Supporting Biological Pathway Curation Through Text Mining

Sophia Ananiadou(✉) and Paul Thompson

School of Computer Science, National Centre for Text Mining, University of Manchester, Manchester, UK
{sophia.ananiadou,paul.thompson}@manchester.ac.uk

Abstract. Text mining technology performs automated analysis of large document collections, in order to detect various aspects of information about their structure and meaning. This information can be used to develop systems that make it much easier for researchers to locate information of relevance to their needs in huge volumes of text, compared to standard search mechanisms. With a focus on the challenging task of constructing biological pathway models, which typically involves gathering, interpreting and combining complex information from a large number of publications, we show how text mining applications can provide various levels of support to ease the burden placed on pathway curators. Such support ranges from applications that provide help in searching and exploring the literature for evidence relevant to pathway reactions, to those which are able to make automated suggestions about how to construct and update pathway models.

Keywords: Text mining · Semantic search · Biological pathway curation · Event extraction · Named entity recognition

1 Introduction

Pathways are key to understanding biological systems. However, the construction of pathway models is dependent upon a complete and accurate representation of these systems, which requires that all relevant molecular species are captured, together with their physical interactions and chemical reactions.

Typically, complete information about the complex mechanisms involved in pathways is highly fragmented amongst many different information sources, making it necessary to assimilate knowledge from a plethora of heterogeneous sources, including not only the scientific literature, but also databases and ontologies. As an example, a manually constructed model of the *mToR* signaling network identified 964 species connected by 777 reactions, and reconstruction of this network required information to be gathered from a total of 522 publications [1].

The complex nature of reconstructing pathways involves not only *finding* appropriate information, but also evaluating, interpreting and distilling details from many different sources in order to construct a coherent and accurate model [2]. For example, it is not sufficient simply to locate statements in literature that provide evidence of relevant

L. Kalinichenko et al. (Eds.): DAMDID/RCDL 2016, CCIS 706, pp. 59–73, 2017.
DOI: 10.1007/978-3-319-57135-5_5

reactions. Rather, it is also important to take into account the contextual details that accompany them (e.g., the degree of confidence or certainty expressed towards a finding) in order to determine their suitability for inclusion within the model. Furthermore, due to constant advances in scientific knowledge and the accompanying rapid growth in the literature, pathway models are not static objects. Instead, they must frequently be refined, verified and updated.

The various intricate stages of reconstructing pathway models typically require considerable amounts of manual effort by domain experts [3, 4]. Indeed, the slow and laborious nature of gathering and interpreting relevant information represents a major barrier to creating and maintaining pathway models, and the overwhelming volume of information that has to be reviewed can result in important details being overlooked.

In this article, we provide an overview of how text mining (TM) methods can support curators of pathways by allowing them to explore the vast body of scientific literature from a *semantic* perspective. This can significantly increase the ease and efficiency with which they can pinpoint, make sense of and use relevant information to construct and update pathways models. As such, TM has the potential to considerably increase the speed and reliability of knowledge discovery [5, 6].

2 Searching the Literature

In order to search for literature evidence that is relevant to support pathway curation, researchers will typically submit a query to a search engine, to try to retrieve a suitable set of publications. However, the querying mechanisms available in most search engines are poorly aligned with the information needs of the researcher.

The goal is usually to discover evidence about reactions (e.g., binding, phosphorylation) that involve different types of concepts (e.g., genes/proteins, chemical compounds, subcellular components). In other words, the researcher is looking for textual evidence for highly specific pieces of *knowledge* that describe relationships amongst concepts.

In standard search engines, querying facilities are usually restricted to locating documents that contain particular sets of words and/or phrases *somewhere* within them. However, isolated words and phrases do not in themselves convey knowledge. Documents have highly complex structures, and knowledge is conveyed according to the *meaning* of phrases and how they are *related* to other phrases in the document.

Consider that a researcher wishes to discover proteins that bind to the protein *Mad1*. Using a standard search engine, a possible query would be *Mad1 AND bind*. Although this query is likely to retrieve *some* relevant documents, the fact that it is not possible to specify *how* these search terms should related to each other means that there is a very high probability of retrieving many documents where there is no specific relation between the terms *Mad1* and *bind*. This implies a great deal of tedious "sifting" through irrelevant documents to isolate those of interest.

A further issue is that language usage is creative and unpredictable. Both concepts and the relationships in which they are involved can be expressed in text in a variety of different ways, using synonyms, abbreviations, acronyms, paraphrases, etc. For

example, *Mad1* could appear in text as *MAD-1, mitotic arrest deficient-like protein 1* or *MAX dimerization protein 1*, while binding relationships could be expressed using words or phrases such as *recruit, forms a complex with*, etc. Trying to account for all such possible variants in a query would be virtually impossible, which leads to inevitable overlook of certain relevant documents.

