BIOINFORMATICS

Vol. 27 ISMB 2011, pages i111–i119
doi:10.1093/bioinformatics/btr214

Discovering and visualizing indirect associations between biomedical concepts

Yoshimasa Tsuruoka[1,*], Makoto Miwa[2,3,4], Kaisei Hamamoto[3], Jun'ichi Tsujii[3,4,5] and Sophia Ananiadou[3,4]

[1]School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), Nomi, [2]Department of Computer Science, The University of Tokyo, Tokyo, Japan, [3]The National Centre for Text Mining (NaCTeM), [4]School of Computer Science, The University of Manchester, Manchester, UK and [5]Microsoft Research Asia, Beijing, China

ABSTRACT

Motivation: Discovering useful associations between biomedical concepts has been one of the main goals in biomedical text-mining, and understanding their biomedical contexts is crucial in the discovery process. Hence, we need a text-mining system that helps users explore various types of (possibly hidden) associations in an easy and comprehensible manner.

Results: This article describes FACTA+, a real-time text-mining system for finding and visualizing indirect associations between biomedical concepts from MEDLINE abstracts. The system can be used as a text search engine like PubMed with additional features to help users discover and visualize indirect associations between important biomedical concepts such as genes, diseases and chemical compounds. FACTA+ inherits all functionality from its predecessor, FACTA, and extends it by incorporating three new features: (i) detecting biomolecular events in text using a machine learning model, (ii) discovering hidden associations using co-occurrence statistics between concepts, and (iii) visualizing associations to improve the interpretability of the output. To the best of our knowledge, FACTA+ is the first real-time web application that offers the functionality of finding concepts involving biomolecular events and visualizing indirect associations of concepts with both their categories and importance.

Availability: FACTA+ is available as a web application at http://refine1-nactem.mc.man.ac.uk/facta/, and its visualizer is available at http://refine1-nactem.mc.man.ac.uk/facta-visualizer/.

Contact: tsuruoka@jaist.ac.jp

1 INTRODUCTION

Text search engines such as PubMed are crucial in everyday research activities in biomedical sciences as a significant fraction of biomedical knowledge is still accessible only in textual form. We have previously developed FACTA, a text-mining system designed to help researchers find direct associations between biomedical concepts in an interactive fashion (Tsuruoka *et al.*, 2008). It is capable of producing ranked lists of important biomedical concepts, e.g. genes, diseases and chemical compounds, which are considered relevant to the query according to their co-occurrence statistics. The system has also been used as a search engine (Kemper *et al.*, 2010) to link biomolecular pathways to textual evidence.

This article describes three new classes of functionality that are introduced to extend and improve FACTA. The first extension is the use of biomolecular events as semantic metadata used for search. Semantic metadata derived from text to index digital documents for retrieval purposes have been used in systems like SUISEKI (Blaschke and Valencia, 2002), iHOP (Hoffmann and Valencia, 2005), Chilibot (Chen and Sharp, 2004), GoWeb (Dietze and Schroeder, 2009), Hanalyzer (Leach *et al.*, 2009), Semantic MEDLINE (Kilicoglu *et al.*, 2008), MEDIE (Miyao *et al.*, 2006), EBIMed (Rebholz-Schuhmann *et al.*, 2007) and KLEIO (Nobata *et al.*, 2008). Automatic extraction of events has a broad range of applications in biology (Ananiadou *et al.*, 2010), ranging from support for the creation and annotation of pathways (Kemper *et al.*, 2010) to automatic population or enrichment of databases. This novel event extension allows users to specify a concept which involves a biomolecular event.

The second extension is to help users discover indirectly associated concepts. Discovering hidden, previously unknown and potentially useful associations between biomedical concepts such as diseases and chemical compounds from the literature is a long-standing goal in biomedical text-mining (Swanson and Smalheiser, 1997). The pioneering work of Swanson (1986) hypothesized the role of fish oil in clinical treatment of Raynaud's disease, combining different pieces of information from the literature, and the hypothesis was later confirmed with experimental evidence. Among various approaches to the automatic generation of such hypotheses (Frijters *et al.*, 2010; Weeber *et al.*, 2003, 2005; Wren *et al.*, 2004), we adopt a simple approach using two-step associations. More specifically, we attempt to discover indirect associations by combining two known associations which are obtained from direct co-occurrence statistics. In this work, we give a probabilistic interpretation to the strengths of the discovered indirect associations, which allows the system to rank them in the order of their expected amount of information.

The third extension is visualization. The output format of FACTA was a tabular format—the associated concepts found by the system are categorized, ranked and presented in multiple columns. Although the tabular format may be useful enough in most cases, visualizing the output can help the user grasp the results more intuitively. The visualization component we have introduced in FACTA+ uses a technique called *treemapping* (Shneiderman, 2009), which enables the user to easily understand the relative importance of many different concepts.

This article is organized as follows. Section 2 describes our machine learning approach to detecting biomolecular events in text. Section 3 presents our algorithm for discovering indirect

*To whom correspondence should be addressed.

associations by using co-occurrence statistics. Section 4 describes the functionality of visualizing associations detected by the text-mining components. Concluding remarks are given in Section 5.

2 RECOGNIZING BIOMOLECULAR EVENTS

The first extension we have introduced in FACTA is the ability to detect biomolecular events mentioned in MEDLINE articles, thereby allowing the user to issue queries including such events. For example, FACTA+ allows the user to specify the documents that contain the word 'ERK2' and also mention positive regulation events, by using the query 'ERK2 GENIA:Positive_regulation'. This extension is motivated by the fact that biomolecular events have recently received considerable attention as an important type of information in biomedical text-mining (Ananiadou *et al.*, 2010; Bjorne *et al.*, 2010; Miwa *et al.*, 2010).

In this work, our definition of biomolecular events follows that of the GENIA event corpus (Kim *et al.*, 2008), in which events are basically characterized by verbs or nominalized verbs. For example, the sentence 'In Escherichia Coli, glnAP2 may be activated by NifA.' contains one event specified by the verb 'activated', with its arguments being 'In Escherichia Coli', 'glnAP2' and 'NifA'. In the GENIA event definition, every event is represented with a *trigger* and their arguments. Table 1 shows some examples of the events in the corpus with the trigger words italicized. For example, 'express' is the trigger word for the gene expression event in the first row in the table.

