LitPathExplorer: A Confidence-based Visual Text Analytics Tool for Exploring Literature-Enriched Pathway Models

Axel J. Soto, Chrysoula Zerva, Riza Batista-Navarro and Sophia Ananiadou *

National Centre for Text Mining, School of Computer Science, University of Manchester, Manchester, M1 7DN, United Kingdom

* To whom correspondence should be addressed.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Pathway models are valuable resources that help us understand the various mechanisms underpinning complex biological processes. Their curation is typically carried out through manual inspection of published scientific literature to find information relevant to a model, which is a laborious and knowledge-intensive task. Furthermore, models curated manually cannot be easily updated and maintained with new evidence extracted from the literature without automated support.

Results: We have developed LitPathExplorer, a visual text analytics tool that integrates advanced text mining, semi-supervised learning and interactive visualization, to facilitate the exploration and analysis of pathway models using statements (i.e., events) extracted automatically from the literature and organized according to levels of confidence. LitPathExplorer supports pathway modellers and curators alike by: 1) extracting events from the literature that corroborate existing models with evidence; 2) discovering new events which can update models; and 3) providing a confidence value for each event that is automatically computed based on linguistic features and article metadata. Our evaluation of event extraction showed a precision of 89% and a recall of 71%. Evaluation of our confidence measure, when used for ranking sampled events, showed an average precision ranging between 61% and 73%, which can be improved to 95% when the user is involved in the semi-supervised learning process. Qualitative evaluation using pair analytics based on the feedback of three domain experts confirmed the utility of our tool within the context of pathway model exploration.

Availability: LitPathExplorer is available at http://nactem.ac.uk/LitPathExplorer_BI/

Contact: sophia.ananiadou@manchester.ac.uk

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Deep understanding of biological processes is of central importance to many researchers who use computational modelling when undertaking research in the areas of biotechnology, cosmetics, health and medicine. To support these activities, researchers often rely on the availability of up-to-date pathway models to perform in silico experiments which elucidate the complex mechanisms that underpin several biological processes and their interactions.

Due to their size and complexity, pathway models are typically neither complete nor error-free, and even curated models become rapidly out of date, and thus require further revisions (Kelder et al., 2012; Matsuoka et al., 2010). Additionally, curators and actual consumers of these models need to contrast and review large collections of research articles, which is a costly and time-consuming task (Bakalov et al., 2012). Methods based on text mining (TM) have provided major contributions to this field, as they can process large literature collections in order to find and extract information from text to support pathway reconstruction (Ananiadou et al., 2010; Oda et al., 2008; Kemper et al., 2010; Chen and Sharp, 2004).

Even though TM and statistical models provide a means for automating and reducing the burden of this work, a number of further challenges exists. The automatic identification of events in natural language is a difficult task, as language ambiguity and variance prevent TM methods from
obtaining perfect accuracy. Ideally, different degrees of confidence should be assigned when this information is automatically extracted (Kılıçgoğlu et al., 2015; Malhotra et al., 2013). Moreover, the volume of information generated when mining scientific text collections is typically too large for manual analysis. TM results are often combined with other resources (Raja et al., 2017), e.g., omics data—which can be heterogeneous and with varying levels of quality—and as such, the combined information needs to be filtered and organized for the consumption of expert users (Bastian et al., 2015). Furthermore, when the user needs to make sense of how different pieces in the model interact with each other, visual methods for exploratory analysis are preferred (Kerren and Schreiber, 2014) as they allow the domain expert to organize, browse and inspect information on demand.

With this scenario in mind, we have developed a tool that aims to address these challenges. We identified four major objectives: a) to enable flexible search and exploration of biomolecular pathway networks through the provision of different views of the data, as well as various interactive functionalities; b) to support corroboration of information in a given pathway model using evidence from the scientific literature; c) to facilitate the discovery of new events that are not yet part of a given model, and also quantify the confidence of the events to be used or integrated within the model; and d) to allow curators and users to become active participants of the analytical process, in which they can explore, inspect and revise the automatically extracted information. Feedback provided via revisions will also be used in learning—in an iterative manner—how the confidence value of events can be computed. This type of visual exploratory analysis, also known as visual analytics (Keim et al., 2010), moves the focus towards the user, so that domain experts become a necessary element of the analytical process rather than acting as mere consumers of the results generated by a statistical black box.

