W416–W421 *Nucleic Acids Research, 2014, Vol. 42, Web Server issue* *Published online 09 May 2014*
doi: 10.1093/nar/gku368

OncoSearch: cancer gene search engine with literature evidence

Hee-Jin Lee[1], Tien Cuong Dang[1], Hyunju Lee[2] and Jong C. Park[1,*]

[1]Department of Computer Science, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 305-701, Republic of Korea and [2]School of Information and Communications, Gwangju Institute of Science and Technology, 123 Cheomdangwagi-ro, Buk-gu, Gwangju 500-712, Republic of Korea

Received February 22, 2014; Revised April 11, 2014; Accepted April 15, 2014

ABSTRACT

In order to identify genes that are involved in oncogenesis and to understand how such genes affect cancers, abnormal gene expressions in cancers are actively studied. For an efficient access to the results of such studies that are reported in biomedical literature, the relevant information is accumulated via text-mining tools and made available through the Web. However, current Web tools are not yet tailored enough to allow queries that specify how a cancer changes along with the change in gene expression level, which is an important piece of information to understand an involved gene's role in cancer progression or regression. OncoSearch is a Web-based engine that searches Medline abstracts for sentences that mention gene expression changes in cancers, with queries that specify (i) whether a gene expression level is up-regulated or down-regulated, (ii) whether a certain type of cancer progresses or regresses along with such gene expression change and (iii) the expected role of the gene in the cancer. OncoSearch is available through http://oncosearch.biopathway.org.

INTRODUCTION

For cancer research, it is essential to identify various genes that are involved in oncogenesis and to understand how the genes affect cancers. While a large amount of information about such genes is reported in the literature, the vastness of the biomedical literature makes it necessary to use databases and Web-tools that enable efficient access to relevant information. The largest databases on genes such as Entrez Gene (1), GeneCards (2) and UniProtKB (3) contain gene-disease associations along with other information about genes. Databases that are specialized for cancer-related genes include: TGDB (http://www.tumor-gene.org/) for information about genes that are targets for cancer-causing mutations; TSGene (4) for prospective tumor suppressor genes; and DDEC (5), DDPC (6), DDOC (7) and CCDB (8) for the genes related to restricted types of cancers, or esophageal, prostate, ovarian and cervix cancers, respectively.

In addition to the databases above, researchers can also use Web-based text-mining tools to search for the genes implicated in cancers and to retrieve textual evidence about the gene–cancer relations. General purpose tools such as PolySearch (9), Facta (10) and Génie (11) allow queries about whether or not a gene is related to a disease in any way. Other Web tools that are specialized for genes and diseases allow queries including molecular context of gene-disease association. MeInfoText (12) and PubMeth (13) can collect genes that show the methylated status in cancers. BioContext (14) and DigSee (15) identify several types of molecular events such as 'gene expression', 'regulation', 'phosphorylation' and 'localization'. These two systems can determine whether an identified molecular event is related to a certain type of cancer or not.

Although the tools above provide essential information about gene–cancer relations including how a gene is changed when the gene is implicated in a type of cancer, the provided information is not yet comprehensive enough to explain the role of the involved gene in the cancer. In order to fully understand how a gene affects a cancer, we need information about how the cancer is changed along with the change in the gene. For instance, suppose that an increase in the expression level of a gene is associated to a certain type of cancer. In order to determine the gene as an oncogene or a suppressor of the cancer, which is essential for designing targeted cancer therapy (16), we should know if the cancer progressed or regressed along with the increase in the expression level of the gene. Therefore, tools that can identify cancer changes from text would facilitate cancer research much further.

In this paper, we present a novel Web-based search engine, or OncoSearch, which identifies from text not only how genes change but also how cancers change. OncoSearch searches Medline abstracts for sentences that mention gene expression changes in cancers with queries that specify (i) whether a gene expression level is up-

*To whom correspondence should be addressed. Tel: +82 42 3503541; Fax: +82 42 3507841; Email: park@cs.kaist.ac.kr

regulated or down-regulated, (ii) whether a certain type of cancer progresses or regresses along with the gene expression change and (iii) the expected role of the gene in the cancer. The retrieved sentences are ranked by the confidence scores produced by text-mining modules and provided via three different views. Note that the system focuses on expression levels of genes among many other properties of genes, as expression levels are one of the most actively studied gene properties. OncoSearch is anticipated to enhance the understanding of the gene–cancer relations.

