Discovery of Gene-disease Associations from Biomedical Texts

Wen-Juan Hou*, Bo-Yuan Kuo

Department of Computer Science and Information Engineering, National Taiwan Normal University, Taiwan

Copyright © 2016 by authors, all rights reserved. Authors agree that this article remains permanently open access under the terms of the Creative Commons Attribution License 4.0 International License

Abstract Due to the ever-expanding growth of biomedical publications, biologists have to retrieve up-to-date information from vast literatures to ensure they do not neglect certain significant publications. It becomes more and more important to deal with the extraction problem from the biomedical texts in an automatic way. The paper focuses on automatically identifying the relationships between human genetic diseases and genes from the biomedical literatures. The experimental data is retrieved from Mendelian Inheritance in Man (MIM) literatures of morbid in Online Mendelian Inheritance in Man (OMIM) database. We propose a hybrid method combining the rule learning and the statistical techniques. To collect the corpus used in the research, the first step is to find the sentences that include both the related human genetic diseases and genes mentioned from the morbid file, and they are regarded as the correct sentences. In the second step, the sentences that neither have the related human genetic diseases nor the genes mentioned from the morbid file are randomly selected, and they are regarded as the incorrect sentences. Next, the Memory-Based Shallow Parser is utilized to analyze these sentences to get some information in order to find rules in the following step. Then, some learning rules are obtained with a rule learner, ALEPH system. These generated rules are applied to catch the pairs of human genetic diseases and genes within one sentence. In the following, the study proposes a statistical approach, called Z-score method, to determine whether the pairs are valid or not. Finally, the experiments are made with considering some constraints and different numbers of rules. Furthermore, the evaluation metrics in the experiments are precision, recall rates, and F-scores.

Keywords Gene-disease Association, Biomedical Text Mining, Statistical Method, Rule Learning

1. Introduction

At present, there has been a fast-increasing amount of biomedical publications spreading in the World Wide Web.

For example, up to now, more than 21 million biomedical articles are available in MEDLINE.¹ For such an expanding-growing rate of the literatures, if biologists have to manually read through all the retrieved texts to find the information they actually need, it will become very difficult to keep up with the new findings. As a consequence, a substantial attention of research interest focuses on developing methods for automatically processing the biological and medical scientific literature – a process often referred to as biomedical text mining. Currently predicting new gene-disease associations has long been a principal goal in computational biology [1]. The goal of this paper is to solve the above issue, i.e., discovering the candidate genes likely to be associated with the specific diseases.

Recent studies have proposed several approaches to investigating the relationships between genes and diseases. Some previous studies use protein-protein interactions to predict gene-disease relationships [2, 3, 4]. Some researchers compute the similarity values between genes and diseases based on Gene Ontology (GO) [5] or Disease Ontology (DO) [6] terms [7, 8, 9]. Other controlled vocabularies such as MeSH [10] have already been utilized for linking proteins to disease terminologies [11, 12, 13]. Some other information like gene expressions [14, 15, 16], protein/genome sequences [17, 18, 19] and positional information [20] are also served as the important evidences to relating genes and diseases. Besides, there are some researchers who utilize the text mining techniques to extract gene-disease associations from the biomedical literatures [13, 21, 22]. Machine learning methods are applied by some researchers [1, 23, 24]. The network-based approaches to analyzing relationships between genes and diseases are proposed in [25, 26, 27]. These works demonstrate that associating genes with diseases is an active area of researches as it can lead to better understanding of diseases and it can reduce both time and expenditure in developing effective drugs and treatment.

The experimental data in this paper is obtained from Online Mendelian Inheritance in Man (OMIM) [28] and MEDLINE. OMIM is one of the most well-known databases

¹ <http://www.nlm.nih.gov/pubs/factsheets/medline.html>

that contain gene-disease annotations. The full-text overviews in OMIM contain information on all known mendelian disorders and over 12,000 genes. It is a comprehensive knowledge base of human genes and genetic diseases. Nowadays, one of the most comprehensive textual sources of biomedical information is MEDLINE. It includes more than 19 million citations from more than 7,300 different publications dating from 1966, and it continues updating weekly by U.S. National Library of Medicine (NLM). It is valuable for the biologists to mine information from the MEDLINE documents.

