BIOINFORMATICS

Vol. 26 ISMB 2010, pages i374–i381
doi:10.1093/bioinformatics/btq221

PathText: a text mining integrator for biological pathway visualizations

Brian Kemper[1], Takuya Matsuzaki[1], Yukiko Matsuoka[2,3], Yoshimasa Tsuruoka[4,5], Hiroaki Kitano[2,6], Sophia Ananiadou[4,7] and Jun'ichi Tsujii[1,4,7,∗]

[1]Department of Computer Science, University of Tokyo, [2]The Systems Biology Institute, Tokyo, [3]JST ERATO Kawaoka Infection-induced Host-Response Network Project, Tokyo, Japan, [4]National Centre for Text Mining, UK, [5]School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa, [6]Okinawa Institute of Science and Technology, Okinawa, Japan and [7]School of Computer Science, University of Manchester, UK

ABSTRACT

Motivation: Metabolic and signaling pathways are an increasingly important part of organizing knowledge in systems biology. They serve to integrate collective interpretations of facts scattered throughout literature. Biologists construct a pathway by reading a large number of articles and interpreting them as a consistent network, but most of the models constructed currently lack direct links to those articles. Biologists who want to check the original articles have to spend substantial amounts of time to collect relevant articles and identify the sections relevant to the pathway. Furthermore, with the scientific literature expanding by several thousand papers per week, keeping a model relevant requires a continuous curation effort. In this article, we present a system designed to integrate a pathway visualizer, text mining systems and annotation tools into a seamless environment. This will enable biologists to freely move between parts of a pathway and relevant sections of articles, as well as identify relevant papers from large text bases. The system, PathText, is developed by Systems Biology Institute, Okinawa Institute of Science and Technology, National Centre for Text Mining (University of Manchester) and the University of Tokyo, and is being used by groups of biologists from these locations.

Contact: brian@monrovian.com.

1 INTRODUCTION

The core of systems biology is the biochemical/signaling network. Signaling and metabolic pathways are an increasingly important part of organizing knowledge in systems biology and are often represented through collective interpretations of facts scattered throughout literature (Heiner *et al*., 2004; Kell and Oliver, 2004; Luciano and Stevens, 2007; Ye and Doak, 2009).

Because of the very integrated nature of pathways, they require substantial human effort to construct. Biologists have to read a large number of published papers, interpret them and construct a pathway (Ananiadou *et al.,* 2006). The curation of a constructed pathway also requires monitoring of recent publications in order to maintain relevance. Furthermore, since biologists may have different interpretations of the same set of facts, a biologist wants to read original papers based on which a pathway is constructed to ensure it is done in a manner consistent with his or her interpretation. The biologist would like to see the biological context, stated in original papers, from which the constructed pathway abstracts away (Kell, 2006; Kell and Oliver, 2004).

PathText (http://www.pathtext.org) is an integrated environment for combining standards compliant (Finney and Hucka, 2003; Hucka *et al.,* 2003) biological pathway models and original papers relevant to selected parts of the pathway, through the use of text mining (TM) technology (Ananiadou *et al.,* 2006) and tools to facilitate the creation of manual annotations.

Unlike existing pathway building platforms, such as WikiPathways (Pico *et al.,* 2008), the Edinburgh Pathway Editor (Sorokin *et al.,* 2006), and PathCase (Elliott *et al.,* 2008), PathText brings together the strengths of different TM tools in a unified and extensible framework. Some of the existing pathway editors offer the functionality of linking parts of a pathway with literature at a very coarse (e.g. PubMed ID) level. Perhaps one of the most similar to PathText in terms of the richness of TM functionality combined with pathway visualization is Pathway Studio (Nikitin *et al.,* 2003), which is a commercial tool for building and analyzing biological pathways. It integrates an automated text processing tool called MedScan to extract biological interactions from scientific literature using natural language processing (NLP) technology.

MedScan employs a full syntactic parser to analyze the semantic and lexical structure of an English sentence and finds relations between any types of objects including proteins, small molecules, protein functional classes, cell processes and diseases (Daraselia *et al.,* 2004; Yuryev *et al.,* 2006).

PathText distinguishes itself from other pathway editing tools by providing a seamless combination of advanced TM technologies, including deep syntactic analysis of individual sentences (MEDIE), named entity recognition and disambiguation of acronyms (KLEIO), and real time co-occurrence searches (FACTA). It provides a flexible interactive environment which allows a biologist to navigate from pathway visualization to TM, to retrieve articles recently published which are potentially relevant, to browse them, and to associate them with relevant parts of pathways.

PathText has been used by biologists to construct signaling pathways of more than 1000 nodes with 400 links. In the following sections, we provide the overall architecture of PathText, describe the manual annotation and TM components, describe how they are integrated with the overall system, and discuss future directions.

∗To whom correspondence should be addressed.

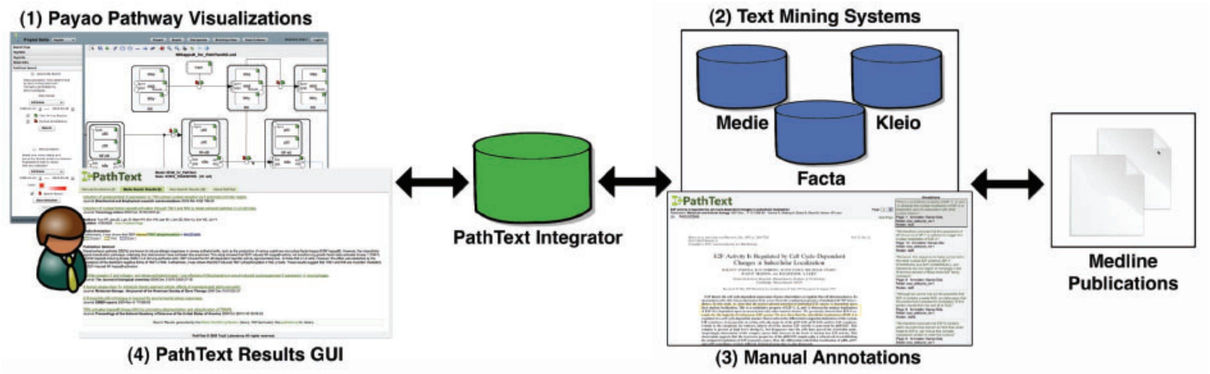


Fig. 1. PathText system architecture.