Recognizing the complete information on events involves the task of detecting triggers and their arguments, and there is a large body of previous work tackling this problem (Airola *et al.*, 2008; Divoli and Attwood, 2005; Huang *et al.*, 2004; Miwa *et al.*, 2009; Miyao *et al.*, 2009). However, we are not concerned with the task of detecting arguments in this article, since FACTA+ only uses information on abstract-level occurrences of concepts.

Since every event is represented with a trigger, what we need for event recognition is a component that can accurately detect triggers in text. Perhaps the most straight-forward approach to detecting trigger words in text is to use a dictionary, but pure dictionary-matching is not suitable for event recognition, since trigger words are often very ambiguous. For example, as seen in Table 1, the word 'express' is used as a trigger word for the gene expression event, but the word 'express' is a very common verb and used in many different meanings. Therefore, including the word 'express' in the dictionary would produce many false positive matchings.

We use a machine learning-based approach to sidestep this ambiguity problem, and use the GENIA event corpus (Kim *et al.*, 2008) as training data. More specifically, we used the data released for the BioNLP'09 shared task (Kim *et al.*, 2009) for training and testing our machine learning models. This shared task data is derived from the GENIA event corpus and contains annotations on nine event types concerning protein biology, which are a subset of the biomolecular events defined in the GENIA event ontology.

The machine learning models trained on the shard task data are used to recognize event triggers in text and their event types, and FACTA+ simply regards the detection of a trigger as an occurrence of the corresponding event in the abstract. Although this simple approach has a risk of producing false positives—because we disregard some semantically important types of information such as modality and negation (Garten *et al.*, 2010; Krallinger, 2010; Nawaz *et al.*, 2010), we leave it for future work.

2.1 Related work

The most straight-forward approach to detecting trigger words is to use a dictionary. Buyko *et al.* (2009) created a dictionary by manually curating and extending a lexicon derived from the original GENIA corpus with the help of researchers with a background in biology. A disambiguation step is performed by considering the co-occurrence statistics between each trigger word with event types in a training corpus. This disambiguation is used for some dictionary-based approaches [e.g. Kilicoglu and Bergler (2009), MacKinlay *et al.* (2009)]. Vlachos *et al.* (2009) extracted frequent triggers using a one-sense-per-term assumption, and performed soft matching (using lemmas and stems) to alleviate the problem of potential variability of trigger words. Vlachos (2010) extended the extracted dictionary by incorporating 'light' and 'ultra-light' triggers, which represent the discriminative modifiers of triggers. Kaljurand *et al.* (2009) extracted the dictionary from a training corpus, and disambiguated the trigger words by considering two kinds of co-occurrence statistics: one between each token and token considered to be a trigger and one between an event structure (event type and argument combination) and the trigger. Kilicoglu and Bergler (2009) manually cleaned the dictionary by removing ambiguous triggers, and also added variations of prefixes and nominal forms of verbs to the dictionary. Van Landeghem *et al.* (2009) built two separated manually cleaned dictionaries for unary events and other events. Cohen *et al.* (2009) selected triggers by iteratively testing manually constructed patterns.

Another popular approach is to use machine learning. Björne *et al.* (2009) and Miwa *et al.* (2010) used a multi-class support vector machines (SVMs) to detect and disambiguate trigger words. Morante *et al.* (2009) detected and disambiguated trigger words

Table 1. Examples of event-describing phrases

Event type	Phrase
Gene expression	Although resting Jurkat cells *express* **Fas**, ...
Positive regulation	**Fas** mRNA was *induced* approximately 10-fold in ...
Binding	... responses induced by **CD40** *engagement*.
Phosphorylation	Differential expression and *phosphorylation* of **CTCF**, a **c-myc** ...
Regulation	*Transcriptional regulation* of the **alpha2 integrin** gene in ...
Negative regulation	..., a specific *inhibitor* of **MAPK kinase 1**, ...

The terms in bold are protein names, and the italicized words are event triggers.

using IB1 memory-based classifier. MacKinlay *et al.* (2009) combined the outputs from a dictionary-based look-up tagger and a conditional random field (CRF)-based tagger.

Some other approaches detected events without a specialized module for trigger detection. Riedel *et al.* (2009) and Poon and Vanderwende (2010) detected events using Markov logic networks (MLNs). Neves *et al.* (2009) used the case-based reasoning, which finds 'case-solution' of a token including event, trigger, and argument types by retrieving the most similar, frequent case in the training data. Hakenberg *et al.* (2009) extracted shortest link paths on parse tree in events as queries, and also created regular expression-based patterns for regulation events. They grouped similar terms together manually, and applied both queries and patterns to the development and test datasets to detect triggers and arguments.

2.2 Detecting trigger words

To detect trigger words, we use a CRF model (Lafferty *et al.*, 2001). CRF models are log-linear probabilistic models for predicting sequences, which are widely used in biomedical text-mining as the machine learning model for named entity recognition (Okanohara *et al.*, 2006; Settles, 2004). The task of detecting trigger words can be performed with a CRF model by converting the task to a sequence prediction problem, in which the trigger sequences are represented with the 'IOB2' representation (Sang and Veenstra, 1999). In this representation, the beginning word of a trigger is given the 'B' tag. The following words are given the 'I' tag. The other words in the sentence are given the 'O' tag. The task of the CRF model is then to predict an 'IOB' sequence for a given sentence. In this work, the 'IOB' tags are combined with the nine different types of biomolecular events defined in the BioNLP'09 shared task data (Available at http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/SharedTask/).

2.3 Joint learning

In this work, we propose to use a model that performs the joint learning to recognize event triggers and protein names simultaneously. The motivation for our joint learning approach is that the presence of a protein name often indicates the presence of a trigger word in its vicinity. It should be noted that, unlike the shared task setting, we cannot use the information from gold-standard annotations for protein names, because we need to process the whole MEDLINE corpus for FACTA+.

The joint CRF model uses three additional tags: 'B-Protein', 'I-Protein' and 'Filler'. Table 2 shows an example of an IOB tag sequence for the sentence 'CD44 activated the transcription factor AP-1'. Note that the trigger word 'activated' is followed by a protein name but there is a gap between them. The tags assigned to 'CD44', 'AP', '-' and '1' are the ones added to recognize protein names. The 'Filler' tags are introduced to represent the regions that reside between the protein names and trigger words belonging to the same event. The filler tags enable the CRF model to propagate information from the existence of trigger words to non-adjacent protein names. In other words, the fact that a trigger word is followed by a protein name is captured by two transition features: (i) transition from 'B-Positive_regulation' to 'Filler' and (ii) transition from 'Filler' to 'B_Protein'.

