Sequential Pattern Mining to Discover Relations between Genes and Rare Diseases

Nicolas Béchet[1], Peggy Cellier[2], Thierry Charnois[1], Bruno Cremilleux[1], Marie-Christine Jaulent[3]
[1] GREYC Université de Caen Basse-Normandie 14032 Caen Cedex, France, {prenom.nom}@unicaen.fr
[2] IRISA, INSA Rennes, 35042 Rennes cedex, France, peggy.cellier@irisa.fr
[3] CRC Jussieu, 15 rue de l'école de médecine, 75006 Paris, France, marie-christine.jaulent@crc.jussieu.fr

Abstract

Orphanet provides an international web-based knowledge portal for rare diseases including a collection of review articles. However, reviews and literature monitoring are manual. Thus, new documentation about a rare disease is a time-consuming process and automatically discovering knowledge from a large collection of texts is a crucial issue. This context represents a strong motivation to address the problem of extracting gene–rare diseases relationships from texts. In this paper, we tackle this issue with a cross-fertilization of information extraction and data mining techniques (sequential pattern mining under constraints). Experiments show the interest of the method for the documentation of rare diseases.

1 Introduction

Rare diseases are a major public health issue. A rare disease (RD) is a disease affecting fewer than 1 in 2,000 persons. There are between 6,000 and 8,000 RDs affecting about 30 million people in Europe and much more in the rest of the world. For most of RDs, information is often scattered and scientists need to share their research to work more efficiently. For that purpose, Orphanet[1] provides an international web-based knowledge portal for rare diseases including a collection of review articles on RDs which is expert-authored and peer-reviewed. However, reviews and literature monitoring are done manually and rely on a number of annotators dealing with rare genetic pathologies. Thus, obtaining new documentation about a RD is a time-consuming process and automatically discovering knowledge from a large collection of texts is a crucial issue. This context represents a strong motivation to address the problem of extracting gene–RD relationships from text collection such as the PubMed repository dealing with more than 21 million biomedical publications. In this paper, we tackle this issue with a cross-fertilization of information extraction and data mining techniques.

Natural Language Processing (NLP), and information extraction in particular, aim to provide accurate parsing to extract specific knowledge such as named entities (e.g., gene, disease) and relationships between the recognized entities (e.g., gene–gene interactions [9], disease-treatment relations [1]). Those NLP approaches require rules such as regular expressions for surface searching [7] or syntactic patterns [10, 6]. When the rules are handcrafted, those methods are then time consuming and also very often devoted to a specific corpus [8].

In contrast, machine learning methods such as support vector machines or conditional random fields [9] are based on automatic processes and then are less time consuming than NLP methods. Although they provide good results on accuracy, they still suffer from limitations. Their outcomes are not really understandable by an end-user, nor they can be used as linguistic patterns in NLP systems. Furthermore, the annotation process of training corpora requires a substantial investment of time, and cannot be reused in other domains (new corpora must be annotated for new domains) [8].

Recent works take advantage of an hybridization of data mining and NLP techniques. Data mining enables the discovery of implicit, previously unknown, and potentially useful information from data [5]. For instance, in [3] a method is proposed to automatically discover linguistic rules to extract relationships between named entities in new corpora. That approach is not supervised and does not need syntactic parsing nor external resources except the training corpus. It relies on extraction of frequent sequential patterns where a sequence is a list of literals called *items*, and an item is a word (or its lemma) within textual data.

In this paper, we show the usefulness of sequential patterns in information extraction tasks and discovery of gene–RD relationships. We propose a new method based on sequential patterns of itemsets in order to extract more ex-

[1] www.orphanet.org

pressive linguistic patterns than ones extracted with single-items as in [3]. It means that a word can be represented by a set of features conveying several pieces of information (e.g., words, lemma) and not only a single information. That is a major step because in many applications we need to combine different levels of abstraction (e.g., words, lemma, part of speech tags) and express information according to different generic levels. For instance, in Section 3.2, we will see the interest of a pattern such as $\langle (mutation\ NNS)\ (IN)\ (isocitrate\ NN)\ (GENE)\ (occur\ VBP)\ (DISEASE)\rangle$ which combines the lemma and category levels. Moreover, the user can easily lead the search according to his interest and/or the application target thanks to constraints. We provide several examples of constraints in the context of RD enabling us to discover relevant linguistic patterns and thus gene–RD relationships. We have conducted some experiments to extract gene-RD relationships from PubMed articles.