In what follows, we explain the three main components of OncoSearch: (i) a Web-based user query interface and a display tool to present search results, (ii) a text-mining process to identify gene expression changes and cancer changes from text and to store them into a database and (iii) a search process for sentences that describe gene–cancer relations as specified by a given query.

RESULTS

OncoSearch is unique in that the system identifies information about three query concepts, or gene expression change (GE), cancer change (CC) and expected gene class (GC), from a sentence. Given a sentence that describes gene expression changes in cancers, OncoSearch analyzes the sentence and annotates it with an appropriate type for each query concept. The relevant types for the three concepts are: {'up-regulated', 'down-regulated'} for GE, {'progression', 'regression'} for CC and {'biomarker', 'oncogene', 'tumor suppressor gene'} for GC. Table 1 shows the definitions of the query concept types, where cancerous properties of cells include self-sufficiency in growth signals, insensitivity to anti-growth signals, tissue invasion and metastasis, limitless replicative potential, sustained angiogenesis, evasion of apoptosis, tumor-promoting inflammation, destruction avoidance and deregulation of cellular energetics (17). When users submit a query, they select query concept types of their interest and OncoSearch will retrieve the sentences annotated with the selected types. Note that we adopted the query concepts and their types from the work by Lee and colleagues (18), who published an annotated corpus of gene–cancer relations.

In this section, we introduce example queries that show the biological significance of the tool and explain the Web interface of the system.

Example queries

Example queries include the following:

(a) genes that are likely to work as oncogenes or tumor suppressor genes in prostate cancer;
(b) literature reports about oncogenic activity of CTNNB1 across various types of cancers;
(c) genes that are up-regulated in breast cancer;
(d) literature reports about down-regulated expression of CDKN2A in lung cancer.

For query (a), OncoSearch returns 422 candidate genes and retrieves 1207 sentences that support the classification of the genes as oncogenes or tumor suppressor genes. The 422 retrieved genes include known oncogenes and tumor suppressor genes such as TP53, AKT1, HRAS and BCL10. Queries similar to (a) are useful when one wants to distinguish cancer-causing genes from the genes that show alterations since they are affected by the cancers. Such distinction is important for understanding the molecular mechanism of oncogenesis and designing new therapies (19).

For query (b), OncoSearch retrieves 109 sentences that support the classification of CTNNB1 as an oncogene. 93 of the sentences describe that an increased expression level of CTNNB1 induces further progression of 28 types of cancers, and the remaining 16 sentences describe that a decreased expression level of CTNNB1 leads to regression of nine types of cancers. Although CTNNB1 is not registered as an oncogene (or as a proto-oncogene) in one of the *de facto* standard gene databases, or UniProtKB (3), mutation in CTNNB1 is known to induce several types of cancers including colorectal cancer, medulloblastoma and ovarian cancer (http://www.ncbi.nlm.nih.gov/gene/1499). In 2013, the gene is designated as an oncogene by Vogelstein and colleagues (20). Note that all the 109 retrieved sentences are from 90 biomedical articles that are published before 2013.

For query (c), OncoSearch returns 1871 genes up-regulated in 55 sub-types of breast cancer. The 1871 retrieved genes include well-known biomarkers of breast cancer such as ERBB2, MYC, EGFR and VEGFA (21). For query (d), OncoSearch retrieves 134 sentences that describe that down-regulation of CDKN2A is associated to seven sub-types of lung cancer. Queries similar to example (d) can be used to validate the results of large-scale experiments on gene expressions such as microarray experiments.

Web interface

An OncoSearch input is composed of genes, cancers and query concept types. Given an input, OncoSearch retrieves sentences that describe expression changes of the queried genes in the queried cancers, where the expression changes, cancer changes and the expected roles of the queried genes are as specified by the input query concept types. For instance, if a user inputs 'breast cancer' and 'ABL1', selecting 'oncogene' for GC, 'up-regulated' for GE and 'progression' for CC, OncoSearch will retrieve six sentences that state that up-regulated ABL1 brought further progression of breast cancer, where the sentences support the classification of ABL1 as an oncogene. While it is mandatory to select at least one type per query concept, users can leave either genes or cancers not specified. When no gene is given, the system searches all known genes for the given cancer type or types. When no cancer type is given, the system searches all known cancer types for the given genes. The query concept CC is categorized as an advanced search option, since the test users selected both of the two CC types for most of the queries.