Table 2. Joint learning of event triggers and protein names

Word	Tag
CD44	B-Protein
activated	B-Positive_regulation
the	Filler
transcription	Filler
factor	Filler
AP	B-Protein
-	I-Protein
1	I-Protein
.	O

Table 3. Feature templates used in the CRF tagger

Word unigram	w_{i-5}, w_{i-4}, w_{i-3}, w_{i-2}, w_{i-1}, w_i w_{i+1}, w_{i+2}, w_{i+3}, w_{i+4}, w_{i+5}	& y_i
Word bigram	$w_{i-1}w_i$, w_iw_{i+1}	& y_i
Word trigram	$w_{i-1}w_iw_{i+1}$	& y_i
Substrings	substrings of w_i (up to length 10)	& y_i
Word shape	S(w_i)	& y_i
Tag bigram	True	& $y_{i-1}y_i$

w_i is the current word. y_i is the current tag. Word shape S(w_i) is produced by converting capital letters into 'A', small letters into 'a' and numerals into '#'.

2.4 Experiments

We present experimental results to evaluate the performance of our joint learning approach. We compare our joint learning approach against two baseline approaches (models). The three CRF models used in the experiments are as follows.

(1) Triggers Only
A model limited to recognize only trigger words. The training data contains only the annotations on trigger words. Since there are nine different types of events in the data, this model considers $19(=2\times9+1)$ different possible tags for each word.

(2) Joint
A model to recognize protein names and trigger words jointly. However, the training data for this model does not include the Filler tag. This model considers $21(=19+2)$ different possible tags for each word.

(3) Joint + Filler
A model to recognize protein names and trigger words jointly. The training data also include the Filler tag as described in the previous subsection. This model considers $22(=21+1)$ different possible tags for each word.

We trained these CRF models using the training data (consisting of 800 MEDLINE abstracts) in the BioNLP'09 shared task corpus, and evaluated them using its development data (consisting of 150 abstracts). The corpus was preprocessed with simple rule-based scripts to perform sentence segmentation and tokenization. The feature templates used in our CRF models are shown in Table 3.

Table 4. Accuracy of trigger detection

	Triggers Only			Joint			Joint + Filler		
	Precision	Recall	F-score	Precision	Recall	*F*-score	Precision	Recall	*F*-score
Gene expression	70.9	60.8	65.5	74.9	57.4	65.0	77.9	66.4	71.7
Transcription	66.7	39.4	49.5	62.5	37.9	47.2	67.5	40.9	50.9
Protein catabolism	93.8	79.0	85.7	93.8	79.0	85.7	93.8	79.0	85.7
Localization	86.4	47.5	61.3	82.8	60.0	69.6	85.2	57.5	68.7
Binding	64.0	26.7	37.6	67.5	31.1	42.6	72.8	32.8	45.2
Phosphorylation	68.6	63.2	65.8	75.8	65.8	70.4	76.7	60.5	67.7
Regulation	57.8	19.0	28.6	54.5	21.9	31.2	50.0	13.9	21.7
Positive regulation	64.5	33.6	44.1	62.0	33.8	43.7	65.2	35.4	45.8
Negative regulation	61.3	30.7	40.9	58.2	30.7	40.2	61.8	28.0	38.5
Micro Average	67.2	38.4	48.9	67.1	39.1	49.4	70.5	40.4	51.4

The features include word n-grams, substrings and the shape of the current word and tag transitions.

Table 4 shows the results. The first nine rows in the table correspond to the nine types of biomolecular events defined in the corpus, and the bottom row shows the micro-average of the scores. Our proposed approach (i.e. the 'Joint + Filler' model) significantly outperforms the 'Triggers Only' model. This shows that the contextual information from the protein names is useful in detecting trigger words. It should also be noted that the performance of the Joint model without the filler tag is worse than the 'Joint + Filler' model, suggesting that it is important to explicitly transfer the information on the neighbouring tags in a CRF model.

Our approach consistently improved the performance for detecting event triggers, but the performance of detecting binding and regulation events was not very high. This is because these events can take multiple arguments, and also because regulation events can take other events as arguments. Rich linguistic information is required to deal with such event structures, and such triggers are not our current focus.

Note that the performance figures presented in this table are not comparable to those reported for the BioNLP shared task, since we did not use the gold standard information on the gene/protein names due to our purpose to evaluate the accuracy of trigger detection in a real-world setting where no gold standard annotation for gene/protein names are available.

The machine learning model described above (i.e. 'Joint + Filler') was applied to the whole MEDLINE corpus containing 20 033 079 articles, and the recognized events are indexed by FACTA+ so that it can accept queries including biomolecular events.[1] The number of articles indexed for each event type ranged from 53 262 (Protein catabolism) to 1 537 441 (Regulation).

We have also carried out a small-scale experiment to assess the quality of this indexing for the whole MEDLINE. We asked a bioNLP researcher with biology background to check the 10 latest articles returned by FACTA+ for each event class to see whether they are really relevant to the target class. In other words, the abstract-level precision was manually evaluated for each event class. The result was that 86 out of the 90 abstracts were actually relevant to the target event class.[2] Although the recall of this event indexing is not completely clear, the precision is probably good enough to be used in practice.

3 DISCOVERING INDIRECT ASSOCIATIONS

A common approach to automatic discovery of useful hypotheses is to combine two (or more) known associations, i.e. if concept X is associated with concept Y, and concept Y is associated with concept Z, then the potential association between X and Z is considered as a useful hypothesis unless there is already a known association between X and Z. This approach is often called Swanson's ABC model approach after his seminal work on literature-based hypothesis generation (Swanson, 1990). Figure 1 illustrates this approach in the context of implementing it on FACTA+, where the user provides a starting concept as a *query* to the system. We call the concepts that are directly associated with the query the *pivot concepts*, and the concepts that are indirectly associated with the query through those pivot concepts the *target concepts*.

3.1 Related work

There are a number of publicly available software tools that offer the functionality for discovering indirect associations. Anni 2.0 (Jelier *et al.*, 2008) is a Java client–server application which provides an ontology-based interface to MEDLINE. It can find concepts that have many intermediate concepts in common, thereby allowing the user to discover concepts that do not directly co-occur with the starting concept. Unlike FACTA+, the starting concept is defined as a combination of predefined concepts provided by the system, i.e. free keywords cannot be used to define a concept. BITOLA (Hristovski *et al.*, 2005) is a web application which allows the user to retrieve target concepts using MeSH terms as pivot concepts. It can also incorporate biomedical knowledge (e.g. chromosomal

[1]The 'Protein' tags output by the joint CRF model were not used for indexing. In FACTA+, protein names were separately indexed and normalized by dictionary matching.

