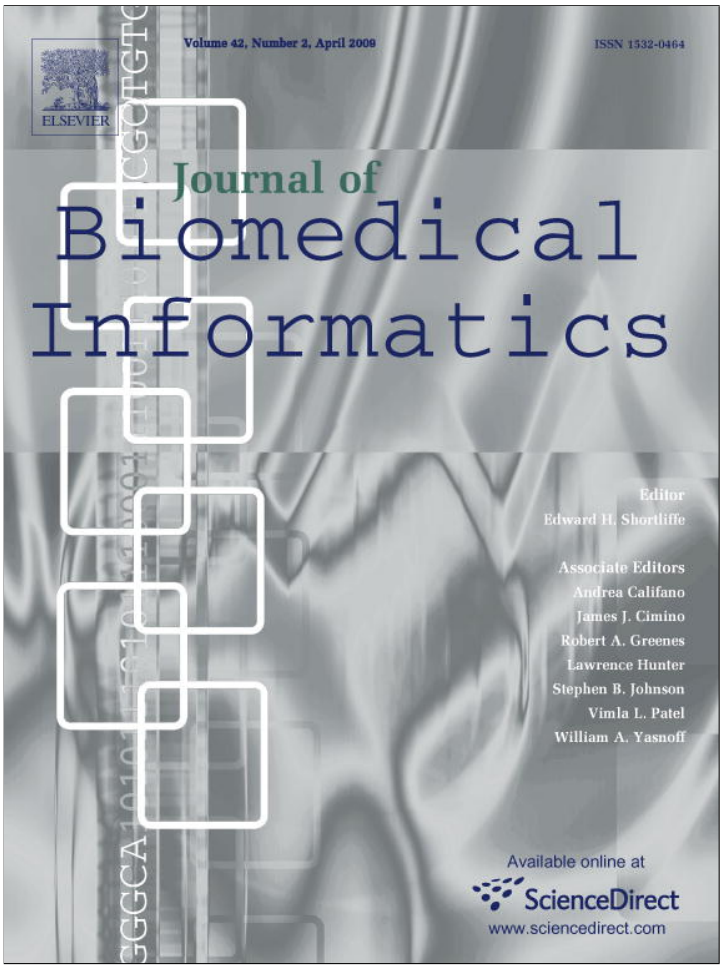

Journal of Biomedical Informatics 42 (2009) 219–227

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Literature mining method RaJoLink for uncovering relations between biomedical concepts

Ingrid Petrič [a,*], Tanja Urbančič [a,b], Bojan Cestnik [b,c], Marta Macedoni-Lukšič [d]

[a] University of Nova Gorica, School of Engineering and Management, Vipavska 13, SI-5000, Nova Gorica, Slovenia
[b] Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[c] Temida, d.o.o., Dunajska 51, 1000 Ljubljana, Slovenia
[d] University Children's Hospital, University Medical Center, 1000 Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 25 July 2007
Available online 19 August 2008

Keywords:
Literature mining
Knowledge discovery
Hypotheses generation
Biomedical articles
Autism

ABSTRACT

To support biomedical experts in their knowledge discovery process, we have developed a literature mining method called RaJoLink for identification of relations between biomedical concepts in disconnected sets of articles. The method implements Swanson's ABC model approach for generating hypotheses in a new way. The main novelty is a semi-automated suggestion of candidates for agents *a* that might be logically connected with a given phenomenon *c* under investigation. The choice of candidates for *a* is based on rare terms identified in the literature on *c*. As rare terms are not part of the typical range of information, which describe the phenomenon under investigation, such information might be considered as unusual observations about the phenomenon *c*. If literatures on these rare terms have an interesting term in common, this joint term is declared as a candidate for *a*. Linking terms *b* between literature on *a* and literature on *c* are then searched for in the closed discovery to provide additional supportive evidence for uncovered connections. We have applied the method to the literature on autism and have used MEDLINE as a source of data. Expert evaluation has confirmed that the discovered relations might contribute to a better understanding of autism.