Given a query, OncoSearch outputs sentences annotated with a gene, a type of cancer and the respective types of GE, CC and GC. The system provides the sentences via three different views, or the Results, the Summary and the Graph views. The Results view shows all the retrieved sentences ranked by their confidence scores. For each sentence, the system provides a link to an Entrez Gene page of the anno-

Table 1. Definitions of query concept types

Type	Definition
Gene expression change (GE)	
Up-regulated	The expression of a gene is increased.
Down-regulated	The expression of a gene is decreased.
Cancer change (CC)	
Progression	The cell or tissue acquires cancerous properties as the gene expression level changes; some cancerous properties of the cell or tissue are strengthened as the gene expression level changes.
Regression	The cell or tissue loses some cancerous properties as the gene expression level changes; some cancerous properties of the cell or tissue are weakened as the gene expression level changes.
Expected gene class (GC)	
Oncogene	A gene that causes cells to acquire cancerous properties, or a gene that strengthens cancerous properties of cells.
Tumor suppressor gene	A gene that causes cells to lose cancerous properties, or a gene that weakens cancerous properties of cells.
Biomarker	A gene that is not identified as an oncogene or a tumor suppressor gene but shows an altered expression level in cells that show cancerous properties when compared to the expression level of the gene in normal cells. The term indicates not only those genes that affect cancers but also those that are affected by cancers.[a]

[a]Note that the usage of the term biomarker in this paper is different from its general usage, which refers to oncogenes and tumor suppressor genes as well.

tated gene and a link to a National Cancer Institute (NCI) thesaurus (22) page of the annotated cancer. When the gene is classified as either an oncogene or a tumor suppressor gene, links to UniProtKB (3), TSGene (4) and the Vogelstein cancer gene list (20) are provided if the gene is registered also as an oncogene or a tumor suppressor gene either in the databases or in the list. The Summary view shows groups of sentences, where the sentences with the same gene, the same cancer type and the same types of query concepts are grouped together. Using the Summary view, users can quickly browse through the retrieved results. In addition, users can select the kinds of annotated information that are of their interest and regroup the sentences. For example, a user may choose to regroup the sentences based only on the types of GE, and would get two groups of sentences, one with 'up-regulated' GE and the other with 'down-regulated' GE. The Graph view visualizes co-occurrences among the retrieved genes. A node in the co-occurrence network, which represents a gene, is colored red and green according to the numbers of sentences that describe up-regulation and down-regulation of the gene, respectively. The system also provides links to the abstracts in which pairs of genes are mentioned together. Last, the retrieved results can be downloaded as a simple text.

Statistics

We first retrieved from PubMed 3 222 366 cancer-related articles (as of 4 November 2013). 152 512 of the retrieved abstracts contain at least one sentence that describes gene expression changes in cancers. We further processed the abstracts and annotated 451 798 sentences with 7555 human genes and 1717 cancer types. Among the 7555 genes, 2295, 1549 and 6779 genes are inferred as candidates of oncogenes, tumor suppressor genes and biomarkers, respectively.

METHODS

Text-mining

We applied a text-mining process to Medline abstracts and built a database of sentences annotated with genes, cancers and the types of GE, CC and GC. We first downloaded cancer-related abstracts via PubMed with a query that consists of 15 cancer-referring terms that include 'cancer', 'tumor' and 'carcinoma' (the full list of the query terms is provided in the supplements page of OncoSearch Web site). Then, we located gene names by using BANNER (23) and normalized the gene names into Entrez Gene IDs by using Moara (24). We also identified cancer names by using a dictionary matching method, where the cancer dictionary consists of the 'synonyms' registered in NCI thesaurus and the lexical variations of the synonyms. After locating the gene names and the cancer names, we tokenized, POS tagged and parsed the sentences in the abstracts by using the Charniak–Johnson parser (25) with a biomedical parsing model (26). We converted the phrase structures produced by the parser into dependency structures by using the Stanford conversion tool (27). Finally, we identified mentions of gene expression changes by using Turku Event Extraction System (TEES) (28). The process above showed 73.47% precision; gene expression change mentions identified from 72 out of 98 randomly selected sentences were found correct. The test dataset is provided in the OncoSearch Web site. Note that for the articles registered to PubMed before 18 November 2008, we used a preprocessed dataset or EVEX (29), which is the result of applying the same process as described above to the articles.

