Enhancing Data Integration with Text Analysis to Find Proteins Implicated in Plant Stress Response

Keywan Hassani-Pak[1], Roxane Legaie[2], Catherine Canevet[1], Hugo A. van den Berg[2], Jonathan D. Moore[2], Christopher J. Rawlings[1]

[1] Centre for Mathematical and Computational Biology, Rothamsted Research, UK
[2] Warwick Systems Biology Centre, University of Warwick, UK

Summary

High throughput genomic studies can identify large numbers of potential candidate genes, which must be interpreted and filtered by investigators to select the best ones for further analysis. Prioritization is generally based on evidence that supports the role of a gene product in the biological process being investigated. The two most important bodies of information providing such evidence are bioinformatics databases and the scientific literature. In this paper we present an extension to the Ondex data integration framework that uses text mining techniques over Medline abstracts as a method for accessing both these bodies of evidence in a consistent way. In an example use case, we apply our method to create a knowledge base of *Arabidopsis* proteins implicated in plant stress response and use various scoring metrics to identify key protein-stress associations. In conclusion, we show that the additional text mining features are able to highlight proteins using the scientific literature that would not have been seen using data integration alone. Ondex is an open-source software project and can be downloaded, together with the text mining features described here, from www.ondex.org.

1 Introduction

Systems approaches to biology research often use high throughput gene expression profiling experiments and these frequently identify large numbers of genes that are significantly differentially expressed. A major challenge when working with such long lists of potential candidate genes is to interpret their likely importance. This requires that each gene is considered in its biological context by comparing it with what is known and can be assembled from public information about gene function. It is generally recognised that the most important sources of contextual information, needed for understanding the biological role of a particular gene, are public bioinformatics databases and the scientific literature. The scale of high-throughput systems biology approaches requires that automated integration methods be developed that can bring together data from a range of information resources. Successful data integration methods must solve a range of significant technical problems and address two important challenges usually referred to as syntactic and semantic heterogeneities among source data sets [1]. The increasing availability of scientific literature in machine readable form has also created the possibility of mining source text for information that would be relevant to interpretation of experimental results. Methodological developments in data integration and text mining have, however, largely been conducted independently. In this paper we wish to demonstrate that significant benefits arise from combining data integration with text mining. We will motivate this assertion using a case study involving analysis of genes identified from gene expression studies of plant stress responses, but first we will provide some background to both data integration and text mining methods.

There are four main approaches to data integration: **portals**, which provide integration through links between data and are therefore more suitable for navigating among data than

doi:10.2390/biecoll-jib-2010-121

querying [2, 3]; **mashups**, which aggregate information rather than integrate them [4]; **mediators**, which distribute queries over separate sources and integrate results in a middle layer [5, 6]; and **warehouses**, which integrate data sources at schema level. A number of different warehouse systems for biological data have been developed including BN++ [7], PathSys [8], Biozon [9] and Ondex [10]. Both Ondex and Biozon use a graph-based representation of the data and exploit this, for example, to infer indirect relationships between nodes of the network which can lead to new discoveries from the data. For example, using a graph-based reasoning approach, Swanson [11] identified new relationships between migraine headache and magnesium deficiency, which have subsequently been validated experimentally.

A full introduction to text mining methods is not appropriate here, but up to date reviews are available [12, 13]. In recent years, stand-alone text mining systems have been developed which can integrate 'facts' from the scientific literature with entries in biological databases. Many of them have been developed to support manual curation of a structured database of information extracted from the literature. It seems that none of the available data integration tools offer text mining as part of their services, even though it has been suggested that this would be extremely valuable [14]. In this paper we explore the potential advantages of including standard text analysis methods with a data integration system. We have implemented these methods as a text mining plugin module for the Ondex data integration framework. A biological use case demonstrates how it can help to identify key proteins involved in stress response such as induced by infection with *Botrytis cinerea* or treatment with ethylene in the model plant *Arabidopsis thaliana*.

2 Implementation & Methods

The original implementation of the Ondex system [10] contained prototype text mining features. However, to benefit from recent developments including the Ondex plugin architecture and workflow engine (the Ondex Integrator), we opted to re-develop the text mining features as a software plugin, which could be invoked with other Ondex mapping methods used for data integration. The text mining plugin identifies biological terms as 'concepts' in Medline abstracts, and creates links (relations) between the publication and the concepts which occur within it. Co-occurring concepts are then exploited to build a weighted association network, which simplifies the graph structure. This can then be visualised and subjected to other graph analysis methods in Ondex (formerly called the OVTK). It is therefore possible to present integrated text mining results in a way that is better suited for users to explore results and arrive at new insights. We will demonstrate this in a biological application case.

2.1 Text mining with Ondex

We currently limit our text analysis to titles and abstracts from Medline. Our approach, however, is general and could be employed with full documents if these were available. Before invoking the text mining plugin, the Medline parser must be used to create publication concepts in the Ondex system. This is generally achieved as part of a workflow executed by the Ondex Integrator. Publication concepts have attributes including PubMed ID (PMID), Digital Object Identifier (DOI), title, abstract, authors, journal, year, Medical Subject Headings (MeSH) and Chemical terms.

The Ondex text mining plugin consists of three components which must be executed sequentially within a workflow: Section 2.1.1 explains how the plugin assembles a core knowledge base by adding publications to an existing integrated dataset; Section 2.1.2

describes how it links new Publication concepts to other concepts in the graph; Section 2.1.3 introduces how co-occurrences are used to create weighted association networks.