1. Introduction

Scientific progress can be accelerated by knowledge exchange that fosters new discoveries. Since an abundant quantity of scientific articles is accessible on-line, the usage of large bibliographic databases can support the process. In biomedicine, databases such as MEDLINE [1] provide enormous collections of texts that could be used for knowledge discovery. However, finding the right information in broad data collections requires a great deal of skill and time. Consequently, the need for tools and techniques for processing the vast amount of data available on the Internet grows rapidly.

Large advances of text mining and knowledge discovery technologies [2] facilitate sharing of knowledge and experience among researchers from different, yet related fields of sciences. This is especially important in interdisciplinary sciences such as biomedicine, since such disciplines need the expertise at the intersections of their component sciences [3]. Finding evidence in the biomedical literature to support previously overlooked relations between biomedical concepts represents new knowledge. When uncovered relations are interesting from medical point of view and can be verified by medical experts, they can contribute to a better understanding of diseases and related phenomena.

The idea of performing literature-based discovery of possible relations between previously disjoint concepts was first presented by Swanson [4,5]. He designed the ABC model to facilitate the discovery of hypotheses by linking findings across scientific literature. The ABC model embodies a search for new indirect relations between two disjoint sets of records (*A* and *C*) via intermediate words and phrases, *B*, that are common to *A* and *C*. According to Swanson, *AB* relations and *BC* relations should have already been separately reported in the published literature, but not considered together. Since our method is based on the same idea, the ABC model approach is described in more detail in the next section.

In this article we present RaJoLink, a method that implements Swanson's ABC model approach in a new way. We use the notations *A*–*C* (uppercase symbols) to represent a set of terms (e.g., literature, or set of records, or list of terms), while *a*–*c* (lowercase symbols) represent a single term. The main novelty of the presented method is a semi-automated suggestion of candidates for agents *a* that might be logically connected with a given phenomenon *c* under investigation. In RaJoLink, the choice of candidates for *a* is based on rare terms identified in the literature on *c*. If literatures on these rare terms have an interesting term in common, this

* Corresponding author. +386 5 331 5240.
E-mail addresses: ingrid.petric@p-ng.si (I. Petrič), tanja.urbancic@p-ng.si (T. Urbančič), bojan.cestnik@temida.si (B. Cestnik), marta.macedoni-luksic@mf.uni-lj.si (M. Macedoni-Lukšič).

1532-0464/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.08.004

joint term is declared as a candidate for *a*. Linking terms *b* between literature on *a* and literature on *c* are then searched for in the closed discovery approach to provide additional supportive evidence for uncovered connections.

The reasoning underlying the selection of candidates *a* is the following: If there are some rare terms that appear in literature on *c*, let us have a look at all available records about these rare terms. If these records have an interesting joint term *a* in the intersection, let us check if it has some logical connections with *c*. Concentrating on rare terms increases the probability that the suggested candidates have not yet been explored in terms of their connections with *c*.

The need to support the process of communicating research findings across the disciplines has been emphasized also in the context of autism research, which is carried out in different fields, such as behavioural psychology, genetics, biochemistry, brain anatomy and physiology [6,7]. The complexity and heterogeneity of autism spectrum disorders, as well as several distinct possible causes pose significant challenges to autism researchers who try to explore causes and to identify phenomena that may lead to autism. These facts have motivated us to select autism as the testing domain for discovering hidden knowledge of value within the scientific literature in different fields.

The proposed method was introduced in [8,9] where our study in mining the literature on autism was presented. In this paper we decouple the method from the problem domain, which enables more general description. Besides, we describe and explain the steps of the method in more detail. The paper is structured as follows. In Section 2 we give a brief overview of Swanson's ABC model approach and related text mining applications, which use the ABC model for discovering complementary concepts in disjoint biomedical articles. In Section 3 we describe the RaJoLink method for identifying implicit and previously unknown connections on the basis of rare terms. In Section 4 we present the application of the RaJoLink method to the literature on autism. Experimental results given in Section 5 are followed by the evaluation in Section 6. Section 7 brings conclusions, opens questions and gives directions for further work.

