Searching Through Scientific PDF Files Supported by Bi-clustering of Key Terms Matrices

Rafal Łancucki(✉), Pawel Foszner, and Andrzej Polanski

Institute of Informatics, Silesian University of Technology, Gliwice, Poland
lancucki@onet.pl

Abstract. We describe an original approach for exploring corpora of pdf format scientific texts in the area of bio-medical research, created over a wide topic of interest, e.g., cancer, thyroid cancer, biological process etc. Our methodology is based on indexing large lists of appropriate key-terms and additionally performing bi-clustering of term occurrence matrices. In our approach the position of phrase inside text (abstract or text) is not considered, but we include statistics based on occurrences frequency. We treat documents as a bags of words and the results are processed toward unique list of values. Bi-clustering is used to achieve separating character of lists of key-terms, characterizing sub-types of the studied category, e.g., different cancers or different sub-classes of a given cancer. We prove usefulness of the algorithm by searching for lists of genes characteristic for cancer types.

Keywords: Rule-based · Text mining · Bi-clustering

1 Introduction

Text searching is a constantly growing area of the data science, due to large fields of applications [16]. In this study we are interested in searching through scientific papers, an important sub-area of text searching. Browsing through scientific texts is particularly important in bio-medical applications, where arranging and ordering knowledge coming from experimental and in-silico researches has significant implications e.g., to development of diagnostic tools, therapies, personalized medical protocols etc.

There are numerous publications devoted to examples of useful strategies for bio-medical text searches [12,13,17,19,23,26]. Scientific text searching applications clearly are constantly contested by general purposes browsers, with its most prominent and competitive Internet search engine Google™. Formulating many specialized text searching tasks make, nevertheless, the field of task-specific scientific texts search engines an interesting and viable research area.

We faced a problem of retrieving gene statistics from publications selected by different cancer types. We tried to find specific cancer-type and gene correlation to improve machine-learning results.

A. Gruca et al. (eds.), *Man-Machine Interactions 5*,
Advances in Intelligent Systems and Computing 659, DOI 10.1007/978-3-319-67792-7_15

In this study we propose an approach for exploring large corpora of pdf format scientific texts bases on indexing large lists of key-terms and additionally supported by bi-clustering of occurrence matrices. In the text mining we used rule-based approach of full body documents. For this research the position of phrase inside text (abstract or text) is not considered, however text inside references part of the document is ignored. We treat documents as a bags of words and the results are processed toward unique list of values. Those lists are used then as an input for bi-clustering to improve the quality.

2 Related Work

The biomedical literature text mining seems to be very actively explored area [14, 15, 20]. Many different techniques are used to extract information from document corpora. Web-based access and more general approach (not focused only on one type of the research) is often explored [2, 3, 18]. Because collecting document corpora is not an easy task some solutions only focus on the retrieving data [1, 25]. From longer time discussion whether only abstract (i.e. [18]) or full body texts (i.e. [4, 6]) should be used was started. Abstracts seems to provide more accurate results but also important details could be easily skipped. Growing number of the rule-based natural language processing systems [2, 3, 8, 10, 11] shows potential of this approach. Recent work is concentrated on the semantic analysis of the text [21, 22]. However our combination of the text mining using rule-based approach, statistical analysis and the bi-clustering seems to be unique.

3 Data

Real data set was taken from TCGA database [24] (The Cancer Genome Atlas). We have retrieved all variant calling format (VCF) files corresponding to patients with one of three types of cancer—head & neck, prostate and thyroid. For our analyses we pick only VCF files which contain gene sequence variations annotated with MuTect [5]. For each patient, we looked at these source files for somatic mutations. The final data set was composed of 3 patient groups (according to cancer type). Such data matrix consist of gene—patient information where rows representing genes and columns representing patients. Data showing the relationship of these two dimensions is the number of somatic mutations of a given gene in a given patient genome. As the following statistics show—the patient classes in the selected set are very well balanced:

- Head & neck: 510 patients
- Prostate: 505 patients
- Thyroid: 504 patients

The total number of genes relevant for the above tumors is 43754. Mutation numbers for a single gene vary from 0 to 2130

The analysis was divided into three stages (1) text mining, (2) bi-clustering and at the end (3) classification. Only the rows (genes) of data matrix were given at the input of the first stage, where they constituted a reference gene collection to the text search. The second stage needed complete data in order to find valid bi-clusters. And the last classification stage work only on conclusions (gene signatures) produced by previous stages.