We present LitPathExplorer, a visual analytics tool for exploring literature-enriched pathway models. To the best of our knowledge, this is the first tool which combines visual interactive analysis with state-of-the-art text mining and semi-supervised learning methods to facilitate pathway reconstruction and revision according to new evidence from the literature. We have evaluated the accuracy of our methods quantitatively, as well as evaluated the suitability of our tool qualitatively. The Supplementary Material contains further details including the presentation of a use case in which we show how our tool can be used for the exploration, corroboration and discovery of biomolecular reactions.

2 Related work

A number of efforts towards supporting pathway model reconstruction and revision using text mining methods have been previously reported. Many were focussed on the extraction of binary relationships between molecular entities, which then formed the basis for the automatic generation of reaction networks (Chen and Sharp, 2004; Fleenor et al., 2011; Raja et al., 2013; Barbosa-Silva et al., 2011; Mandoli and Chakrabarti, 2015; Usié et al., 2011). Certain approaches were driven by analysis of textual co-occurrences of names of genes, proteins, and small molecules (Fleenor et al., 2011; Barbosa-Silva et al., 2011; Usié et al., 2011), while others exploited handcrafted linguistic patterns (Chen and Sharp, 2004; Raja et al., 2013). Going beyond the extraction of binary relationships, other approaches extracted more complex, typed n-ary relationships, i.e., events, from text, but without the capability to link extracted events with existing target pathway reactions in models (Razhensky et al., 2004; Schmidt et al., 2012; Ravikumar et al., 2014; Ponn et al., 2015). MET, one of the systems evaluated as part of BioCreative V (Dai et al., 2016), supports the extraction of cancer-related events from literature. A reaction network can be automatically constructed, but on a per-document basis. To enable users to view literature-derived evidence and potential extensions to reference pathway models, Czarnecki et al. (2012) and PathText2 (Miwa et al., 2013a) extracted events and subsequently linked them to metabolic and signalling pathway reactions, respectively. The former gathered names of enzymes and metabolites contained in a pathway model of interest and then extracted metabolic reactions—in the form of events—from papers in which they were mentioned. PathText2 utilized reactions from signalling pathway models to query three text mining-driven search systems, in order to retrieve and rank relevant documents. A compilation and benchmarks of different approaches in this area can be obtained from the BioCreative and BioNLP shared task series (Hirschman et al., 2005; Pyyysalo et al., 2015).

In terms of interactive visualization of biomolecular models, early approaches were heavily focussed on the provision of options for editing the model manually rather than on analytical capabilities. BioLuke (Salamonson et al., 1999) is a stand-alone tool specifically for the visualization and manual modification of metabolic pathways. Similarly, Patika (Demir et al., 2002) provides an environment for network visualization, but with a particular focus on cellular events. It provides a sophisticated search functionality and a basic means for binary uncertainty visualization, in which a question mark is overlaid on reactions not verified by experts. Some other visualization attempts (Breitkreutz et al., 2003; Han et al., 2004) focussed on optimizing structure and presentation for very large biomedical networks. In terms of more recent work, WikiPathways (Kutmon et al., 2015b) is a popular crowdsourced effort dedicated to the collaborative curation of biological pathways. For the analysis of pathway networks, it relies on PathVisio (Kutmon et al., 2015a), which computes different statistics based on inputs from plug-ins connected to this tool. SIGNOR (Perfetto et al., 2015) is a repository of causal reactions aimed at storing manual annotations specifically focussed on signalling networks. Its other purpose is to serve as a tool for designing new experiments by making manually annotated data available. While all of the above tools provide graphical means for visualizing pathways, they are lacking in terms of exploratory means for searching, filtering and inspecting the information contained in the pathways in a convenient manner. RenoDoI (Vehlow et al., 2015) and MINERVA (Gawron et al., 2016) are exceptions in the sense that they offer a means for interactive exploration of large biological networks. However, they do not provide native support for automatically linking networks with textual evidence.

Visual tools that aim to integrate interactive visualization of pathways along with the display of supporting textual evidence from the literature have been also proposed. Pathway Studio (Nikitin et al., 2003) was one of the pioneers in providing detailed visualization of pathway reactions with links to supporting evidence in publications. Although a confidence score is provided for each reaction in the pathway, this is determined simply based on the number of evidence passages. Cytoscape (Su et al., 2014) is an open-source collaborative effort for network analysis that, in contrast to Pathway Studio, has the capacity to connect with other external data sources and software packages through plug-ins. For instance, the AgilentLiteratureSearch package1 allows the incorporation of relations extracted via literature search, although these are limited to certain network configurations and are based only on the co-occurrence of entity mentions. Similarly, CABIN (Singhal and Domico, 2007) includes confidence information determined from external experimental evidence. However, this confidence is not visually encoded in the network, but rather in other types of visualizations such as heat maps and histograms.

