From text to pathway: corpus annotation for knowledge acquisition from biomedical literature

Jin-Dong Kim[†], Tomoko Ohta[†], Kanae Oda[†] and Jun'ichi Tsujii[†,‡,§]

[†] *University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan*
[‡] *University of Manchester, Oxford Road, Manchester, M13 9PL, UK*
[§] *National Centre for Text Mining, 131 Princess Street, Manchester, M1 7DN, UK*
E-mail: {jdkim, okap, k.oda, tsujii}@is.s.u-tokyo.ac.jp

We present a new direction of research, which deploys Text Mining technologies to construct and maintain data bases organized in the form of pathway, by associating parts of papers with relevant portions of a pathway and vice versa. In order to materialize this scenario, we present two annotated corpora. The first, Event Annotation, identifies the spans of text in which biological events are reported, while the other, Pathway Annotation, associates portions of papers with specific parts in a pathway.

Keywords: Bioinformatics; Text Mining; Pathway; Corpus; Annotation.

1. Introduction

The importance of pathway as a means of integrating biological knowledge into a coherent system has been increasingly recognized by the community of biologists.[1] Due to the research on formal frameworks and ontologies for pathway representation such as SBML,[2] BioPAX,[3] PSI MI,[4] SBO,[5] etc., pathways have become not only graphical means of representing biological systems, but also structured data bases for storing biological knowledge, to be continuously maintained in order to keep abreast with new relevant discoveries.

However, the rapidly growing amount of literature in the field makes it extremely difficult to identify the relevant new discoveries, which should lead to revisions of the relevant portions of the pathways. Furthermore, starting from a graphical depiction of a rather small biological system, some of the current pathways, which are used as organized knowledge bases, have become a huge collection of nodes and links.[6,7] Thus, it has become increasingly difficult, if not impossible, to associate discoveries in the literature with the relevant portions of such large pathways.

On the other hand, the recent progress of text mining (TM) technologies has made it possible to perform many tasks[8–10] including: (1) identifying biological entities that appear in papers,[11] (2) extracting interactions among proteins and other biological entities,[12,13] (3) retrieving text in which specific biological entities are involved in specific types of events,[14,15] and (4) classifying literature into distinct classes, like *relevant* or *non-relevant* to a given topic.[16]

We present a new direction of research that deploys these TM technologies to construct and maintain data bases organized in the form of pathways, by associating parts of papers with relevant portions of pathways, and vice versa. In order to materialize this scenario, we have been constructing a corpus, GENIA Pathway corpus, which associates portions of papers with specific parts in a pathway. Since we have also completed another GENIA annotation, Event Annotation, the main objective of this paper is to analyze the two corpora to discuss how we can integrate events in papers with an organized whole of a pathway.

Section 2 introduces the overall construction of the GENIA corpus, while Section 3 focuses on Event Annotation, which we have recently completed. Section 4 explains the two pathway corpora that we are constructing. One is confined to the GENIA corpus. The other one, centered on a specific pathway, collected a set of all relevant sentences from full-text papers deemed to be relevant to the pathway. Section 5 reports the results of feasibility studies that links these two different streams of work, and discusses how an event recognition program can be used for pathway construction.

2. GENIA corpus

The event and pathway annotation presented here builds on our earlier work in compiling the GENIA corpus[17] and annotating it with linguistic features[18] and biological terms.[17] The documents in the corpus come from the PubMed database, which covers a broad range of domains in bio-medicine. Since we are interested in providing semantically rich annotation for text mining in molecular biology, we have focused on a much smaller, semantically homogeneous subject domain: biological reactions concerning transcription factors in human blood cells. We used the search query, *"Humans"[MeSH] AND "Blood Cells"[MeSH] AND "Transcription Factors"[MeSH]* to retrieve a set of articles, and then chose 2,000 of these articles for our annotation.

3. Event annotation

While biological entities are related with each other in various ways, we have focused on dynamic relations and have defined the GENIA event ontology: a simplified and modified version of the Gene Ontology (GO). By "dynamic", we mean that at least one of the biological entities in the relationship is affected, with respect to its properties or its location, in the reported context.

Figure 1 shows the hierarchy of the GENIA event classes. Those in dotted boxes represent the classes that we have newly created or modified to better support the text annotation. Other classes are taken from GO as they are. The number of annotation instances made to the GENIA corpus is shown in parenthesis next to the class names.