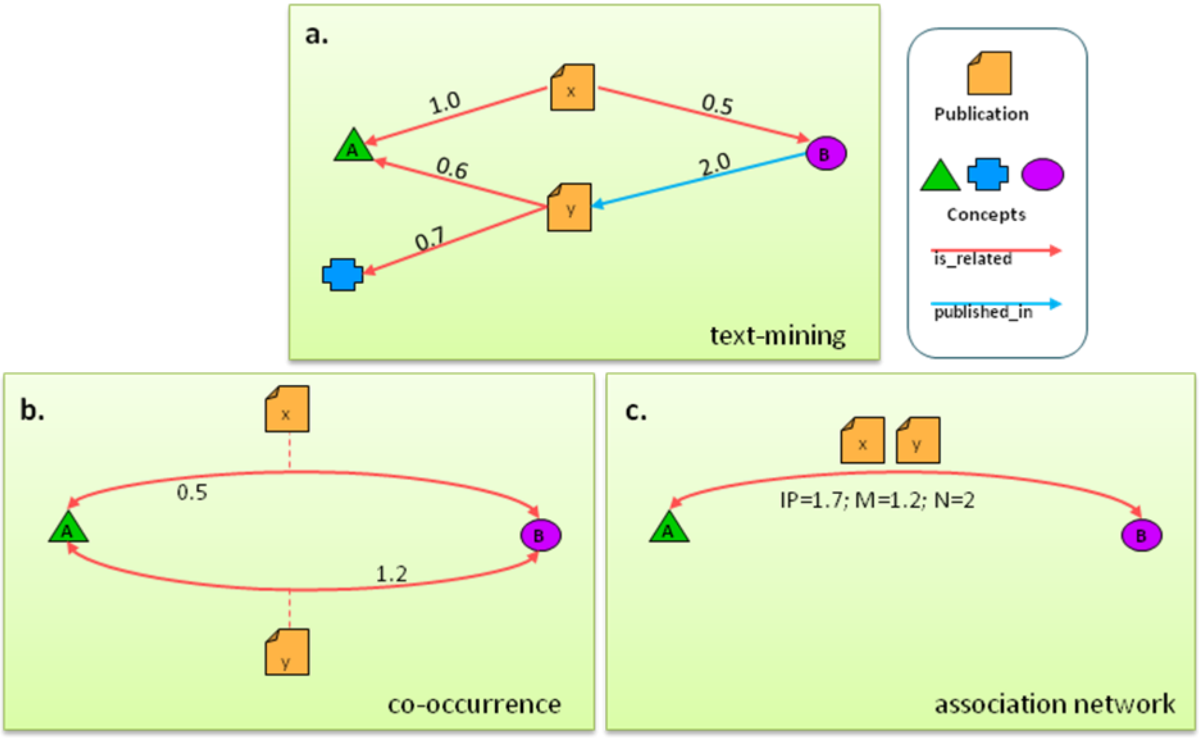

Figure 1: a) The outcome of the text mining and data integration step. The tf-idf score on a text mining based relation (red) indicates the relevance of a publication to a concept. The blue relation is imported from a manually curated database and is processed with a fixed high score. b) The co-occurrence step (here for triangle and circle) creates a direct link between the co-cited concepts within each publication (qualifier of the relation) and calculates the score as the product of the two original tf-idf scores. c) The final association step combines all *is_related* scores of A and B. Various metrics are calculated to determine the strengths of the association (see Section 2.1.3).

2.1.1 Corpus retrieval

Three alternative strategies are available for adding a collection of related publications (the corpus) to a pre-existing Ondex knowledge base. The first is to search PubMed[1] for keywords (e.g. “Arabidopsis”) and to use the Ondex Medline parser to import PubMed’s results saved in XML format. Secondly, if the Ondex graph already contains concepts with PMIDs, the Medline parser’s eFetch parameter can be used to add title, abstract, MeSH terms etc. to each publication using NCBI Web services (updating approximately 1000 publications/minute). Thirdly, Medline licensees who have a locally installed copy can parse all or parts of the Medline XML files.

To conduct efficient text analysis, it is necessary to create indices using the Lucene [15] service which is built into Ondex. Firstly, abstracts are normalised by converting everything to lower-case, then filtering-out non-alphanumeric characters and stop words (e.g. “the”, “of”, “a”). Secondly, text is tokenised into words and added to the index. Typographical variants are thereby excluded, so that words like *Kcnip3* or *KCNIP3* are stored as *kcnip3* and

[1] http://www.ncbi.nlm.nih.gov/pubmed

KCNIP-3 or *kcnip_3* are stored as *kcnip 3* in the index. The indexing and search query normalisation are completed automatically by Ondex and do not require any user interaction.

2.1.2 Mapping publications to concepts in the knowledge base

With the ontologies, databases and publications loaded into Ondex, the text mining plugin can be used to create the required links between concepts based on evidence from the literature. The aim of this step is to recognise a biological concept (name or synonym) embedded in the title or abstract of a publication. This is known as Named Entity Recognition (NER) for which Ondex uses information retrieval and dictionary-based methods. The NER approach implements several Lucene query strategies to search publications in an Ondex knowledge base. The standard search method considers ***exact*** occurrence of the query term in the abstract or title of the publication. Three other search methods have been implemented: ***fuzzy search*** also matches documents that contain terms similar to the specified query term based on the Levenshtein algorithm and ***proximity search*** supports finding patterns i.e. ordered words appearing within a specific distance of one another. The ***AND search*** considers exact occurrences of query terms unlimited by any order or distance.

Having recognised a concept in the publication, a relation of type *is_related* is created, indicating that the given publication is somehow related to the identified concept (Figure 1a). In order to provide evidence for the relation, each abstract is split into sentences and each sentence containing the matching query is stored as evidence text. Furthermore, the product of term frequency and inverse document frequency (tf-idf) is used as a weighting scheme to define the specificity of the publication-concept mapping. Robertson gives a very good review on tf-idf weighting schemes [16]. Our Lucene implementation of tf-idf gives a higher weight to publications containing the query in their title rather than in their abstract. This recognises the assumption that reference to a concept in a title will be more informative than in an abstract. If several synonyms of a concept are found in the publication, the *is_related* relation is annotated with the highest tf-idf score.