2. Background and related studies

The MEDLINE database can serve as a rich source of hidden relations between biomedical concepts, as shown in 1986 by Swanson [4]. According to his proposal, some distinct unrelated literatures could be linked to each other by arguments that they treat. For instance, if a literature *A* (i.e. a set of all available records about *a* in the database serving as a source of data) reports about a term *a* being in association with a term *b*, and another literature *C* associates a term *c* with a term *b*, we can thus assume the literature *B* as an unintended implicit potential connection between the literatures *A* and *C*. For literature-based discovery, Smalheiser and Swanson [10] designed the ARROWSMITH system based on MEDLINE search. Their major focus was on the hypothesis testing approach [11], which Weeber et al. [12] defined as a *closed* discovery process (left model in Fig. 1), where both *a* and *c* have to be specified at the start of the process.

As reported by Weeber [13], Swanson's first literature-based hypothesis that dietary fish oil might benefit patients with Raynaud's disease [4], was a coincidence. Thereafter, Swanson studied the literatures of both target concepts, namely fish oil (*a*) and Raynaud's disease (*c*) for finding the linking terms (*b*) with already having this hypothesis in mind. On the other hand, an *open* discovery process (the right model in Fig. 1), is characterized by the absence of prespecified target concepts. If we are investigating a subject denoted with the term *c*, the open discovery starts with having only the term *c* and the corresponding set of articles in which term *c* appears (called also literature *C*), without knowing the target term *a*, which is discovered later as a result of this process.

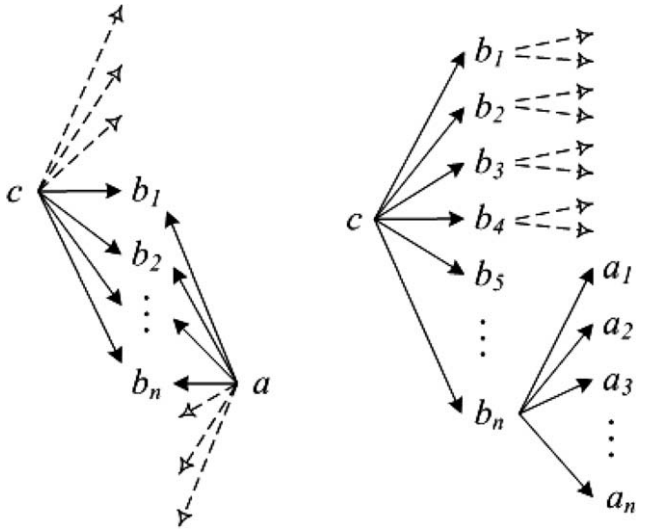

Fig. 1. Closed (left model) versus open (right model) discovery process as defined by Weeber et al. [11].

Several other text mining and hypothesis generating systems were developed following the early work of Swanson. Lindsay and Gordon [14] tested Swanson's discoveries of connecting Raynaud's disease to fish oil and migraine to magnesium deficiency by using different lexical statistics, such as word frequency counts. The authors tried to find relevant words on top of ranked lists in their open discovery approach and thus replicated Swanson's first two discoveries. However, their relative frequency statistic fail in suggesting magnesium and extensive analysis has to be based on human knowledge and judgement rather than automated procedures. Weeber [13] points out that the expert knowledge is indispensable in the literature-based discovery to choose among possible results and to determine potentially contradicting information. Weeber et al. [12] simulated the same two Swanson's discoveries with Natural Language Processing techniques by searching biomedical Unified Medical Language System concepts in texts. They developed a system for generating new hypotheses from the literature, called Literaby [13]. The Srinivasan and colleagues' [15] open discovery approach, on the other hand relies nearly completely on Medical Subject Headings (MeSH). But although they reduced the amount of manual effort and intervention, an important part of the process depends almost entirely on the user.