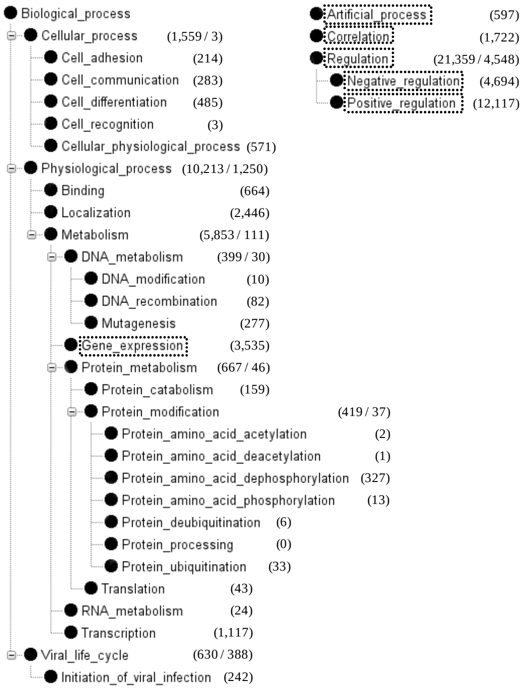

Biological_process
Cellular_process (1,559 / 3)
Cell_adhesion (214)
Cell_communication (283)
Cell_differentiation (485)
Cell_recognition (3)
Cellular_physiological_process (571)
Physiological_process (10,213 / 1,250)
Binding (664)
Localization (2,446)
Metabolism (5,853 / 111)
DNA_metabolism (399 / 30)
DNA_modification (10)
DNA_recombination (82)
Mutagenesis (277)
Gene_expression (3,535)
Protein_metabolism (667 / 46)
Protein_catabolism (159)
Protein_modification (419 / 37)
Protein_amino_acid_acetylation (2)
Protein_amino_acid_deacetylation (1)
Protein_amino_acid_dephosphorylation (327)
Protein_amino_acid_phosphorylation (13)
Protein_deubiquitination (6)
Protein_processing (0)
Protein_ubiquitination (33)
Translation (43)
RNA_metabolism (24)
Transcription (1,117)
Viral_life_cycle (630 / 388)
Initiation_of_viral_infection (242)
Artificial_process (597)
Correlation (1,722)
Regulation (21,359 / 4,548)
Negative_regulation (4,694)
Positive_regulation (12,117)