The above review only includes published work that is of the most relevance to LitPathExplorer; an extended list of other related visualization tools (and their respective properties) is provided in the Supplementary Material. None of the efforts that we have reviewed combines all of the functionalities of LitPathExplorer in a single tool, namely: 1) a flexible way

1 http://apps.cytoscape.org/apps/agilentliteraturesearch
for querying and exploring pathway networks; 2) automatic extraction of
text passages that corroborate and support the discovery of new reactions;
3) multiple coordinated views to inspect different aspects of information
with different types of visualizations; 4) visual cues to indicate erroneous,
missing or contradictory information, and 5) computation and aggregation
of multiple features for the calculation of confidence, facilitated by a semi-
supervised model that learns from user input.

3 Methods
In this section, we present the methods underpinning LitPathExplorer that
support the exploration of pathway reactions based on textual evidence
from the scientific literature.

3.1 Text mining and confidence computation
We firstly describe the automatic processing behind LitPathExplorer,
which includes the extraction of events from the literature and the computa-
tion of event confidence based on different textual and article metadata
features.

3.1.1 Entity and event extraction
To extract events from the literature and link them with pathway react-
ions, we employed methods for named entity recognition (NER) and event
extraction. These were integrated into a unified processing pipeline using
the Web-based text mining workbench Argo (Rak et al., 2012), which ena-
bles the development and execution of reconfigurable, modular NER and
event extraction workflows.

For event extraction, we apply the machine learning-based EventMine
(Miya et al., 2012) trained on merged gold standard corpora (Miwa et al.,
2013b). EventMine finds trigger words indicating events (e.g., binding,
phosphorylates, inhibits), which are assigned event types (e.g., binding,
Phosphorylation, Negative regulation). Subsequently, these are linked
with their respective event participants, i.e., the identified entities, or other
events (if the event in question is complex, i.e., it itself contains events).
The automatically extracted events are then mapped to reactions in pathway
models.

We propose two modes of linking literature-derived events with path-
way models: corroboration and discovery. An event corroborates an
existing reaction in a model only if the following criteria are met: 1) the
event type corresponds to the type of reaction, and 2) each of the entities
participating in the event corresponds to a participant in the reaction. As
diverse representation standards have been adopted by the systems bio-
logy community in encoding pathway models, it was necessary to define
mappings between the types recognized by EventMine and those contain-
ed in a given model. Building upon the work of Ohta et al. (2013), we
defined mappings between the types specified by two different standards
(i.e., BioPAX and Pathway Studio) and EventMine types, which have been
provided as part of our Supplementary Material. Meanwhile, we consider
an event as discovered if it does not meet the criteria for corroboration
described above, but it involves at least one entity that is part of a reaction
in the model of interest. These discovered events are still considered to
be relevant, since they serve as potential extensions to the model. Both
corroborating and discovered events are linked to the model together with
the supporting information, in the form of the sentence text that contains
the extracted event.

3.1.2 Language-derived confidence quantification
We applied several methods for the calculation of language-derived confi-
dence values based on evidence sentences associated with pathway model
reactions. Firstly, each of the extracted events is analysed to assign values
to the following attributes: polarity, i.e., whether or not the event is nega-
ted (and therefore whether or not it occurs), and certainty, which indicates
whether any speculation is expressed towards the event. Determining the
value of each of these event attributes was cast as a binary classifica-
tion task that made use of linguistic features from the events extracted.
The classification was carried out through the combination of a random
forest (RF) classifier and a rule-based model, both developed based on
the GENIA Metaknowledge corpus (Thompson et al., 2011). The events
contained in this corpus have been annotated in terms of polarity (whose
value is either positive or negative), and certainty level whose value is
one of the following: considerably speculative, somewhat speculative and
non-speculative. For our own purposes, we consider the first two of these
labels to correspond to the uncertain value, and the last one to cor-
respond to certain. The RF classifier was trained on a rich set of lexical,
semantic and syntactic features, while the rule-based model relies on a set
of rules that capture syntactic dependencies between linguistic cues and
event triggers, automatically induced based on patterns in annotated text
(Zetz et al., 2017). An enumeration of these features can be found in the
Supplementary Material.

Initially, the predicted confidence value for an event in a given sentence
is either 1 for certain events or 0.5 for uncertain events. As explained in
Section 3.2.4, the user can set the confidence to any value within the
[0, 1] range, where the sign represents the polarity value for the event within
the sentence. An article-level language-derived certainty value for each event
is obtained by averaging the certainty values of all event instances that
map to the same representation (type, participants and roles) within
the same article. The sign of this article-level certainty-value represents
the article-level polarity of the event.