We observed that some researches also investigate extracting biological relations of interest rather than focusing on gene-disease relationships. Most of them apply the following four different kinds of approaches. First, one method utilizes the manually generated template-based/rule-based approaches which apply patterns generated by domain experts to extract concepts connected by a specific relation from text [29, 30, 31]. Secondly, other studies employ an automatic template/rule-learning method which creates similar templates/rules automatically by generalizing patterns/rules from text surrounding concept pairs known to have the relationship of interest [32, 33, 34]. Thirdly, the statistical method is proposed that identifies relationships by looking for concepts that are found with each other more often than would be predicted by chance [35]. Finally, Natural Language Processing-based (NLP-based) methods try to parse the text in order to generate the structure from which relationships can be extracted [36, 37, 38]. Among these approaches, we are interested in combining rule learning and statistical methods to help with relationship extraction to genes and diseases. We know that rules have several advantages, including that they are easy for humans to interpret, represent knowledge modularly and can be applied using tractable inference procedures. From the rule learning technique, we can obtain the rules relevant to gene-disease associations automatically. Furthermore, this study introduces the statistical method to improve the results returned by the rule learning technique.

2. Architecture Overview

Figure 1 illustrates the overall architecture of our approach to identifying gene-disease relationships. First, we preprocess each document in the corpus. In this phase, we make use of the OMIM database which is a popular resource in the biomedical domain. Next, we select a set of positive sentences and a set of negative sentences for the rule learning purpose. We refer to the morbid data from the OMIM which records some correct relationships between genes and diseases. Therefore, the morbid data can be served as a gold standard for evaluation. In the following, a memory-based shallow parser (MBSP) is adopted to produce the tagged sentences with the parsing information.² Then a rule learner, the ALEPH system, is used to learn the relationships

between genes and diseases so that a set of rules is obtained.³ For the test documents, we first preprocess the full texts to identify the genes and diseases appearing in the documents. Then the rules learned from the previous steps are used to generate the candidate associative gene-disease pairs. After that, we apply the statistical approach, Z-score method, to filter out some candidate gene-disease pairs. Finally, gene-disease associations in the corpus are extracted.

3. Experimental Data

3.1. Training Data

As mentioned before, the morbid text records some correct relationships between genes and diseases. From the OMIM website, we refer to the phenotype and gene MIM (Mendelian Inheritance in Man) number in the morbid text, and then download 302 MEDLINE documents. With this method, there are 2,532 sentences that contain both disease and gene names. For avoiding the bias in the learning phase, we have to select the positive and negative sentences with the equal number. Hence 1,000 sentences with correct gene-disease associations are selected to be the positive examples for learning, and 1,000 sentences with incorrect gene-disease pairs are chosen to be the negative examples.

3.2. Testing Data

To do outside test, the testing data must be different with the training data in the experiment. From the morbid text, we refer to 108 phenotype and gene MIM numbers, and then retrieve 108 MEDLINE full texts because there are 919 sentences with correct gene-disease associations and this number is similar to the sentence number 1,000 for learning.

4. Methods

The proposed methods are divided into six steps: (1) preprocessing the corpus, (2) selecting correct and incorrect sentences for learning, (3) parsing the sentences, (4) rule learning, (5) making candidate gene-disease associations using the learned rules, and (6) using the statistical method to filter out the candidate pairs.

4.1. Corpus Preprocessing

Our preprocessing procedure for each document consists of gene name tagging and disease name tagging. Since the OMIM database contains lots of gene-disease annotations, we extract the gene name list and disease name list from the OMIM. Finally, there are 13,943 genes and 5,592 disease names in the lists, respectively.

² <http://www.clips.ua.ac.be/pages/MBSP>

³ <http://www.cs.ox.ac.uk/activities/machlearn/Aleph/aleph.html>

4.2. Correct and Incorrect Sentence Selection for Learning

With the morbid data, the relevant diseases and genes can be annotated in the corpus. As mentioned in Section 3.1, two parts for rule learning are found. The first part contains 1,000 positive sentences where the associated genes and diseases appear in the same sentence. The second part includes 1,000 negative sentences where the gene-disease pairs of appearing in the same sentence do not match the gene-disease pairs annotated by the morbid data.