Some other researchers concentrated on literature mining and knowledge extraction from biomedical databases, too. Among them, Hristovski and Peterlin constructed a literature-based biomedical discovery support system, called BITOLA, and found the evidence for associations between several genes and diseases [16]. Another literature-based discovery system, named LitLinker, was presented by Yetisgen-Yildiz and Pratt [17], who combined knowledge-based methodologies with statistical methods to capture new connections between diseases and chemicals, genes or molecular sequences from biomedical literature. Both systems, BITOLA and LitLinker, like in Srinivasan and colleagues' case, also use MeSH descriptors as a representation of the MEDLINE documents, instead of using title or abstract words. Here, problems arise since some significant terminology from the subject content of a document may not be covered because MeSH indexers normally use only the most specific vocabulary terms to describe the topic discussed in a document [18].

3. RaJoLink method

In the approach presented in this paper we have expanded the Swanson's ABC model for literature mining by suggesting how terms *a* can be determined in a semi-automatic way. Although our literature-based discovery approach combines open and closed discovery processes (Fig. 2), the main focus is on the open discovery with a goal to identify the term *a*, which is represented in the upper half of Fig. 2.

The main emphasis in text mining investigation has been to directly exploit co-occurrence based relationships between MEDLINE documents [15]. The open discovery process in RaJoLink is, on the contrary, based on identifying rare terms within the literature *C*. Rare terms are our innovative pathways to an unknown term *a*. Each rare term in the RaJoLink method refers to a term that is exceptional for the literature on term *c*. More precisely, a term is rare if it appears in less than or equal to *n* records, *n* being a parameter that can be varied in the experiments. In our case, *n* was set to 1, because we are interested in only novel connections. The rationale to motivate our choice of *n* is that if a piece of information appears rarely in the set of articles, not many researchers are acquainted with it, so it might be worth exploring it further. Note that the role of rare terms in our approach does not correspond to the role of terms *b* as used in Swanson's ABC model for relating two disjoint literatures (literature *C* and literature *A*). In our case, rare terms are used to generate the term *a* as a joint term that is shared by two or more individual literatures on the selected rare terms.

The reasons for our focus on rare items within the literature *C* lay in the associationist creativity theory [19] with particular regard to the kinds of context-crossing associations, called bisociations [20]. Bisociation involves the literal processes of the mind when making entirely new connections among concepts from contexts or categories of objects that are normally considered separate categories. Throughout the history of science, this mechanism has been the essence of innovative insights and paradigm shifts. However, no comprehensive ICT methodology has yet been developed on this basis. Therefore we firmly believe that our methodology contributes to this particular approach to scientific discovery, which is based on an existing, but hitherto not computationally implemented notion of bisociation.

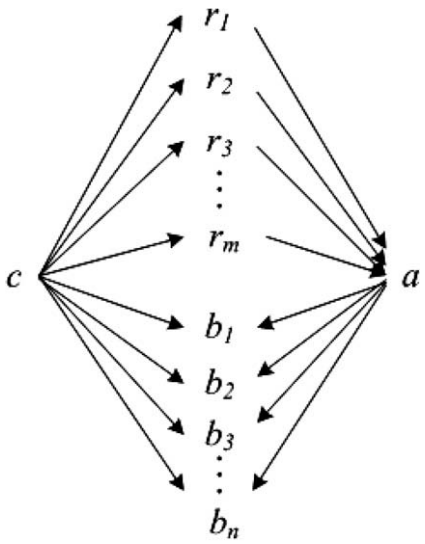

Fig. 2. Combined open and closed discovery process in the RaJoLink method. The upper half of the figure corresponds to the open discovery (identifying rare terms *r* and finding a joint term *a*) and the lower half to the closed discovery (searching for linking terms *b*).

Mednick [19] defined creative thinking as the ability to generate new combinations of distant associative elements (e.g., words). He explained that thinking of concepts, which are not strictly related to the elements under research, inspires unexpected useful connections between elements and thus considerably improves creative process. Actually, marginal observations are not necessarily characterized by mistakes or inaccuracies but may provide an indication of valuable information [21]. From this point of view, creative thinking constantly involves a process of evoking latent possibilities to discover new useful information and unforeseen knowledge.