2.1.3 Using co-occurrence to build weighted association networks

The third component of the plugin is a transformation step to simplify the Ondex knowledge base to combine text-based evidence and reduce graph complexity. This makes subsequent analysis and visualisation more effective.

If more than one concept is linked to the same publication (abstract or title) then the concepts are connected by a new relation (*is_related*). Because a bag-of-words approach is used, the direction and semantics of these new co-occurrence relations cannot be specified. For example, if two concepts A and B are cited in the same publication, an undirected ternary relation of type *is_related* would be created with the publication as qualifier, and the product of the two original tf-idf scores assigned as the combined tf-idf score of the relation (Figure 1b). Here we assume that for a given document d, tf-idf$_{A,d}$ and tf-idf$_{B,d}$ are independent scores. Thus, the combined tf-idf represents the relevance of a document d for the query "*a* AND *b*", with *a*, *b* being possible concept names of A and B.

It is often the case that two concepts may be cited together more than once in different publications. Therefore in order to simplify the graph, all *is_related* relations between two concepts A and B are combined into one binary relation. For each pair A-B, the following quantities were calculated and assigned to the relation: (i) the inner product of the scores (**IP** = Σ_i x_i y_i) where the index i ranges over the co-citations of the pair at hand, with x_i being the A score and y_i being the B score in the ith co-citation; (ii) **M** = Max$_i$(x_i y_i), with i ranging over the co-citations of the protein-stress pair at hand; (iii) **N** = number of documents in which A and B were co-cited. An illustration of all three metrics is shown in Figure 1c.

In cases where the data integration step has used a source containing curated literature references (e.g. such as in the UniProt database) these links are captured using the relation *published_in*. During the derivation of co-occurrence data, these *published_in* relations are processed in an equivalent way to the *is_related* terms but a fixed pseudo-tf-idf value is assigned. To reflect the confidence in the curated *published_in* relations, this pseudo-tf-idf is set at 2.0.

2.2 An application case: stress responses in *Arabidopsis thaliana*

It is believed that plants respond to abiotic and biotic stresses via core response networks in which transcriptional changes are a fundamental component. In the PRESTA (Plant Responses to Environmental Stress in *Arabidopsis*) project at the Universities of Warwick, Essex and Exeter such networks are being investigated to better understand how plants respond to environmental stresses, with the aim of informing future crop improvement programmes. In experimental work by the PRESTA project, high resolution time series expression profiling of *Arabidopsis* leaves infected by the pathogen *Botrytis cinerea,* had identified hundreds of significantly differentially expressed genes. Some of these genes are known to effect susceptibility to infection by *Botrytis*, but many others have not previously been considered important in stress response and documented as such in the reference databases used in *Arabidopsis* research. Manual searching of scientific literature suggested that many genes could have plausible indirect links to *Botrytis* infection and other stress responses and might be relevant to the search for core gene networks for biotic and abiotic stress. The large number of genes of interest, however, meant that an automated method was desirable. The initial study reported here shows the feasibility of our approach. Future research will employ expert review and curation to assess whether automatic methods can outperform wholly manual literature searching and comparison with other internet-based *Arabidopsis* databases.

The application case requires the creation of a knowledge base of *Arabidopsis* genes and proteins implicated in both biotic and abiotic stresses. This provides the context for exploring possible regulatory networks involved in stress response. The following databases were used to build an Ondex knowledge base that could be augmented with the new text mining plugin.

UniProtKB (release 15.8) was searched for *Arabidopsis* (TaxID: 3702) proteins and the subset of reviewed proteins[2] (8582 distinct proteins) was downloaded as an XML file. Using the UniProt parser, this subset was loaded into Ondex and served as the *Arabidopsis* protein name dictionary to be used for text mining. This set of UniProt proteins contained 13,502 verified links to published papers. When curated links such as those from UniProt are used (represented as *published_in* relations), the text mining plugin incorporates them into the subsequent co-occurrence analyses (see Section 2.1.3). Some of the proteins of the dataset are enzymes and were therefore annotated with Enzyme Classifications (EC). The EC identifiers provide a basis for integrating these data with pathway databases such as KEGG or AraCyc. PubMed was used to retrieve all Medline articles that contained the keywords “Arabidopsis thaliana” in their abstract, title or MeSH header. On August 28th 2009, this resulted in 28653 articles being retrieved. This PubMed subset was downloaded in XML format and loaded into Ondex using the Medline parser.

[2] http://www.uniprot.org/uniprot/?query=taxonomy%3A3702+AND+reviewed%3Ayes

To enable recognition of plant stress terminology, a Plant Stress Ontology was defined (Figure 2). This constituted the second dictionary used in the text mining steps. The ontology encompasses 33 concepts related to biotic and abiotic stresses, relevant to the PRESTA project, such as *Bacteria*, *Fungus*, *Ethylene* and *Drought*. A tab-delimited file parser imported the ontology into Ondex.