4.3. Parsing the Sentences

Before parsing the sentences, we clear the gene and disease tags of the sentences made in the previous stage and get 2,000 sentences with no tags. Then we make use of the memory-based shallow parser (MBSP) to do the parsing. Taking the sentence “HR54 and RAD54L may cause breast cancer.” as an example, the parsing result is shown in Figure 2. The explanation for MBSP tags is shown in Table 1.

Figure 1. System architecture

```

HR54/NNS/I-NP/O/NP-SBJ-1/O/hr54
and/CC/O/O/O/O/O/and
RAD54L/NN/I-NP/O/O/O/RAD54L
may/MD/I-VP/O/VP-1/O/may
cause/VB/I-VP/O/VP-1/O/cause
breast/NN/I-NP/O/NP-OBJ-1/O/breast
cancer/NN/I-NP/O/NP-OBJ-1/O/cancer

```

Figure 2. Example sentence with MBSP parsing tags. NNS: plural noun, NP: noun phrase, SBJ: subject, CC: coordinating conjunction, NN: singular or mass noun, MD: modal, VP: verb phrase, VB: base form verb, OBJ: object

Table 1. Explanation for MBSP parsing tags

WORD	Part-of-speech	Chunk	PNP	Relation	Anchor	Lemma
HR54	NNS	NP	O	NP-SBJ-1	O	hr54

```

[Rule1 = cause] [Pos cover = 51 Neg cover = 3]
MAX((pos-neg)*pos)= 2448
Rule1 S1 = 0.4000000
s(s1).s(s6).s(s18).s(s39).s(s57).s(s91).s(s129).s(s131).s(s148).s(s169).s(s170).s(s184).s(s257).s(s271).s
(s285).s(s288).s(s293).s(s295).s(s303).s(s311).s(s319).s(s326).s(s351).s(s361).s(s365).s(s366).s(s370).s
(s374).s(s376).s(s392).s(s401).s(s429).s(s447).s(s468).s(s470).s(s471).s(s474).s(s482).s(s483).s(s491).s
(s507).s(s531).s(s555).s(s621).s(s626).s(s673).s(s693).s(s714).s(s885).s(s890).s(s892).

```

Figure 3. An example of the learned rule

Oshiro et al. (1995) reported a 32-year-old Japanese woman with combined partial 17-alpha-hydroxylase/17,20-lyase deficiency caused by mutation in the CYP17A1 gene (609300.0008).

Figure 4. A sentence with a gene-disease pair where 17-alpha-hydroxylase/17,20-lyase is a disease name, and CYP17A1 is a gene name

In Table 1, MBSP determines that HR54 is a plural noun (NNS). Its chunk belongs to a noun phrase (NP). It is not a prepositional noun phrase (PNP). Its relation tag is NP-SUBJ-1, which means that it is the subject of the sentence. It is not an anchor. Finally, the word lemma hr54 is a canonical lexical form of HR54. In this study, we adopt the parsing information that includes the part-of-speech, chunk and SVO relation (Subject-Verb-Object relation).

4.4. Rule Learning

The inductive logic programming (ILP) [39] is the machine learning framework for our rule learning algorithm. In the ILP framework, a hypothesis H (i.e. a set of rules) is induced from examples E with the background knowledge B such that H implies examples E with the background knowledge B . Given B and E , we use an ILP system, ALEPH system to learn H , so that some candidate rules are generated. After the learning step, ALEPH will output the information about learning, i.e., the numbers of positive and negative sentences covered by the rules and the rule forms.

To calculate the score of each rule, we propose a formula as follows.

$$S = (Pos - Neg) \times \frac{Pos}{Max_i((Pos - Neg) \times Pos)}, \text{ for all rule } i \quad (1)$$

where S is the score, Pos and Neg represent the numbers of positive and negative sentences covered by the rule, respectively. The denominator is used for normalization.