4 Text Mining

Text mining was used to produce the distinct gene list for each cancer type. Complete algorithm workflow is presented on the Fig. 1.

4.1 Collecting Source Documents

The source document corpus was collected using the PubMed database. For the purpose of this research document corpora were collected, using 3 different search phrases:

- Thyroid cancer
- Prostate cancer
- Head and neck cancer

The same list was later used during text-mining process for scanning engine. All publicly available pdf documents were downloaded and used for the text mining. The text corpora contained 47311 documents for the prostate cancer, 55153 for head and neck cancer and 31243 for thyroid cancer. Documents were downloaded from ftp server of PubMed using specialized software. This software is capable to run PubMed query, fetch the results and download available publicly documents.

4.2 Gene List

The gene list is prepared basing on the "The Cancer Genome Atlas" source. From patients data with one of disease considered in this research all genes with somatic mutations were collected. Joint list of genes was used as an input for the text mining. The list contained 36.117 genes.

4.3 Scanning Document Corpus

For this task 2 list were used. One list containing all diseases names and second one with the narrowed gene list. For each document corpus the same cancer type and gene lists were used. Scanning algorithm is capable to detect when the cancer type name or gene is inside bibliography/references area. Such results are eliminated to reduce the noise in the source data. Scanning engine uses iTextSharp library to access to the source document and at the end of the process

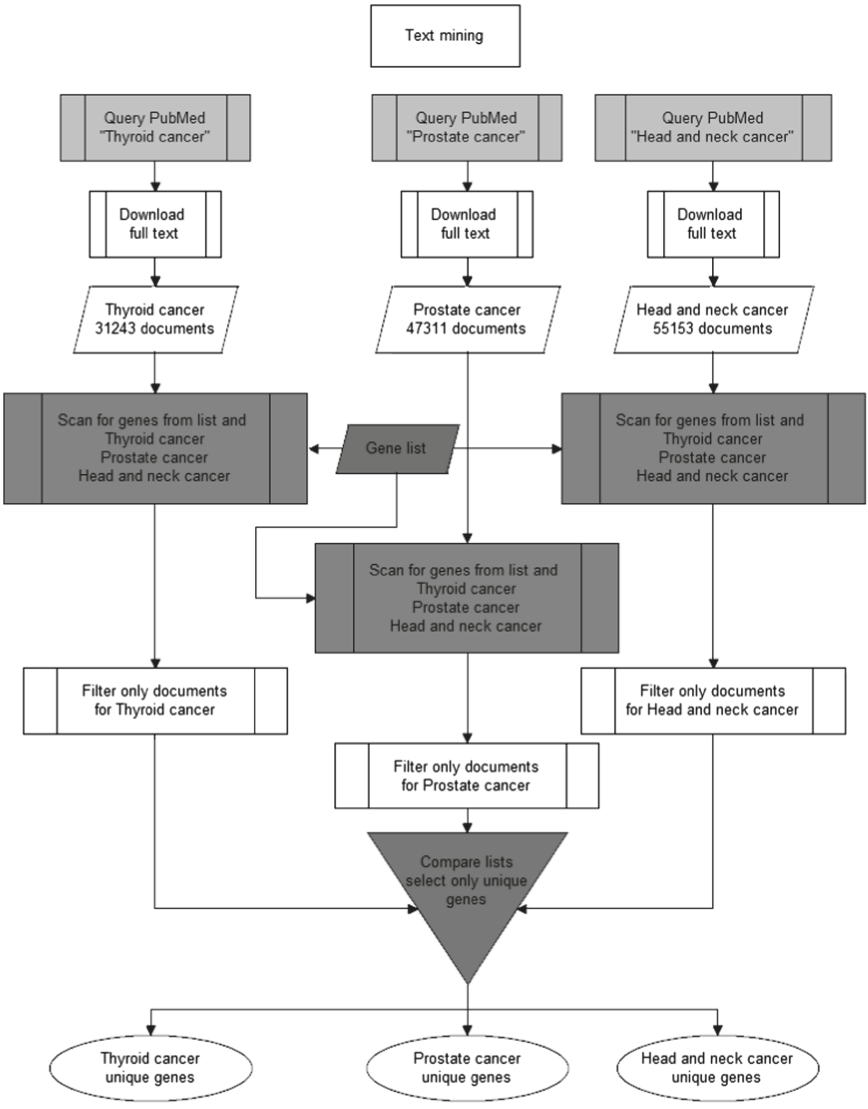

Fig. 1. Algorithm workflow

exports results into .csv file. Scanning results are later processed in a way that only documents describing specific cancer type are taken into account. This means document can contain nothing but only one specific cancer type and no other cancer types (See Tables 1 and 2).