Furthermore, many search terms can be ambiguous. For example, *ER* is a common abbreviation for *estrogen receptor*, but the same abbreviation may be used to refer to *endoplasmic reticulum* (a cellular subunit) and *emergency room*, amongst others. In the absence of sense-based filtering, searches for such terms will retrieve a large number of irrelevant documents.

3 Text Mining

Text mining methods can offer solutions to the issues above, by aiming to recognise various aspects of the semantic structure of documents [7]. TM accounts for the fact that (a) words and phrases have specific meanings, (b) particular meanings can be expressed in different ways, and (c) words and phrases can be related to each other in many different ways to convey complex information, such as reactions that are relevant to pathway curation. The results of TM analyses can be exploited in different applications that make it easier for researchers to pinpoint, filter and explore information that is of direct relevance to them. Such applications aim to minimise the tedious task of reading through irrelevant information, and reduce the likelihood of overlooking potentially vital information.

TM systems usually consist of a complex pipeline of different tools, which perform various levels of analysis that are required to gain an understanding of the information expressed in text. The functionalities of these tools range from low-level tasks, such as splitting a text into sentences and identifying the individual words within them, to more complex tasks, such as determining the parts-of-speech of individual words (e.g. nouns, verbs), grouping words into phrases and identifying structural (syntactic) relationships between these phrases, etc.

Typically, various tools are available that can carry out each individual task. This means that there are potentially many ways in which different tools could be combined to create TM processing pipelines. Since different tools may interact with each other in different ways, alternative pipelines with the same overall goal may perform with varying levels of accuracy. Accordingly, it can be advantageous to consider various combinations of different tools in order to construct an optimal TM system.

Until fairly recently, this could be a complex task, since different tools are implemented using various programming languages and have different input and output formats. The Argo TM platform [8] offers a solution to such issues, by providing a large (and continually growing) library of *interoperable* tools [9], which employ standardised input and output formats (based on the standards defined by the Unstructured Information Management Architecture (UIMA) [10]). This allows tools in the library to be flexibly combined into pipelines that carry out various different types of textual analysis (via a web-based, graphical user interface). Each pipeline can be evaluated and compared

against others to find the best solution [11]. This makes it straightforward to build systems that perform various complex analyses of text, such as those described in the subsequent sections.

4 Concept Recognition

As outlined above, typical literature search goals revolve around discovering information about specific *concepts* rather than words. A researcher who searches for *estrogen receptor* will normally be interested in finding out information about a specific protein, regardless of the various ways in which it may be denoted in text (e.g. *oestrogen receptor, estrogen-receptor, ER* etc.).

The application of two well-established TM methods, i.e., named entity recognition (NER) and normalisation, can help to facilitate searching at the level of concepts rather than words. NER involves automatically identifying words and phrases in text that denote concepts of interest, and assigning a semantic label according to the concept category that they represent (e.g., *protein, chemical compound*, etc.). In the normalisation step, each phrase identified by NER is automatically *grounded* to a unique concept (usually by linking it to a domain specific database of known concepts). Accordingly, normalisation effectively identifies and links together all possible ways in which a concept could be expressed in text. Given the many potential variant expressions for a concept, combined with the fact that certain expressions may have multiple senses, the normalisation process may include disambiguation of acronyms according to their context [12] and comparison of phrases according to various types of surface-level and semantic similarities between them [13].

4.1 Kleio

The results of applying NER and normalisation to a collection of texts can form the basis of semantically-oriented search systems, such as Kleio[1] [14], which facilitates enhanced search over MEDLINE abstracts. In Kleio, semantic restrictions can be placed on query terms (such that, e.g., only documents in which *ER* refers to a protein are retrieved), and also to ensure that documents that use different ways of referring to the same concept are also automatically retrieved. This provides an important step towards reducing both overlook and overload of information.

Other functionality provided in Kleio further demonstrates the power of semantic TM analysis. Using NER and normalisation results for several different concept types, obtained by processing the whole of the MEDLINE, a *faceted* search mechanism makes it easier to explore and filter the results of an initial query according to their semantic content. Faceted search presents other concepts that frequently co-occur in the same documents as concepts of interest (e.g., proteins, symptoms, drugs, etc.). Frequent co-occurrences are likely to indicate interesting relationships.

[1] http://www.nactem.ac.uk/Kleio/.

Figure 1 illustrates a search in Kleio for the protein *Mad1*. Opening the *Protein* facet reveals other proteins mentioned in the same documents as *Mad1*, the most common being *Mad2*. By selecting *Mad2* from the list, it is possible to "drill down" to documents in which the two proteins are mentioned together, and thus determine the nature of the relation that holds between them. The text snippets displayed, which mention a *Mad1–Mad2 complex*, provide evidence of their binding ability.