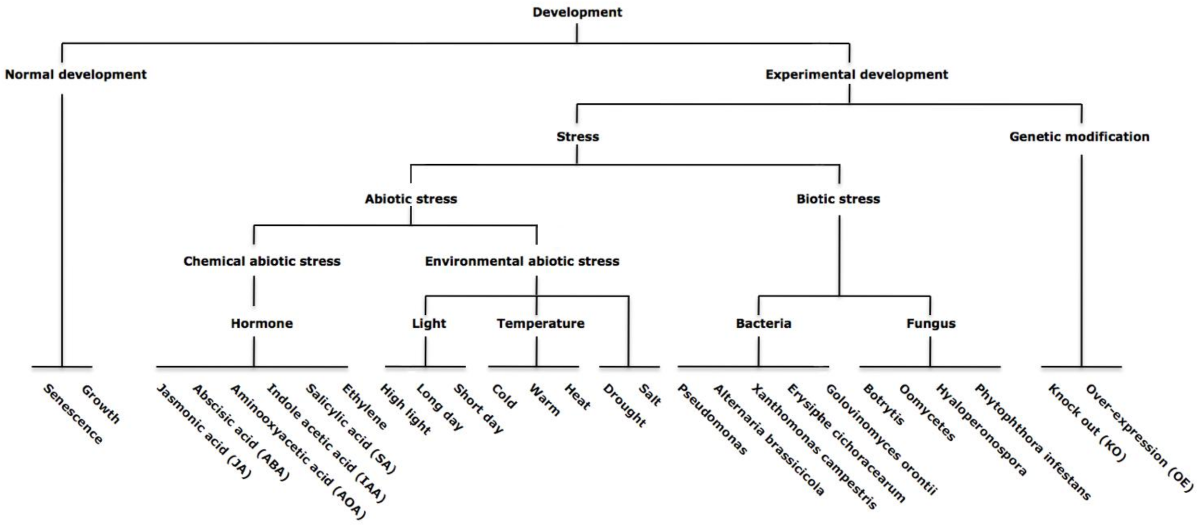

Figure 2: An ontology of biotic and abiotic causes of plant stress responses.

Figure 3 is based on the Ondex metagraph visualisation to illustrate the steps taken in the Ondex Integrator to create the knowledge base. At the outset some *Arabidopsis* proteins from UniProt are already cross-referenced to publications. These are as shown by the orange relations in Figure 3a. The NER step of the text mining plugin then linked proteins and stresses to individual publications (see blue *is_related* relations on Figure 3b). The NER step also assigned a tf-idf score and evidence sentences (which were extracted from the publications) to each relation. In the final step, co-occurring protein-stress pairs (Figure 3c) were identified and connected with *is_related* relations. Publications serve as qualifiers/context (dotted lines) of the association. All *is_related* relations between protein-stress pairs are annotated with three different scores presented in Section 2.1.3.

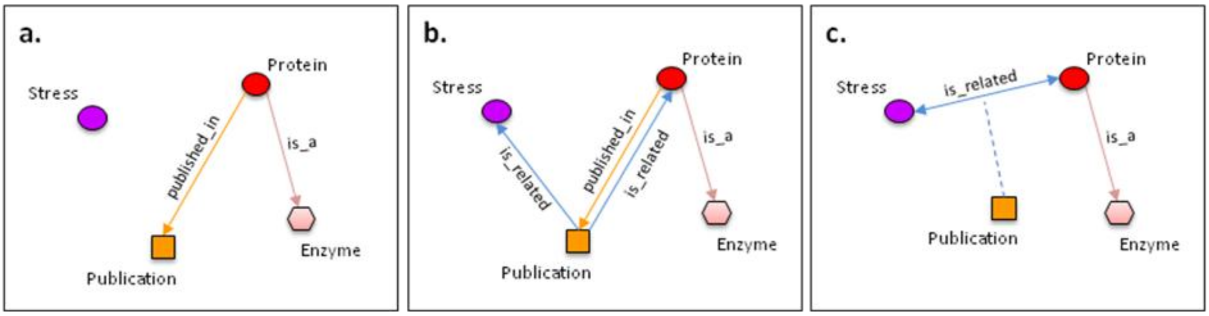

Figure 3: Three steps towards a protein-stress association network using the Ondex text mining plugin. a) Input data, b) Named Entity Recognition (NER) and c) from co-occurrence to weighted association networks.

Additionally to the three IP, M and N metrics calculated for each protein-stress pair, an association measure (P-value) was derived, defined as $\mathbf{P}=\mathrm{Prob}[\Sigma_i\ x_i\ y_{p(i)} \geq S]$ where p(i) represents a random permutation of the index values. For 10 co-citations or less, P was calculated by listing all permutations exhaustively; for 11 or more co-citations, P was calculated by sampling 80000 random permutations. Low P-values are an indication of strong association. Both IP and M are stable against the number of co-citations, whereas P can never be lower than 1/(n!) where n is the number of co-citations.

This weighted association network will provide the basis for identification of key proteins associated with particular stresses in *Arabidopsis*.

3 Results

We have developed a text mining plugin for the Ondex Integrator which can link concepts (entities) from any data source based on information from related scientific publications. In our application case, two data sources were incorporated into a knowledge base: a Plant Stress Ontology and *Arabidopsis* proteins from UniProtKB (reviewed). Our aim was to establish the links between the proteins and plant stresses by text mining publications about *Arabidopsis* (the corpus) and use relations from text mining to extend the knowledge base. After exporting results of the text-analysis, Ondex was used for visualisation and further analysis.

3.1 Mapping concepts to the corpus

In total 52,430 protein/stress entities were recognised in 19,884 publications from our corpus; approximately 2.6 per publication. Tf-idf scores in our network indicate how significant a protein or stress concept is to a document in the *Arabidopsis* corpus. In order to understand how tf-idf scores are distributed, a histogram was plotted for the whole dataset (Figure 4). A long-tailed distribution is observed, which is characterised by a peak at approximately 0.2 and a long tail with a maximum value of approximately 15.0. The data are a very close fit to a mixed distribution of two lognormal components. We can explain this observation by the fact that queries occurring in the title of a document receive an enhanced tf-idf score. Choosing a cut-off value of 0.956, separates the data very clearly (specificity=sensitivity=98.85%) into these two subpopulations (abstract vs title). Thus at this cut-off value, we have a 98.85% chance of classifying a tf-idf score correctly as the result of a query matching either abstract or title of the document.