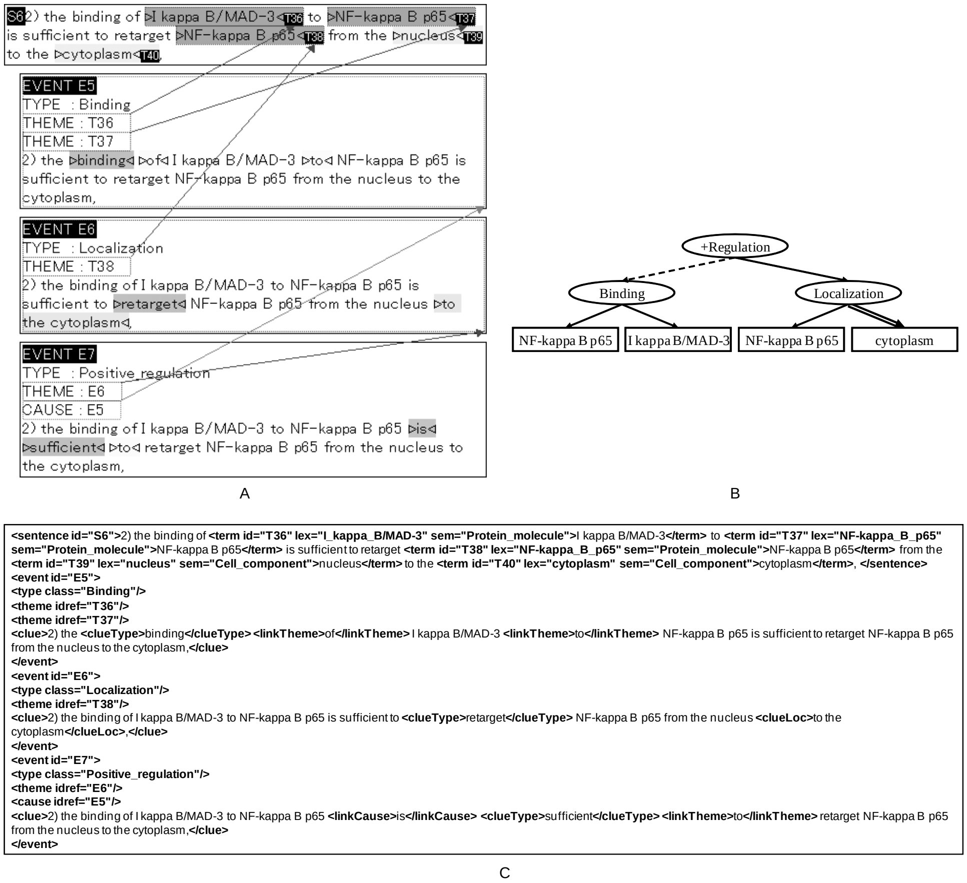
Fig. 1. GENIA event ontology

3.1. *Annotation scheme*

Figure 2A shows a screen snapshot of our annotation tool. There are four regions within the figure, each outlined by a box. The top box contains a sentence, which is undergoing annotation. Biological entities, which have been annotated during term annotation, are shown in color on the screen. Each term is assigned a term Id (T36∼T40 in the example of Figure 2A). The remaining three boxes display event annotations, which are attached to the sentence.

In the GENIA framework, an individual event is identified by its type and the *theme*: an entity or entities whose properties are affected by the event. The type of an event is selected from among the classes of the GENIA event ontology, and the theme is selected from among the entities annotated to the given sentence. Each event is also assigned a unique Id, e.g. E5∼E7 in Figure 2A. In the figure, The first (E5) and the second (E6) event annotation represent the binding of the two entities, T36 (“I kappa B/MAD-3”) and T37 (“NF-kappa B p65”), and the localization of the protein T38 (“NF-kappa B p65”), respectively.

One of our annotation principles requires annotators to mark-up text spans that belong to the corresponding annotation. We call the text expressions or the words in such text spans *clue expressions* or *clue words*. In order to allow the mark-up for clue expressions, the original sentence without term annotation is copied inside



S6 2) the binding of ▷I kappa B/MAD-3◁ T36 to ▷NF-kappa B p65◁ T37 is sufficient to retarget ▷NF-kappa B p65◁ T38 from the ▷nucleus◁ T39 to the ▷cytoplasm◁ T40,

EVENT E5
TYPE : Binding
THEME : T36
THEME : T37
2) the ▷binding◁ ▷of◁ I kappa B/MAD-3 ▷to◁ NF-kappa B p65 is sufficient to retarget NF-kappa B p65 from the nucleus to the cytoplasm,

EVENT E6
TYPE : Localization
THEME : T38
2) the binding of I kappa B/MAD-3 to NF-kappa B p65 is sufficient to ▷retarget◁ NF-kappa B p65 from the nucleus ▷to the cytoplasm◁,

EVENT E7
TYPE : Positive_regulation
THEME : E6
CAUSE : E5
2) the binding of I kappa B/MAD-3 to NF-kappa B p65 ▷is◁ ▷sufficient◁ ▷to◁ retarget NF-kappa B p65 from the nucleus to the cytoplasm,

```
<sentence id="S6">2) the binding of <term id="T36" lex="I_kappa_B/MAD-3" sem="Protein_molecule">I kappa B/MAD-3</term> to <term id="T37" lex="NF-kappa_B_p65"
sem="Protein_molecule">NF-kappa B p65</term> is sufficient to retarget <term id="T38" lex="NF-kappa_B_p65" sem="Protein_molecule">NF-kappa B p65</term> from the
<term id="T39" lex="nucleus" sem="Cell_component">nucleus</term> to the <term id="T40" lex="cytoplasm" sem="Cell_component">cytoplasm</term>, </sentence>
<event id="E5">
<type class="Binding"/>
<theme idref="T36"/>
<theme idref="T37"/>
<clue>2) the <clueType>binding</clueType> <linkTheme>of</linkTheme> I kappa B/MAD-3 <linkTheme>to</linkTheme> NF-kappa B p65 is sufficient to retarget NF-kappa B p65
from the nucleus to the cytoplasm,</clue>
</event>
<event id="E6">
<type class="Localization"/>
<theme idref="T38"/>
<clue>2) the binding of I kappa B/MAD-3 to NF-kappa B p65 is sufficient to <clueType>retarget</clueType> NF-kappa B p65 from the nucleus <clueLoc>to the
cytoplasm</clueLoc>,</clue>
</event>
<event id="E7">
<type class="Positive_regulation"/>
<theme idref="E6"/>
<cause idref="E5"/>
<clue>2) the binding of I kappa B/MAD-3 to NF-kappa B p65 <linkCause>is</linkCause> <clueType>sufficient</clueType> <linkTheme>to</linkTheme> retarget NF-kappa B p65
from the nucleus to the cytoplasm,</clue>
</event>
```