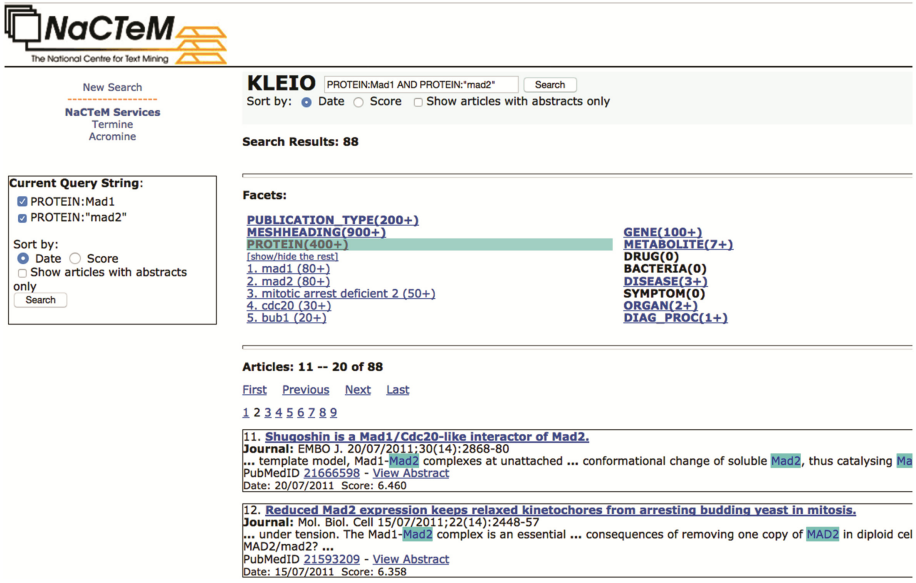

Fig. 1. Faceted search in Kleio

4.2 FACTA+

FACTA+[2,3] [15, 16] uses a type of TM analysis similar to Kleio, and has the aim of making it easy to find and visualise associations that occur between different concepts mentioned in MEDLINE abstracts. An additional useful feature of FACTA+ is the ability to specify that abstracts retrieved by a search should additionally contain mention of a *relationship* of a given type, e.g., *binding* or *positive regulation*, which can help to further filter the search results. A machine learning module determines the different ways in which each type of relation may be described. For example, searching for a binding relationship will automatically retrieve documents that include various different phrases that can denote binding, such as *recruitment, crosstalk, engagement, binding* and *interaction*.

In FACTA+, it is possible to discover both directly associated concepts (i.e., concepts which are mentioned together in the same abstracts) and indirectly associated

[2] http://www.nactem.ac.uk/facta/.
[3] http://www.nactem.ac.uk/facta-visualizer/.

concepts. This latter functionality aims to account for the fact that associations may exist between concepts even if they are never mentioned together in the same document. The discovery of potential indirect associations works on the assumption that if two concepts *A* and *B* frequently occur together in documents, and if concept *B* frequently occurs together with a third concept, *C* (possibly in a distinct set of documents), then there is a possibility that concepts *A* and *C* also share some kind of association, via the "pivot" concept *B*.

Figures 2 and 3, respectively, illustrate direct and indirect associations in FACTA+ that are retrieved by a search for documents that mention *E-cadherin,* as well as a negative regulation relation. As can be seen in Fig. 2, the directly associated diseases are almost exclusively different forms of cancer. However, indirect associations (Fig. 3) reveal potential links with other diseases. For example, Parkinson's disease and Alzheimer's disease have a possible association with *E-cadherin* via the pivot gene concept *CASS4*. Thus, it could be hypothesised that E-cadherin may be a potential candidate drug target for nervous system disorders.

