lond)* 2005, **96**:501-505.
27. Sharma P, Sharma N, Deswal R: **The molecular biology of the low-temperature response in plants.** *Bioessays* 2005, **27**:1048-1059.
28. Wasternack C: **Jasmonates: an update on biosynthesis, signal transduction and action in plant stress response, growth and development.** *Ann Bot (Lond)* 2007, **100**:681-697.
29. Broekaert WF, Delaure SL, De Bolle MF, Cammue BP: **The role of ethylene in host-pathogen interactions.** *Annu Rev Phytopathol* 2006, **44**:393-416.
30. Loake G, Grant M: **Salicylic acid in plant defence–the players and protagonists.** *Curr Opin Plant Biol* 2007, **10**:466-472.
31. Pieterse CM, van Loon LC: **Salicylic acid-independent plant defence pathways.** *Trends Plant Sci* 1999, **4**:52-58.
32. Kachroo A, Kachroo P: **Salicylic acid-, jasmonic acid- and ethylene-mediated regulation of plant defense signaling.** *Genet Eng (N Y)* 2007, **28**:55-83.
33. Kell DB, Oliver SG: **Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era.** *Bioessays* 2004, **26**:99-105.
34. Ananiadou S, Kell DB, Tsujii J: **Text mining and its potential applications in systems biology.** *Trends Biotechnol* 2006, **24**:571-579.
35. Swanson DR: **Fish oil, Raynaud's syndrome, and undiscovered public knowledge.** *Perspect Biol Med* 1986, **30**:7-18.
36. Ding J, Berleant D, Nettleton D, Wurtele E: **Mining MEDLINE: abstracts, sentences, or phrases?.** *Pac Symp Biocomput* 2002, 326-337.
37. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C: **STRING 8–a global view on proteins and their functional interactions in 630 organisms.** *Nucleic Acids Res* 2009, **37**:D412-416.
38. Saric J, Jensen LJ, Ouzounova R, Rojas I, Bork P: **Extraction of regulatory gene/protein networks from Medline.** *Bioinformatics* 2006, **22**:645-650.
39. Krallinger M, Rodriguez-Penagos C, Tendulkar A, Valencia A: **PLAN2L: a web tool for integrated text mining and literature-based bioentity relation extraction.** *Nucleic Acids Res* 2009, **37**:W160-165.
40. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res* 1999, **27**:29-34.

doi:10.1186/1471-2105-11-70
Cite this article as: Barbosa-Silva *et al.*: LAITOR - Literature Assistant for Identification of Terms co-Occurrences and Relationships. *BMC Bioinformatics* 2010 **11**:70.