[2]Note that the precision scores reported in Table 4 are calculated by entity-level exact matching, which is much more strict than abstract-level precision.

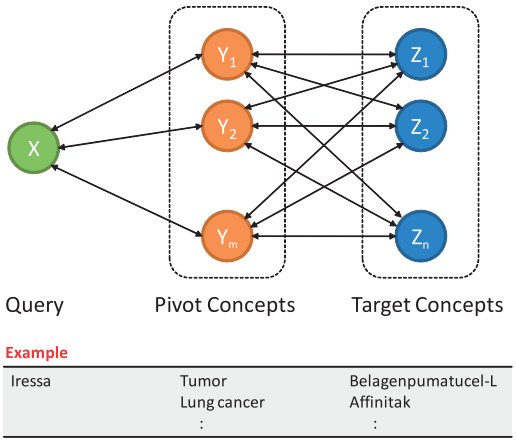


Fig. 1. Finding indirectly associated concepts.

location) to improve the precision of the output. BITOLA requires the user to specify each pivot concept manually. In other words, the pivot concepts that are not selected by the user are not used for computing the association strengths between the query concept and the target concepts, whereas FACTA+ retrieves target concepts by considering all potential pivot concepts of a specified class. Arrowsmith (Smalheiser *et al.*, 2009) is perhaps a more well-known tool for literature-based hypotheses generation, but it is designed to find concepts or terms that interlink two distinct concepts—it is more similar, in our terminology, to finding pivot concepts given a query and a target concept. CoPub Discovery (Frijters *et al.*, 2010) is a web application based on a co-occurrence database in CoPub (Frijters *et al.*, 2008). It allows the user to discover not only target concepts by considering all potential pivot concepts of a specified class, also but pivot concepts given a query and a target concept. It employed two concept classes Gene and BioConcept for pivot concepts, and used regular expressions to search for keywords in MEDLINE and linking a query to the starting concept, while FACTA+ employs six concept classes, event-based detection and flexible keyword matching. CoPub Discovery shows the ranking of the target concepts, whereas FACTA+ incorporates a visualization for the indirect associations.

As for the scoring scheme for ranking indirect associations, Yetisgen-Yildiz and Pratt (2009) describe a comprehensive overview of existing approaches and compare the performance of four different criteria (Association Rules, TF-IDF, Z-score and Mutual Information Measure) using a cut-off date technique.

3.2 Interface

Like many text search engines, FACTA+ accepts arbitrary keywords, concept identifies or their boolean combination as a query to specify the starting concept (i.e. concept X in Fig. 1). The system first retrieves pivot concepts using co-occurrence statistics from the literature, and then produces target concepts that are scored and ranked in accordance to their association strengths with the query and pivot concepts. One of the distinct features of our system is that it achieves real-time responses in most cases while allowing the user to use a very flexible query as the input to the system.

Currently, FACTA+ accepts the following six classes of biomedical concepts as pivot and target concepts: human genes/proteins, diseases, symptoms, drugs, enzymes and chemical compounds. The user can choose one of these classes as pivot and target concepts when performing a search.

As an example of indirect associations, the search result for the input query 'E-cadherin and GENIA:Negative_regulation' returned by FACTA+ is shown in Figure 2. The first row in the table shows that the query concept is indirectly associated with a disease 'acute respiratory failure' through multiple genes including tumour suppressor candidate 3. The visualized version of this search result is shown in Figure 4.

E-cadherin is a cell adhesion molecule involved with the binding between a cell and other cells or extracellular matrix. The search results shown in Figures 2 and 4 indicate that E-cadherin is associated with multiple nervous system disorders (e.g. Alzheimer's disease, Parkinson's disease, epilepsy) via several proteins/genes, even though E-cadherin itself rarely appears with such disorders (see also Fig. 3 for direct associations). This indirect search result suggests that E-cadherin could be a potential candidate of drug target for nervous system disorders.

3.3 Ranking

Since the number of indirectly associated concepts can be vast, it is important to rank them properly when they are presented to the user. On the one hand, the ranking criterion should favour 'hidden' associations, because the associations that can be easily observed in the existing literature are not interesting to the users who are seeking for novel associations. In fact, such 'known' associations can be browsed by using the existing functionality of FACTA.

On the other hand, the indirect associations output by the system need to be plausible. To incorporate these two factors, FACTA+ defines a ranking score for each target concept Z_j as follows:

$$(\text{score}) = R(X,Z_j) \times \{-\log D(X,Z_j)\}, \tag{1}$$

where $D(X,Z_j)$ is the strength of direct association between concept X and Z_j, and $R(X,Z_j)$ is the reliability of the indirect association between the two. Notice that this score can translate into the expected amount of *information* (in the information theoretic sense) if $R(X,Z_j)$ and $D(X,Z_j)$ are given as probability values. The term $-\log D(X,Z_j)$ takes on a large value if the strength of association between X and Z_j is weak. In other words, this term represents how hidden or surprising the association is.