Figure 3 shows an example of the learning result. The first

line identifies that the rule is for the verb “cause.” There are 51 positive sentences and 31 negative sentences in the training data. The second line computes the value of denominator in Equation (1). The third line shows the score S . In the following lines show the 51 positive sentence IDs: s1, s6, ..., and s892.

4.5. Rule-based Approach

In this phase, the testing data are first preprocessed as described in Section 4.1. For finding the sentences with correct gene-disease pairs, we filter out the sentences with negative meanings. A sentence including some negative words (e.g., not, neither, never) will be considered as negative. Furthermore, anaphora resolution is done by a simple heuristic: the preceding gene or disease name is extracted. Then we apply the rules learned previously so that some gene-disease pairs are retrieved. Figure 4 shows a sentence where a gene-disease pair can be retrieved by the rule “cause” in Figure 3.

4.6. Statistical Approach

Referring to the statistical method proposed by Al-Mubaid and Singh [40], we try to find the relatedness between genes and diseases. First, we extract the disease names appearing in the gene-disease pairs by the above rule-based approach. Then we retrieve 20 MEDLINE abstracts that mention about the extracted diseases in the OMIM and call it as the *interest set* S_1 . Next, we randomly select 182 MEDLINE abstracts

that do not contain the extracted diseases, and call it as the *control set* S_2 . Furthermore, a gene list ($G_1, G_2, \dots, G_{7939}$) extracted from the morbid file is used to count the numbers of genes mentioned the *interest set* and *control set*, respectively. With the above data and values, the *expectation* (ex) and *evidence* (ev) values are calculated for each gene. The equations are stated as follows.

$$ex(G_i) = (tf_t(G_i) / |S_1 + S_2|) \times |S_1| \quad (2)$$

$$ev(G_i) = tf_1(G_i) \quad (3)$$

where G_i is the i th gene in the gene list, tf means term frequency and $tf_t(G_i)$ is the sum of the numbers of G_i mentions in the interest set (S_1) and the control set (S_2). $tf_1(G_i)$ is the number of occurrences (mentions) of G_i in S_1 . In the above equations, the *expectation* measures how many G_i is normally expected to appear in S_1 ; while the *evidence* determines how many G_i has actually appeared in S_1 . It is obvious that the larger of the difference between ex and ev is, the more of the likelihood that G_i and the disease have a significant association is. Thus, we compute the normalized difference as shown in Equation (4).

$$f(G_i) = (ev(G_i) - ex(G_i)) / tf_t(G_i) \quad (4)$$

After computing $f(G_i)$ for each gene, we sort the genes according to their f values and use the *Z-score* metric to determine the significant f values.4

$$Z - score(G_i) = (f(G_i) - mean(f)) / SD(f) \quad (5)$$

where $mean(f)$ is the average of all f values of all genes and $SD(f)$ is the standard deviation of f values. Here the *Z-score* measures how many standard deviations of each f value are greater than the mean f value, for all genes, to indicate the statistical significance. In our experiments, we consider a gene as having the significant association with the disease if it gets *Z-score* of 1.0 or more.

5. Results

5.1. Evaluation Metrics

We use the standard precision, recall and F-score evaluation measures. Precision and recall are defined as follows.

$$\text{Precision} = TP / (TP + FP) \quad (6)$$

$$\text{Recall} = TP / (TP + FN) \quad (7)$$

$$F - \text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives. In this experiment, true positives are the sentences with correct gene-disease pairs proposed by our system. False positives are the sentences with incorrect gene-disease pairs proposed by our system. False negatives are the sentences with gene-disease pairs in the answer key, but our system does not propose. F-score measures the performance that includes precision and recall rates.

5.2. Results and Discussion

The experimental results are listed in Table 2. Column “ID” represents the experiment number. Column “Experiment” is the experiments we made. Column “Positive” is the number of correct gene-disease pairs the system proposes. Column “Negative” is the number of incorrect gene-disease pairs the system proposes. Columns “Precision,” “Recall” and “F-score” stand for the values computed with Equations (6), (7) and (8), respectively.