After identifying gene expression change mentions and cancer names from the abstracts, we selected only the sentences that contain at least one cancer name and at least one mention of gene expression change. For each of such sentences, we identified the respective types of GE, CC and GC. When a sentence contains more than one mention of gene expression change or cancer name, we identified query concept types for each pair of cancer name and gene expression change mention. The types tb: GC Sensitivity of the three query concepts, GE, CC and GC, are identified as follows. First, the type of GE is deterministically identified from the event types provided by TEES. When the event type is 'positive_regulation' and 'negative_regulation', the type of GE is determined to be 'up-regulated' and 'down-regulated', respectively. Second, the type of CC is identified by a Max-

Table 2. Inference rules for Gene Class (GC) types

#	GE	CC	PT	GC
1	Up-regulated	Progression	Causality	Oncogene
2	Up-regulated	Regression	Causality	Tumor suppressor gene
3	Down-regulated	Progression	Causality	Tumor suppressor gene
4	Down-regulated	Regression	Causality	Oncogene
5	*	*	Observation	Biomarker

The asterisk denotes all the relevant types of the corresponding concept.

Table 3. Sensitivity of the gene class inference

Data source	# Registered	# Inferred	Sensitivity (%)
UniProtKB—oncogene	231	109	47.19
UniProtKB —TSG	163	70	42.94
Vogelstein—oncogene	74	45	60.81
Vogelstein—TSG	64	33	51.56
All—oncogene	301	150	49.83
All—TSG	226	102	45.13

The table shows, for each data source, the number of genes registered as oncogenes or tumor suppressor genes in the data source, the number of genes which are inferred by OncoSearch as the same classes as in the data source and the sensitivity of the gene class inference. Note that we counted only the human genes. 'all' data source represents the union of the genes in UniProtKB and the Vogelstein list. TSG is an abbreviation for 'tumor suppressor gene'.

Table 4. Comparison to other similar Web tools

Tool	Gene change	Gene–cancer relation	Gene class	Data source	Human curation	Gene co-occurrence
MeInfoText (12)	Methylation	Positive Negative	n/a	Abstract	No	n/a
PubMeth (13)	Methylation	Genes that are reported to be methylated in certain types of cancers	n/a	Abstract	Yes	n/a
BioContext (14)	Gene expression	Cancer cells and tissues as anatomical contexts of gene changes	n/a	Abstract	No	n/a
	Transcription Protein catabolism Localization Phosphorylation Binding Regulation Positive regulation Negative Regulation			Full-text		
DigSee (15)	Gene expression	Positive	n/a	Abstract	No	Interactive graph
	Transcription Protein catabolism Localization Phosphorylation Binding Regulation	Negative				
OncoSearch	Gene expression	Progression	Oncogene	Abstract	No	Interactive graph
		Regression	TSG			
		Unidentifiable	Biomarker			

The table shows, for each tool, (i) the supported types of gene changes, (ii) the supported types of gene–cancer relations, (iii) the supported types of gene classes, (iv) the data sources used, (v) whether the tool's database is manually examined by human experts or not and (vi) how the information on gene co-occurrence is provided. While MeInfoText, DigSee and OncoSearch label each gene–cancer pair with one of the predefined types, PubMeth and BioContext collect gene–cancer pairs that suit predefined criteria. TSG is an abbreviation for 'tumor suppressor gene'.

Ent classifier. We trained the classifier using a corpus provided by Lee and colleagues (18), or CoMAGC, since we adopted the query concepts from their work. The classifier achieved accuracies of 79.78% on 10-fold cross validation on CoMAGC and 73.03% on 152 randomly chosen test sentences, where accuracy is defined as the proportion of correctly classified results among the classification results of all test data. Last, the type of GC is identified by applying deterministic inference rules on top of the GE type, CC type and the type of an additional concept, or 'proposition type (PT)'. PT indicates whether the causality between the gene and the cancer is claimed in the sentence or not, and the type of PT can be either 'causality' or 'observation'. We identified the PT type by using another MaxEnt classifier, which

is also trained on CoMAGC. This second MaxEnt classifier achieved accuracies of 85.71% on 10-fold cross validation and 89.69% on random test sentences. The test sentences are provided in OncoSearch Web site. Features used by the two MaxEnt classifiers are as follows:

- **Surface tokens:** all tokens in the sentence;
- **Keywords:** gene name, cancer name and the event keyword of the gene expression change mention as provided by TEES;
- **Context words:** tokens before and after the keywords;
- **Paths on dependency parse trees:** tokens, dependency types and the [token-dependency_type-token] units in the paths that connect the event keyword and the cancer names;
- **Keywords order:** whether the gene name and the event keyword occur before or after the cancer names (used only for PT classification).