Having disjoint literatures *A* and *C*, we are looking for linking terms *b* that are mentioned in both, the literature *A* as well as in the literature *C*. Pairs of documents with the same *b* are subject to closer inspection in order to find out whether by putting statements about *b* in these two articles together supports the hypothesis about a meaningful relation between *a* and *c*. In this manner, the closed discovery process in RaJoLink is based on the original Swanson's model. However, we contributed an innovative approach also to the closed discovery. In fact, our search for linking terms *b* is done in a semi-automated way that significantly reduces manual work and efficiently points to meaningful relations between the concepts *a* and *c* as described in [8,9] and in the Section 3.3 of this paper.

We have named the method RaJoLink after its key procedural elements: *Ra*re terms, *Jo*int terms, and *Link*ing terms. The method consists of three steps as presented in Fig. 3. Step *Ra* and step *Jo* together implement the open discovery process, while step *Link* implements the closed discovery. In the continuation we present these steps in more detail.

3.1. Step Ra

The aim of this step is to identify rare pieces of information in a set of documents about the term *c* (the literature *C*) in order to increase the chance of discovering useful implicit relations, which are still unpublished in the literature. Let us denote the total number of records in the input file with *N*, and with $n(T)$ the number of records that contain term *T*. In practice, the attention is focused on terms that rarely appear in the input set. A term appears rarely in the input set of records if it appears in a relatively small portion of them. Typically, we take that term *T* is rare if $n(T)$ equals 1. However, note that such constraint is quite sensitive to adding new articles to the input set of records. The term rareness or commonness in text corpus may change by adding new text to the existing input corpus. An infrequent term will become more frequent if the text that was added to the input file contains such term.

In order to make the search for rare terms effective, the preprocessing step also includes lemmatization, exclusion of words from a stoplist and filtering according to MeSH classification [18]. The presented method first compares three-word terms from the input text with the 2008 MeSH terms. If a multiple word term is found among MeSH terms, it is added to the list of terms for the further statistical analysis of strings. Afterwards, the same is executed with two-word terms from the input text. Multiple word terms that are not found in MeSH thesaurus are treated as individual words. We use the second-level categories from the 2008 MeSH tree structure (e.g., Behaviour and Behaviour Mechanisms—F01, Psychological Phenomena and Processes—F02, Mental Disorders—F03, Behavioural Disciplines and Activities—F04) to classify terms from the input text collection. Each of the second-level categories belongs to one of the top-level categories in the MeSH hierarchy, which are: Anatomy—A, Organisms—B, Diseases—C, Chemicals and Drugs—D, Analytical, Diagnostic and Therapeutic Techniques and Equipment—E, Psychiatry and Psychology—F, Biological Sciences—G, Natural Sciences—H, Anthropology, Education, Sociology and Social Phenomena—I,