Fig. 2. Example of event annotation

each of the annotation boxes. In Figure 2, the words “binding” and “retarget” are marked-up as clue words for the event type *Binding* and *Localization*.

Additionaly, clue expressions for the locational, temporal, or experimental information are also marked-up. The text span “to the cytoplasm,” in the event annotation E6, is an example of clue expressions indicating the location of the event.

The last event, E7, represents the causal relation between E5 and E6. That is, the binding event (E5) of the two proteins “causes” the localization event (E6) of one of the two proteins. In the GENIA event ontology, the three classes, *Positive_regulation*, *Negative_regulation*, and *Regulation*, are used to represent causal relations between events or entities; e.g. promotion, inhibition, up-/down-regulation. The events of those classes are identified by its type, its theme and its *cause*: an event or an entity that positively or negatively affects the event. Note that, although the expression “is sufficient to” is hardly a linguistic expression for causality, the annotator recognized it as such in this sentence.

To assist the reader in understanding these relationships, we present Figure 2B: a graphical depiction of the example from Figure 2A. In this representation, entities from the GENIA term ontology are shown in rectangular boxes, while entities from the GENIA event ontology are shown in circles. The solid, dotted, and double arrows indicate the link between an event, and its theme, cause, and location, respectively. Figure 2C shows the XML representation of the three event annotations. This format will be used for public distribution of the event-annotated corpus.

3.2. *Annotation results*

This new annotation was made on half of the GENIA corpus, consisting of 1,000 Medline abstracts. It contains 9,372 sentences from which 36,114 events are identified. The quality and the size of the annotated corpus make it one of the best and largest corpus, in comparison with similar attempts. The event-annotated corpus and the full specification of the annotation scheme will be publicly available in XML at http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/.

4. Pathway corpus

A pathway is a detailed graphical representation of a biological system, which encompasses a set of mutually related events.[6] It integrates pieces of information on biological events scattered in many scientific publications into a coherent system, thereby facilitating the discussion among a large group of biologists, developing consensus for what actually happens in a biological system.

As a prototype of biological knowledge, we have constructed the NF-κB pathway, modeling the lifecyle of the NF-κB protein. For the pathway representation we use *Systems Biology Mark-up Language* (SBML), which is becoming a de facto standard for biological model representation.[2] In SMBL, a pathway or biological model is a collection of chemical reactions, and a reaction is characterized by its reactants, products, modifiers, and kinetic laws - which describes how quickly it takes place. Since our focus is on the construction of event networks describing pathways, we omit the kinetic laws.

We couple a pathway with a collection of evidence sentences that support reactions in the pathway. Designed to support the development of NLP-based TM systems for pathway construction, we call the collection a *pathway corpus*. We have constructed the NF-κB pathway&corpus in two versions; the full-text version and the GENIA version, described in the following sections.

4.1. *NF-κB pathway and corpus, the full-text version*

The full-text version of the NF-κB pathway is constructed based on a set of full-text papers. The papers were collected using a traditional keyword-based search. To raise the reliability, we only considered papers cited by at least two other papers. Because the NF-κB pathway is a well-studied pathway, we could find a lot of reliable review