After this filtering document corpus contains only documents describing one type of disease and at least one gene from the gene list. In the next step only documents with significant amount of cancer type (ten or more occurrences) in

Table 1. Sample of the processed document for thyroid cancer

File name	Disease	Occurrences	Gene	Gene occurrences
PMC4780491.pdf	thyroid cancer	92	TYK2	1
PMC4780491.pdf	thyroid cancer	92	STAT6	1
PMC4780491.pdf	thyroid cancer	92	STAT3	19

Table 2. The sample of the ignored document for thyroid cancer

File name	Disease	Occurrences	Gene	Gene occurrences
PMC4156403.pdf	thyroid cancer	87	HR	13
PMC4156403.pdf	thyroid cancer	87	MAG	3
PMC4156403.pdf	prostate cancer	2	HR	13
PMC4156403.pdf	prostate cancer	2	MAG	3

Table 1 Associations of levels of IL-1Ra in cancer

Elevated levels associated with greater disease severity
Postsurgical sepsis [75]
Survival ovarian arcinoma
Chronic fatigue in breast cancer [76]
Tumor extent in bone sarcoma [77]
Pelvic metastasis in cervical cancer [78]
Tumor load in childhood leukemia [79]
Malignant histiocytosis [80]
Hairy cell leukemia [81]
Tumor size and metastases and colorectal cancer [82–84]
Testicular cancer-related fatigue [85]
Thyroid cancer [86]
Pancreatic cancer [87]
Estrogen receptor breast cancer [43]
Elevated levels associated with lesser disease severity
Pancreatic carcinoma [88]
[a]Colorectal carcinoma [46]
Metastatic gastric cancer [89]
[c]Ovarian carcinoma [90]
[d]Lung carcinoma [44]
Low levels associated with greater disease severity

Fig. 2. Sample of ignored document

text are selected for further processing. This is required to eliminate documents where cancer type is used only in comparison tables or just referenced. We had to add this limitation after reviewing several documents found in the first generation of results. We found out after manual reviewing that up to few references of the cancer type could appear inside document referring different type. It applies mostly to 2 different kind of documents. First group are documents comparing different cancer types. The second group describes other cancer type but it also compares influence of some factor to the other cancer types. Example of ignored article—[7]—part containing references to cancer types is shown on the Fig. 2. In this particular research we focused on minimizing impact of potential wrong input values. The output is list of gene for each of cancer types from cancer type list.

Because lists are not unique (many genes appears on list for each disease) next step in preparing paths is to eliminate duplicated genes from the result set (See Table 3).

Table 3. Results after filtering duplicated values

Thyroid cancer	Prostate cancer	Head and neck cancer
BRAF	AR	*EGFR*
RET	PC	HR
EGFR	HR	EGF
GAPDH	GAPDH	NHS
HR	PTEN	TP53
PAX8	*EGFR*	GAPDH
PIK3CA	ERG	CP
KRAS	EGF	PC
TPO	SDS	CD44

The final result from text mining contains only distinct values divided per cancer type. We identified 168 genes for head and neck cancer, 1013 for prostate cancer and 137 for thyroid cancer. Genes are unique for each cancer type in a way described above. Those list of values were used for bi-clustering as an input.

5 Bi-clustering and Classification

5.1 Bi-clustering

In order to elevate the results obtained by text mining—additional signatures were produced using structural bi-clustering method [9]. The selected method, unlike other bi-cluster methods, is a semi-supervised method. The dimension in which knowledge about classes is introduced is in our case the dimension

representing patients. This will make full use of the potential and information contained in the source data, which should result in better bi-clusters at the output. As the input to algorithm we provide data described in Sect. 3. We expect to receive bi-clusters consisting of sub-groups of genes having similar mutation pattern for the subgroup of patients with a certain type of cancer each. In such a case supervised dimension will be a set of patients and latent knowledge witch will be extracted is the set of genes.