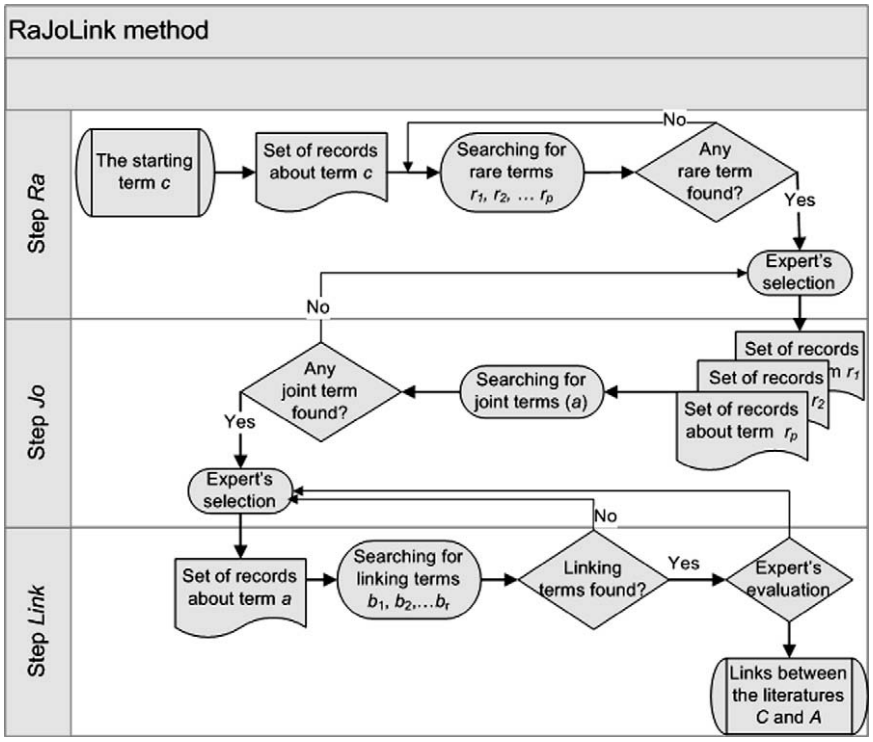

Fig. 3. Flow chart showing the procedures of the RaJoLink method.

Technology, Industry, Agriculture—J, Humanities—K, Information Science—L, Named Groups—M, Health Care—N, Publication Characteristics—V, Geographicals—Z. To implement filtering of terms there are alternative options that can be chosen by the user, which can be either one or more top-level and/or second-level categories from the MeSH hierarchy. Words that are of high interest in the medical science are included in Medical Subject Headings. Focusing on such words can narrow down the search space and, thus, speed-up and improve the inference process. On the other hand, words that are not part of the MeSH, are automatically added to the second-level general category, which we named Various—V05 and added to the second-level categories V01, V02, V03, and V04 of the top-level MeSH category: Publication Characteristics—V. For this reason we use the top-level category also named *Various* (Fig. 4) instead of the originally named top-level MeSH category (i.e. Publication Characteristics).

Lemmatization is used to eliminate various forms of a single word. We employed the lemmatizing software from the LemmaGen library developed by Juršič and colleagues [22]. Stoplists contain words that are predictably of no interest and should, therefore, be excluded from the input records. In this way, all the terms that are not subject-oriented should be ignored. We use a list of 571 English stop words (i.e. generic words such as *a*, *able*, *about*, *above*, and *according*).

For each term T that appears in the set of records, $n(T)$ is calculated using frequency statistics based on Bag of Words (BoW) text representation [23], wherefore we employed the Txt2Bow utility from the TextGarden library [24]. As previously stated, all terms with $n(T)$ equal to 1 are selected as rare. Let us denote such terms with r. After that, a subject expert has to indicate interesting rare terms $r_1, r_2, \ldots r_p$ out of them, with regard to the background knowledge about the subject of interest. Note that p can be regarded as a parameter of the method. When rare terms cannot be found by using this minimum frequency of term's occurrence, it is necessary to return back to the beginning of this step and repeat the process by taking into account larger number of input documents or by choosing higher value of the parameter $n(T)$. The appearance of the selected rare terms in the literature C, however, means that they had already been reported in the context related to term c. Therefore, the rare terms represent only intermediate results towards the new knowledge discovery. Neither should they be considered to have the same property as b terms, although the rare