3.1.3 Metadata extraction
In addition to the language-derived confidence values, other features were
extracted for each scientific publication whose text mentions a corrobora-
ting or discovered event. Bibliographic metadata such as title, journal and
publication date were extracted using Entrez Programming Utilities,3 a
publicly available application programming interface (API) for accessing
various databases including PubMed. Citation counts, which were also
obtained from the Entrez API, were exploited to quantify the authority
of the publication based on other publications. Another feature that we
extracted in order to measure a paper’s popularity is the Altmetric score,4
which has been shown to represent a complementary view to citation-based
metrics (Costas et al., 2015). This complementarity derives from the fact
that Altmetric finds mentions of scientific articles within other platforms,
e.g., citations in Wikipedia and in public policy documents, mainstream
media coverage, bookmarks on reference managers, and mentions in social
media. Finally, we used as an indirect measure of reputation, an arguable
yet commonly used metric, i.e., the impact factor of the journal in which
the paper was published (Saha et al., 2003).5

3.1.4 Event confidence computation
One of the objectives of LitPathExplorer is the quantification of confidence
in the validity of a reaction with respect to a model. Such quantification
encompasses multiple aspects (e.g., evidence in the literature, reputa-
tion of the sources), which should be taken into account and also made

4 http://api.alexmetric.com/
accessible to the user to inform decision-making. We also observed that confidence can be measured at multiple levels of granularity, i.e., at a sentence, document and corpus level. Finally, confidence values are: 1) user-dependent, i.e., different users may quantify the certainty of an event differently based on their own experience or knowledge; and 2) dynamic, as new experimental results are being published continuously, which may or may not necessitate revisions to the pathway model. Therefore, our approach to quantifying confidence for each event consists of firstly building a multi-attribute vector $v$ using our extracted features, i.e., citation counts, language-derived confidence, etc. (the specific details of which can be found in the Supplementary Material), and subsequently normalizing the vector by dividing each feature by its maximum value, such that $|v_\alpha| = 1$. In the absence of any user supervision, the initial value of the event confidence score is computed by taking the average over all attributes.

As different features have varying degrees of correlation with the actual likelihood of the event taking place (e.g., the number of papers mentioning an event sometimes may be more important than the language or wording used in those papers), a simple average of these features may not be optimal when estimating confidence scores. To this end, we developed an approach, which is underpinned by a neural network (NN), that learns how to combine these features based on user input. Initially, the NN is trained to approximate the initial average. When a user overrides an aggregated confidence value—and thus revises the event—the NN is retrained such that its weights are modulated to approximate or target the value specified by the user for the event instance in question. In this way, rather than imposing an arbitrary means for aggregating event confidence values, the system learns from user feedback how confidence should be quantified.

A two-layer feed-forward NN is thus employed, where sigmoid functions are used in the activation of the NN nodes. An important design decision dealt with the issue of how to ensure that the network prioritizes user-specified values without overfitting the model. Training on mini-batches was thus carried out, ensuring that user-revised events were always included in each mini-batch along with other non-revised events (which use the average-based confidence values as target values for training). In addition, a max-norm constraint on the weights of the network nodes was used as a regularizer. Further details about the NN architecture, model selection, and training procedure can be found in the Supplementary Material.

3.2.2 Network Viewer

Our graph-based visualization represents the main component of LitPath-Explorer, located on the left side of the interface (Fig. 1-B). It visualizes the subset of the model that matches the user’s query, based on a node-link representation: nodes represent entities and events, while links connect event nodes with the entities that take part in the event. Different glyph icons are used to differentiate between entity and event nodes. In turn, different colours are used to represent different types of events available in the model. The confidence values described in Section 3.1.4 are also visually encoded in the Network Viewer. The size of an event node is proportional to the confidence value computed for the corresponding event. The same analogy applies to the width of the links that connect event nodes, which are proportional to the confidence value for the event.

As we explained in Section 3.1.1, pathway models are enriched with biomolecular events extracted from scientific literature. Events that are not part of the model but involve at least one entity present in the model are referred to as discovered events. Such events are shown in transparent colours (the specific hue is dependent upon event type) and linked by dashed lines. Moreover, events with evidence sentences that seem contradictory are shown with a glowing yellow border, to enable users to quickly identify them.