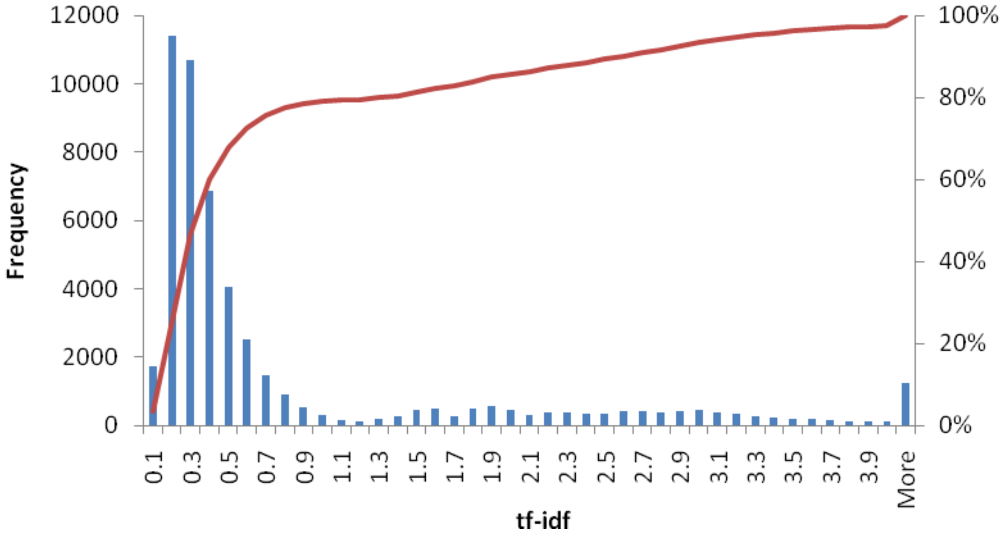

Figure 4: Histogram (blue) and cumulative frequency distribution (red) of tf-idf scores from 52430 protein/stress entities found in the *Arabidopsis* corpus by named entity recognition. For details see text.

3.2 Weighted association networks

When creating co-occurrence networks, a particular protein and stress are connected via one or several publications. In order to make the protein-stress relations more evident, the structure of the text mining based network was transformed to an association (co-citation) network.

The resulting association network, after filtering out unconnected nodes, contained 3145 proteins linked to 32 stresses by 10,777 relations. In other words, 36.7% of reviewed *Arabidopsis* proteins from UniProt (3145/8582) were co-cited with at least one term from the Plant Stress Ontology. On average, each co-cited protein was related to approximately 3.4 stresses and each stress related to 337 proteins.

Three different confidence scores were assigned to protein-stress associations: (i) a tf-idf based association score (**IP**) ranging between 0.01 and 347.26; (ii) the tf-idf score of the top hit document containing the protein-stress pair (**M**) ranging between 0.01 and 26.86; (iii) the number of documents in which two entities are co-cited (**N**) with a minimum of 1 and maximum of 600 co-citations. The highest IP score was found between "Light" and "Phytochrome A" (photoreceptor) while the lowest one appeared between "Hormone" and "ADP-glucose synthase" (a protein known to be regulated by hormones in rice cells). In the majority of cases the tf-idf based score is numerically lower than the co-citation number. However there are also opposite cases, for example, the association between ACBP4 and ethylene has only one co-citation but a tf-idf-based score of 13.51.

Comparison of the three scoring metrics over all 10,777 protein-stress pairs showed that IP and N are strongly correlated (r=0.79), and the correlation-coefficient between IP and M is r=0.53. In Figure 5 is shown that the association score IP is correlated with N (co-citation number) but with a large variance especially for small N, e.g. for N=5 the IP score is 0.1>IP>13.0. The M score on the other hand does not seem to be correlated with N, in the sense that a similar proportion at each N is greater than a given value. Further work will explore these relationships in more detail with a view to determining optimal cut-offs and whether the scores can be combined in a way that could improve the precision and recall of the text mining methods.

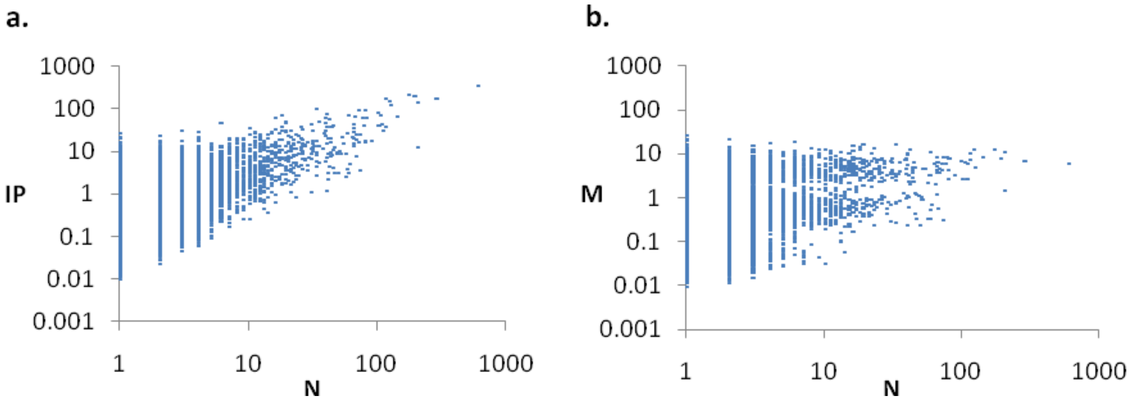

Figure 5: Comparison of IP scores vs number co-citations N (a); and the M scores vs N for all 10,777 protein-stress pairs (b). For N=5 the IP varies from 0.1 – 13 and the M from 0.05 – 10.