Bi-clustering approach to finding gene signatures is vulnerable to random initial conditions. For this reason, a number of signatures were generated, each time using different initial conditions. For further analysis following signature were selected:

1. Gene signature obtained by text mining approach (Sect. 2)
   - Size of that signature were equal to 1318 genes
2. 80 signatures coming from bi-clustering approach
   - Average size: 264
   - Min size: 222
   - Max size: 301
3. 80 signatures obtain by union of (1) and (2)
   - Average size: 1565
   - Min size: 1522
   - Max size: 1617
4. 80 signatures obtain by join of (1) and (2)
   - Average size: 17
   - Min size: 2
   - Max size: 31

Size of all attributes in input data is 43754.

5.2 Classification

Classification is a data mining and artificial intelligence technique which is widely used in a support of medical diagnosis. The technique is a supervised machine learning which allows for precise goal definition. In this case specific cancer diagnosis. Improving the quality of this technique should validate the resulting signatures (in other words, the gene signatures will differentiate the selected patient groups).

As a point of reference to the classification results with obtained signatures was input matrix without any filtration. A small number of classification algorithms guaranteed finite calculation time for data on so many dimensions. We choose Naive Bayes approach, random forests, random trees and nearest neighbors algorithm with $k = 3$ and 1. Other parameters of mentioned methods were left unchanged from WEKA defaults (release version 3.6). Classification with signatures were performed on the same set of algorithms. In summary—we managed to carry 1,210 classification tasks (80 signatures with 3 different versions + text mining signature + row data benchmark all performed on on 5 different classification algorithms).

The best classification rate for row data were obtained for random forests, and it was: 63,3%. The best results for our signatures (version calculated on union approach (3)) generates classification rate equal to 73,31% (kNN). All 1205 results fitted into normal distribution generate Gaussian bell with mean 66,82% and standard deviation equal to 4,34%. Summary of all results with and without signatures is shown in Table 4. The results show that the obtained gene signatures differentiate the patient groups.

Table 4. Results that shows classification with and without obtained signatures

	kNN k = 1	kNN k = 3	Naive Bayes	Random forests	Random trees
All genes	61,23%	63,12%	62,34%	63,3%	60,75%
Only signatures	72,43%	73,31%	69,27%	69,54%	71,49%

6 Conclusions

In this paper we explore combination of text-mining and bi-clustering in the area of cancer data analysis. We proved that combination of text-mining results and bi-clustering could improve the results of classification. And by significantly reducing the dimensionality of data open it up to previously inaccessible (due to large time and memory consumption) computing algorithms. In this study we achieved 10% result improvement comparing to the classification on the same input data (but with whole range of attributes). Which means that reduced signatures where significant from the point of view of the analyzed data and removed attributes was only introducing noise.

Conducted experiments and the results obtained with them allow us to assume that the proposed methods are correct. To fully assert this assumption further work should focus on testing a broader spectrum of data and extension of text mining stage for deeper analysis of the text.

Acknowledgements. This work has been supported by the following grants: Silesian University of Technology, Institute of Informatics, Statute project, (BK/RAU-2/2016) to PF, National Science Centre, OPUS grant (2016/21/B/ST6/02153) to RL and AP.

References

1. Agarwal, S., Yu, H.: Figure summarizer browser extensions for PubMed Central. Bioinformatics **27**(12), 1723–1724 (2011)
2. Barbosa-Silva, A., Fontaine, J.F., Donnard, E.R., Stussi, F., Ortega, J.M., Andrade-Navarro, M.A.: PESCADOR, a web-based tool to assist text-mining of biointeractions extracted from pubmed queries. BMC Bioinform. **12**(1), 435 (2011)
3. Barbosa-Silva, A., Soldatos, T.G., Magalhães, I.L., Pavlopoulos, G.A., Fontaine, J.F., Andrade-Navarro, M.A., Schneider, R., Ortega, J.M.: Laitor-literature assistant for identification of terms co-occurrences and relationships. BMC Bioinform. **11**(1), 70 (2010)