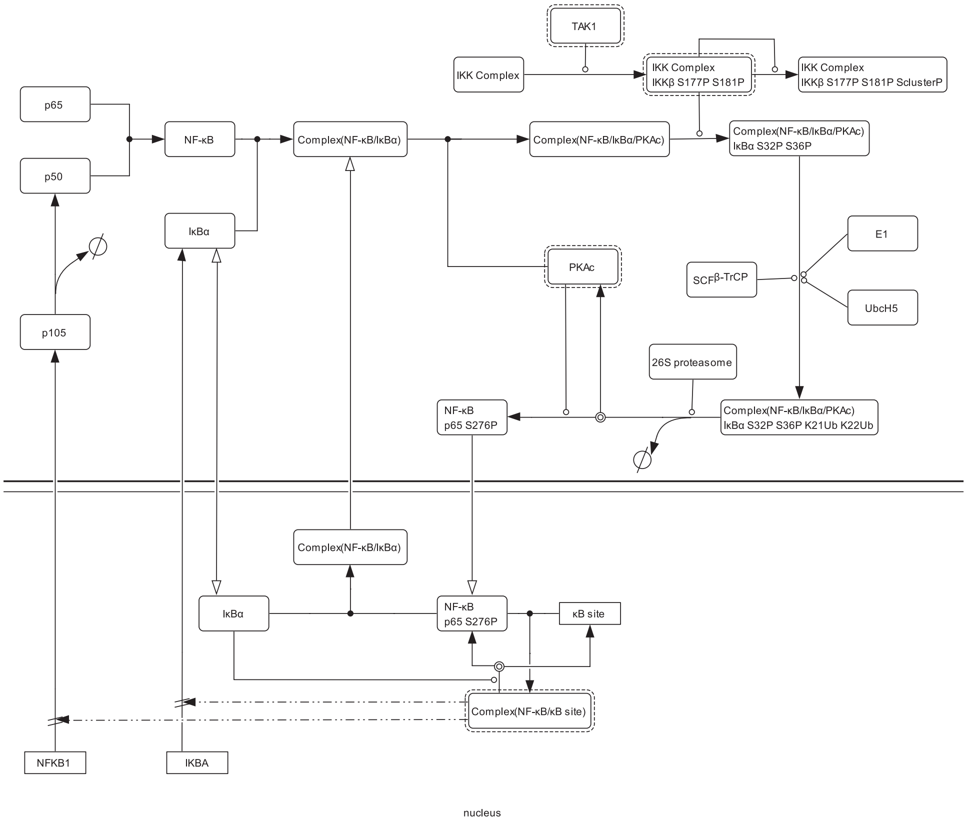


Fig. 3. The full-text version of NF-κB pathway

papers concerning its signaling.[19,20] The full-text version of the NF-κB Pathway was constructed based on the set of searched papers.

During the construction, evidence sentences supporting the pathway were collected and associated with the relavant portion of the pathway. As the result, we collected 467 sentences from the full text of 62 key papers and constructed the NF-κB pathway based on the evidence sentences. Figure 3 shows the full-text version of the NF-κB pathway.

4.2. *NF-κB pathway and corpus, the GENIA version*

As already mentioned, a pathway is a network representation of a course of focused events, which are supported by a collection of evidence texts. A pathway is thus subject to the availability of evidence texts. The GENIA version of the NF-κB pathway was constructed to explicitly address the correspondence between a pathway and a pathway corpus.