Graphs are intuitive representations, but have the disadvantage of not being able to scale up well. Dynamic generation of graphs, i.e., drawing them on demand, and interactive capabilities through which users can inspect and modify graph components, have therefore become a means for bypassing this scalability limitation (Kerren and Schreiber, 2014). However, alternative layouts may be selected, including those for bypassing this scalability limitation (Kerren and Schreiber, 2014). The visual analytics mantra, initially proposed by Shneiderman (1996), highlights the need for a sequence of “overview”, “filter” and “details on demand” as part of the process for allowing exploratory search and visual analysis. These “details on demand” are prompted by different user actions that can be performed on the graph. When hovering the mouse over a node, additional information about the event or entity is shown, and their connected nodes are highlighted. Clicking on a node allows details to be viewed in the Inspector panel (Section 3.2.3), while shifting-clicking on an event node expands the graph to show other events that are connected to this entity. This functionality complements query-based search and facilitates the visual exploration of the network by querying different parts of the network ‘on demand’.

Several options for filtering and manipulating the network are available at the bottom of the Network Viewer panel. Maximum and minimum confidence values can be set to filter out events that do not meet the specified range. This functionality has a major role in an expert’s analysis in order to identify interesting entities and verify their reactions. The layout of the network is initially determined by the well-known force-directed layout algorithm (Fruchterman and Reingold, 1991), Kerren and Schreiber, 2014). However, alternative layouts may be selected, including those which organize and sort the event nodes by confidence, or which aim to separate entity nodes from event nodes. Finally, options for fusing nodes together, as well as for showing/hiding discovered or negated events, are also available. The option for fusing nodes is particularly useful in eliminating redundancies inherent in the model (e.g., duplicated reactions or the same entity being duplicated in different reactions) from the view.

3.2.3 Inspector

The Inspector panel (Fig. 1-C) facilitates closer inspection of entities and events in the model. The nodes selected in the Network Viewer can be visualized through this panel, along with a sequential list of all currently visible entities and events that are connected to the selected node.
The user can interact with this panel to inspect the confidence value computed for every event. Additionally, a bar is displayed to visually encode this value. As domain experts need to understand the reasons behind such confidence assignments, they can expand the bar plot to show a breakdown of the feature values that were used to calculate the confidence value of the selected event. If further information is required, it is possible to inspect supporting evidence found in the literature, as explained in Section 3.2.4. A final but nevertheless important functionality allows a user to modify an event’s current confidence value based on his or her own judgement/knowledge. Such a modification will initiate changes in the visual representation of the network (i.e., size of event nodes) as well as in relearning other confidence values based on the user input (i.e., the neural network will use the newly assigned confidence value as a labelled instance, retrain the model and update other event confidence values). The bars showing the breakdown of the confidence value for a user-modified event are then blurred to indicate that the value was manually assigned. More details about the retraining procedure and parameters can be found in the Supplementary Material.

3.2.4 Text Analyzer

The last panel is the Text Analyzer (Fig. 1-D), which visualizes textual evidence that has been found in scientific literature for the specific event selected in the Inspector panel. The information shown in this panel is organized into two different levels. At the first level are scientific articles mentioning the specific event, which are shown together with their metadata. At the second level are the sentences containing specific pieces of text describing biomolecular events. Words referring to the entities involved in an event of interest are shown in bold, while trigger words denoting the type of event are underlined. The language-derived confidence value can be inspected and modified, both for each sentence and for an article.

Being able to inspect sentences helps the user to check the language that was used in reporting a particular biomolecular reaction as well as its surrounding context. However, analyzing each sentence in this manner can easily turn into a tedious task when several—possibly, contradictory—sentences from different publications exist. For this reason, we have also implemented word tree visualization (Wattenberg and Viégas, 2008; Görg et al., 2010), which supports the visual inspection of multiple sentences at the same time (Fig. 2). This visualization builds a suffix-tree (or prefix-tree) based on a given focus word, such that pertinent sentences are arranged in a tree showing all the sentences finishing (or starting) with the focus word. In addition, a sentence’s position within the tree’s vertical arrangement, as well as the shade of its font colour, indicate its language-derived confidence value. The user can change the focus word as well as switch between visualizing a suffix or prefix tree. To complement word tree visualization, LitPathExplorer also offers trigger visualization which shows the different trigger words that have been used in the literature to describe a given event.
This provides a different analytical perspective on the text, as it allows a user to quickly grasp the frequency as well as the confidence value assigned to an event involving entities of interest. A screenshot of this visualization is shown in the Supplementary Material. Other interactive features available in this panel include redirection to the PubMed entry of the article and modification of sentence and article-level confidence values.