4. Chiang, J.H., Shin, J.W., Liu, H.H., Chin, C.L.: Genelibrarian: an effective gene-information summarization and visualization system. BMC Bioinform. **7**(1), 392 (2006)
5. Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., Getz, G.: Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. **31**(3), 213–219 (2013)
6. Cohen, K.B., Johnson, H.L., Verspoor, K., Roeder, C., Hunter, L.E.: The structural and content aspects of abstracts versus bodies of full text journal articles are different. BMC Bioinform. **11**(1), 492 (2010)
7. Dinarello, C.A.: Why not treat human cancer with interleukin-1 blockade? Cancer Metastasis Rev. **29**(2), 317–329 (2010)
8. Divoli, A., Attwood, T.K.: Bioie: extracting informative sentences from the biomedical literature. Bioinformatics **21**(9), 2138–2139 (2005)
9. Foszner, P., Polanski, A.: Structured bi-clusters algorithm for classification of dna microarray data. In: ITBI 2016, pp. 161–171. Springer, Kamien Slaski (2016)
10. Fundel, K., Küffner, R., Zimmer, R.: Relex-relation extraction using dependency parse trees. Bioinformatics **23**(3), 365–371 (2007)
11. Hur, J., Schuyler, A.D., Feldman, E.L., et al.: Sciminer: web-based literature mining tool for target identification and functional enrichment analysis. Bioinformatics **25**(6), 838–840 (2009)
12. Jenssen, T.K., Lægreid, A., Komorowski, J., Hovig, E.: A literature network of human genes for high-throughput analysis of gene expression. Nat. Genet. **28**(1), 21–28 (2001)
13. Keerrthega, M., Thenmozhi, D.: Identifying disease-treatment relations using machine learning approach. Procedia Comput. Sci. **87**, 306–315 (2016)
14. Krallinger, M., Leitner, F., Valencia, A.: Analysis of biological processes and diseases using text mining approaches. Bioinform. Methods Clin. Res. 341–382 (2010)
15. Krallinger, M., Valencia, A., Hirschman, L.: Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome Biol. **9**(2), S8 (2008)
16. Marciniak, M., Mykowiecka, A. (eds.): Aspects of Natural Language Processing. Lecture Notes in Computer Science, vol. 5070. Springer, Heidelberg (2009)
17. Matos, S., Campos, D., Pinho, R., Silva, R.M., Mort, M., Cooper, D.N., Oliveira, J.L.: Mining clinical attributes of genomic variants through assisted literature curation in egas. Database **2016**, baw096 (2016)
18. Papanikolaou, N., Pavlopoulos, G.A., Pafilis, E., Theodosiou, T., Schneider, R., Satagopam, V.P., Ouzounis, C.A., Eliopoulos, A.G., Promponas, V.J., Iliopoulos, I.: Biotextquest+: a knowledge integration platform for literature mining and concept discovery. Bioinformatics **30**(22), 3249–3256 (2014)
19. Ravikumar, K.E., Wagholikar, K.B., Li, D., Kocher, J.P., Liu, H.: Text mining facilitates database curation-extraction of mutation-disease associations from biomedical literature. BMC Bioinform. **16**(1), 185 (2015)
20. Rodriguez-Esteban, R.: Biomedical text mining and its applications. PLoS Comput. Biol. **5**(12), e1000,597 (2009)
21. Šarić, J., Jensen, L.J., Ouzounova, R., Rojas, I., Bork, P.: Extraction of regulatory gene/protein networks from medline. Bioinformatics **22**(6), 645–650 (2006)
22. Taha, K., Yoo, P.D.: Predicting the functions of a protein from its ability to associate with other molecules. BMC Bioinform. **17**(1), 34 (2016)

23. Verspoor, K.M., Heo, G.E., Kang, K.Y., Song, M.: Establishing a baseline for literature mining human genetic variants and their relationships to disease cohorts. BMC Med. Inform. Decis. Mak. **16**(1), 68 (2016)
24. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., Cancer Genome Atlas Research Network: The cancer genome atlas pan-cancer analysis project. Nat. Genet. **45**(10), 1113–1120 (2013)
25. Yu, H., Hatzivassiloglou, V., Friedman, C., Rzhetsky, A., Wilbur, W.J.: Automatic extraction of gene and protein synonyms from medline and journal articles. In: Proceedings of the AMIA Symposium, pp. 919–923. American Medical Informatics Association, San Antonio, TX (2002)
26. Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., Vongsangnak, W., Shen, B.: Biomedical text mining and its applications in cancer research. J. Biomed. Inform. **46**(2), 200–211 (2013)