In Table 2, experiments 1-4 use the rule-based approach and numbers 6, 9, 18 and all identify the numbers of applied rules. Figure 5 shows nine rules used in experiments 2 and 6. Experiments 5-8 employ both rule-based and statistical methods (i.e. Z-score method). Experiment 9 only uses the Z-score method (with the threshold > 1) to filter out the candidate gene-disease pairs.

Comparing experiment 1 to 5, 2 to 6, 3 to 7, and 4 to 8, it shows that combining the Z-score method can enhance the precision rates and F-scores while the recall rates are decreased a little bit. It demonstrates that Z-score can filter out some incorrect gene-disease pairs which may be mentioned in the same sentence. Comparing experiments 5-8 to 9, the precisions, recalls and F-scores are higher when rules are involved. It shows that the learned rules are helpful for the gene-disease association. In short, both the rule-based and statistical approaches are useful because experiments 7 and 8 get the best performance.

4 http://en.wikipedia.org/wiki/Standard_score

Table 2. Experimental results

ID	Experiment	Positive	Negative	Precision	Recall	F-score
1	6 Rules	211	168	55.67%	60.46%	57.97%
2	9 Rules	196	137	58.85%	63.43%	61.05%
3	18 Rules	207	162	56.09%	66.99%	61.06%
4	All Rules	207	162	56.09%	66.99%	61.06%
5	6 Rules+Z-score	211	79	72.76%	60.29%	65.94%
6	9 Rules+Z-score	193	60	76.28%	60.50%	67.48%
7	18 Rules+Z-score	202	73	73.45%	65.37%	69.18%
8	All Rules+Z-score	202	73	73.45%	65.37%	69.18%
9	Z-score	163	179	47.66%	46.57%	47.10%

Rule 1: <gene> cause <disease>
Rule 2: <gene> be involved in <disease>
Rule 3: <disease> be associated with <gene>
Rule 4: <disease> affect by <gene>
Rule 5: <gene> result in <disease>
Rule 6: <gene> regulator of <disease>
Rule 7: <gene> determine <disease>
Rule 8: <gene> be expressed in <disease>
Rule 9: relationship to <gene> with <disease>

Figure 5. The 9 rules used in experiments 2 and 6

6. Conclusions

The paper aims to automatically identify the associations of the human genetic diseases and genes from the biomedical literatures. We first use the MBSP system to get the parsing information. Then we apply a rule learning algorithm to get the learned rules. In the following, the Z-score method is proposed to filter out the incorrect gene-disease pairs. The best results are 73.45% precision, 65.37% recall and 69.17% F-score, which are achieved by combining the learned rules and the Z-score method. It shows the proposed methods in this paper are effective in identifying the gene-disease associations. In future, we want to try to do more elaborate anaphora resolution to enhance the recall rate. Furthermore, enlarging the amount of the training sentences is the other feasible direction.

Acknowledgements

Research of this paper was partially supported by Ministry of Science and Technology, Taiwan, under the contract MOST 103-2221-E-003-014.

REFERENCES

- [1] Singh-Blom, U.M., Natarajan, N., Tewari, A., Woods, J.O., Dhillon, I.S., Marcotte, E.M., Prediction and validation of gene-disease associations using methods inspired by social network analyses, *PLoS ONE*, 8(5), e58977, 2013.
- [2] Gottlieb, A., Magger, O., Berman, I., Rupp, E., Sharan, R., PRINCIPLE: a tool for associating genes with diseases via network propagation, *Bioinformatics*, 27(23), pp. 3325–3326, 2011.
- [3] Keith, B.P., Robertson, D.L., Hentges, K.E., Locus heterogeneity disease genes encode proteins with high interconnectivity in the human protein interaction network, *Frontiers in Genetics*, 5, 434, 2014.
- [4] Navlakha, S., Kingsford, C., The power of protein interaction networks for associating genes with diseases, *Bioinformatics*, 26(8), pp. 1057–1063, 2010.
- [5] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., Gene Ontology: tool for the unification of biology, *Nature Genetics*, 25(1), pp. 25–29, 2000.
- [6] Kibbe, W.A., Osborne, J.D., Wolf, W.A., Smith, M.E., Zhu, L., Lin, S., Chisholm, R.L., The Disease Ontology and browser, *Proceedings of 2006 caBIG Annual Meeting*, Arlington, USA, April 9-11, 2006, pp. 56–57.
- [7] Schlicker, A., Lengauer, T., Albrecht, M., Improving disease gene prioritization using the semantic similarity of Gene Ontology terms, *Bioinformatics*, 26(18), pp. i561–i566, 2010.
- [8] Singleton, M.V., Guthery, S.L., Voelkerding, K.V., Chen, K., Kennedy, B., Margraf, R.L., Durtsche, J., Eilbeck, K., Reese, M.G., Jorde, L.B., Huff, C.D., Yandell, M., Phevor combines multiple biomedical ontologies for accurate