We first collected 561 abstracts from GENIA corpus that had the MeSH term

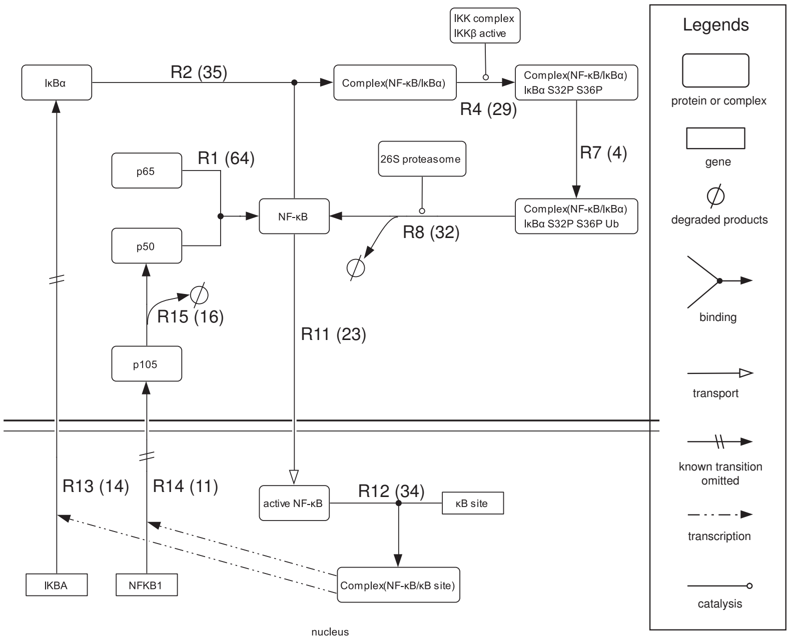


Fig. 4. The GENIA version of NF-κB pathway

"NF-kappa B" as their indexing term, and limited the source of evidence texts to this set of abstracts. We then manually examined each of the 5,223 sentence in the abstracts to collect evidence sentences for reactions in the full-text version of the NF-κB pathway. After collecting all of the evidence sentences, from the pathway, we removed the reactions without any evidence sentence from the GENIA corpus, and reorganized the pathway using only the remaining reactions to produce the GENIA version of NF-κB pathway.

Figure 4 shows the GENIA version of the NF-κB pathway. Note that the difference between the two versions of the NF-κB pathway comes from the availability of literature. Some of the evidence sentences are given in Table 1. The Id of reactions supported by the sentences are given in square brackets. Graphical depiction of the event annotation is given to some of the sentences.

Important characteristic features of the GENIA version of the NF-κB pathway-corpus include the following:

- Every sentence has been manually examined and tagged with the Id of reactions supported by the sentence.
- Every sentence that states events comes with event annotations.

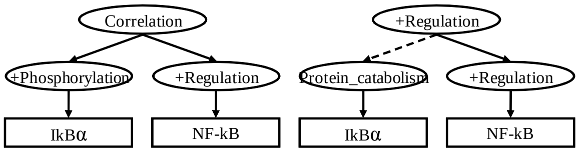
Due to the first feature, the corpus can be used as a gold standard for the development and evaluation of evidence sentence retrieval systems. The second feature

Table 1. Sentences describing reactions in Figure 4.

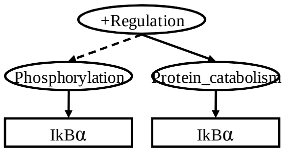
(1) Associated with its inhibitor, I kappaB, **NF-kappaB** resides as an inactive form in the cytoplasm. [R2]

(2) Upon stimulation by various agents, I kappaB is proteolyzed and **NF-kappaB** translocates to the nucleus, where it activates its target genes. [R8, R11]

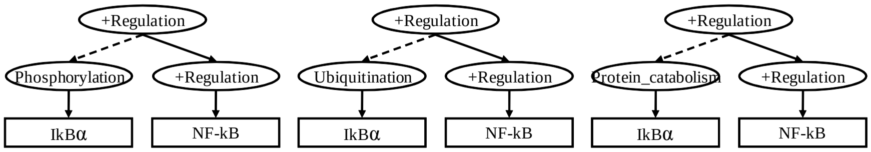
(3) *Activation* of NF-kappa B correlates with *phosphorylation* of I kappa B-alpha and requires the *proteolysis* of this inhibitor. [R4, R8]



(4) The present study demonstrates that tumor necrosis factor alpha-induced *degradation* of I kappa B alpha in human T cells is preceded by its rapid *phosphorylation* in vivo. [R4, R8]



(5) NF-kappa B *activation* involves signaled *phosphorylation*, *ubiquitination*, and *proteolysis* of I kappa B. [R4, R7, R8]



(6) IkappaB alpha *phosphorylation* on **Ser-32** and **Ser-36** is followed by its *degradation* and NF-kappaB *activation*. [R4, R8]

(7) The proteolytic degradation of the post-translationally modified I-kappa B is known to be mediated by the **26S proteasome complex**. [R8]

(8) During normal T-cell activation, IkappaBalpha is rapidly phosphorylated, ubiquitinated, and degraded by the **26S proteasome**, thus permitting the release of functional NF-kappaB. [R4, R7, R8]

enables the comparison of event representation between natural language and pathway representations.

5. Discussion: Pathway construction from Event annotation

The event annotation is intended to be used for the development of an ER (event recognition) program. While the results of ER can be used for various NLP-based TM such as intelligent text retrieval, question answering, etc., one of the major challenges is to use them to associate text fragments with the relevant part of pathways or to use them to semi-automatically construct pathways.