The rest of the paper is organized as follows. Preliminaries about sequential pattern mining are given in Section 2. Section 3 presents the method to extract relations between genes and rare diseases in biomedical texts. Finally, experiments described in Section 4 show the interest of the method for the documentation of RDs.

2 Preliminaries: Sequential Pattern Mining

Sequential pattern mining is a data mining technique introduced in [2] to find regularities in a sequence database. There are a lot of algorithms to extract sequential patterns [12, 13, 14].

In sequential pattern mining, an *itemset* I is a set of literals called *items*. For example, $(a\ b)$ is an itemset with two items: a and b. A *sequence* S is an ordered list of itemsets, denoted by $s = \langle I_1 \dots I_m\rangle$. For instance, $\langle (a)\ (a\ b\ c)\ (a\ c)\ (d)\rangle$ is a sequence of four itemsets. A sequence $S_1 = \langle I_1 \dots I_n\rangle$ is *included* in a sequence $S_2 = \langle I'_1 \dots I'_m\rangle$ if there exist integers $1 \leq j_1 < ... < j_n \leq m$ such that $I_1 \subseteq I'_{j_1}$,..., $I_n \subseteq I'_{j_n}$. The sequence S_1 is called a *subsequence* of S_2, and we note $S_1 \preceq S_2$. For example, $\langle (a)(a\ c)\rangle$ is included in $\langle (a)(a\ b\ c)(a\ c)(d)\rangle$. A sequence database SDB is a set of tuples (sid, S), where sid is a sequence identifier and S a sequence. For instance, Table 1 depicts a sequence database of four sequences. A tuple (sid, S) *contains* a sequence S_1, if $S_1 \preceq S$. The *support*[2] of a sequence S_1 in a sequence database SDB, denoted $sup(S_1)$, is the number of tuples in the database containing S_1. For example, in Table 1 $sup(\langle (a\ b)(c)\rangle) = 2$, since Sequences 1 and 3 contain $\langle (a\ b)(c)\rangle$. A *frequent sequential pattern* is a sequence such that its support is greater or equal to a given support threshold $minsup$.

[2]Note that the *relative support* is also used: $sup(S_1) = \frac{|\{(sid,S)\ |\ (sid,S) \in SDB \wedge (S_1 \preceq S)\}|}{|SDB|}$.

Sequence identifier	Sequence
1	$\langle (a)\ (a\ b\ c)\ (a\ c)\ (d)\rangle$
2	$\langle (a\ d)\ (c)\ (b)\rangle$
3	$\langle (a\ b)\ (d)\ (b)\ (c)\rangle$
4	$\langle (a\ d)\ (b)\ (b)\rangle$

Table 1. Example of a sequential database.

The set of frequent sequential patterns can be very large. Pattern condensed representations, such as closed sequential patterns [13], have been proposed in order to eliminate redundancy without loss of information. A frequent sequential pattern S is closed if there is no other frequent sequential pattern S' such that $S \preceq S'$ and $sup(S) = sup(S')$. For instance, with $minsup = 2$, the sequential pattern $\langle (a\ b)\rangle$ from Table 1 is not closed because $sup(\langle (a\ b)\rangle) = sup(\langle (a\ b)(c)\rangle)$ and $\langle (a\ b)\rangle \preceq \langle (a\ b)(c)\rangle$.

The constraint-based pattern paradigm [4] brings useful techniques to express a user's interest in order to focus on the most promising patterns. A very widespread constraint is the frequency. However, it is possible to define many other useful constraints such as the gap constraint. A sequential pattern with a gap constraint $[M, N]$, denoted by $P_{[M,N]}$, is a pattern such as at least M itemsets and at most N itemsets are allowed between every two neighbor itemsets, in the matched sequences. For instance, in Table 1, $P_{[0,2]} = \langle (a)(c)\rangle$ and $P_{[1,2]} = \langle (a)(c)\rangle$ are two patterns with gap constraints. $P_{[0,2]}$ matches three sequences (1, 2 and 3) whereas $P_{[1,2]}$ matches only two sequences (1, 3). Indeed, in Sequence 2 there is no itemset between the itemset that contains a and the itemset that contains c.