identification of disease-causing alleles in single individuals and small nuclear families, *The American Journal of Human Genetics*, 94(4), pp. 599–610, 2014.

- [9] Yilmaz, S., Jonveaux, P., Bicep, C., Pierron, L., Smail-Tabbone, M., Devignes, M.D., Gene-disease relationship discovery based on model-driven data integration and database view definition, *Bioinformatics*, 25(2), pp. 230–236, 2009.
- [10] Nelson, S.J., Schopen, M., Savage, A.G., Schulman, J.L., Arluk, N., The MeSH translation maintenance system: structure, interface design, and implementation, *Proceedings of the 11th World Congress on Medical Informatics*, San Francisco, USA, September 7-11, 2004, 11(Pt 1), pp. 67–69.
- [11] Mottaz, A., Yip, Y.L., Ruch, P., Veuthey, A.L., Mapping proteins to disease terminologies: from UniProt to MeSH, *BMC Bioinformatics*, 9(Suppl 5), S3, 2008.
- [12] Ono, T., Kuhara, S., A novel method for gathering and prioritizing disease candidate genes based on construction of a set of disease-related MeSH terms, *BMC Bioinformatics*, 15, 179, 2014.
- [13] Tsafnat, G., Jasch, D., Misra, A., Choong, M.K., Lin, F.P.Y., Coiera, E., Gene-disease association with literature based enrichment, *Journal of Biomedical Informatics*, 49, pp. 221–226, 2014.
- [14] Ala, U., Piro, R.M., Grassi, E., Damasco, C., Silengo, L., Oti, M., Provero, P., Di Cunto, F., Prediction of human disease genes by human-mouse conserved coexpression analysis, *PLoS Computational Biology*, 4(3), e1000043, 2008.
- [15] Gaulton, K.J., Mohlke, K.L., Vision, T.J., A computational system to select candidate genes for complex human traits, *Bioinformatics*, 23(9), pp. 1132–1140, 2007.
- [16] Ma, X., Lee, H., Wang, L., Sun, F., CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data, *Bioinformatics*, 23(2), pp. 215–221, 2007.
- [17] Adie, E.A., Adams, R.R., Evans, K.L., Porteous, D.J., Pickard, B.S., SUSPECTS: enabling fast and effective prioritization of positional candidates, *Bioinformatics*, 22(6), pp. 773–774, 2006.
- [18] Bloss, C.S., Scott-Van Zeeland, A.A., Topol, S.E., Darst, B.F., Boeldt, D.L., Erikson, G.A., ... Torkamani, A., A genome sequencing program for novel undiagnosed diseases, *Genetics in Medicine*, 2015.
- [19] George, R.A., Liu, J.Y., Feng, L.L., Bryson-Richardson, R.J., Fatkin, D., Wouters, M.A., Analysis of protein sequence and interaction data for candidate disease gene prediction, *Nucleic Acids Research*, 34(19), e130, 2006.
- [20] Maver, A., Peterlin, B., Positional integratonic approach in identification of genomic candidate regions for Parkinson's diseases, *Bioinformatics*, 27(14), pp. 1971–1978, 2011.
- [21] Hou, W.J., Chen, H.Y., Rule extraction in gene-disease relationship discovery, *Gene*, 518(1), pp. 132–138, 2013.
- [22] Yu, S., Vooren, S.V., Tranchevent, L.C., Moor, B.D., Moreau, Y., Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text mining, *Bioinformatics*, 24(16), pp. i119–i125, 2008.
- [23] Mordelet, F., Vert, J.P., ProDiGe: Prioritization Of Disease Genes with multitask machine learning from positive and unlabeled examples, *BMC Bioinformatics*, 12(1), 389, 2011.