4 Results

This section reports on the performance of the various methods underpinning LitPathExplorer, and on a user-centred evaluation of the tool. The reader is referred to the Supplementary Material for a use case on how a domain expert can use the tool to navigate through a model and find reactions for further experimentation.

To support these evaluation activities, we exploited the following resources: a pathway model in the BioPAX format, which encapsulates reactions involving the Ras protein, and a corpus of 12,660 full papers on breast cancer. The former was the output of querying PathwayCommons (Cerami et al., 2010) for one- and two-hop reactions centred on Ras. The corpus, meanwhile, is a resource that was constructed in-house as part of our work for the Big Mechanism program (Cohen, 2015), in which we automatically generated propositions pertaining to the expression of breast cancer genes by chemical compounds, which were then experimentally validated using specific cell lines. The full papers in the corpus were retrieved from the PubMed Central Open Access repository using “breast cancer” and its synonyms as keywords, combined with names of breast cancer cell lines, e.g., “T-47D”, “MCF-7” (and their variants). The methods for event extraction, model mapping and confidence computation described in Section 3.1 were applied to these documents, the results of which formed the basis of the currently available version of our visualization tool.

4.1 Quantitative Evaluation

In this section, we focus on evaluating the performance of the methods which underpin LitPathExplorer. We firstly report on the precision and recall of our event extraction workflow independently of any pathway model (i.e., based only on text). We then assess the ability of TM to discover events relative to a given model. Finally, we evaluate our proposed approach for quantifying the confidence of events, in the context of determining events that are most suitable for updating a pathway model, before and after user supervision is taken into account.

Using Argo, we developed an event extraction workflow (described in Section 3.1) that was evaluated in the context of the Shallow Reading task of the Big Mechanism program (Cohen, 2015). Its performance was measured against a small corpus of cancer biology article sentences, which contains 154 gold standard (i.e., manually annotated) events corresponding to signalling pathway events. Performance was computed based on two metrics: (1) precision, i.e., the number of automatically extracted events which are correct, and (2) recall, i.e., the number of gold standard events which were extracted. There are three points to note regarding the scoring scheme that was employed by the Shallow Reading task organizers. Firstly, in counting correct events, half a point was given to partially matching events. For example, if an extracted event has the same trigger word as a gold standard event but has missing participants, a partial score of 0.5 would be given. Secondly, in calculating precision, the task organizers manually checked each of the extracted events to give credit to ones which are not in the gold standard but may also be considered as correct. Lastly, 23 of the 154 gold standard events were optional: points were awarded to automatically extracted events that matched optional events; however, no penalty was given for missing them.

Our workflow extracted a total of 115 events. Out of these, 95 were fully correct and 15 were given half a point yielding a precision score of 102.5/115 = 89.13%. Meanwhile, our workflow was able to extract 89 out of 131 non-optional events in the gold standard, as well as 17 out of the 23 optional ones, resulting in 71.62% recall (i.e., 106 out of 148).

Evaluating the accuracy of LitPathExplorer with respect to a model while in discovery mode requires manually checking the validity of discovered events, which is a costly and time-consuming process. In order to avoid this expensive type of evaluation, we reverse-engineered the construction of the model: we removed a subset of its reactions (i.e., a subnetwork) and then evaluated the extent to which we could recover them from the literature. In order to systematically remove subnetworks, we firstly grouped reactions according to their distance from an entity of interest. Distance was defined in terms of hops: the minimum number of event nodes between two entity nodes. In our evaluation, for example, we removed the subnetworks comprised of reactions that were within one hop and two hops away from the Ras protein. We refer to these subnetworks as $S_1$ and $S_2$, respectively.

After removing $S_1$ and $S_2$, we re-ran our text mining workflow described in Section 3.1.1 on the same corpus of 12,660 papers that formed the basis of our currently available version of LitPathExplorer. We first analyzed the set of discovered events that are one hop away from Ras, which we refer to as $D_1$. The goal was to compare the discovered events, $D_1$, with reactions in the removed subnetwork, $S_1$. We then repeated this procedure for the two-hop subnetwork, i.e., comparing discovered events $D_2$ with $S_2$.