Because of the large size of this integrated graph, it was found useful to filter associations using confidence scores. In this case, the difficulty is to remove false positives while retaining true positives, improving signal-to-noise while retaining sensitivity. Figure 6 shows all protein-stress pairs that are co-cited five times or more. The co-citation number is the simplest way to potentially reduce noise in such association networks. Jenssen *et al.* [17] examined the accuracy and type of interactions found among genes mentioned more than once or more than five times together and found a decrease in the number of false positives as the number of co-occurrences increased. We found that sorting and filtering by IP and M metrics is in general more accurate than by simple co-citation frequency as both IP and M consider the frequency of terms in the corpus. However, none of the metrics seems to be superior overall, and the selection of the best metric may depend on the use case. Considering several metrics at the same time when analysing protein-stress associations seems to be the method of choice to highlight key associations and filter noise.

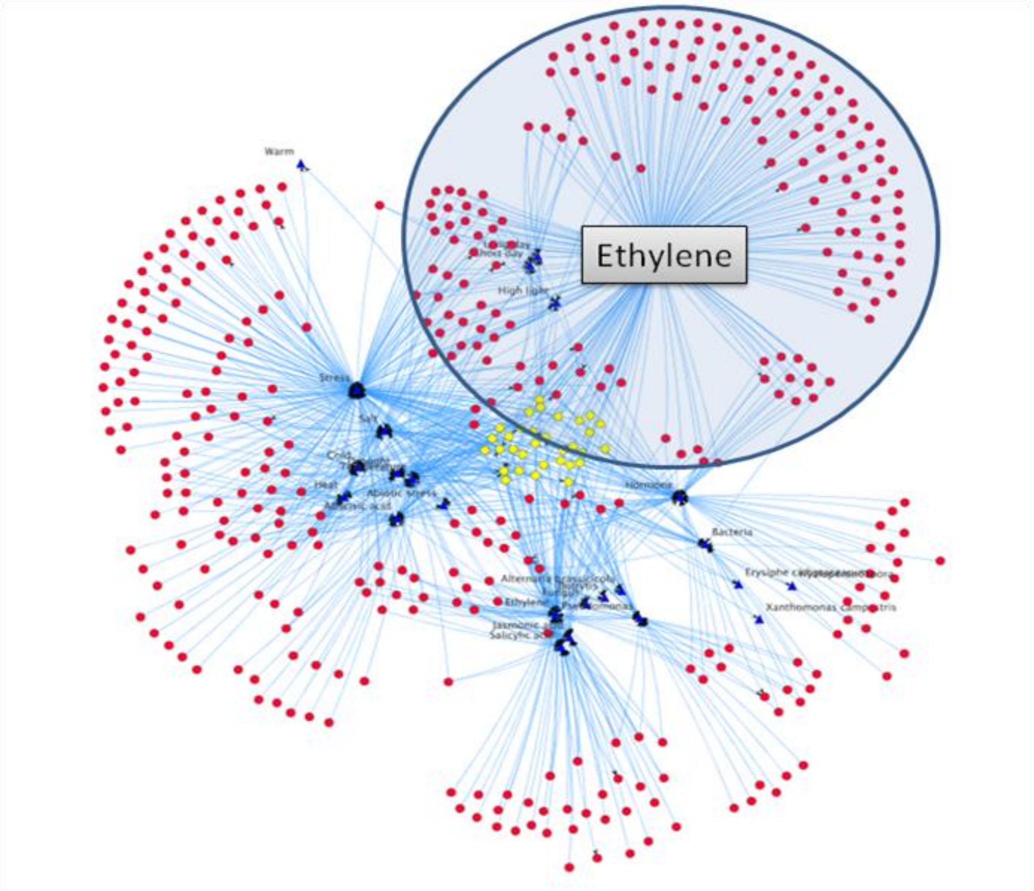

Figure 6: An example protein-stress association network based on 5 or more co-citations. The network contains 444 proteins (red circles), 25 stresses (blue triangles) and 1133 relations (blue edges). Proteins in the centre of the network (yellow circles) are implicated in several plant stress responses. The Ethylene association sub-network is highlighted.

3.3 Validation of ethylene-protein associations

To test the validity of the association (co-citation) network, we decided to focus on ethylene, a major plant hormone included in our Plant Stress Ontology. The ethylene association network (see ellipse in Figure 6) contained 533 proteins and the same number of relations, with IP scores ranging from 0.016 to 202.35 and M scores from 0.016 to 13.91. To compare our predictions with a manually curated set, we chose the Arabidopsis Hormone Database (AHD) [18] as a gold standard. In AHD, 31 proteins are related to ethylene response based on manual curation of the published literature. Our text mining derived ethylene network contained 22 of the AHD proteins, a recall rate of 71.0%. The 9 proteins omitted from the network are not in UniProtKB as “reviewed” (yet), and thus not included in our protein dictionary. In a second validation we compared our predictions to all 166 ethylene related proteins of AHD (including those extracted from GO) and achieved a recall rate of 44.8%.