3 Discovering Relations between Genes and Rare Diseases

In this section, we present our global approach to discover relations between genes and rare diseases (Section 3.1). Then, we define constraints to extract sequential patterns (Section 3.2).

3.1 Global Process

Figure 1 provides a global view of the process. There are two main steps in the method: extraction and validation of sequential patterns as linguistic patterns and their application to discover gene–RD relationships.

The sequence database is built from a training corpus. Sequences are sentences of the training corpus having at least one rare disease and one gene. Thanks to the POS tagging step, each word is replaced by an itemset containing information about the word: the lemma of the word and

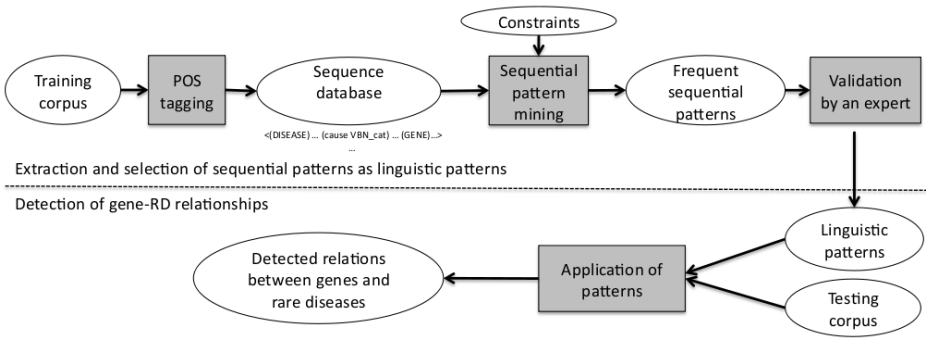

Figure 1. Global view of the method to extract gene–RD relationships.

its grammatical categories (Part-Of-Speech (POS) tags[3]). Table 2 gives an excerpt of a sequence database with three sequences [4]. For instance, in Sequence 1 the verb "conclude" is replaced by the itemset $(conclude\ VBP)$, i.e. its lemma "conclude" and its grammatical category: a verb, non-third person singular, present tense. Then, all gene names are replaced by the general item *GENE* and in the same way all disease names are replaced by the item *DISEASE*. Note that unlike machine learning based approaches, the training corpus have not annotated relations (e.g. gene-RD relations). Once the sequence database built, sequential patterns (see Section 3.2) under constraints are extracted. This part of the method is detailed in the next section. In order to exclude redundancy between patterns, we used closed sequential patterns instead of frequent sequential patterns. In addition, a word is described as set of features, providing different kinds of information. This word representation allows to combine levels of abstraction, and to build generic patterns (i.e. patterns having only grammatical categories, as $\langle(NNS)(IN)(NN)(GENE)\ (VBP)\ (DISEASE)\rangle$) and more specific (i.e. patterns having grammatical categories and lemma, as $(mutation\ NNS)\ (in\ IN)\ (isocitrate\ NN)\ (GENE)\ (occur\ VBP)\ (DISEASE)\rangle$. Then, an expert filters out patterns which are not relevant as linguistic patterns to discover relations between genes and RD in texts. The validated patterns are then applied on the testing corpus to discover new gene–RD relationships.