The sentences in Table 1 and the pathway in Figure 4 demonstrate the difference between the natural language expressions and biology-oriented representations. In this section, the difference will be characterized and the required processes to bridge the gap will be discussed.

5.1. *Finding instances from continuants*

Pathway representation is entity-centered, while language organizes information in a predicate-centered manner. To explain the difference, we apply definitions introduced in Ref. 21 that distinguishes between *continuants* and *instances.* A continuant is an entity which endures, or continues to exist throughout time, while undergoing different sorts of changes, including changes in location. We use the term *biological entity* to refer to an *instance* of a continuant at a specific time, which is also bound to a specific biological context.

The pathway representation is entity-centered since it gives an independent status to each of the biological entities or instances of the same continuant. The major players in this type of representation, e.g. nodes in a graphical representation, are biological entities that correspond to continuants in specific biological contexts. Events are expressed as directed edges between nodes, indicating the transition between biological entities.

In the pathway in the Figure 4, the reaction R11 represents translocation of NF-κB from cytoplasm to nucleus. In pathway representation, the same continuant NF-κB appears as different nodes before and after the event. These nodes denote instances of the same continuant in different biological contexts. Since these instances have different properties, it is natural that a pathway representation captures them as different nodes.

On the contrary, natural language text does not usually make explicit such distinctions among instances of the same continuant with different properties or in different contexts. In the sentence (2), the event corresponding to R11 is expressed simply as "NF-kappa B translocates to the nucleus," which indicates that NF-κB is involved in the localization event.

To construct a pathway representation that is entity-centered, distinct entities in different biological contexts must be captured at the mention of a continuant and its surrounding context in natural language expressions.

The same applies across sentences. In the sentence (1), which is followed by the sentence (2), the textual expression "NF-kappa B" refers to the continuant NF-κB, which is involved in the binding with IκB, thus suggesting the existence of two different instances of the continuant before and after the binding. Meanwhile, in the sentence (2), the preceding clause of the same textual expression "NF-kappa B," indicates a different context, in which occurs proteolysis of IκB. This suggests that a completely different set of instances from those of sentence (1) have to be introduced.

5.2. *Integration of fragmentary evidences*

While a pathway organizes a course of reactions that are carefully integrated, individual papers, and especially research papers, usually focus on a couple of reactions of the author's interest and on the causal relations between them.

In the pathway in Figure 4, the sequence of reactions R4, R7 and R8 represents how NF-κB is activated. The sentences (3), (4), (5), and (6) are evidence sentences supporting the reactions. With the exception of (5), all other sentences support the reactions only partially. For example, sentence (3) implies a causal relationship between the two events: "activation of NF-κB" and "phosphorylation of IκBα;" however, the direction of the causality is not mentioned. The sentence also states that "proteolysis of IκBα" causes "activation of NF-κB," which corresponds to the reaction R8. The order of the three events, "phosphorylation of IκBα," "proteolysis of IκBα," and "activation of NF-κB," can be determined when we consider another sentence (4) where the direction of the causal relation between "phosphorylation of IκBα," and "proteolysis of IκBα" is expressed. This exemplifies that we have to consider events in more than one sentence collectively in order to recover the integrated organization of events in a pathway representation.

The sentence (5), which mentions all three reactions: R4, R7, and R8, is from a review paper which integrates publications regarding the NF-κB pathway. Review papers, by nature, have similar properties with pathways in that they tend to pursue comprehension.

The sentence (6) is from a paper that was published later than the papers of all other sentences, and provides a novel, detailed information about the specific residue, which is phosphorylated (Ser-32 and Ser-36). The sentence (7) and (8) are the only sentences supporting the involvement of 26S proteasome complex in the proteolysis event, which means that without the two sentences, the GENIA version of the NF-κB pathway could not include the node of 26S proteasome.

6. Conclusion

In order to link the results of event recognition with pathways, we have to resolve the essential differences between the two representations. In this paper, we formulated one of the major differences as entity-centered vs. event-centered. We showed that looking at the linking problem as the problem of transforming event-centered representation to an entity-centered one helps us to formulate the technical problems in a clear manner. Another major obstacle is the underspecificity of information in text. We showed that this inevitably leads us to the problem of reconstructing actual event sequences by gathering pieces of information from more than one papers.