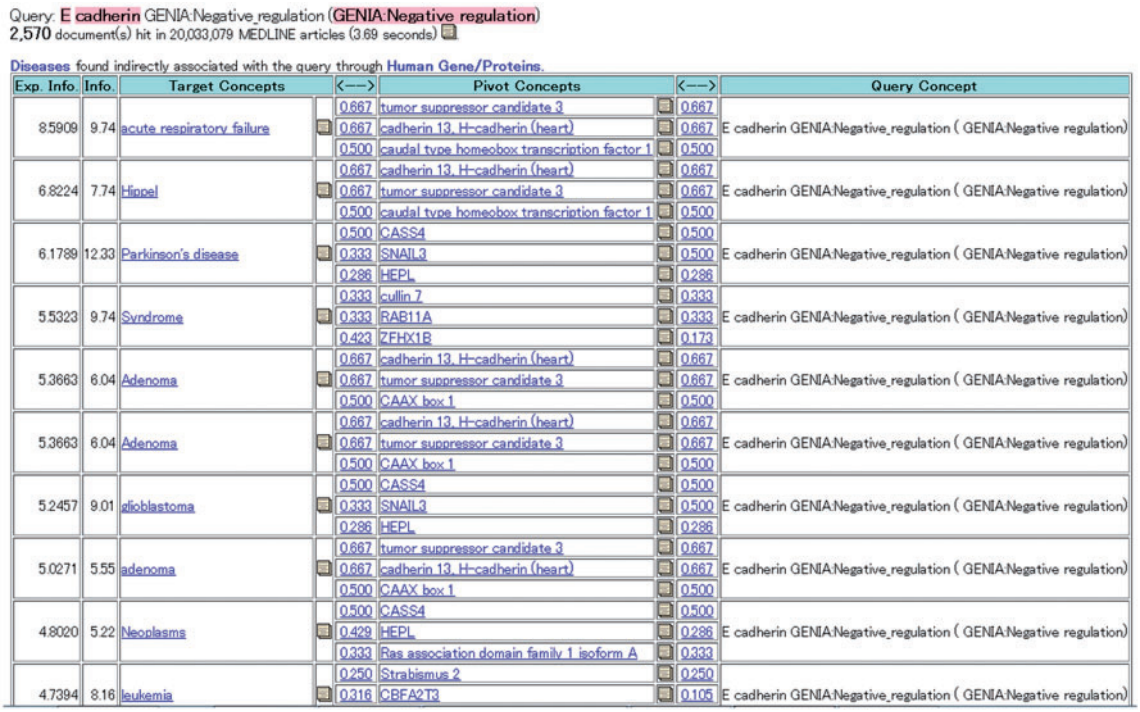
If we assume that $D(\cdot,\cdot)$ are given as probabilities, and that all associations connecting X with Z_j are independent, the reliability term can be computed as follows:

$$R(X,Z_j) = 1 - \prod_{i=1}^{m} \{1 - D(X,Y_i)D(Y_i,Z_j)\}, \tag{2}$$

since the probability that the connection between X and Z_j is true is given by the probability that there is at least one true path connecting X with Z_j. Now the remaining question is how $D(\cdot,\cdot)$ are computed as probabilities. In FACTA+, we approximate them using conditional probabilities:

$$D(V,W) = \max\{P(V|W), P(W|V)\}, \tag{3}$$

where $P(V|W)$ is the (smoothed) conditional probability of the occurrence of concept V given that of concept W within the same document. $P(W|V)$ is defined likewise. We take the maximum of



Query: E cadherin GENIA:Negative_regulation (GENIA:Negative regulation)
2,570 document(s) hit in 20,033,079 MEDLINE articles (3.69 seconds)

Diseases found indirectly associated with the query through Human Gene/Proteins.

Exp. Info.	Info.	Target Concepts	<--->	Pivot Concepts	<--->	Query Concept
8.5909	9.74	acute respiratory failure	0.667	tumor suppressor candidate 3	0.667	E cadherin GENIA:Negative_regulation (GENIA:Negative regulation)
			0.667	cadherin 13, H-cadherin (heart)	0.667	
			0.500	caudal type homeobox transcription factor 1	0.500	
6.8224	7.74	Hippel	0.667	cadherin 13, H-cadherin (heart)	0.667	E cadherin GENIA:Negative_regulation (GENIA:Negative regulation)
			0.667	tumor suppressor candidate 3	0.667	
			0.500	caudal type homeobox transcription factor 1	0.500	
6.1789	12.33	Parkinson's disease	0.500	CASS4	0.500	E cadherin GENIA:Negative_regulation (GENIA:Negative regulation)
			0.333	SNAIL3	0.500	
			0.286	HEPL	0.286	
5.5323	9.74	Syndrome	0.333	cullin 7	0.333	E cadherin GENIA:Negative_regulation (GENIA:Negative regulation)
			0.333	RAB11A	0.333	
			0.423	ZFHX1B	0.173	
5.3663	6.04	Adenoma	0.667	cadherin 13, H-cadherin (heart)	0.667	E cadherin GENIA:Negative_regulation (GENIA:Negative regulation)
			0.667	tumor suppressor candidate 3	0.667	
			0.500	CAAX box 1	0.500	
5.3663	6.04	Adenoma	0.667	cadherin 13, H-cadherin (heart)	0.667	E cadherin GENIA:Negative_regulation (GENIA:Negative regulation)
			0.667	tumor suppressor candidate 3	0.667	
			0.500	CAAX box 1	0.500	
5.2457	9.01	glioblastoma	0.500	CASS4	0.500	E cadherin GENIA:Negative_regulation (GENIA:Negative regulation)
			0.333	SNAIL3	0.500	
			0.286	HEPL	0.286	
5.0271	5.55	adenoma	0.667	tumor suppressor candidate 3	0.667	E cadherin GENIA:Negative_regulation (GENIA:Negative regulation)
			0.667	cadherin 13, H-cadherin (heart)	0.667	
			0.500	CAAX box 1	0.500	
4.8020	5.22	Neoplasms	0.500	CASS4	0.500	E cadherin GENIA:Negative_regulation (GENIA:Negative regulation)
			0.429	HEPL	0.286	
			0.333	Ras association domain family 1 isoform A	0.333	
4.7394	8.16	leukemia	0.250	Strabismus 2	0.250	E cadherin GENIA:Negative_regulation (GENIA:Negative regulation)
			0.316	CBFA2T3	0.105	

Fig. 2. A screen-shot of FACTA+ search results for indirect associations. The links and icons in the table give the user a quick access to the textual evidence (snippets) of the associations.

the conditional probabilities of both directions to avoid missing any association detectable from direct co-occurrence statistics.

3.4 Indexing

To achieve real-time responses, we pre-compute the association statistics between all predefined concepts (i.e. between all possible Ys and Zs) and store them on memory—this may sound prohibitively expensive but is possible because we need to store only the information about the pairs of concepts that actually co-occur in at least one document. The number of pairwise associations indexed for achieving real-time discovery of indirect associations was 49 620 438.

The indexing for the predefined concepts was performed by dictionary matching. More specifically, we employed the longest matching method using six different dictionaries built for the aforementioned six classes of concepts. The number of unique concepts indexed for the whole MEDLINE was 107 060. One of the major difficulties in the indexing process was the problem of semantic ambiguity of the terms. We used a simple rule-based method to alleviate the problem of acronym ambiguity, but the ambiguity problems with terms like protein family names are difficult to solve and left for future work.