2 OVERALL ARCHITECTURE OF PATHTEXT

PathText provides a user friendly interface by which one can freely move from pathways to relevant articles, or to parts of full papers annotated by biologists during the model construction phase. A detailed illustration of the user interface is provided on the PathText website (http://www.pathtext.org). The central component of PathText is the Integrator, which coordinates interaction between (i) Payao pathway visualizations, (ii) TM systems, (iii) manual annotations and (iv) the PathText results GUI (see Fig. 1).

2.1 Payao pathway visualizations

Users of the PathText system access pathway model visualizations through the Payao Web 2.0 (Matsuoka *et al*., 2008) community-based collaborative web-service platform for modeling biological networks, a framework designed to enable a community to work on models concurrently. Payao reads the models in SBML (Systems Biology Markup Language, http://sbml.org) format, displays them with CellDesigner (http://celldesigner.org), a process diagram editor (http://celldesigner.org) which complies with the process description notation defined by SBGN (Systems Biology Graphical Notation, http://sbgn.org) language, and provides an interface for model enrichment (adding tags and comments to models) for the access-controlled community members (see Fig. 2).

2.2 TM systems

The PathText Integrator reads information from a pathway model file in SBML and SBGN, and generates queries for TM systems tailored to the specific requirements of each. These queries are sent to each of the TM systems through a web service and results from each are stored in the repository of the PathText Integrator.

2.3 Manual annotations

PathText includes tools to assist biologists in creating manual annotations linking specific sections of original publications with nodes in a pathway model. These links are also stored in the repository of the PathText Integrator.

2.4 PathText results GUI

A Payao user interacts with the PathText system by selecting document icons overlaid on the pathway visualization. These links are presented not only for individual species (proteins, genes, etc.), but the reactions linking them as well. The PathText Integrator presents matching results through a new tab in the user's web browser, making PathText feel as if it were a part of Payao.

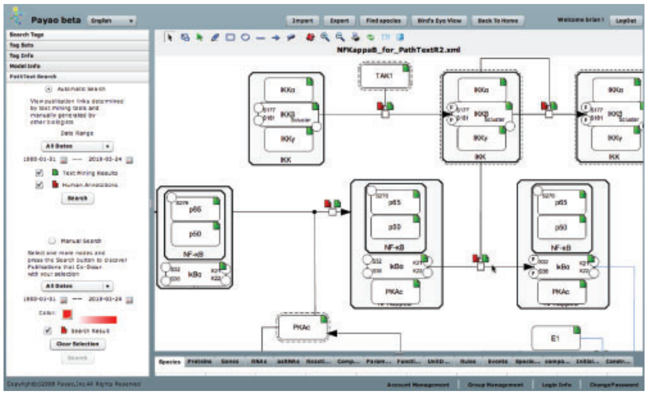
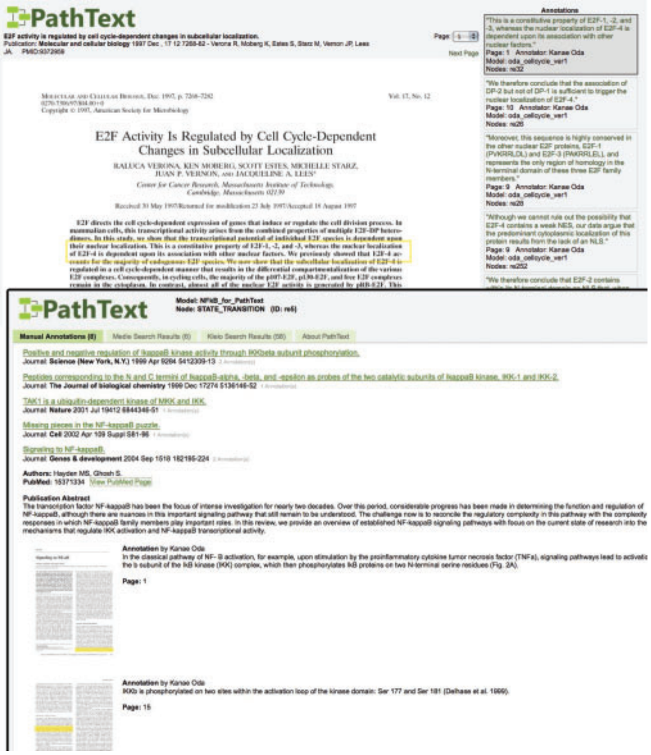


Fig. 2. Payao interface with integrated PathText control panel (far left) and PathText search results displayed as document icons over the corresponding parts of the pathway. Red and green icons correspond to sets of manually curated articles and ones retrieved by the TM systems. The shades of the color of these icons reflect the number of articles in the sets. By clicking one of these colored icons, a user can move another window to see the articles.

3 MANUAL ANNOTATION/MODEL CURATION

One of the crucial deficiencies in currently constructed biological pathways is their lack of detailed linkage with original articles. While some models embed the identifiers of articles such as PubMed IDs or their bibliographic information, some do not have any links. Even for models with uniquely identified papers, biologists who want to check the evidence provided in the original papers spend a significant amount of time retrieving the papers and then identifying the parts (e.g. paragraphs, sentences, tables, figures, etc.) relevant to the specific nodes or links in a model that they are interested in.

PathText provides a tool for manual annotation of full papers that can be used during the model curation phase. The tool is web based and assists biologists to create manual annotations, connecting nodes and links in a pathway to arbitrary rectangular regions in papers.