Upon examination of the results, the extracted events in $D_1$ obtained perfect precision, i.e., 100%, while we were able to recover 30.43% of the reactions in the one-hop pathway subnetwork, $S_1$. Results for the two-hop subnetwork are lower: 46% of the reactions in $S_2$ were found in $D_2$ while only 17.1% of the $S_2$ reactions were matched by events automatically extracted from the corpus. However, some points are worth noting in interpreting these results. While the corpus is focussed on breast cancer, with which Ras is known to be associated, not all of the reactions in the Ras-centric pathway model are necessarily covered by our set of papers. Thus, some of the reactions in the model may be impossible to recover from this corpus. Moreover, our results are in line with those reported by analogous systems on the same pathway network (Cohen, 2015). Also, as stated at the beginning of the paper and by Cohen (2015), this pathway network is neither complete nor error-free, and even biologists have disagreed on the validity of the relationships. Finally, we noticed that the difficulty of extracting complex (higher-order) events from text accounted for around 10% of the recall loss (i.e., only six complex events extracted from the whole corpus were discovered and matched to the model).

An important metric that is useful in evaluating the system is a measure of how the confidence values of the discovered events correlate with the likelihood of them being part of the removed subnetwork. By ranking the discovered events by confidence, we can measure average precision (AP), a standard metric in information retrieval. Intuitively, this metric evaluates
whether the events with the highest confidence values are the ones that appear in the model or not. While the pathway model cannot be taken as a “gold standard”, we can consider it a good approximation of such. The AP for the $D_2$ events using the average-based confidence is 63%, which suggests that the extracted attributes are useful as predictors of the validity of these events. This value corresponds to the beginning of the solid line in Fig. 3 (i.e., when no user revisions are made on the event confidence values).

Yet, when the neural network is trained on up to half of the $D_2$ events—emulating a user adjusting the event scores to 1 or 0 based on their validity—we see that the AP on the other half of the events drastically increases with the number of user revisions (solid line in Fig. 3). This figure also shows that when we take all of the events found in the literature into account (dotted line), and evaluate them on a stratified sample, the initial AP (no NN training) is similar to that for $D_2$. However, after several user revisions, AP increases up to 95% when evaluated on unlabelled (i.e., non-trained) events. When all the events are considered and no stratified sample is taken (dashed line), we can see that the initial AP is low due to the high imbalance between corroborated and discovered events. Yet as more events are labelled, the AP gets consistently higher and rises to 80%. Further experiments using other NN configurations and supervised models can be found in the Supplementary Material. Finally, a scatter plot of the confidence values for events in $D_2$ is shown in Fig. 4. The plot illustrates that after using the NN for learning based on user feedback (top), the system assigns higher confidence values to corroborating events than when no learning is applied (bottom).

4.2 Qualitative evaluation

We evaluated LitPathExplorer by running a small user study with three domain experts (Users 1, 2 and 3). Firstly, we gave them an overview of the tool, in which we explained its different panels, the meaning of the different visualization encodings and the available interactions. We then asked each participant to help in analysing a pre-defined use case. This activity follows what is known as pair analytics (Arias-Hernandez et al., 2011), in which a domain expert and tool developer sit together to perform an analytical task. This is a commonly adopted method for evaluating the exploratory features of a visual analytics tool independently of its usability aspects. Throughout the exercise, we encouraged users to ask questions, provide feedback and even interrupt the flow of discussions when necessary. The discussions were documented by an independent observer. Finally, the participants used the tool on their own use cases without our supervision. They were then asked to provide their feedback by filling out a questionnaire, the results of which are presented in the Supplementary Material.

Overall, the users expressed positive feedback on their experience while using LitPathExplorer and none of them encountered any difficulty in grasping the objectives of the tool. Each user agreed that combining the text of sentences from the literature with model visualization is very helpful when trying to make sense of a pathway model. In this context, User 2 added that it would be desirable to additionally extract details of laboratory/experimental conditions from the text if they are mentioned, while User 3 mentioned that it would be desirable if, in addition to linking to the PubMed entry of the article, the tool highlighted the specific evidence sentences in context. The discovery mode of the tool was unanimously praised, even when noisy or unwanted reactions were shown, since they can be organized by confidence and easily inspected in any case. They liked the ability to enable or disable discovery mode, and that discovered events can be overlaid on the model. User 2 suggested the provision of an option to hide pathway reactions in the model, so that only text-mined events will be displayed. This functionality is indeed very relevant to her own use case, as she is primarily looking for reactions that have not yet been reported.

In terms of confidence handling, the users liked the fact that they can inspect the different feature values provided by LitPathExplorer and that they can calibrate event confidence values if necessary. The search and filtering options were heavily used during the pair analytics session, especially the one for filtering by confidence. Users concurred on the advantages of querying a subset of the model and then filtering the network or expanding a node on demand. Users 1 and 2 also commented on the possibility of specifying a list of entity names for filtering and searching as opposed to defining those interactively one at a time. This is useful, for instance, when the user wishes to apply a filter based on a black list of compounds that are unavailable in the laboratory. This functionality was incorporated into a post-evaluation version of LitPathExplorer.