Table 1 shows the top 10 proteins (sorted by M score) from our analyses that are linked to ethylene but were not found in AHD. The four different confidence scores N, M, IP and P display the strength of the association. The smaller the P-value the more likely is the protein-ethylene association. However, the P-value calculation depends on N, and makes only sense for larger N, if N=1 then P=1. To classify these as true or false positives, evidence sentences from the literature were manually inspected. This can easily be done within Ondex as the text-mining graph contains the publications (titles and abstracts) and annotates relations with evidence sentences extracted from the publication. For example:

- *the interaction of ACBP4 and AtEBP may be related to AtEBP-mediated defence possibly via ethylene and/or jasmonate signalling.* [PMID: 18836139]
- *protein phosphatase 2C ABI1 modulates biosynthesis ratio of ABA and ethylene.* [PMID: 19705149]
- *a specific interaction of ETR1 with the histidine-containing transfer protein AHP1, supporting the idea that a phosphorelay module is involved in ethylene signalling* [PMID: 18384742]

Preliminary analyses of this manually validated subset of text mining based associations that are not included in AHD indicate that the inner product (IP) and the maximum scores (M) are highly significant (P<0.0001) correlates of a physiologically meaningful link between the protein and the stress.

Table 1: Top 10 ethylene related proteins (sorted by M) that the text mining analysis predicted but were not found in AHD. Each row displays the top hit PubMed ID and the year of publication. The next three columns show the weights of the protein-stress association according to different scoring metrics (see Section 2.1.3). The P-value calculation is described in Section 2.2. The last column indicates whether the association is correct or not according to expert evaluation.

ACCESSION	NAME	PUBMED	YEAR	N	M	IP	P	TRUE
AT3G05420	ACBP4	18836139	2008	1	13.51	13.51	1.00	yes
AT1G31812	ACBP6	18836139	2008	2	11.57	17.14	0.50	yes
AT3G03190	ATGSTF6	14617075	2003	7	7.36	15.75	0.25	yes
AT4G26080	ABI1	19705149	2009	10	6.66	12.22	0.39	yes
AT3G21510	AHP1	18384742	2008	3	6.60	6.70	0.17	yes
AT1G75040	PR-5	15988566	2005	12	5.18	5.47	0.07	yes
AT2G45820	Remorin	9159183	1997	4	5.04	6.77	0.86	no
AT3G11410	PP2CA	19705149	2009	1	5.00	5.00	1.00	yes
AT1G09570	Phytochrome A	8703080	1996	11	4.79	8.47	0.19	no
AT1G04240	IAA3	19213814	2009	3	4.54	5.14	0.67	yes

4 Discussion

A major challenge for those working with high-throughput ‘omics datasets is to contextualise new results by comparing them with prior information, either from structured databases or the scientific literature. The development of data integration and text mining methods have, however, largely been conducted independently. In this paper we demonstrate that there are significant benefits to being able to combine data integration and text mining. The text mining plugin we developed extends the data integration framework Ondex with text mining functionality. It gives Ondex the capability to create improved links between life science data sources using the scientific literature, extending those achieved by simple cross-referencing. It is flexible and computationally undemanding, and therefore practical to integrate into workflows for high-throughput analysis of ‘omics data. For example, it could be used in other application cases where high-throughput ‘omics data are being analysed, such as:

i. **Genomics**: The combination of sequence homology and automated literature analysis could help refine predictions of gene function for unknown genes [19].
ii. **Proteomics**: Protein co-occurrences extracted from literature can add rich contextual data to an interaction network and improve protein function prediction [20].
iii. **Transcriptomics**: Enriching gene co-expression data with relationships derived from literature can contribute to a better understanding of co-expression patterns by identifying terms that are significantly overrepresented in the collection [17].

Our text mining plugin features information retrieval techniques and dictionary-based NER to recognise concepts (names and synonyms) in an application case related corpus. In cases where the data integration step has used a source containing curated literature references (e.g. such as in the UniProt database) these links are processed in an equivalent way to the text mining based relations. Using co-occurrence relationships between concepts of interest the plugin can generate weighted association networks. Our work demonstrates that basic text mining brings value to data integration. We have not attempted at this stage to encompass ‘state of the art’ text mining method which address issues such as identifying semantic links between mapped concepts and term disambiguation [21]. In the future, however, we plan to investigate such approaches and access more advanced text mining services from Ondex. These should extract more information from the literature and have increased sensitivity and specificity.

To the best of our knowledge, this is the first report of combining text analysis and basic text mining methods with data integration. Our plugin can be linked in to Ondex workflows to enhance large-scale data integration. Integrated text mining graphs can be exported to open exchange formats, such as RDF or OXL, and viewed and analysed using graph visualisation tools such as Cytoscape [22] or Ondex itself. Graph-based representations facilitate users’ understanding of integrated results. It offers a more user-friendly way of displaying text mining results compared to systems such as EBIMed [23] or Kleio [24] where only tables and links are given for navigation. Visualisation of highly connected nodes (hubs), indirect relationships and evidence text on edges can significantly support semi-automated knowledge discovery.

Our application case demonstrated the usage of the text mining plugin to integrate co-occurrence of *Arabidopsis* proteins with a Plant Stress Ontology. Focusing on protein-stress association (co-occurrence) networks from *Arabidopsis*, we showed that workflows incorporating text mining produced meaningful results consistent with manually curated data. We identified proteins involved in ethylene response in *Arabidopsis*. Validating against the manually curated *Arabidopsis* Hormone Database (AHD) showed that our method produced a recall of 71.0%. Moreover, we identified many more significant associations that are not present in AHD (yet) but were considered plausible by experts. The challenge lays in distinguishing significant information from noisy text mining data. Improving the signal-to-noise ratio is a demanding task, and benefits from domain-expert curation, at least in setting thresholds. A crucial requirement is therefore to calculate appropriate confidence scores on association pairs. We used four different scoring metrics IP, M, N and P for the protein-stress associations and made a first attempt to compare them. Many alternative weighting schemes for the definition of association strength between two concepts exist, and it is easy to envisage how a weighted mean of multiple scores may compensate for the weaknesses in any particular scoring scheme. In future work we will investigate whether other scoring methods such as mutual information, hypergeometric distribution or asymmetric co-occurrence fraction [20] bring added value and how the thresholds for an optimal signal-to-noise ratio can be chosen. Until then, interactive filters combined with visualisation methods in Ondex provide domain experts with tools to explore the results from both text mining and data integration in an intuitive and semi-automated way.