E2F activity is regulated by cell cycle-dependent changes in subcellular localization.
Publication: **Molecular and cellular biology** 1997 Dec , 17 12 7268-82 - Verona R, Moberg K, Estes S, Starz M, Vernon JP, Lees JA PMID:9372958

Molecular and Cellular Biology, Dec. 1997, p. 7268–7282 Vol. 17, No. 12
0270-7306/97/$04.00+0
Copyright © 1997, American Society for Microbiology

E2F Activity Is Regulated by Cell Cycle-Dependent Changes in Subcellular Localization

RALUCA VERONA, KEN MOBERG, SCOTT ESTES, MICHELLE STARZ, JUAN P. VERNON, and JACQUELINE A. LEES*

Center for Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Received 30 May 1997/Returned for modification 23 July 1997/Accepted 18 August 1997

E2F directs the cell cycle-dependent expression of genes that induce or regulate the cell division process. In mammalian cells, this transcriptional activity arises from the combined properties of multiple E2F-DP heterodimers. In this study, we show that the transcriptional potential of individual E2F species is dependent upon their nuclear localization. This is a constitutive property of E2F-1, -2, and -3, whereas the nuclear localization of E2F-4 is dependent upon its association with other nuclear factors. We previously showed that E2F-4 accounts for the majority of endogenous E2F species. We now show that the subcellular localization of E2F-4 is regulated in a cell cycle-dependent manner that results in the differential compartmentalization of the various E2F complexes. Consequently, in cycling cells, the majority of the p107-E2F, p130-E2F, and free E2F complexes remain in the cytoplasm. In contrast, almost all of the nuclear E2F activity is generated by pRB-E2F. This

Annotations

"This is a constitutive property of E2F-1, -2, and -3, whereas the nuclear localization of E2F-4 is dependent upon its association with other nuclear factors."
Page: 1 **Annotator:** Kanae Oda
Model: oda_cellcycle_ver1
Nodes: re32

"We therefore conclude that the association of DP-2 but not of DP-1 is sufficient to trigger the nuclear localization of E2F-4."
Page: 10 **Annotator:** Kanae Oda
Model: oda_cellcycle_ver1
Nodes: re28

"Moreover, this sequence is highly conserved in the other nuclear E2F proteins, E2F-1 (PVKRRLDL) and E2F-3 (PAKRRLEL), and represents the only region of homology in the N-terminal domain of these three E2F family members."
Page: 9 **Annotator:** Kanae Oda
Model: oda_cellcycle_ver1
Nodes: re28

"Although we cannot rule out the possibility that E2F-4 contains a weak NES, our data argue that the predominant cytoplasmic localization of this protein results from the lack of an NLS."
Page: 9 **Annotator:** Kanae Oda
Model: oda_cellcycle_ver1
Nodes: re252

"We therefore conclude that E2F-2 contains

Model: NFkB_for_PathText
Node: STATE_TRANSITION (ID: re6)

Positive and negative regulation of IkappaB kinase activity through IKKbeta subunit phosphorylation.
Journal: **Science (New York, N.Y.)** 1999 Apr 9284 5412309-13 2 Annotation(s)

Peptides corresponding to the N and C termini of IkappaB-alpha, -beta, and -epsilon as probes of the two catalytic subunits of IkappaB kinase, IKK-1 and IKK-2.
Journal: **The Journal of biological chemistry** 1999 Dec 17274 5136146-52 1 Annotation(s)

TAK1 is a ubiquitin-dependent kinase of MKK and IKK.
Journal: **Nature** 2001 Jul 19412 6844346-51 1 Annotation(s)

Missing pieces in the NF-kappaB puzzle.
Journal: **Cell** 2002 Apr 109 Suppl S81-96 1 Annotation(s)

Signaling to NF-kappaB.
Journal: **Genes & development** 2004 Sep 1518 182195-224 2 Annotation(s)

Authors: Hayden MS, Ghosh S.
PubMed: 15371334 View PubMed Page

Publication Abstract
The transcription factor NF-kappaB has been the focus of intense investigation for nearly two decades. Over this period, considerable progress has been made in determining the function and regulation of NF-kappaB, although there are nuances in this important signaling pathway that still remain to be understood. The challenge now is to reconcile the regulatory complexity in this pathway with the complexity responses in which NF-kappaB family members play important roles. In this review, we provide an overview of established NF-kappaB signaling pathways with focus on the current state of research into the mechanisms that regulate IKK activation and NF-kappaB transcriptional activity.

Annotation by Kanae Oda
In the classical pathway of NF- B activation, for example, upon stimulation by the proinflammatory cytokine tumor necrosis factor (TNFa), signaling pathways lead to activatio
the b subunit of the IkB kinase (IKK) complex, which then phosphorylates IkB proteins on two N-terminal serine residues (Fig. 2A).

Page: 1

Annotation by Kanae Oda
IKKb is phosphorylated on two sites within the activation loop of the kinase domain: Ser 177 and Ser 181 (Delhase et al. 1999).

Page: 15

Fig. 3. PathText manual annotation tool and search results GUI. In the top screen, the first page of an article is shown with a yellow rectangular region which is judged relevant by a biologist. In the right column, a list of annotations in this article with the name of annotator, her/his comment, etc. is shown. By clicking one of them, a user can jump to the corresponding part of the article. In the bottom screen, a list of full articles which are manually annotated. By clicking one of the articles, a user can see pages of the article in which annotations are made (the bottom half of the screen).

It is used not only at the initial model curation phase, but also for the monitoring phase in later stages of the process.

In Figure 3, we see an example of PathText manual annotation. First, a curator identifies an initial set of articles by searching through MEDLINE via the automatic search or manual search mode (see Section 6). Then the curator collects the full-paper version of the retrieved abstracts judged to be relevant and submits them to the PathText Integrator where they are converted (often from PDF) into images and stored in the PathText repository. The curator then views the publications using PathText to annotate the relevant parts in the model with selected rectangular regions on the images of the full papers. A rectangular region is stored in the Integrator along with the text inside the region, publication details and the identifier of the node or link in the model annotated by the region in the full paper. PathText enables biologists to see all the full papers, and the parts judged as relevant to a specific part of a model by other biologists, by simple operations such as clicking an icon attached to links or nodes in a network.

4 TM SYSTEMS

In addition to the manual annotation tool for biologists to create annotated links between models and full papers, the PathText Integrator includes a method to retrieve new papers relevant to specific parts of a model through the use of TM systems. The only way of linking parts of a model with such an implicit set of papers is in the form of queries, by which each of the individual TM systems retrieves a set of text. Because the results returned by each TM system have their own semantic annotations, the Integrator needs to interpret the annotations in retrieved text to identify the portions of text relevant to the model and visualize them.