[3]DT: Determiner, IN: Preposition or subordinating conjunction, JJ: Adjective, NN: Noun, RB: Adverb, VB: Verb. The complete list of part-of-speech tags is available at http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[4]1. *"We conclude that VCP is essential for maturation of ubiquitin-containing autophagosomes and that defect in this function may contribute to IBMPFD pathogenesis."* 2. *"Somatic mutations in isocitrate dehydrogenase 1 (IDH1) and IDH2 occur in gliomas and acute myeloid leukaemia (AML)."* 3. *"Osteogenesis imperfecta is normally caused by an autosomal dominant mutation in the type I collagen genes COL1A1 and COL1A2."*

3.2 Constraints to Model Linguistic Knowledge

The sequential pattern mining step is done under constraints which enables to model some linguistic knowledge and gives prominence to the most significant patterns by filtering the specific ones. The goal is to retain sequential patterns which convey linguistic regularities (e.g., gene–rare disease relationships). Moreover, the patterns are closed patterns to avoid redundancy (cf. Section 2).

We have previously introduced two usual constraints: the frequency and the gap. We define in this section other useful constraints in order to discover relations between genes and rares diseases. The *membership* constraint enables to filter out sequential patterns that do not contain some selected items. For example, we express that the extracted patterns must contain at least two items: $GENE$ and $DISEASE$. The *min length* constraint is useful to remove sequential patterns that are too small with respect to the number of itemsets (number of words) to be relevant linguistic patterns. The *max scope* constraint allows to set the maximal number of itemsets between the first itemset and the last itemset of a sequential pattern in the original sequences. Finally, the *association* constraint expresses that all sequential patterns that contain the verb item (VB) must contain a lemma in the same itemset. It means that a pattern $\langle(GENE)(VB)(DISEASE)\rangle$ is not correct with respect to the association constraint whereas $\langle(GENE)(encode\ VB)(DISEASE)\rangle$ is correct. That constraint enables to prune too generic patterns (i.e. containing only grammatical categories). Indeed, the interesting information is the lemma associated with the VB item which characterizes the gene-RD relation. All these constraints are defined in the mining step. Moreover, constraints such as maximal scope and gap can also be used in the application step, possibly with different scope and gap values (cf. Section 4.1).

In addition, we tackle a challenge of extracting closed

Sequence identifier	Sequence
1	$\langle(we\ PP)\ (conclude\ VBP)\ (that\ IN)\ (GENE)\ (be\ VBZ)\ (essential\ JJ)\ (for\ IN)$ $(maturation\ NN)\ (of\ IN)\ (ubiquitin\ NN)\ (contain\ VBG)\ (autophagosomes\ NNS)$ $(and\ CC)\ (that\ DT)\ (defect\ NN)\ (in\ IN)\ (this\ DT)\ (function\ NN)\ (may\ MD)$ $(contribute\ VB)\ (to\ TO)\ (DISEASE)\ (pathogenesis\ NN)\rangle$
2	$\langle(somatic\ JJ)\ (mutation\ NNS)\ (in\ IN)\ (isocitrate\ NN)\ (dehydrogenase\ NN)$ $(1\ CD)\ (GENE)\ (and\ CC)\ (GENE)\ (occur\ VBP)\ (in\ IN)\ (glioma\ NNS)$ $(and\ CC)\ (acute\ JJ)\ (myeloid\ JJ)\ (leukaemia\ NN)\ (DISEASE)\rangle$
3	$\langle(DISEASE)\ (be\ VBZ)\ (normally\ RB)\ (cause\ VBN)\ (by\ IN)\ (an\ DT)\ (autosomal\ JJ)$ $(dominant\ JJ)\ (mutation\ NN)\ (in\ IN)\ (the\ DT)\ (type\ NN)\ (i\ NN)$ $(collagen\ NN)\ (gene\ NNS)\ (GENE)\ (and\ CC)\ (GENE)\rangle$

Table 2. Excerpt of a sequential database from medical texts with three sequences.

sequential patterns of itemsets instead of sequential patterns of single-items as in [3]. A word is represented by a set of features conveying several pieces of information and not only a single information. It allows to combine different levels of abstraction and express information according to different generic levels. For example, a generic pattern as $\langle(NNS)\ (IN)\ (NN)\ (GENE)\ (VBP)\ (DISEASE)\rangle$ with only grammatical information can be extracted, but a more specific pattern as $\langle(mutation\ NNS)\ (IN)\ (isocitrate\ NN)\ (GENE)\ (occur\ VBP)\ (DISEASE)\rangle$ which combines lemma and grammatical information can also be extracted.