- [24] Yang, P., Li, X., Chua, H.N., Kwoh, C.K., Ng, S.K., Ensemble positive unlabeled learning for disease gene identification, *PLoS ONE*, 9(5), e97079, 2014.
- [25] Ganegoda, G.U., Wang, J., Wu, F.X., Li, M., Prediction of disease genes using tissue-specified gene-gene network, *BMC Systems Biology*, 8(Suppl 3), S3, 2014.
- [26] Valentini, G., Paccanaro, A., Caniza, H., Romero, A.E., Re, M., An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods, *Artificial Intelligence in Medicine*, 61(2), pp. 63–78, 2014.
- [27] Zhou, M., Wang, X., Li, J., Hao, D., Wang, Z., Shi, H., Han, L., Zhou, H., Sun, J., Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network, *Molecular BioSystems*, 11, pp. 760–769, 2015.
- [28] Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C., Valle, D., McKusick, V.A., Online Mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders, *Nucleic Acids Research*, 30(1), pp. 52–55, 2002.
- [29] Feldman, R., Regev, Y., Finkelstein-Landau, M., Hurvitz, E., Kogan, B., Mining biomedical literature using information extraction, *Current Drug Discovery*, 2(10), pp. 19–23, 2002.
- [30] Fundel, K., Kuffner, R., Zimmer, R., RelEx–relation extraction using dependency parse tree, *Bioinformatics*, 23(3), pp. 365–371, 2007.
- [31] Yu, H., Hatzivassiloglou, V., Friedman, C., Rzhetsky, A., Wilbur, W.J., Automatic extraction of gene and protein synonyms from MEDLINE and journal articles, *AMIA 2002 Annual Symposium Proceedings*, San Antonio, USA, November 9-12, 2002, pp. 919–923.
- [32] Gopalakrishnan, V., Lustgarten, J.L., Visweswaran, S., Cooper, G.F., Bayesian rule learning for biomedical data mining, *Bioinformatics*, 26(5), pp. 668–675, 2010.
- [33] Kim, J.H., Mitchell, A., Attwood, T.K., Hilario, M., Learning to extract relations for protein annotation, *Bioinformatics*, 23(13), pp. i256–i263, 2007.
- [34] Liu, Y.C., Cheng, C.P., Tseng, V.S., Discovering relational-based association rules with multiple minimum supports on microarray datasets, *Bioinformatics*, 27(22), pp. 3142–3148, 2011.
- [35] Lindsay, R.K., Gordon, M.D., Literature-based discovery by lexical statistics, *Journal of the American Society for Information Science*, 50(7), pp. 574–587, 1999.
- [36] Friedman, C., Kra, P., Yu, H., Krauthammer, M., Rzhetsky, A., GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles, *Bioinformatics*, 17(Suppl. 1), pp. S74–S82, 2001.
- [37] Jelier, R., Jenster, G., Dorssers, L.C.J., van der Eijk, C.C., van Mulligen E.M., Mons, B., Kors, J.A., Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes, *Bioinformatics*, 21(9), pp. 2049–2058, 2005.

- [38] Xu, R., Wang, Q., Automatic construction of a large-scale and accurate drug-side-effect association knowledge base from biomedical literature, *Journal of Biomedical Informatics*, 51, pp. 191-199, 2014.
- [39] Muggleton, S., Readt, L.D., Inductive logic programming theory and methods, *The Journal of Logic Programming*, 9, p. 629–679, 1994.
- [40] Al-Mubaid, H., Singh, R.K., A new text mining approach for finding protein-to-disease associations, *American Journal of Biochemistry and Biotechnology*, 1(3), pp. 145–152, 2005.