We integrate three TM systems (MEDIE, KLEIO and FACTA) in PathText, each of which has different characteristics, its own strengths and weaknesses. The crucial considerations are how to maximize the effectiveness of the system by exploiting the characteristics of these TM systems, and how to reduce the burden of a user in interaction with these TM systems.

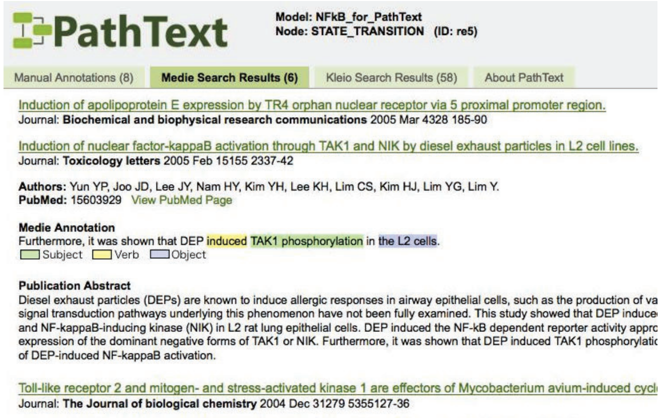
A pathway represents a specific biological context in which species (genes, proteins, enzymes, etc.) and relations among them (phosphorylation, binding, degradation, etc.) occur. When a user clicks a specific node in a pathway, she/he is most likely to be not interested in the gene, protein, enzyme, etc. in general that the node represents. Instead, they are interested in the gene, protein or enzyme in the specific context. Such rich contextual information should be used in query formation to improve the precision. MEDIE, the query language of which is highly expressive, provides the means by which such rich contextual information is embedded in a query. The integrator in PathText stores the associations between nodes/relations and complex query formulas. The complex queries and their associations are established in advance during the model curation phase. On the other hand, while such fixed queries embedded in a pathway model in advance are effective, a biologist wants to navigate a set of documents rather freely. The general query format which MEDIE provides is, though expressive, too complex. Interactive query formation provided by KLEIO, including rich semantic annotations and facet-search based on them, is ideal for such free navigation of document sets.

The other functionality, which we have found extremely useful, is to allow a user to formulate a query on the fly by using visualization of pathway. That is, she/he clicks an arbitrary number of nodes in a pathway and then, the system returns a set of documents in which the species those nodes represent co-occur, together with distribution statistics of the other species in the pathway. FACTA provides this functionality with its specialized indexing.

4.1 MEDIE

MEDIE (http://www-tsujii.is.s.u-tokyo.ac.jp/medie/search.cgi) (Miyao *et al.,* 2006) is an intelligent search engine to retrieve biomedical relational information from a large textbase created from MEDLINE. Figure 4, shows MEDIE text mining search results in PathText. The textbase stores the whole MEDLINE abstracts, with annotations represented by XML-like text markup for both the metadata of the articles provided by NLM (MeSH terms, publication date, etc.), and analysis results by various NLP modules. The annotated text is indexed for efficient structure search by extended region algebra (ERA) (Masuda and Tsujii, 2008). A query with high precision can be formulated by using NLP analyses results and the region algebra.

The NLP modules used for the annotation include (but not limited to) a deep syntactic analyzer, an event expression recognizer (EER) and a term recognizer. The syntactic analyzer, Enju



PathText

Model: NFkB_for_PathText
Node: STATE_TRANSITION (ID: re5)

Induction of apolipoprotein E expression by TR4 orphan nuclear receptor via 5 proximal promoter region.
Journal: **Biochemical and biophysical research communications** 2005 Mar 4328 185-90

Induction of nuclear factor-kappaB activation through TAK1 and NIK by diesel exhaust particles in L2 cell lines.
Journal: **Toxicology letters** 2005 Feb 15155 2337-42

Authors: Yun YP, Joo JD, Lee JY, Nam HY, Kim YH, Lee KH, Lim CS, Kim HJ, Lim YG, Lim Y.
PubMed: 15603929 View PubMed Page

Medie Annotation
Furthermore, it was shown that DEP induced TAK1 phosphorylation in the L2 cells.
🟩 Subject 🟨 Verb 🟦 Object

Publication Abstract
Diesel exhaust particles (DEPs) are known to induce allergic responses in airway epithelial cells, such as the production of va
signal transduction pathways underlying this phenomenon have not been fully examined. This study showed that DEP induce
and NF-kappaB-inducing kinase (NIK) in L2 rat lung epithelial cells. DEP induced the NF-kB dependent reporter activity appro
expression of the dominant negative forms of TAK1 or NIK. Furthermore, it was shown that DEP induced TAK1 phosphorylatic
of DEP-induced NF-kappaB activation.

Toll-like receptor 2 and mitogen- and stress-activated kinase 1 are effectors of Mycobacterium avium-induced cycl
Journal: **The Journal of biological chemistry** 2004 Dec 31279 5355127-36

Fig. 4. MEDIE text mining search results shown in the PathText GUI. A sentence which contains a biological event is shown together with the whole abstract in which the sentence appears. The biological event correspond to a link in a pathway which the user clicks. The event in a sentence is shown with the verb, the subject and object, each of which is highlighted with different colors.

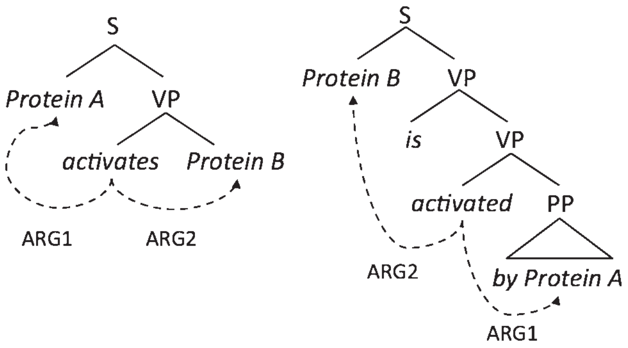
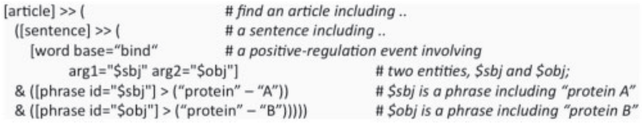