There exist in the literature many algorithms to extract sequential patterns (e.g. [12, 14]) or closed sequential patterns (e.g. [13]). But, to the best of our knowledge, there is no algorithm mining closed sequential patterns of itemsets under constraints. We address this issue by designing an algorithm mining sequential patterns of itemsets under the previously defined constraints. Details of the algorithm are not given here because it is out of the scope of the paper.

4 Experiments

4.1 Settings

We created a corpus from the PubMed database using HUGO[5] dictionary and Orphanet dictionary to query the database to get sentences having these two kinds of entities. 17,527 sentences have been extracted in this way and we labelled the gene and RD names thanks to the two dictionaries. For instance, the sentence "*<**disease**>**Muir-Torre syndrome**<**disease**> is usually inherited in an autosomal dominant fashion and associated with mutations in the mismatch repair genes, predominantly in <**gene**>**MLH1**<**gene**> and <**gene**>**MSH2**<**gene**> genes.*" contains one recognized RD, and two recognized genes. From these 17,527 sentences, we randomly extract 200 sentences as a testing corpus, the remaining sentences being the training corpus.

[5] www.genenames.org

The 200 sentences of the testing corpus have been evaluated by an expert in order to identify sentences having gene–RD relationships. Note that a sentence of this corpus can have multiple gene–RD relationships or conversely none. From the 200 sentences of this corpus, 189 gene–RD relationships have been identified by an expert and 132 couples of gene-RD are not in relation.

Sequential Pattern Extraction Sequences of the SDB are the sentences of the training corpus and are built as described in Section 3.1. We carry out a POS tagging of the sentences thanks to the TreeTagger tool [11]. The algorithm was running with the following different constraint characteristics to extract the closed sequential patterns:

• *The minimal frequency (minsup).* Three values of minimal frequency have been experimented: 0.5% (88 sequences), 0,2% (35 sequences), and 0,05% (8 sequences).

• *The gap.* We have conducted experiments without and with gap value (chosen empirically at [0,10]).

• *The maximal scope.* We set a maximal scope value at 20 to reduce the number of extracted patterns: we assume that the maximal number of itemsets between the first itemset and the last itemset of patterns having gene–RD relationships is almost 20 (corresponding to 20 words in the sentence).

• *The minimal length.* The aim of this constraint is to limit the number of generic patterns. We tested (a) this constraint with a value set to 4 and (b) without this constraint.

• *The membership.* Patterns must contain at least 3 items: one gene, one RD, and one noun or one verb (expressing the linguistic relation).

• *The association.* We want for each verb and noun its lemma and its grammatical category.

Furthermore, we consider only binary relations, i.e. between only one gene and one disease.

Applying linguistic patterns We have applied on the testing corpus the patterns extracted from the training corpus under the gap and maximal scope constraints (called "application gap" and "application scope"). Constraint values used during the application of patterns are: "no gap" or

minsup	gap	min length	# of val. pat.	# of pat.
0.50%	[0,10]	all	6,346	24,888
0.50%	[0,10]	4	6,310	22,794
0.50%	no gap	all	6,193	23,823
0.50%	no gap	4	6,156	22,084
0.20%	[0,10]	all	54,512	133,533
0.20%	[0,10]	4	54,429	126,777
0.20%	no gap	all	56,404	138,175
0.20%	no gap	4	56,290	130,579
0.05%	[0,10]	all	416,786	1,530,085
0.05%	[0,10]	4	416,533	1,493,914

Table 3. Number of patterns according to the gap and the minsup constraints

minsup	gap	recall	precision	F-measure
0.50%	[0.10]	0.37	0.67	0.48
0.50%	no gap	0.46	**0.69**	0.55
0.20%	[0.10]	0.50	0.65	0.56
0.20%	no gap	0.53	0.64	0.58
0.05%	[0.10]	**0.65**	0.66	**0.65**

Table 4. Experimental results for different gap and minsup values

[0,10] ; maximal scope is set at 20 or is not used (i.e. ∞).