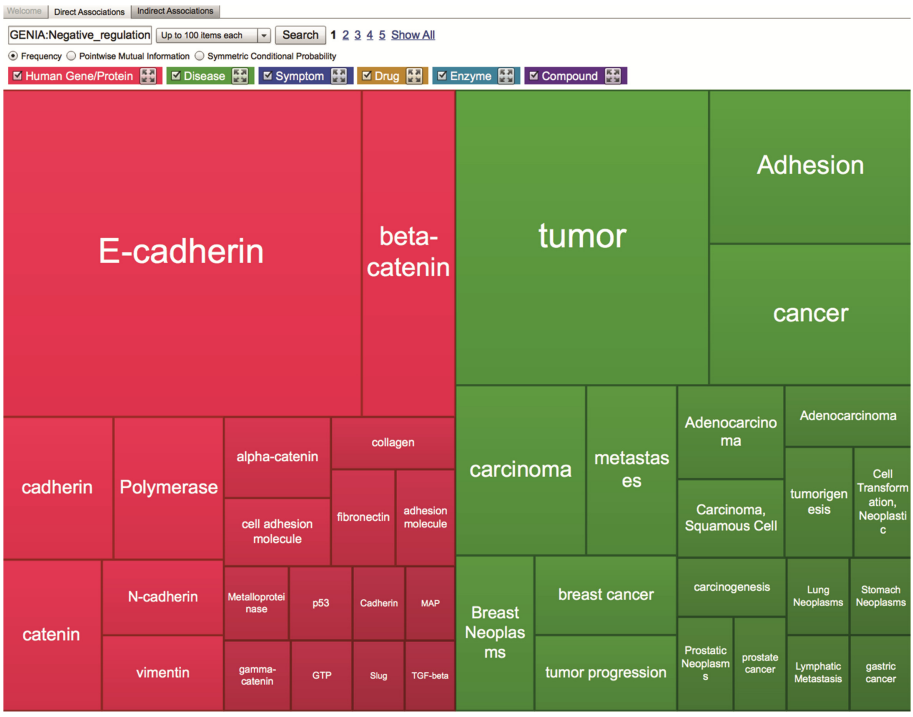

Fig. 2. Direct associations in FACTA+

Fig. 3. Indirect associations in FACTA+

5 Relationship Recognition

Although systems such as Kleio and FACTA+ can make it far easier than a standard search engine to find potential relationships between concepts, they are still essentially "shallow" approaches, which only consider the semantics of individual words and phrases, without considering the exact nature of the relationship (if any) that holds between them. Whilst co-occurrence is used as an indicator of likely relationships, it is still necessary to carefully read the retrieved documents to determine whether a relationship of interest is actually described. For a complex task such as pathway reconstruction, in which many possible reactions have to be considered, this still implies a large amount of work on the part of the researcher.

Since constructing pathway models generally involves searching for very specific types of relationships, much work could be saved through the ability to identify more definite evidence of particular types of relationships that involve concepts of interest. However, this requires a more complex analysis of sentence structure, such that relationships between phrases can be identified.

5.1 MEDIE

MEDIE[4] [17] is a further search engine over MEDLINE abstracts, which combines NER [18] and normalisation with the results of a domain-specific syntactic parser [19]. MEDIE exploits the fact that in many cases, relationships between concepts are specified within the boundaries of a single sentence, and conveyed by means of a verb, where the concepts that constitute the "participants" in the relationship correspond to the syntactic arguments of the verb.

Syntactic analysis permits search queries in MEDIE to be *structured*. Instead of a search query consisting of a set of terms that must co-occur somewhere within documents, MEDIE allows specification of *how* the search terms should be related to each other. Queries in MEDIE are specified in terms of subject-verb-object triples – users can specify values for one or more of these slots, depending on the specificity of their query.

In Fig. 4, values have been specified for both the *subject* and *verb* slots, i.e., *p53* and *bind*, respectively, in order to discover what binds to p53. The query specifically finds sentences in which the word *bind* (or one of its inflections) occurs as a verb and *p53* occurs as its syntactic subject. In the retrieved results, the *object* phrase is highlighted in each relevant sentence, such that proteins that bind to p53 can be readily identified. Deep syntactic analysis allows MEDIE to find relationships described using a variety of sentence structures which may involve, e.g., the passive use of the verb. It can also handle sentences in which the participants in the relation do not necessarily occur in direct proximity to each other, e.g. *Furthermore, **p53** promotes cisplatin-induced apoptosis by directly **binding** and counteracting **Bcl-x(L)** antiapoptotic function.*

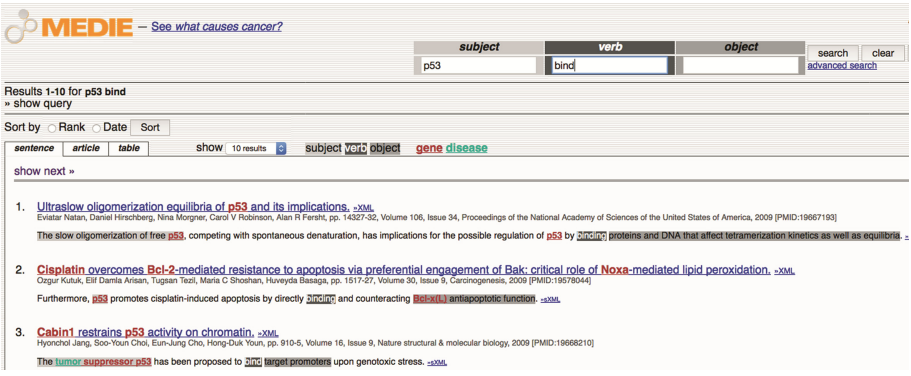

Fig. 4. Structured searching in MEDIE

5.2 Beyond Simple Relations

Although syntactic analysis is a vital part of accurately determining relationships that hold between concepts, MEDIE can only find relationships that are centered around verbs. However, relationships described in other ways, particularly using

[4] http://www.nactem.ac.uk/medie/.

nominalisations, such *interaction* or *phosphorylation*, are highly prevalent in biomedical text [20]. Accordingly, failure to take such relationships into account risks overlooking potentially vital information.

Additionally, MEDIE is restricted to finding relations between *pairs* of concepts. However, such behaviour cannot account for the fact that, in constructing pathway models, it can be important to identify relations that involve multiple participants. For example, it can be vital to take into account the cellular context of a signaling event, such as cell type and localization; such information frequently occurs within the textual context of relationships.