4 VISUALIZATION

The ranked list of found concepts is often used for showing the results and is used in FACTA+ and other services like Arrowsmith, and it is useful for displaying the details of the extracted results, but lists are not well suited to grasp the big picture since the ranking scores are not intuitive to use. We, therefore, developed a visualization component to make the found concepts easy to understand. The primary technique used in this component is treemapping (Shneiderman, 2009), a method for visualizing hierarchical data by using a set of rectangles. Figures 3 and 4 show the visualization of directly and indirectly associated concepts for the query 'E-cadherin and GENIA:Negative_regulation'. In these visualizations, each rectangle represents an individual concept, and its area is proportional to the score given to the concept. The component is built using Adobe Flash because it enables rich graphical expressions suitable for visualization like gradient fill and smooth animation.

4.1 Related work

One of the most well-known sites which use treemapping is Newsmap (http://newsmap.jp/). Newsmap visualizes popular Google News stories. Stories are represented as rectangles and they are categorized into countries or topics. Users can browse the full text of the story by clicking a rectangle.

4.2 Visualizing direct and indirect associations

FACTA+ has two visualization components: one for presenting directly associated concepts (e.g. Fig. 3) and the other indirectly associated concepts (e.g. Fig. 4). The functionality and usage of these visualization components are explained in this subsection.

The directly associated concepts to a user query are visualized as rectangles. The concepts are grouped into six categories (human

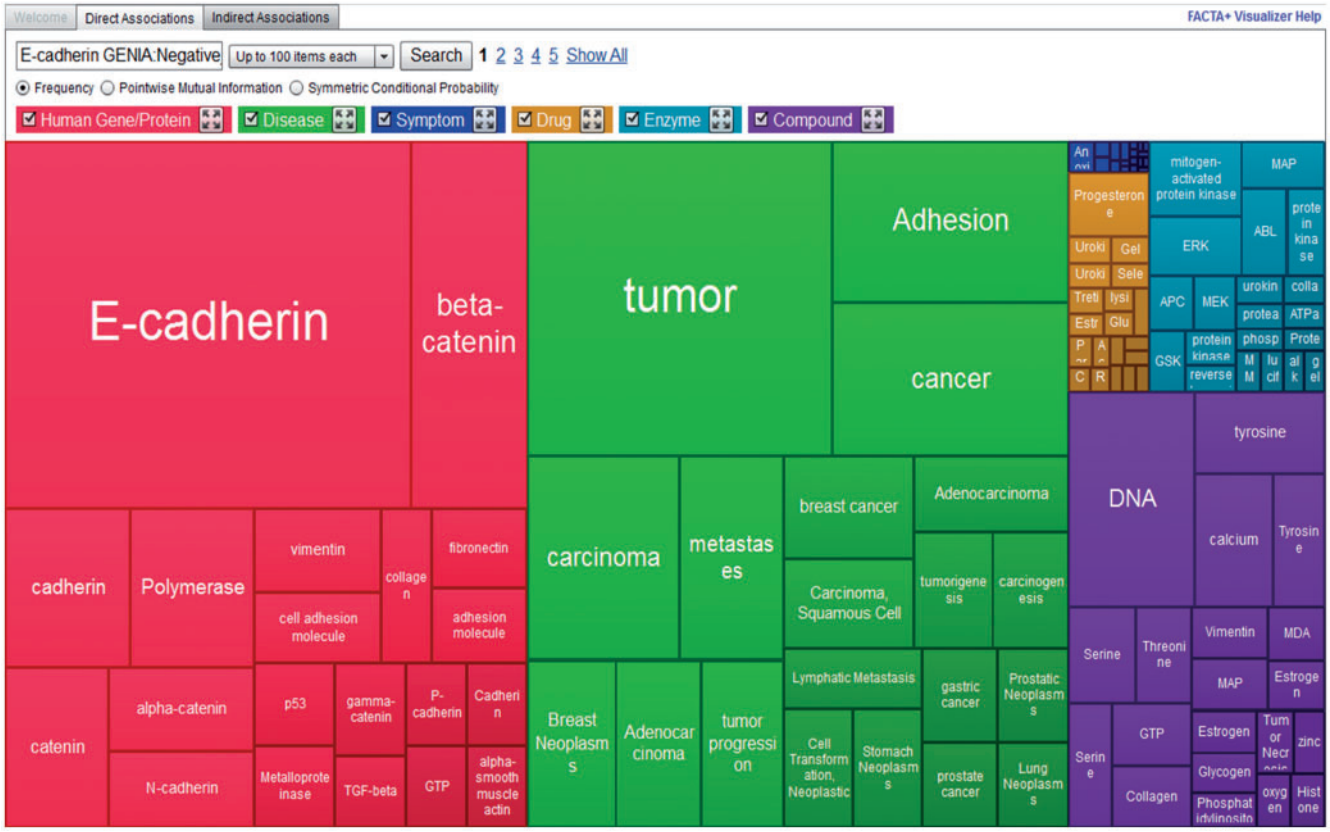


Fig. 3. Visualization of directly associated concepts using treemapping.

genes/proteins, diseases, symptoms, drugs, enzymes and chemical compounds) and each category forms a parent rectangle. The number of concepts shown is limited at first, but more results become visible by using a pager control. Users can easily grasp the relative importance of each concept using the size of each rectangle. The rectangles are arranged to maintain the similar aspect ratios to make the rectangles visually recognizable. Users can also focus on a particular set of categories by using check boxes. The layout of rectangles is recalculated instantaneously. This is done using a smooth animation so that the change is visually traceable.

The method for visualizing indirectly associated concepts is slightly different to the one for directly associated concepts. Pivot concepts co-occurred with the query are shown on the left-hand side, and target concepts co-occurred with the pivot concepts are shown on the right-hand side. In addition to the treemapping, we introduced a 'link' visualization between pivot concepts and target concepts. Links represent co-occurrences between pivot and target concepts, and the width of each link indicates the strength of its association. When users point the mouse cursor on a particular pivot concept, links from the concept to its corresponding target concepts appear. Similarly, when users point a target concept, links from the concept to its corresponding pivot concepts appear.

In both directly and indirectly associated concepts views, users can browse underlying documents by clicking a rectangle or a link and select 'view documents'.

5 CONCLUSION

We have presented three extensions which have recently been introduced in FACTA+, a text search engine for MEDLINE.

First, we have proposed a joint learning approach to detecting biomolecular events described in text. The performance of detecting trigger words has been significantly improved by performing the task jointly with protein name recognition.