4.2 Results

Table 3 gives the number of extracted patterns according to the constraint values (# of val. pat. is the number of validated patterns and # of pat. the number of extracted patterns). Minimal frequency constraint ($minsup$) is the constraint having the most important impact on the number of extracted patterns.

Results in Table 4 show the impact of the gap and the frequency constraints on the detection of gene–RD relationships. Decreasing $minsup$ value improves the recall and the f-measure (few patterns are produced when the $minsup$ value is high, see Table 3). In contrast, the gap [0,10] decreases the recall and the f-measure whereas the precision is stable. Indeed, without gap constraint, more generic patterns are extracted which substantially improves the recall. Finally the best f-measure is obtained with the lowest $minsup$ 0.05%.

Table 5 gives the impact of the minimal length constraint which slightly improves the precision.

Then we study the impact of applying the constraints during the pattern application step. Results are given in Table 6. *Application gap* and *application scope* (denoted by *App. gap and App. sc.* in Table 6) are respectively the gap and the scope constraints during the pattern application step (the gap and the minimal length during the pattern extraction step are not taken into account). The scope is fixed to 20. The app. scope constraint strongly degrades recall, and also the f-measure. Otherwise, the app. gap slightly increases precision, but decreases the recall.

minsup	min. length	recall	precision	F-measure
0.50%	all	0.37	0.67	0.48
0.50%	4	0.36	0.68	0.47
0.20%	all	0.50	0.65	0.56
0.20%	4	0.48	0.67	0.56
0.05%	all	0.65	0.66	0.65
0.05%	4	0.64	0.66	0.65

Table 5. Impact of the min. length constraint

minsup	App. gap	App. sc.	rec.	prec.	F-measure
0.50%	[0,10]	all	0,33	0,68	0,44
0.50%	[0.10]	20	0.25	0.68	0.37
0.50%	no gap	all	0.37	0.67	0.48
0.50%	no gap	20	0.26	0.68	0.37
0.20%	[0.10]	all	0.48	0.66	0.55
0.20%	[0.10]	20	0.35	0.66	0.46
0.20%	no gap	all	0.50	0.65	0.56
0.20%	no gap	20	0.36	0.66	0.46
0.05%	[0.10]	all	0.60	0.66	0.63
0.05%	[0.10]	20	0.41	0.65	0.50
0.05%	no gap	all	0.65	0.66	0.65
0.05%	no gap	20	0.41	0.65	0.50

Table 6. Results on constrained patterns

4.3 Discussion

Table 7 summarizes the impact of the constraints. The minimum frequency ($minsup$) and the minimum length are the most relevant constraints respectively improving recall and precision. This means that for a better precision, we have to tune the minimum length constraint, and for a better recall, we can use a lower value of minsup. The maximum precision obtained with the proposed method is 0.69 and the better recall is 0.65. These results are very close to the results of other approaches in literature for similar tasks (.e.g. [1]). However, patterns are automatically extracted unlike [1] where patterns are handcrafted. The method allowed to extract relevant patterns, for instance: $\langle(DISEASE)(be\ VBP))(JJ)(IN)(factor\ NN)(GENE)\rangle$, $\langle(DISEASE)\ (be\ VBZ)\ (JJ)\ (DT)\ (dominant\ JJ)\ (cause\ VBN)\ (by\ IN)\ (DT)\ (GENE)\rangle$, or $\langle(JJ)\ \ (DISEASE)\ (superoxide\ NN)\ (dismutase\ NN)\ (GENE)\rangle$.

We present, in the following, a qualitative analysis of the errors. Some false negatives (affecting recall) can be explained by the human expertise. Indeed, the expert vali-

Constraints	recall	precision
minsup	Increase	No impact
min. length	Decrease	Increase
gap	Decrease	No impact
App. gap	Decrease	No impact
App. scope	Decrease	No impact

Table 7. Impact of the constraints on the discovering of gene–rare disease relationships

dated only patterns with notion of causality (i.e. gene *cause* rare disease). That means a sentence as "*We report on a case of B-ALL of L3 morphology with MYC- IGH translocation*" cannot be discovered by the extracted patterns because important terms of the sentence are too generic and do not express a causality. For instance, terms "report", "case", "morphology" are too generic.