Fig. 5. A semantic relation expressed in different textual expressions: these two tree structures represent two sentences with different voices (active and passive), which essentially describe the same event. The identity of the event is captured by the predicate-argument structures (PAS, shown by dotted lines) of these sentences. Enju computes such predicate-argument structures of sentences, and one can formulate a query based on PAS in the region algebra to retrieve the sentences which contain events of ‘protein A activates protein B’.

parser (Miyao *et al.,* 2009), produces a syntactic and semantic analysis of the text, based on a linguistic formalism called HPSG (Pollard and Sag, 1994). A relational concept, such as ‘protein A activates protein B’, can be precisely described as a query which specifies the semantic structure given by the Enju parser as constraint (see Figs 5 and 6). This is the main strength of MEDIE compared to other publicly available TM modules which use Boolean formula of keywords or concepts for query formulation. Boolean formulas basically specify co-occurrence of concepts or words as constraint for retrieval. One can only specify co-occurrence of protein A, protein B and the verb ‘to activate’ in the same textual unit (usually an abstract) as constraint, which results in a large number of false positives.

Units of retrieval in MEDIE are finer than those in other TM modules. They can be individual sentences in abstracts, or even phrases. Furthermore, the ERA allows us to specify constraints on context from constraints on units of retrieval. That is, we can formulate a query for retrieving sentences which contain a specific biological event (e.g. *Protein A* activates *protein B*) and which appear in abstracts with certain keywords or other biological events reported. This separation of units of retrieval and context is extremely useful for PathText to specify constraints embodied by the neighboring parts of a pathway network.



```
[article] >> (                    # find an article including ..
  ([sentence] >> (                # a sentence including ..
    [word base=“bind“             # a positive-regulation event involving
        arg1="$sbj" arg2="$obj"]                 # two entities, $sbj and $obj;
  & ([phrase id="$sbj"] > (“protein” – “A”))     # $sbj is a phrase including “protein A”
  & ([phrase id="$obj"] > (“protein” – “B”)))))  # $obj is a phrase including “protein B”
```

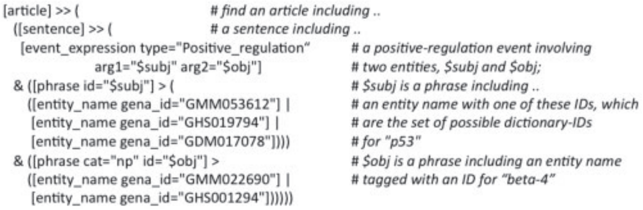
Fig. 6. An example of an ERA query: >> is a operator. [R1] >> [R2] means a text span tagged by R1 should contain another text span tagged by R2. $sbj and $obj are variables. They are used to express the dotted lines in Figure 5. For example, $sbj is used to equate the phrase which plays the role of arg1 in a sentence with the phrase which contains the word ‘protein-A’.

The semantic representation produced by Enju also works as an intermediate language which bridges the gap between a search query and the textual expression in the article. That is, a single semantic relation is represented in the same way in the semantic representation level, across various different textual expressions, and hence retrieved with a single query. See Figure 5 for an example of a semantic relation ‘protein A activates protein B’, expressed differently in surface textual form, in which the semantic subject (ARG1) and semantic object (ARG2) of the verb ‘activate’ is represented in the same way in the semantic representation level (indicated by the dashed arrows).

The EER and the term recognizer further enhance the search capability of MEDIE by introducing another level of abstraction of semantics. They map surface textual expressions of biological events or technical terms to the corresponding concept identifiers defined in ontologies. The EER recognizes biological molecular events mentioned in text and map them to identifiers of event types defined in terms of Ontology (Ashburner *et al.,* 2000). The current version of EER distinguishes 35 event types in GO, which include binding, positive/negative regulation, etc. Using the annotations by Enju and the EER, we can retrieve sentences in which a biological event of ‘positive regulation of protein A by protein B’ is reported, even though they may be expressed in diverse surface expressions like ‘A activates B’, ‘B is induced by A’, and so on. The domain specific lexical knowledge, like the synonymy of ‘activate’ and ‘induce’ in the molecular biology domain, was collected from the GENIA Event corpus (Kim *et al.,* 2008).

The term recognizer detects gene, protein and disease names in the text, and assign unique database IDs to differently expressed entity names (i.e. synonyms). The gene/protein IDs are taken from a gene/protein meta-DB, Gena (Koike and Takagi, 2004) and the disease IDs are taken from UMLS. By combining the annotations given by the term recognizer with ones by the parser and the EER, we can recognize a biological event in even when an entity (protein or gene) involved there is mentioned in different names.

The index for MEDIE is based on the ERA. In the ERA, we can specify a semantic relation encoded as the topological relations (e.g. a text span includes another text span, a text span follows another text span, etc.) among textual spans and annotations. Structural relations



```
[article] >> (                        # find an article including ..
  ([sentence] >> (                    # a sentence including ..
   [event_expression type="Positive_regulation"        # a positive-regulation event involving
               arg1="$subj" arg2="$obj"]                # two entities, $subj and $obj;
  & ([phrase id="$subj"] > (                            # $subj is a phrase including ..
    ([entity_name gena_id="GMM053612"] |                # an entity name with one of these IDs, which
     [entity_name gena_id="GHS019794"] |                # are the set of possible dictionary-IDs
     [entity_name gena_id="GDM017078"])))               # for "p53"
  & ([phrase cat="np" id="$obj"] >                      # $obj is a phrase including an entity name
    ([entity_name gena_id="GMM022690"] |                # tagged with an ID for "beta-4"
     [entity_name gena_id="GHS001294"])))))
```

Fig. 7. An ERA query for <subject, verb, object>=<p53, activate, beta-4>.

can also be directly represented by linking variables. For example, to retrieve the sentences that mention a binding event between 'protein A' and 'protein B', we formulate a query that has three key phrases: 'protein A', 'protein B' and 'bind' among which a semantic relation 'protein A binds to protein B' holds. The query in the ERA is shown in Figure 6.