Second, we have presented a method for detecting indirect associations. The associations are ranked by the level of novelty and reliability, which are estimated by combining the strengths of multiple known associations that are directly observable from co-occurrence statistics in the literature.

Third, we have implemented a novel visualization component which provides an intuitive overview of the discovered concepts and their associations. Each concept is represented with a coloured rectangle; the colour shows the category and the area indicates the importance. Each association is displayed with a link whose width indicates the importance.

The three classes of functionality described in this article are implemented in FACTA+. The system accepts concept identifiers, arbitrary keywords and their boolean combinations as a query and immediately produces a ranked list (or its visualization) of concepts that are indirectly associated with the query. FACTA+ is implemented in C++ and currently running on a single Linux server with 32 GB of memory. The service is available at http://refine1-nactem.mc.man.ac.uk/facta/, and the visualization

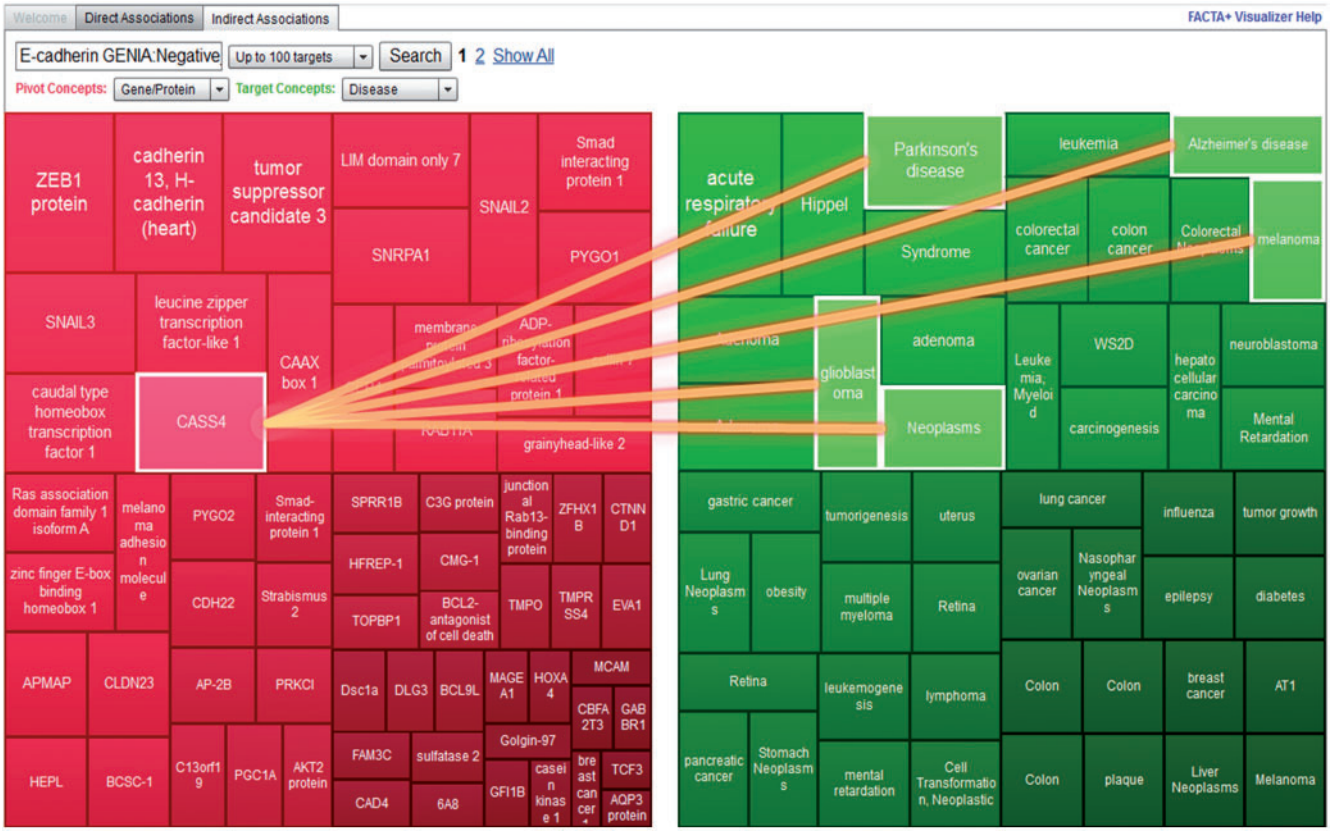


Fig. 4. Visualization of indirectly associated concepts using treemapping and links.

system is available at http://refine1-nactem.mc.man.ac.uk/facta-visualizer/.

ACKNOWLEDGEMENTS

We are grateful to the anonymous referees for their insightful comments on the earlier version of this article. Thanks also to T. Ohta, X. Wang, and S. A. Iqbal for their valuable feedback and comments.

Funding: Biotechnology and Biological Sciences Research Council (BBSRC BB/G013160/1). The UK National Centre for Text Mining is funded by the UK Joint Information Systems Committee (JISC).

REFERENCES

Airola,A. *et al*. (2008) All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC bioinformatics*, **9** (Suppl. 11), S2.

Ananiadou,S. *et al*. (2010) Event extraction for systems biology by text mining the literature. *Trends in Biotechnol.*, **28**, 381–390.

Björne,J. *et al*. (2009) Extracting complex biological events with rich graph-based feature sets. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, Association for Computational Linguistics, pp. 10–18.

Bjorne,J. *et al*. (2010) Complex event extraction at PubMed scale. *Bioinformatics*, **26**, i382–i390.

Blaschke,C. and Valencia,A. (2002) The frame-based module of the suiseki information extraction system. *IEEE Intell. Syst.*, **17**, 14–20.

Buyko,E. *et al*. (2009) Event extraction from trimmed dependency graphs. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, Association for Computational Linguistics, pp. 19–27.

Chen,H. and Sharp,B. (2004) Content-rich biological network constructed by mining pubmed abstracts. *BMC Bioinformatics*, **5**, 147.

Cohen,K.B. *et al*. (2009) High-precision biological event extraction with a concept recognizer. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, Association for Computational Linguistics, pp. 50–58.

Dietze,H. and Schroeder,M. (2009) Goweb: a semantic search engine for the life science web. *BMC Bioinformatics*, **10** (Suppl. 10), S7.

Divoli,A. and Attwood,T.K. (2005) BioIE: extracting informative sentences from the biomedical literature. *Bioinformatics*, **21**, 2138–2139.