Some false positive cases (affecting precision) can be explained by errors in named entity recognition. We have discovered some sentences having a gene identified as a disease. For instance, in the sentence "*One of the most versatile defence mechanisms against the accumulation of DNA damage is nucleotide excision repair, in which, among others, the Xeroderma pigmentosum group C (XPC) and group A (XPA) proteins are involved.*". In this sentence, the Xeroderma pigmentosum was identified has a disease instead of a gene. Is it also possible that false positive cases come from errors in the expertise of the testing corpus. Indeed, some sentences have been judged negative by the expert whereas they are positive. For instance, sentence "*Small granular SOD1-immunoreactive inclusions were found in spinal motoneurons of all 37 sporadic and familial ALS patients studied, but only sparsely in 3 of 28 neurodegenerative and 2 of 19 non-neurological control patients.*" has been judged negative.

Finally, some false positives are due to negative forms, which is a usual NLP problem. For instance, the sentence "*None of these patients had ATP13A2 sequence variants likely to be causal for their disease, suggesting that mutations in this gene are not common causes of Kufs disease*" is detected as positive by our method whereas the expert tagged it as negative.

5 Conclusion

We have proposed a new method based on sequential pattern mining to automatically discover relations between genes and diseases in biomedical texts. Indeed, the extracted patterns are used as information extraction rules. The method needs a training corpus where genes and rare diseases are identified, but does not require a priori knowledge on relations between genes and rare diseases. Moreover, the discovered linguistic patterns are understandable by a human, and they can by easily changed or excluded if necessary. We have experimented the method to discover gene-RD relations from PubMed articles. Results show the interest of the method and the role of the constraints, and then leading to enhance documentation about rare diseases.

References

[1] A. B. Abacha and P. Zweigenbaum. A hybrid approach for the extraction of semantic relations from medline abstracts. In *Computational Linguistics and Intelligent Text Processing*, LNCS, pages 139–150. Springer, 2011.

[2] R. Agrawal and R. Srikant. Mining sequential patterns. In *Int. Conf. on Data Engineering*. IEEE, 1995.

[3] P. Cellier, T. Charnois, and M. Plantevit. Sequential patterns to discover and characterise biological relations. In *Computational Linguistics and Intelligent Text Processing*, LNCS, pages 537–548. Springer, 2010.

[4] G. Dong and J. Pei. *Sequence Data Mining*. Springer, 2007.

[5] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus. Knowledge discovery in databases: An overview. In *Knowledge Discovery in Databases*. AAAI/MIT Press, 1991.

[6] K. Fundel, R. Küffner, and R. Zimmer. RelEx - relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371, 2007.

[7] C. Giuliano, A. Lavelli, and L. Romano. Exploiting shallow linguistic information for relation extraction from biomedical literature. In *Proc. of the Conf. of the European Chapter of the Association for Computational Linguistics*. The Association for Computer Linguistics, 2006.

[8] J. R. Hobbs and E. Riloff. Information extraction. In N. Indurkhya and F. J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL, 2010.

[9] M. Krallinger, F. Leitner, C. Rodriguez-Penagos, and A. Valencia. Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biology*, 2008.

[10] F. Rinaldi, G. Schneider, K. Kaljurand, M. Hess, and M. Romacker. An environment for relation mining over richly annotated corpora: the case of genia. *BMC Bioinformatics*, 7(S-3), 2006.

[11] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proc. of the Int. Conf. on New Methods in Language Processing*, Manchester, UK, 1994.

[12] R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. In P. M. G. Apers, M. Bouzeghoub, and G. Gardarin, editors, *EDBT*, volume 1057 of *LNCS*, pages 3–17. Springer, 1996.

[13] X. Yan, J. Han, and R. Afshar. Clospan: Mining closed sequential patterns in large databases. In D. Barbará and C. Kamath, editors, *SDM*. SIAM, 2003.

[14] M. J. Zaki. SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning Journal*, 42(1/2):31–60, Jan/Feb 2001. special issue on Unsupervised Learning.