5 Conclusion

The addition of the text mining plugin has given Ondex new semantic integration features, opening up the wealth of previously inaccessible knowledge in free text. Through the application case study we have demonstrated successful integration of text mining into bioinformatics workflows that enhance high-throughput ‘omics research. While our case study did not use state of the art text mining methods, we have demonstrated that basic information retrieval and text analysis methods combined with visualisation, adds practical value to the existing data integration tools in Ondex. This is a report of work in progress and more detailed analysis of the methods are ongoing - as well as exploration of other text mining tools. Ondex is an open-source software project and can be downloaded, together with the text mining features described here, from www.ondex.org.

Acknowledgements

The authors gratefully acknowledge their thanks to the Ondex team for developing the data integration framework and to the PRESTA team for sharing their time and data.

Rothamsted Research receives grant in aid from the Biotechnology and Biological Sciences Research Council (BBSRC) and this work was supported by the BBSRC Systems Biological Approach to Biological Research (SABR) initiative through grants to Rothamsted Research [BB/F006039/1] and to Warwick University [BB/F005806/1].

References

[1] Lysenko A, Hindle MM, Taubert J, Saqi M, Rawlings CJ: Data integration for plant genomics--exemplars from the integration of Arabidopsis thaliana databases. *Briefings in bioinformatics* 2009, 10(6):676-693.

[2] Etzold T, Ulyanov A, Argos P: SRS: information retrieval system for molecular biology data banks. *Methods Enzymol* 1996, 266:114-128.

[3] Baxevanis AD: Searching the NCBI databases using Entrez. *Curr Protoc Bioinformatics* 2006, Chapter 1:Unit 1 3.

[4] Dowell RD, Jokerst RM, Day A, Eddy SR, Stein L: The distributed annotation system. *BMC Bioinformatics* 2001, 2:7.

[5] Donelson L, Tarczy-Hornoch P, Mork P, Dolan C, Mitchell JA, Barrier M, Mei H: The BioMediator system as a data integration tool to answer diverse biologic queries. *Stud Health Technol Inform* 2004, 107(Pt 2):768-772.

[6] Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock MR, Wipat A *et al*: Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* 2004, 20(17):3045-3054.

[7] Küntzer J, Blum T, Gerasch A, Backes C, Hildebrandt A, Kaufmann M, Kohlbacher O, Lenhof H-P: BN++ - A Biological Information System. *J Integrative Bioinformatics* 2006, 3(2).

[8] Baitaluk M, Qian X, Godbole S, Raval A, Ray A, Gupta A: PathSys: integrating molecular interaction graphs for systems biology. *BMC Bioinformatics* 2006, 7:55.

[9] Birkland A, Yona G: BIOZON: a system for unification, management and analysis of heterogeneous biological data. *BMC Bioinformatics* 2006, 7:70.

[10] Kohler J, Baumbach J, Taubert J, Specht M, Skusa A, Ruegg A, Rawlings C, Verrier P, Philippi S: Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics* 2006, 22(11):1383-1390.

[11] Swanson DR: Migraine and magnesium: eleven neglected connections. *Perspectives in biology and medicine* 1988, 31(4):526-557.

[12] Ananiadou S, Kell DB, Tsujii J: Text mining and its potential applications in systems biology. *Trends in biotechnology* 2006, 24(12):571-579.

[13] Krallinger M, Valencia A, Hirschman L: Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol* 2008, 9 Suppl 2:S8.

[14] Cohen KB, Hunter L: Getting started in text mining. *PLoS Comput Biol* 2008, 4(1):e20.

[15] Gospodneti O, Hatcher E: Lucene in action. Greenwich, CT: Manning; 2005.

[16] Robertson S: Understanding Inverse Document Frequency: On theoretical arguments for IDF. *Journal of Documentation* 2004, 60(5):503-520.

[17] Jenssen TK, Laegreid A, Komorowski J, Hovig E: A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet* 2001, 28(1):21-28.

[18] Peng ZY, Zhou X, Li L, Yu X, Li H, Jiang Z, Cao G, Bai M, Wang X, Jiang C *et al*: Arabidopsis Hormone Database: a comprehensive genetic and phenotypic information database for plant hormone research in Arabidopsis. *Nucleic Acids Res* 2009, 37(Database issue):D975-982.

[19] Scherf M, Epple A, Werner T: The next generation of literature analysis: integration of genomic analysis into text mining. *Briefings in bioinformatics* 2005, 6(3):287-297.

[20] Gabow AP, Leach SM, Baumgartner WA, Hunter LE, Goldberg DS: Improving protein function prediction methods with integrated literature data. *BMC Bioinformatics* 2008, 9:198.

[21] Wang X, Tsujii J, Ananiadou S: Disambiguating the Species of Biomedical Named Entities Using Natural Language Parsers. *Bioinformatics* 2010.

[22] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003, 13(11):2498-2504.

[23] Rebholz-Schuhmann D, Kirsch H, Arregui M, Gaudan S, Riethoven M, Stoehr P: EBIMed--text crunching to gather facts for proteins from Medline. *Bioinformatics* 2007, 23(2):e237-244.

[24] Nobata C, Cotter P, Okazaki N, Rea B, Sasaki Y, Tsuruoka Y, Tsujii Ji, Ananiadou S: Kleio: a knowledge-enriched information retrieval system for biology. In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval.* Singapore, Singapore: ACM; 2008: 787-788.