Analysing the textual context of a relationship can also be important for other reasons. In particular, relationships can have different *interpretations*, e.g., they may not necessarily represent definite information. For example, in the sentence *We hypothesize that unphosphorylated cdr2 interacts with c-myc to prevent c-myc degradation*, the potential interaction between *cdr2* and *c-myc* is a hypothesis whose truth value is unknown. However, in building pathway models, it is important that only *reliable* knowledge is integrated into the model.

Whilst such contextual details could be found by careful reading of documents returned by a system such as MEDIE, the volume of documents that must typically be reviewed to construct pathway models means that researchers could benefit from systems that can automatically identify additional details about relationships. This can permit detailed information about the relationships to be presented to researchers, and/or allow more complex filtering of relationships.

6 Event Recognition

Over recent years, a large amount of research has focused on the extraction of complex information structures from text, known as events (e.g., [21–27]). Importantly in the context of constructing pathway models, events can encode interactions that involve an arbitrary number of concepts, and are thus able to capture different types of contextual participants in interactions that are missing from binary relations.

Another important feature of events is their semantically-oriented representations of relationships, which abstract from the surface structure of the text. Firstly, events are assigned labels according to the general type of information that they convey (e.g., *negative regulation, phosphorylation, carboxylation*). Secondly, the participants in events are assigned specific *semantic role* labels (e.g., *modifier, reactant, product, cause, location*) according to their exact contribution towards the description of the relationship.

Generally, event extraction systems aim to capture all events of a given set of types that occur in a collection of documents, regardless of how they are expressed in text (e.g., using different nouns or verbs). This implies much more complex processing than the syntactic analyses carried out by MEDIE. Although syntactic analysis is still an important part of event extraction, the challenge lies in learning how to map from the surface level structure of the text to the more abstract semantic level event representation. For example, events of type *negative regulation* may be centred around verbs such as

inhibit or *inactivate,* nouns such as *repression* or *loss,* or adjectives like *deficient* or *defective*. Depending on the exact word used, the mapping between syntactic structure and semantic participant roles may be different. For example, a possible way of denoting *positive regulation* events is using the verb *activate,* as in the following sentence: *Furthermore, a previous study has reported that* ***SAHA activates*** *p53*. Here, the *cause* of the positive regulation of *p53* is *SAHA,* which corresponds to the subject of the verb *activate*. However, other patterns will hold when the reaction is described by words belonging to other parts-of-speech. For example, if the positive regulation relationship is denoted by the noun *activation*, then it is likely that the *cause* will instead be preceded by the preposition *by*, i.e., ***activation*** *of p53 by* ***SAHA***.

Despite the many complexities of recognising events automatically, they make it feasible to search for reactions based purely on a high-level semantic representation of the researcher's information needs, which is totally independent of the many ways in which the relevant knowledge may be expressed in documents. Specifically, event-based searching can allow the researcher to specify that they are looking for an event of a particular semantic category, and to place restrictions on the types of semantic participants being sought, without worrying about the many potential ways in which this information could be expressed in text.

6.1 EventMine

In order to extract complex events automatically, we have developed a pipeline-based event extraction system, EventMine [28], which employs a series of classifier modules to capture core event elements: detection of triggers (words or phrases that characterise the event), detection of edges (finding links between pairs of concepts), and complex event detection (combining edges into complex n-ary relations).

Since extracting event representations from text is heavily reliant on learning how to map from syntactic structure to semantic representations, EventMine uses a rich set of features, including those obtained from two different parsers [29, 30]. The system is also very flexible and can be adapted to extract different types of events without the need for task-specific tuning [31]. A further important feature of the EventMine is its incorporation of results from a pre-executed co-reference resolution method [32]. When event participants correspond to semantically empty expressions such as *it* and *that*, their exact interpretation is determined by looking in other sentences.

The results of applying EventMine to MEDLINE are used in an event-centric version of MEDIE (as illustrated in Fig. 5), in which search criteria are specified entirely in terms of structured semantic representations. Searches can be carried out over a number of different event types, and can place restrictions on participants that have different semantic roles. In Fig. 5, an event-based search for binding events involving *p53* shows that various different ways of expressing the binding relationship are recognised, in which the relation is specified using nouns as well as verbs.