Table 2 shows the inference rules for GC types. The rules are also adopted from the work by Lee and colleagues. The rules state: (i) if increased expression level of a gene accompanies cancer progression and there is a causal relation between the change in gene expression and the cancer progression, then the gene is considered an 'oncogene'; (ii) if increased expression level of a gene accompanies cancer regression and the change in cancer is caused by the change in gene expression level, the gene is considered a 'tumor suppress gene'; (iii) if decreased expression level of a gene accompanies cancer regression and there is causality, the gene is considered an 'oncogene'; (iv) if decreased expression level of a gene accompanies cancer progression and there is causality, the gene is considered a 'tumor suppressor gene'; and (v) if change in gene expression level accompanies change in cancer but there is no evidence of causality between the two, the gene is considered a 'biomarker'.

Searching and ranking

Given a user query, OncoSearch searches the database for sentences that describe gene–cancer relations as specified by the query. Gene and cancer names in a query are normalized into Entrez Gene IDs and NCI thesaurus codes, respectively, by using dictionary matching methods. After normalization, NCI codes are expanded to include all of their subtypes. Each sentence is scored with the weighted harmonic mean of the confidence scores provided by TEES, the CC classifier and the PT classifier, with weights 0.5, 0.3 and 0.2, respectively. Sentences that are likely to describe hypothesis or study purpose are penalized. When a sentence contain expressions such as 'We investigated', 'To study' and 'Objective: ', the score of the sentence is multiplied by 0.5. Last, the sentence groups in the Summary view are scored and ranked according to the sum of the scores of the sentences in each group.

DISCUSSION

Given a sentence that describes expression change of a gene in a type of cancer, OncoSearch infers the gene's class based on the content of the sentence. However, we do not claim that the inferred gene classes are definite. Rather, one should interpret the inferred gene classes and corresponding sentences as textual evidence that supports hypothesis on the expected roles of the genes in the cancers. In order to fully understand how a gene functions regarding cancers, one should collect many pieces of such textual evidence and conclude based on the collected evidence.

Comparison of the inferred gene classes to the gene classes registered in other biology databases gives insight into how one should interpret the gene classes provided by OncoSearch. In particular, we measured how much of the cancer-related genes in biology databases are picked out by OncoSearch, or the sensitivity of the system, and how much of the genes inferred by OncoSearch are evidenced by biology databases, or the precision of the system. Table 3 shows the sensitivity of gene class inference when measured against the oncogenes and tumor suppressor genes registered in UniProtKB (we used the genes annotated with the keywords 'proto-oncogene (KW-0656)' and 'tumor suppressor (KW-0043)') (3) and the Vogelstein cancer genes list (20). We consider the sensitivity rates adequate, given the fact that only 6.87% (18/262) of the oncogenes and 3.76% (16/426) of the tumor suppressor genes in UniProtKB are designated as such in the Vogelstein list. On the other hand, the precision of gene class inference is low. Only 6.53% (150/2295) of the inferred oncogenes and 6.58% (102/1549) of the inferred tumor suppressor genes are validated with either UniProtKB or the Vogelstein list. We suspect that such low precision is due to the fact that OncoSearch infers oncogenes and tumor suppressor genes in a less restricted way than other biology databases. For example, TSGene (4), a repository of tumor suppressor genes with literature evidence, is built on 5795 Medline abstracts that explicitly mention 'tumor' and 'suppressor', contrary to OncoSearch, which does not require such explicit mentions for gene classification.

We did not compare the biomarkers inferred by the rules to the biomarkers registered in biology databases, since the meaning of the term as used in biology databases is different from the meaning as used in this paper. In biology databases, a biomarker refers to a molecule that can be used as an indicator of a normal or abnormal process, or of a condition or disease (30). In this context, oncogenes and tumor suppressor genes are often classified as biomarkers. Given the definition of biomarker in this paper, we regard the sentences in the OncoSearch database as providing sufficient evidence to validate the classification of corresponding genes into biomarkers.