Based on the users’ reflection on how they carried out their respective use cases with LitPathExplorer, User 1 provided the most informative feedback. His task involved assembling a Petri net model (Murata, 1989) in order to predict the effects of drugs on cell lines related to cancer. He concluded that by using LitPathExplorer to identify reactions with high confidence accompanied by textual evidence, he managed to improve the accuracy of his predictions. He stated that he was able to predict the effect of most drugs on cell viability based on the TM results. Some other positive feedback from the users touched upon: the clean design of the interface, user revisions on event confidences, provide feedback and even interrupt the flow of discussions when necessary. The discussions were documented by an independent observer. Finally, the participants used the tool on their own use cases without our supervision. They were then asked to provide their feedback by filling out a questionnaire, the results of which are presented in the Supplementary Material.

Overall, the users expressed positive feedback on their experience while using LitPathExplorer and none of them encountered any difficulty in grasping the objectives of the tool. Each user agreed that combining the text of sentences from the literature with model visualization is very helpful when trying to make sense of a pathway model. In this context, User 2 added that it would be desirable to additionally extract details of laboratory/experimental conditions from the text if they are mentioned, while User 3 mentioned that it would be desirable if, in addition to linking to the PubMed entry of the article, the tool highlighted the specific evidence sentences in context. The discovery mode of the tool was unanimously praised, even when noisy or unwanted reactions were shown, since they can be organized by confidence and easily inspected in any case. They liked the ability to enable or disable discovery mode, and that discovered events can be overlaid on the model. User 2 suggested the provision of an option to hide pathway reactions in the model, so that only text-mined events will be displayed. This functionality is indeed very relevant to her own use case, as she is primarily looking for reactions that have not yet been reported.

In terms of confidence handling, the users liked the fact that they can inspect the different feature values provided by LitPathExplorer and that they can calibrate event confidence values if necessary. The search and filtering options were heavily used during the pair analytics session, especially the one for filtering by confidence. Users concurred on the advantages of querying a subset of the model and then filtering the network or expanding a node on demand. Users 1 and 2 also commented on the possibility of specifying a list of entity names for filtering and searching as opposed to defining those interactively one at a time. This is useful, for instance, when the user wishes to apply a filter based on a black list of compounds that are unavailable in the laboratory. This functionality was incorporated into a post-evaluation version of LitPathExplorer.
the wordtree visualization, and being able to fuse entities or events. A few other criticisms were raised, including the tool not being able to display synonyms for entity names, default force-based layout being distracting and lack of a means for enumerating discovered events.

5 Conclusion
We have presented LitPathExplorer, a visual text analytics tool providing: 1) a flexible way for querying and exploring pathway models; 2) integration of automatically extracted events from the literature to corroborate reactions in the model and discover new ones; 3) computation and aggregation of multiple features for the calculation of confidence; and 4) close involvement of the domain expert in the analytical process, where the tool also learns from user feedback using a neural network. Based on our survey of related efforts, our tool is the first to offer the above functionalities.

LitPathExplorer was evaluated in two ways. Firstly, it was evaluated in terms of the performance of its underpinning methods. Based on this, we established that events extracted from a document collection can be discovered with a precision of 89%. Meanwhile, our novel approach to quantifying event confidence with regard to their existence in the pathway model represents a reliable indicator of event validity and that average precision can reach 80% or higher when employing a neural network to learn from user input. Additionally, we recruited three domain experts who employed the tool in addressing their respective use cases. All of the three users provided positive feedback, and more importantly, were unanimous in their evaluation of the tool as a valuable means to help them in their pathway modeling tasks.

As future work, we will incorporate features for predicting confidence that could, for example, take into account experimental information (e.g., co-expression) from biological databases. We will also investigate methods for improving the extraction accuracy of higher-order events. Furthermore, we plan to support the Systems Biology Graphical Notation (Le Novere et al., 2008), and implement functionalities that will allow users to load their own models as well as execute the TM workflow on a corpus of their choice.

Acknowledgements
The authors thank Paul Thompson and John McNaught for proofreading, Dr. Paul Dobson for insightful discussions that helped design our tool, and the three domain experts who helped evaluate it.

Funding
This work has been supported by BBSRC, Enriching Metabolic PATHway models with evidence from the literature (EMPATHY) [Grant ID: BB/M006891/1]; the Defense Advanced Research Projects Agency (Big Mechanism [Grant DARPA- BAA-14-14]); and the Engineering and Physical Sciences Research Council [EP/P010809/1 (CDT)]

References