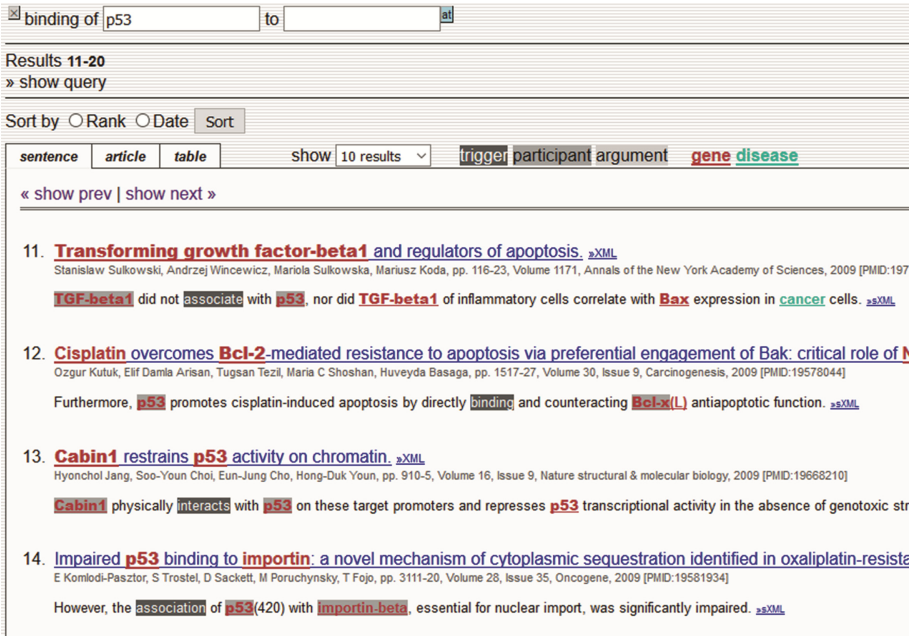

Fig. 5. Event-based querying in MEDIE

6.2 Event Meta-Knowledge

A recent enhancement to EventMine concerns the ability to detect and assign *interpretative* information to extracted events [33]. Values are determined for several different aspects or *dimensions* of interpretation, which we refer to collectively as *meta-knowledge* [34, 35]. By learning from a collection of texts in which event structures are manually annotated with meta-knowledge information [36], EventMine is able to automatically assign to each event its *polarity* (whether the event is negated), *knowledge type* (e.g., whether the event represents a well-known fact, a subject of investigation, an experimental observation or an analysis of experimental results), *certainty level*, *manner* (whether the reaction takes place with high or low intensity) and *source* (whether the event represents new or previously published information).

Such information can be useful in filtering events that describe potentially relevant interactions. For example, when constructing a new pathway model, it is likely to be important to consider *all* events that are considered sufficiently reliable, e.g., those that correspond both to well-known facts and experimental outcomes that are stated with a high degree of confidence, where the association is positive (rather than negated) and possibly excluding interactions that are characterised as being weak. In contrast, if the task is to *update* an existing pathway model to take into account the latest scientific knowledge, then the search may be further narrowed to consider only those events that represent new, confidently expressed experimental knowledge reported within articles published within a certain range of dates.

7 TM Support for Pathway Curation

Many existing pathway models are encoded using machine-readable representation formats such as the Systems Biology Markup Language (SBML) [37, 38] or the Biological Pathway Exchange (BioPAX) [39] format. These models may be exploited in TM applications, based on a mapping that has been defined between these formal models and event structures [40]. This mapping can allow pathway model reactions to be converted automatically into event-based queries, which can be used to find supporting evidence for the reactions in the literature. There is also potential for new events found in the literature to be converted into formal pathway representations, which can then be used to construct/update pathway models semi-automatically.

7.1 PathText^2

The PathText^2 system [41] aims to associate existing pathway model reactions with supporting evidence from the literature. PathText^2 translates pathway reactions encoded in SBML into queries for Kleio, FACTA+ and the both the original and event-based versions of MEDIE. Given that each system identifies associations in a different way, each system may find relationships that are not extracted by the other systems. Thus, the submission of queries to multiple systems is aimed at retrieving a maximal number of documents that contain relevant evidence.

The results returned by each system are combined and presented to the user in a unified interface, ranked according to their relevance to pathway reactions (see Fig. 6). The ranking gives priority to the results of the event-based MEDIE, since experiments have shown this to be the most effective system in retrieving relevant documents. This further reinforces the importance of event extraction.