MEDIE accepts a search query through a WEB-API, in addition to an interactive search UI. The API takes a tuple of <subject, verb, object> as the input, which describes a biological event/relation, such as <p53, activate, beta-4>, and returns a set of articles in which the event/relation is mentioned. The tuple is internally translated to an ERA query, using the same gene/protein dictionary and event expression dictionary used in the above-mentioned NLP modules. For example, the tuple <p53, activate, beta-4> is translated to the following region-algebra query shown in Figure 7. The WEB-API thus hides irrelevant details of the backend database from the viewpoint of the users, such as the annotation schemes used in the NLP modules or the dictionary used in developing it. A specification on the meta-data part such as journal titles can be expressed as additional fields to the subject-verb-object tuple.

4.2 KLEIO

KLEIO (http://www.nactem.ac.uk/software/kleio/) (Nobata *et al.,* 2008) uses the results of named entity recognition to provide a range of semantic search functions. A standard indexing tool, Lucene, is used to generate an index over the terms for proteins, genes, metabolites and medical terms that have been recognized. This is an index of the concepts that are referred to in the text, rather than individual, or canonical word forms. This functionality allows us to retrieve documents that refer to a specific concept, although the surface form used may differ in each case, as in the use of orthographic variants, or acronyms instead of their expansions. In KLEIO, full forms of named entities, including variants, are linked to their acronyms via an acronym recognition and disambiguation process (Okazaki and Ananiadou, 2006). The system also offers document retrieval based on the unique identifier for a concept, providing a link back to the original databases from which the system's dictionary was generated. In addition, by further classifying terms into semantic categories the system allows the user to specify a specific concept, by associating a semantic category with a query term. This can radically reduce the search space. For example, more than 60 000 documents were returned when the word 'cat' was given as a query, due to its ambiguity. However, when the query was modified to specify the desired semantic category for 'cat' e.g. PROTEIN, a more focused result is returned. For the query 'PROTEIN:cat', 200 documents were returned. Moreover, the documents returned by the initial query are dynamically organized into semantic facets based on the named entities recognized both in the query and occurring in the same immediate context in each document retrieved. The user may thus refine the initial query by combining concepts from the offered facets or may pursue the links to the document representations. The documents themselves are presented with concept markup on all the recognized terms.

As with MEDIE, KLEIO stores the whole set of abstracts from MEDLINE together with metadata provided by the National Library of Medicine and augments these data with rich semantic annotations. Semantic annotation in KLEIO is much richer in semantic categories of named entities than those of MEDIE, though it does not have syntactic/semantic annotations of sentences. The normalized identifiers which KLEIO uses are, therefore, not only UniProt identifiers and UMLS identifiers, but also HMDB and DrugBank identifiers for small molecules and metabolites which are crucial for integrating metabolic pathways. Acronyms, which are pervasive in biological papers, are also disambiguated (Okazaki *et al.,* 2010) and normalized into identifiers if the disambiguated results belong to the semantic categories which KLEIO is able to deal with. Because of surface word indexing and richer semantic categories, KLEIO is used as a fall-back system when species in a model are not covered by MEDIE. KLEIO accepts PathText queries through a WEB-API. The API accepts space separated terms as well as Boolean queries, for example 'p65 AND beta4'. KLEIO then returns a set of articles relevant to the query in XML containing PubMed IDs and the abstract highlighted with the terms matching the query.

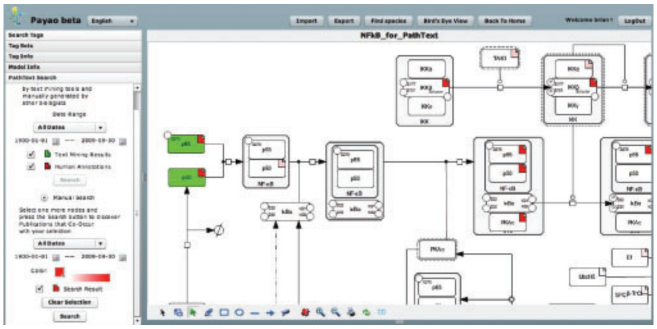
4.3 FACTA

FACTA (http://refine1-nactem.mc.man.ac.uk/facta/) (Tsuruoka *et al.,* 2008) is an information retrieval system with a usage very different from MEDIE and KLEIO. It takes large sets of articles (the whole MEDLINE in the current version of FACTA) to find implicit associations between named entities, by using statistical measures of co-occurrences of entities in the same articles. It can find and show a biologist a list of genes, for example, which would be relevant to a given disease.

FACTA was originally designed as an interactive system to show the user such a list of entities on the fly, and special care was taken to compute the statistical measures very quickly. More specifically, it builds a special data structure called inverted index that allows for efficient access to articles in which a particular set of entities appear. This data structure enables the system to compute co-occurrence statistics on the fly even if the input entities appear in a large number of articles. For example, FACTA can produce a ranked list of genes and proteins that co-occur with 'p53', which is mentioned in more than 46 000 articles, in 0.04 s. Combined with the PathText interface, FACTA allows the user to select an arbitrarily subset of the species in the pathway and immediately find the information about which other species co-occur with them in the literature (Fig. 8) (see Section 6.2).

It should be noted that such co-occurrence statistics cannot be computed off-line, because the user is allowed to specify any combination of the species as the input. This is the reason why PathText needed to integrate the functionality of FACTA, which is tuned for real-time uses with a special index structure. In contrast, retrieval of articles by MEDIE and KLEIO is performed in batch mode (i.e. off-line) and the results are attached to the relevant part of a model in the repository of the Integrator (see Section 6.1).

Unlike MEDIE or KLEIO, FACTA currently uses a simple longest matching algorithm to recognize gene/protein names in



by text mining tools and manually generated by other biologists

Select one more nodes and press the search button to discover Publications that Co-Occur with your selection

Fig. 8. Payao GUI showing PathText manual search results discovered using FACTA. The two green colored nodes are the nodes clicked by a user. FACTA retrieves a set of documents in which these two species co-occur. The red icons show that these species also occur in the retrieved documents. The shades of the red icon reflect the numbers of documents in which the three species co-occur.

the literature. The dictionary was created from BioThesaurus (Liu *et al.,* 2006) with some manual curation efforts including the removal of noisy and highly ambiguous entries.