Frijters,R. *et al*. (2008) CoPub: a literature-based keyword enrichment tool for microarray data analysis. *Nucleic Acids Res.*, **36**, W406–W410.

Frijters,R. *et al*. (2010) Literature mining for the discovery of hidden connections between drugs, genes and diseases. *PLoS Comput. Biol.*, **6**, e1000943.

Garten,Y. *et al*. (2010) Recent progress in automatically extracting information from the pharmacogenomic literature. *Pharmacogenomics*, **11**, 1467–1489.

Hakenberg,J. *et al*. (2009) Molecular event extraction from link grammar parse trees. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, Association for Computational Linguistics, pp. 86–94.

Hoffmann,R. and Valencia,A. (2005) Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, **21** (Suppl. 2), 252–258.

Hristovski,D. *et al*. (2005) Using literature-based discovery to identify disease candidate genes. *Inter. J. Med. Infor.*, **74**, 289–298.

Huang,M. *et al*. (2004) Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics*, **20**, 3604–3612.

Jelier,R. *et al*. (2008) Anni 2.0: a multipurpose text-mining tool for the life sciences. *Genome Biol.*, **9**, R96.

Kaljurand,K. *et al*. (2009) Uzurich in the bionlp 2009 shared task. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, Association for Computational Linguistics, pp. 28–36.

Kemper,B. *et al*. (2010) PathText: a text mining integrator for biological pathway visualizations. *Bioinformatics*, **26**, i374–i381.

Kilicoglu,H. and Bergler,S. (2009) Syntactic dependency based heuristics for biological event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, Association for Computational Linguistics, pp. 119–127.

Kilicoglu,H. *et al*. (2008) Semantic MEDLINE: a web application to manage the results of PubMed searches. In *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008)*, Turku Centre for Computer Science, pp. 69–76.

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

Kim,J.-D. *et al*. (2009) Overview of bionlp'09 shared task on event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, Association for Computational Linguistics, pp. 1–9.

Krallinger,M. (2010) Importance of negations and experimental qualifiers in biomedical literature. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, University of Antwerp, pp. 46–49.

Lafferty,J. *et al*. (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, Morgan Kaufmann Publishers Inc., pp. 282–289.

Leach,S.M. *et al*. (2009) Biomedical discovery acceleration, with applications to craniofacial development. *PLoS Comput. Biol.*, **5**, e1000215.

MacKinlay,A. *et al*. (2009) Biomedical event annotation with crfs and precision grammars. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, Association for Computational Linguistics, Boulder, Colorado, pp. 77–85.

Miwa,M. *et al*. (2009) Protein-protein interaction extraction by leveraging multiple kernels and parsers. *Inter. J. Med. Infor.*, **78**, e39–e46.

Miwa,M. *et al*. (2010) Event extraction with complex event classification using rich features. *J. Bioinformatics Comput. Biol.*, **8**, 131–146.

Miyao,Y. *et al*. (2006) Semantic retrieval for the accurate identification of relational concepts in massive textbases. In *Proceedings of the 21th international Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*, Association for Computational Linguistics, pp. 1017–1024.

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

Morante,R. *et al*. (2009) A memory-based learning approach to event extraction in biomedical texts. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, Association for Computational Linguistics, pp. 59–67.

Nawaz,R. *et al*. (2010) Evaluating a meta-knowledge annotation scheme for bio-events. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, University of Antwerp, pp. 69–77.

Neves,M. *et al*. (2009) Extraction of biomedical events using case-based reasoning. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, Association for Computational Linguistics, pp. 68–76.

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

Okanohara,D. *et al*. (2006) Improving the scalability of semi-markov conditional random fields for named entity recognition. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*, Association for Computational Linguistics, pp. 465–472.

Poon,H. and Vanderwende,L. (2010) Joint inference for knowledge extraction from biomedical literature. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 813–821.

Rebholz-Schuhmann,D. *et al*. (2007) EBIMed–text crunching to gather facts for proteins from MEDLINE. *Bioinformatics*, **23**, e237–e244.

Riedel,S. *et al*. (2009) A markov logic approach to biomolecular event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, Association for Computational Linguistics, pp. 41–49.

Sang,E.F.T.K. and Veenstra,J. (1999) Representing text chunks. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Association for Computational Linguistics, pp. 173–179.

Settles,B. (2004) Biomedical named entity recognition using conditional random fields and rich feature sets. In *COLING 2004 International Joint workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP) 2004*, COLING, pp. 107–110.

Shneiderman,B. (2009) Treemaps for space-constrained visualization of hierarchies. Availabe at http://www.cs.umd.edu/hcil/treemap-history/.

Smalheiser,N.R. *et al*. (2009) Arrowsmith two-node search interface: a tutorial on finding meaningful links between two disparate sets of articles in MEDLINE. *Comput. Methods Program. Biomed.*, **94**, 190–197.

Swanson,D.R. (1986) Fish oil, raynaud's syndrome, and undiscovered public knowledge. *Pers. Biol. Med.*, **30**, 7–18.

Swanson,D.R. (1990) Medical literature as a potential source of new knowledge. *Bull. Med. Libr. Assoc.*, **78**, 29–37.

Swanson,D.R. and Smalheiser,N.R. (1997) An interactive system for finding complementary literatures: a stimulus to scientifc discovery. *Artif. Intell.*, **91**, 183–203.

Tsuruoka,Y. *et al*. (2008) FACTA: a text search engine for finding associated biomedical concepts. *Bioinformatics*, **24**, 2559–2560.

Van Landeghem,S. *et al*. (2009) Analyzing text in search of biomolecular events: a high-precision machine learning framework. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, Association for Computational Linguistics, pp. 128–136.

Vlachos,A. (2010) Two strong baselines for the bionlp 2009 event extraction task. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, Association for Computational Linguistics, pp. 1–9.

Vlachos,A. *et al*. (2009) Biomedical event extraction without training data. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, Association for Computational Linguistics, pp. 37–40.

Weeber,M. *et al*. (2003) Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. *JAMIA*, **10**, 252–259.

Weeber,M. *et al*. (2005) Online tools to support literature-based discovery in the life sciences. *Brief. Bioinformatics*, **6**, 277–286.

Wren,J.D. *et al*. (2004) Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics*, **20**, 389–398.

Yetisgen-Yildiz,M. and Pratt,W. (2009) A new evaluation methodology for literature-based discovery systems. *J. Biomed. Inform.*, **42**, 633–643.