While OncoSearch can identify genes that show altered expression levels in cancers and can classify the genes according to their expected roles in cancers, the system cannot identify genes that are related to cancers via other types of gene changes such as methylation. Since there are a number of Web tools that search biomedical literature for other types of gene changes, users may choose to use such tools in combination with OncoSearch. The characteristics of the tools, in comparison to OncoSearch, are summarized in Table 4.

Last, although the current version of OncoSearch can effectively search for cancer-related genes, we believe that the performance of the system can be improved further. First,

we can enhance the overall precision of the system by improving the performance of the text-mining modules such as TEES and the CCS/PT classifiers. We plan to devise a post-processing method for TEES to filter out false positive results. We also plan to apply semi-supervised learning or transfer learning techniques (31,32) to train CCS/PT classifiers, in order to overcome the fact that the size of the CoMAGC corpus is limited to about 800 sentences. Second, we can improve the overall sensitivity of the system by including other types of data. We plan to include full-texts of biomedical articles as well as abstracts, and to account for gene changes of other types such as mutation.

CONCLUSION

OncoSearch is a novel text mining search engine that searches Medline abstracts for sentences describing gene expression changes in cancers. The system is unique in that it allows a query specifying (i) whether a gene expression level is up-regulated or down-regulated, (ii) whether a type of cancer progresses or regresses along with such gene expression change and (iii) the expected role of the gene in the cancer. We anticipate that OncoSearch will be used to further enhance the understanding of the gene–cancer relations.

FUNDING

National Research Foundation of Korea [MEST, 20110029447]. Open access funding charge: National Research Foundation of Korea [MEST, 20110029447].
Conflict of interest statement. None declared.