FACTA runs on a generic Linux server with 2.2 GHz AMD Opteron processors and 16 GB memory, on which all the inverted indexes are stored.

5 QUERYING OF TM SYSTEMS

The primary function of the PathText Integrator is to gather information from pathway model files, generate queries that can be interpreted by various TM tools and store the results from these queries in a repository.

In the first step of this process, the PathText Integrator reads an SBML model file and generates tables for the various components in the repository. This data includes information such as proteins and gene labels, characteristics for these entities such as phosphorylation and ubiquitination, and details for the reactions that link them together. The Integrator also reads MIRIAM annotations (Le Novère *et al.,* 2005) containing unique identifiers such as UniProt. This data can then be used to generate queries conforming to the disparate formats of each TM application. For example, MEDIE accepts queries in a <subject, verb, object> format. In the case of a reaction contained in the pathway, the Integrator will use a set of rules to determine a valid verb for describing the reaction by interpreting the components of that reaction. For example, if a reaction starts with a non-phosphorylated protein on one side, and contains the same protein with phosphorylation on the other, then the verb 'phosphorylates' can be generated. For generating a query for a species, the system uses unique identifiers, such as UniProt if the system supports them, and text labels for that entity in other cases. PathText also includes a 'manual override' function, where a query for a particular protein or reaction can be manually added to the SBML model file in a PathText Annotation Tag that takes precedence over queries generated by the Integrator.

6 PATHTEXT MODES OF OPERATION

PathText provides rich modes of operation for assisting biologists to create and maintain links between biological models and large collections of articles, through the whole life cycle of a model, e.g. curation of articles and model building, model updating, exploration of articles by TM tools through a model, etc. In addition to the manual annotation tools described above, two distinct methods for interacting with the PathText Integrator and the manual annotations and TM results it contains: automatic and manual Search.

6.1 Automatic search with updating function

Once a model is constructed and annotated with relevant papers, curators or biologists want to retrieve other articles which may be relevant to each part of the model. The queries generated by the Integrator for individual nodes and links are used to retrieve abstracts from MEDLINE. Since these queries are generated in advance, PathText performs text retrieval regardless of actual user requests and results are attached to corresponding parts of a model. This off-line renewal of retrieval results is to be performed on a nightly schedule as new set of abstracts are added to the MEDLINE databases of MEDIE and KLEIO on daily basis. The automatic search mode uses this set of retrieved abstracts. The mode is called 'automatic' since biologists cannot create new queries on the fly. They can only specify a date range. Date rage specification is important for up-dating a model by newly published articles. The Integrator passes the information on how many abstracts are found for each part of a model to Payao, and Payao then displays document icons overlaid on the pathway visualization over the corresponding entities (proteins, genes, reactions, etc.).

Clicking on any of these document icons will open a new tab in the user's browser with the matching PathText results. This results page contains tabbed sections, one for already curated articles by manual annotation and one for each TM system (MEDIE and KLEIO) that has relevant matches to display. While one can easily view parts of full papers relevant in the curated articles, the Integrator interprets the annotations in abstracts given by the TM systems and displays them in the same format including article title, authors, publication date, the name of journals and the abstract with the matching terms highlighted.

6.2 Manual search

The 'manual search' mode allows a biologist to freely explore new articles by using the TM services, currently FACTA, with their own queries. The simplest form of a new query is constructed by clicking several nodes (species) in a model. When receiving a set of identifiers for the selected nodes from Payao, the PathText Integrator invokes FACTA to retrieve a set of abstracts in which the selected species co-occur, and then checks whether these abstracts also contain other species in the model. The numbers of abstracts which contain other species are sent back to Payao with the identifiers of the species. Payao displays document icons on the corresponding entities in the network color-coded according to the numbers of abstracts (see Fig. 5). Clicking on any of these document icons will open up a PathText results tab.

7 CONCLUSION AND FUTURE DIRECTIONS

PathText integrates three knowledge sources indispensable for systems biology, i.e. (i) external databases such as SwissProt, EntreGene, Flybase, HUGO, etc., (ii) text databases such as

MEDLINE and full papers, and (iii) pathways as organized interpretations of biological facts. Since integration of external databases with other knowledge sources has already been attempted and achieved through dictionaries of identifiers of the databases, this article focused on the integration of pathways with literature.

PathText successfully provides integration of text to pathways and is now being used by three biological groups at the Systems Biology Institute, the University of Tokyo (Oda *et al.,* 2008) and the Manchester Centre for Integrative Systems Biology in the UK (Herrgård *et al.,* 2008). The implementation of PathText relies on the software being developed by the consortium members. Payao, which maintains a database of pathway models and provides the software for the pathway visualization interface, is being developed by SBI and OIST. Information is available on the Payao website (http://www.payaologue.org). The two TM systems, KLEIO and FACTA, which annotate MEDLINE abstracts by rich semantic annotations of named entities, are provided by the National Centre for Text Mining (http://www.nactem.ac.uk). The TM system for automatic search, MEDIE, has been developed by the University of Tokyo.

While these subcomponents had their standard WEB-APIs, we added extra functionalities to them for PathText. These extended APIs will be made available with necessary standardization for other groups that are interested in similar knowledge integration.

There are several issues for future development.

(1) The design decisions of the current PathText are highly dependent on a specific type of pathway, i.e. signaling pathway which is being developed at the University of Tokyo. In particular, automatic query generation reflects specific characteristics of this pathway. The biology research group in Manchester, which is engaged in research of metabolic pathways, has significantly different requirements. We need to impose proper modularization of the current PathText Integrator, in order to meet the differing requirements of users.

(2) The current use of biological context in generating queries is rather straightforward. As our pilot study shows (Oda *et al.*, 2008), the biological contexts in a pathway affects relevance judgments by biologists. We need detailed error analyses on the results produced by text retrieval components to generate appropriate queries from a given pathway model.

(3) The three TM subsystems work separately in PathText. The query generation is designed separately for each of them, and mixing of their results is not done at present. The results are only presented in the same display format. In the future, these results will be ranked according to different types of user needs such as displaying contradictory facts extracted from text.

Internal informal evaluation by our collaborating biology teams has demonstrated the usefulness of PathText for their work. The teams are continuing to use the system. Their ongoing feedback is being incorporated into a novel user based evaluation framework which will be used for a formal evaluation in conjunction with a wider community.