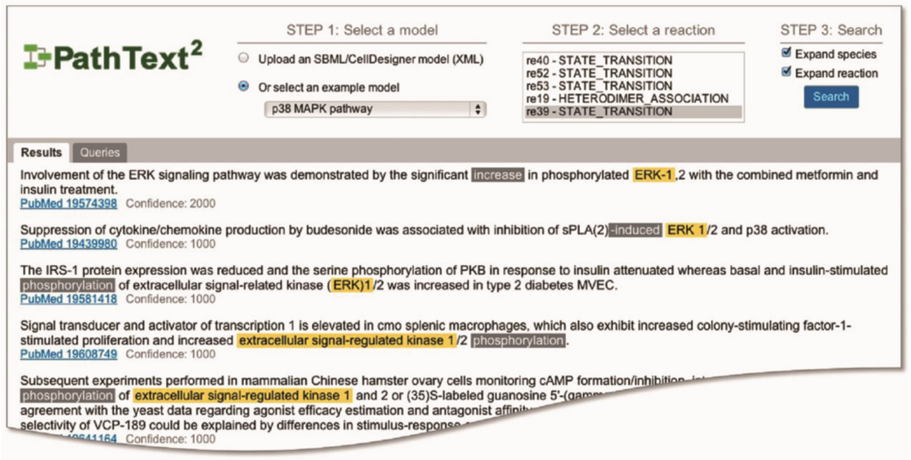

Fig. 6. Literature evidence for pathway reactions in PathText^2

7.2 Big Mechanism

The ability to link between formal pathway models and textual events is also being exploited in the *Big Mechanism* project[5], whose overall goal is to automate the process of intelligent, optimised drug discovery in cancer research. Pathway models will be used as the basis to generate new hypotheses for subsequent testing. TM techniques are being employed to construct, update and verify information in relevant models, to ensure that the information used for hypothesis generation is as accurate as possible. Events are extracted from the literature using EventMine, and are compared to event structures converted from reactions in existing pathway models.

The comparisons allow the existing models to be verified or updated in several ways. For example, events from the literature that match completely with events derived from the model act as corroborative evidence of the validity of these reactions. Other events found in the literature may help to extend the model, e.g., by identifying specific sites of a more general reaction, or by identifying a reaction that is not included in the model at all. By taking into account meta-knowledge, it is possible to identify potential contradictions for existing reactions in model. For example, a reaction included in the model may occur as a negated event in the literature, and thus it may require further investigation. Similarly, an existing reaction may be questioned if only tentative evidence for the reaction can be found in the literature.

8 Conclusion

Through their sophisticated semantic analysis of document collections, advanced TM methods can be used to develop search applications that can considerably increase the ease with which researchers can locate evidence to support tasks such as biological pathway curation, compared to traditional search methods.

In particular, the advances in the accuracy of automatic event extraction are paving the way for the development of systems that can immediately pinpoint information of direct relevance to the researcher and can largely eliminate the issues of information overload and overlook. Through the possibility of specifying information needs in terms of precise, abstract semantic structures, the burden of determining the many potential ways in which the information could be expressed in text can be increasingly shifted from the expert curator to the computer. According to the possibility of automated mapping between event structures extracted from text and formal pathway models, new opportunities are arising to further automate the processes of constructing, updating and validating pathway models. This will ultimately free experts from mundane, tedious tasks while aiding with more intellectually challenging ones.

Acknowledgements. The work described in this article has been supported by the BBSRC-funded *EMPATHY* project (Grant No. BB/M006891/1) and by the DARPA-funded *Big Mechanism* project Grant No. DARPA-BAA-14-14).

[5] http://www.nactem.ac.uk/big_mechanism/.

References

1. Caron, E., et al.: A comprehensive map of the mTOR signaling network. Mol. Syst. Biol. **6**, 453 (2010)
2. Oda, K., et al.: New challenges for text mining: mapping between text and manually curated pathways. BMC Bioinform. **9**(Suppl 3), S5 (2008)
3. Herrgard, M.J., et al.: A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. Nat. Biotechnol. **26**(10), 1155–1160 (2008)
4. Thiele, I., Palsson, B.Ø.: Reconstruction annotation jamborees: a community approach to systems biology. Mol. Syst. Biol. **6**, 361 (2010)
5. Ananiadou, S., McNaught, J. (eds.): Text Mining for Biology and Biomedicine. Artech House, Boston/London (2006)
6. Ananiadou, S., Kell, D.B., Tsujii, J.: Text mining and its potential applications in systems biology. Trends Biotechnol. **24**(12), 571–579 (2006)
7. Ananiadou, S.: Text mining bridging the gap between knowledge and text. In: Selected Papers of the XVIII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2016), vol. 1752, pp. 140–141 (2016). http://ceur-ws.org/
8. Rak, R., et al.: Argo: an integrative, interactive, text mining-based workbench supporting curation. Database: J. Biol. Databases Curation **2012** (2012). bas010
9. Rak, R., et al.: Interoperability and customisation of annotation schemata in Argo. In: Proceedings of LREC, pp. 3837–3842 (2014)
10. Ferrucci, D., et al.: Towards an interoperability standard for text and multi-modal analytics. IBM Research Report RC24122 (2006)
11. Batista-Navarro, R., Rak, R., Ananiadou, S.: Optimising chemical named entity recognition with pre-processing analytics, knowledge-rich features and heuristics. J. Cheminf. **7**(Suppl. 1), S6 (2015)
12. Okazaki, N., Ananiadou, S., Tsujii, J.: Building a high-quality sense inventory for improved abbreviation disambiguation. Bioinformatics **26**(9), 1246–1253 (2010)
13. Alnazzawi, N., Thompson, P., Ananiadou, S.: Mapping phenotypic information in heterogeneous textual sources to a domain-specific terminological resource. PLoS ONE **11**(9), e0162287 (2016)
14. Nobata, C., et al.: Kleio: a knowledge-enriched information retrieval system for biology. In: Proceedings of the 31st Annual International ACM SIGIR, pp. 787–788 (2008)
15. Tsuruoka, Y., Tsujii, J., Ananiadou, S.: FACTA: a text search engine for finding associated biomedical concepts. Bioinformatics **24**(21), 2559–2560 (2008)
16. Tsuruoka, Y., et al.: Discovering and visualizing indirect associations between biomedical concepts. Bioinformatics **27**(13), i111–i119 (2011)
17. Miyao, Y., et al.: Semantic retrieval for the accurate identification of relational concepts in massive textbases. In: Proceedings of ACL, pp. 1017–1024 (2005)
18. Tsuruoka, Y., Tsujii, J.: Bidirectional inference with the easiest-first strategy for tagging sequence data. In: Proceedings of HLT/EMNLP, pp. 467–474 (2005)
19. Hara, T., Miyao, Y., Tsujii, J.: Adapting a probabilistic disambiguation model of an HPSG parser to a new domain. In: Dale, R., Wong, K.-F., Su, J., Kwong, O.Y. (eds.) IJCNLP 2005. LNCS (LNAI), vol. 3651, pp. 199–210. Springer, Heidelberg (2005). doi: 10.1007/11562214_18
20. Cohen, K.B., Palmer, M., Hunter, L.: Nominalization and alternations in biomedical language. PLoS ONE **3**(9), e3158 (2008)
21. Kim, J.-D., et al.: Extracting bio-molecular event from literature—The BioNLP'09 shared task. Computational Intelligence **27**(4), 513–540 (2011)