REFERENCES

1. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **33**(Suppl. 1), D54–D58.
2. Safran,M., Dalah,I., Alexander,J., Rosen,N., Iny Stein,T., Shmoish,M., Nativ,N., Bahir,I., Doniger,T., Krug,H. *et al.* (2005) GeneCards Version 3: the human gene integrator. *Database*, **2010**, baq020.
3. Magrane,M. and Consortium,U. (2005) UniProt Knowledgebase: a hub of integrated protein data. *Database*, **2011**, bar009.
4. Zhao,M., Sun,J. and Zhao,Z. (2005) TSGene: a web resource for tumor suppressor genes. *Nucleic Acids Res.*, **41**, D970–D976.
5. Essack,M., Radovanovic,A., Schaefer,U., Schmeier,S., Seshadri,S., Christoffels,A., Kaur,M. and Bajic,V. (2005) DDEC: Dragon database of genes implicated in esophageal cancer. *BMC Cancer*, **9**, 219–225.
6. Maqungo,M., Kaur,M., Kwofie,S.K., Radovanovic,A., Schaefer,U., Schmeier,S., Oppon,E., Christoffels,A. and Bajic,V.B. (2005) DDPC: Dragon database of genes associated with prostate cancer. *Nucleic Acids Res.*, **39**(Suppl. 1), D980–D985.
7. Kaur,M., Radovanovic,A., Essack,M., Schaefer,U., Maqungo,M., Kibler,T., Schmeier,S., Christoffels,A., Narasimhan,K., Choolani,M. *et al.* (2005) Database for exploration of functional context of genes implicated in ovarian cancer. *Nucleic Acids Res.*, **37**(Suppl. 1), D820–D823.
8. Agarwal,S.M., Raghav,D., Singh,H. and Raghava,G. (2005) CCDB: a curated database of genes involved in cervix cancer. *Nucleic Acids Res.*, **39**(Suppl. 1), D975–D979.
9. Cheng,D., Knox,C., Young,N., Stothard,P., Damaraju,S. and Wishart,D.S. (2005) PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res.*, **36**(Suppl. 2), W399–W405.
10. Tsuruoka,Y., Tsujii,J. and Ananiadou,S. (2005) FACTA: a text search engine for finding associated biomedical concepts. *Bioinformatics*, **24**, 2559–2560.
11. Fontaine,J.-F., Priller,F., Barbosa-Silva,A. and Andrade-Navarro,M.A. (2005) Génie: literature-based gene prioritization at multi genomic scale. *Nucleic Acids Res.*, **39**(Suppl. 2), W455–W461.
12. Fang,Y.-C., Lai,P.-T., Dai,H.-J. and Hsu,W.-L. (2005) MeInfoText 2.0: gene methylation and cancer relation extraction from biomedical literature. *BMC Bioinformatics*, **12**, 471–478.
13. Ongenaert,M., Van Neste,L., De Meyer,T., Menschaert,G., Bekaert,S. and Van Criekinge,W. (2005) PubMeth: a cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Res.*, **36**(Suppl. 1), D842–D846.
14. Gerner,M., Sarafraz,F., Bergman,C.M. and Nenadic,G. (2005) BioContext: an integrated text mining system for large-scale extraction and contextualization of biomolecular events. *Bioinformatics*, **28**, 2154–2161.
15. Kim,J., So,S., Lee,H.-J., Park,J.C., Kim,J.-j. and Lee,H. (2005) DigSee: disease gene search engine with evidence sentences (version cancer). *Nucleic Acids Res.*, **41**, W510–W517.
16. Luo,J., Solimini,N.L. and Elledge,S.J. (2005) Principles of cancer therapy: oncogene and non-oncogene addiction. *Cell*, **136**, 823–837.
17. Hanahan,D. and Weinberg,R.A. (2005) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.
18. Lee,H.-J., Shim,S.-H., Song,M.-R., Lee,H. and Park,J. (2005) CoMAGC: a corpus with multi-faceted annotations of gene-cancer relations. *BMC Bioinformatics*, **14**, 323–339.
19. Haber,D.A. and Settleman,J. (2005) Cancer: drivers and passengers. *Nature*, **446**, 145–146.
20. Vogelstein,B., Papadopoulos,N., Velculescu,V.E., Zhou,S., Diaz,L.A. and Kinzler,K.W. (2005) Cancer genome landscapes. *Science*, **339**, 1546–1558.
21. Ross,J.S., Linette,G.P., Stec,J., Clark,E., Ayers,M., Leschly,N., Symmans,W.F., Hortobagyi,G.N. and Pusztai,L. (2005) Breast cancer biomarkers and molecular medicine: part II. *Expert Rev. Mol. Diagn.*, **4**, 169–188.
22. Sioutos,N., de Coronado,S., Haber,M.W., Hartel,F.W., Shaiu,W.-L. and Wright,L.W. (2005) NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J. Biomed. Informatics*, **40**, 30–43.
23. Leaman,R. and Gonalez,G. (2005) BANNER: an executable survey of advances in biomedical named entity recognition. In: *Proceedings of the Pacific Symposium on Biocomputing,*pp. 652–663.
24. Neves,M., Carazo,J.-M. and Pascual-Montano,A. (2005) Moara: a Java library for extracting and normalizing gene and protein mentions. *BMC Bioinformatics*, **11**, 157–169.
25. Charniak,E. and Johnson,M. (2005) Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In: *Proceedings of the 43rd ACL Association for Computational Linguistics,* pp. 173–180.
26. McClosky,D. (2005) Any domain parsing: automatic domain adaptation for natural language parsing. *PhD Thesis*, Brown University, Department of Computer Science.
27. De Marneffe,M.C., MacCartney,B. and Manning,C.D. (2005) Generating typed dependency parses from phrase structure parses. In: *Proceedings of the LREC 2006*, pp. 449–454.
28. Björne,J., Ginter,F., Heimonen,J., Airola,A., Pahikkala,T. and Salakoski,T. (2005) Extracting complex biological events with rich graph-based features sets. In: *Proceedings of the BioNLP'09 Shared Task on Event Extraction Association for Computational Linguistics*, pp. 10–18.
29. Van Landeghem,S., Hakala,K., Rnnqvist,S., Salakoski,T., Van de Peer,Y. and Ginter,F. (2005) Exploring biomolecular literature with EVEX: connecting genes through events, homology and indirect associations. *Adv. Bioinformatics*, **2012**, 582765.
30. Mishra,A. and Verma,M. (2005) Cancer biomarkers: are we ready for the prime time? *Cancers*, **2**, 190–208.
31. Chapelle,O., Schölkopf,B. and Zien,A. (eds.) (2005) *Semi-supervised Learning*, MIT Press, Cambridge, One Rogers Street, Cambridge, MA.
32. Xu,Q. and Yang,Q. (2005) A survey of transfer and multitask learning in bioinformatics. *JCSE*, **5**, 257–268 .