(4) While the PathText data repository is managed locally, this should be maintained by a global database. Moreover, future plans include expansion to tackle the analysis of full papers not just abstracts. Techniques to achieve such analysis are already being developed by the National Centre for Text Mining within the UKPubMed Central project. However, we note that in the general case, the use of full papers is related with the issue of copyright and access to commercial publishers' collections. PathText is currently only used by biologists who belong to institutions with access rights to full papers. In order for models in PathText to be freely accessed by a wider community, we need to include a proper authentication method for accessing full papers.

ACKNOWLEDGEMENTS

We wish to thank Kanae Oda from Tokyo Medical and Dental University for the development of pathway models and testing of PathText, Chikashi Nobata at the National Centre for Text Mining for his work on the KLEIO system, Rune Saetre from the University of Tokyo for development of the PathText system and Nobuhiro Kikuchi from Mitsui Knowledge Industry for Payao development.

Funding: Biotechnology and Biological Sciences Research Council (BB/E004431/1).

Conflict of Interest: none declared.

REFERENCES

Ananiadou,S. *et al*. (2006) Text mining and its potential applications in systems biology. *Trends Biotechnol.,* **24**, 571–579.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.*, **25**, 25–9.

Bader,G. *et al*. (2006) Pathguide: a pathway resource list. *Nucleic Acids Res.*, **34**, D504–D506.

Berners-Lee,T. *et al*. (2001) The semantic web. *Sci. Amer.,* **2001**, 35–43.

Daraselia,N. *et al*. (2004) Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics*, **20**, 604–611.

Elliott,B. *et al*. (2008) PathCase: pathways database system. *Bioinformatics*, **24**, 2526–2533.

Finney,A. and Hucka,M. (2003) Systems biology markup language: level 2 and beyond. *Biochem. Soc. Trans.,* **31**, 1472–1473.

Funahashi,A. *et al*. (2003) CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *Biosilico*, **1**, 159–162.

Heiner,M. *et al*. (2004) Model validation of biological pathways using Petri nets–demonstrated for apoptosis. *Bio Systems*, **75**, 15–28.

Herrgård,M. *et al*. (2008) A consensus yeast metabolic network obtained from a community approach to systems biology. *Nature Biotechnol*., **26**, 1155–1160.

Hucka,M. *et al*. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics,* **19**, 524–531.

Kell,D.B. (2006) Systems biology, metabolic modelling and metabolomics in drug discovery and development. *Drug Discovery Today*, **11**, 1085–1092.

Kell,D.B. and Oliver,S.G. (2004) Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays,* **26**, 99–105.

Kim,J.D. *et al*. (2008) Corpus annotation for mining biomedical events from literature. *BMC Bioinfomatics,* **9**,10.

Kitano,H. *et al*. (2005) Using process diagrams for the graphical representation of biological networks. *Nat Biotechnol.,* **23**, 961–966.

Koike,A. and Takagi,T. (2004) Gene/protein/family name recognition in biomedical literature. *Biolink-2004*, 9–16.

Le Novère,N. *et al*. (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). *Nature Biotechnol.*, **23**, 1509–1515.

Le Novère,N. *et al*. (2009) The systems biology graphical notation. *Nat. Biotechnol.,* **27**, 735–741.

Liu,H. *et al.* (2006) BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics*, **22**, 103–105.

Luciano,J.S. and Stevens,R.D. (2007) e-Science and biological pathway semantics. *BMC Bioinformatics*, **8**(Suppl. 3), S3.

Masuda,K. and Tsujii,J. (2008) Nested region algebra extended with variables for tag-annotated text search. *CIKM-2008,* 1349–1350.

Matsuoka,Y. *et al*. (2008) Payao: web community tagging system to SBML models. In *Proceedings of The Ninth International Conference on Systems Biology* (ICSB 2008).

Miyao,Y. and Tsujii,J (2008) Feature forest models for probabilistic HPSG parsing. *Comp. Linguistics,* **34**, 35–80.

Miyao,Y. *et al*. (2006) Semantic retrieval for the accurate identification of relational concepts in massive textbases. *COLING-ACL-2006*.

Miyao,Y. *et al*. (2009) Evaluating contributions of natural language parsers to protein-protein interaction extraction. *Bioinformatics,* **25**, 394–400.

Nikitin,A. *et al*. (2003) Pathway studio—the analysis and navigation of molecular networks. *Bioinformatics*, **19**, 2155–2157.

Nobata,C. *et al*. (2008) Kleio: a knowledge-enriched information retrieval system for biology. In Myaeng,S.-H. *et al*. (eds) *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Singapore*. ACM, Singapore, pp. 787–788.

Oda,K. *et al*. (2008) New challenges for text mining: mapping between text and manually curated pathways. *BMC Bioinformatics*, **9**, S5.

Okazaki,N. and Ananiadou,S. (2006) Building an abbreviation dictionary using a term recognition approach. *Bioinformatics*, **22**, 3089–3095.

Okazaki,N. *et al*. (2010) Building a high quality sense inventory for improved abbreviation disambiguation. *Bioinformatics* [E-pub ahead of print: doi: 10.1093/bioinformatics/btq129, March 30, 2010].

Pico,A.R. *et al*. (2008) WikiPathways: pathway editing for the people. *PLoS Biol.*, **6**, e184.

Pollard,C. and Sag,I.A. (1994) Head-driven phrase structure grammar. University of Chicago Press, Chicago.

Sorokin,A. *et al*. (2006) The pathway editor: a tool for managing complex biological networks. *IBM J. Res. Develop.*, **50**, 561–573.

Splendiani,A. (2008) RDFScape: semantic web meets systems biology. *BMC Bioinformatics*, **9**(Suppl. 4), S6.

Tsuruoka,Y. *et al*. (2008) FACTA: a text search engine for finding associated biomedical concepts. *Bioinformatics*, **24**, 2259–2260.

Ye,Y. and Doak,T.G. (2009) A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Comput. Biol*., **5**, e1000465.

Yuryev,A. *et al*. (2006) Automatic pathway building in biological association networks. *BMC Bioinformatics*, **7**, 171.