22. Kim, J.-D., Pyysalo, S., Nedellec, C., Ananiadou, S., Tsujii, J. (eds.): Selected Articles from the BioNLP Shared Task 2011. BMC Bioinformatics, vol. 13, Suppl. 11 (2012)
23. Nédellec, C., Kim, J.-D., Pyysalo, S., Ananiadou, S., Zweigenbaum, P. (eds.): BioNLP Shared Task 2013: Part 1. BMC Bioinformatics, vol. 16, Suppl. 10 (2015)
24. Nédellec, C., Kim, J.-D., Pyysalo, S., Ananiadou, S., Zweigenbaum, P. (eds.): BioNLP Shared Task 2013: Part 2. BMC Bioinformatics, vol. 16, Suppl. 16 (2015)
25. Thompson, P., Iqbal, S., McNaught, J., Ananiadou, S.: Construction of an annotated corpus to support biomedical information extraction. BMC Bioinform. **10**, 349 (2009)
26. Pyysalo, S., et al.: BioInfer: a corpus for information extraction in the biomedical domain. BMC Bioinform. **8**, 50 (2007)
27. Ananiadou, S., et al.: Event-based text mining for biology and functional genomics. Brief. Funct. Genomics **14**(3), 213–230 (2015)
28. Miwa, M., et al.: Event extraction with complex event classification using rich features. J Bioinform. Comput. Biol. **8**(1), 131–146 (2010)
29. Sagae, K., Tsujii, J.: Dependency parsing and domain adaptation with LR models and parser ensembles. In: Proceedings of the CoNLL 2007 Shared Task, pp. 1044–1050 (2007)
30. Miyao, Y., et al.: Evaluating contributions of natural language parsers to protein-protein interaction extraction. Bioinformatics **25**(3), 394–400 (2009)
31. Miwa, M., Ananiadou, S.: Adaptable, high recall, event extraction system with minimal configuration. BMC Bioinform. **16**(Suppl. 10), S7 (2015)
32. Miwa, M., Thompson, P., Ananiadou, S.: Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. Bioinformatics **28**(13), 1759–1765 (2012)
33. Miwa, M., et al.: Extracting semantically enriched events from biomedical literature. BMC Bioinform. **13**, 108 (2012)
34. Nawaz, R., et al.: Meta-knowledge annotation of bio-events. Proc. LREC **2010**, 2498–2507 (2010)
35. Nawaz, R., Thompson, P., Ananiadou, S.: Evaluating a meta-knowledge annotation scheme for bio-events. In: Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, pp. 69–77 (2010)
36. Thompson, P., et al.: Enriching a biomedical event corpus with meta-knowledge annotation. BMC Bioinform. **12**, 393 (2011)
37. Hucka, M., et al.: The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics **19**(4), 524–531 (2003)
38. Hucka, M., et al.: Evolving a lingua franca and associated software infrastructure for computational systems biology: the Systems Biology Markup Language (SBML) project. Syst. Biol. **1**(1), 41–53 (2004)
39. Demir, E., et al.: The BioPAX community standard for pathway data sharing. Nat. Biotechnol. **28**(9), 935–942 (2010)
40. Ohta, T., Pyysalo, S., Tsujii, J.: From pathways to biomolecular events: opportunities and challenges. In: Proceedings of BioNLP 2011 Workshop, pp. 105–113 (2011)
41. Miwa, M., et al.: A method for integrating and ranking the evidence for biochemical pathways by mining reactions from text. Bioinformatics **29**(13), i44–i52 (2013)