These foreseen problems may not be able to be automatically solved by programs, but we believe that even semi-automatic means will substantially reduce the burden of constructing and maintaining large pathway data bases.

Acknowledgments

This work was partially supported by Grant-in-Aid for Specially Promoted Research (MEXT, Japan) and Genome Network Project (MEXT, Japan).

References

1. J. Luciano and R. Stevens, e-Science and biological pathway semantics, *BMC Bioinformatics* **8**, p. S3 (2007).
2. M. Hucka, A. Finney, B. Bornstein, S. Keating, B. Shapiro, J. Matthews, B. Kovitz, M. Schilstra, A. Funahashi, J. Doyle and H. Kitano, Evolving a Lingua Franca and Associated Software Infrastructure for Computational Systems Biology: The Systems Biology Markup Language (SBML) Project, *Systems Biology* **1**, 41 (2004).
3. BioPAX, `http://www.biopax.org/`.
4. L. Martens, S. Orchard, R. Apweiler and H. Hermjakob, Human Proteome Organization Proteomics Standards Initiative: Data Standardization, a View on Developments and Policy, *Mol Cell Proteomics* **6**, 1666 (2007).
5. Systems Biology Ontology, `http://www.ebi.ac.uk/sbo/`.
6. G. D. Bader, M. P. Cary and C. Sander, Pathguide: a Pathway Resource List, *Nucl. Acids Res.* **34**, D504 (2006).
7. Kyoto Encyclopedia of Genes and Genomes, `http://www.genome.ad.jp/kegg/`.
8. S. Ananiadou and J. e. McNaught, *Text Mining for Biology and Biomedicine* (Artech House, 2006).
9. L. Hirschman, J. Park, J. Tsujii, L. Wong and C. Wu, Accomplishments and challenges in literature data mining for biology, *Bioinformatics* **18**, 1553 (2002).
10. S. Ananiadou, D. B. Kell and J. Tsujii, Text mining and its potential applications in systems biology, *Trends in Biotechnology* **24** (2006).
11. A. A. Morgan and L. Hirschman, Overview of BioCreative II Gene Normalization, in *Proceedings of Second BioCreative Challenge Evaluation Workshop*, eds. L. Hirschman, M. Krallinger and A. Valencia (2007).
12. M. Krallinger, F. Leitner and A. Valencia, Assessment of the Second BioCreative PPI task: Automatic Extraction of Protein-Protein Interactions., in *Proceedings of Second BioCreative Challenge Evaluation Workshop*, eds. L. Hirschman, M. Krallinger and A. Valencia (2007).
13. C. Nédellec, Learning Language in Logic - Genic Interaction Extraction Challenge, in *Proceedings of the 4th Learning Language in Logic Workshop (LLL05)*, eds. J. Cussens and C. Nédellec (2005).
14. C. Feng, F. Yamashita and M. Hashida, Automated Extraction of Information from the Literature on Chemical-CYP3A4 Interactions, *Journal of Chemical Information and Modeling* (2007).
15. J. G. Caporaso, J. Baumgartner, William A., D. A. Randolph, K. B. Cohen and L. Hunter, MutationFinder: a high-performance system for extracting point mutation mentions from text, *Bioinformatics* **23**, 1862 (2007).
16. W. Hersh, A. M. Cohen, P. Roberts and H. K. Rekapalli, TREC 2006 Genomics Track Overview, in *Proceeding of the TREC 2006*, (2006).
17. J. D. Kim, T. Ohta, Y. Tateisi and J. Tsujii, GENIA corpus - a semantically annotated corpus for bio-textmining, *Bioinformatics* **19**, i180 (2003), ISSN 1367-4803.
18. Y. Tateisi, A. Yakushiji, T. Ohta and J. Tsujii, Syntax Annotation for the GENIA corpus, in *Proceedings of the IJCNLP 2005, Companion volume*, (2005).
19. M. S. Hayden and S. Ghosh, Signaling to NF-kappaB, *Genes Dev.* **18**, 2195 (2004).
20. S. Ghosh and M. Karin, Missing Pieces in the NF-kappaB Puzzle., *Cell* **109**, S81 (2002).
21. B. Smith, W. Ceusters, B. Klagges, J. Köhler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. L. Rector and C. Rosse, Relations in biomedical ontologies, *Genome Biology* **6**, p. R46 (2005).