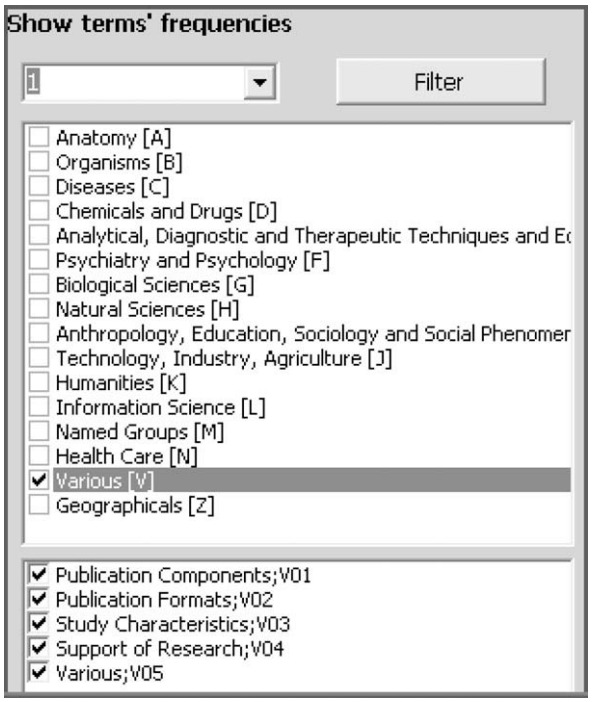

Fig. 4. A screenshot of RaJoLink system showing second-level MeSH categories (V01, V02, V03, VO4) and the general category of terms (V05) within their top-level category (V).

terms individually co-occur with the literature *A* and the literature *C*, but not with *A* and *C* jointly. The reason is that when testing the hypotheses by searching for meaningful linking terms *b* between the literatures *A* and *C*, the focus should be on the most frequent terms, while the rarest ones are interesting in the hypotheses generation phase.

3.2. Step Jo

When the data analyses described in the step *Ra* are completed, new individual sets of records, one for each selected rare term, have to be obtained and further analyzed. For this purpose, a set of articles about each of the rare terms selected in the step *Ra* are automatically retrieved from MEDLINE or extracted from other document source. After preprocessing each set of articles as in step *Ra,* the goal is to find the terms that the textual collections have in common. Again, BoW representation is used for the task. For taking into account multi-word terms, the maximal length of n-grams being 3 can be used as standard set of parameters for Txt2Bow utility [24]. A term from the BoW qualifies as a joint term if it appears in at least two sets of records about the rare terms. At the same time, it has to be absent from the set of records about the term *c*, generated in step *Ra*. The intermediate output of this step is, therefore, joint terms $a_1, a_2, \ldots, a_q$ as the intersection of the literatures on the rare terms. If joint terms are not obtained via the rare terms selected in the step *Ra*, it is necessary to return back to the results of the previous step and broaden or change the actual selection of rare terms and repeat the process.

The expert role is crucial also in the step *Jo*. Based on the expert's opinion, one of the proposed joint terms is selected for further investigation to be done in the step *Link*. On the basis of the selected joint term new hypotheses are formulated and subsequently tested following the Swanson's ABC model for closed discovery.

3.3. Step Link

In order to provide explanation for hypotheses generated in the step *Jo*, our method searches for links between the literature on joint term *a* and the literature on term *c*. This step is equivalent to Swanson's closed discovery [5]. Nevertheless, our closed discovery approach contains another unique aspect of our method in comparison to the literature-based discovery investigated by others. It is the focusing on neighbouring documents in the documents' similarity graph (Fig. 5), which is defined over the combined dataset consisting of literatures *A* and *C* as we extensively analyzed in our previous work [8,9].

Within the whole corpus of the textual dataset consisting of literatures *A* and *C*, which acts as input for step *Link*, each text document represents a single record. After preprocessing similar to that used in steps *Ra* and *Jo*, a single document is represented by a set of words using BoW representation. The appearance of co-occurring words is employed as a measure of content similarity. Its computation is performed with OntoGen, which was designed by Fortuna and colleagues for interactive data-driven construction of topic ontologies [25]. The content similarity is based on the textual description of documents and is measured using the standard TF∗IDF (term frequency inverse document frequency) weighting method [26]. This way, all the records are sorted according to similarity and the content related documents are obtained by comparing neighbouring documents from the list (Fig. 5). Since the search for the linking terms can be combinatorially complex [5,27], focusing on neighbouring documents can be used as a heuristic guidance to speed-up the process and alleviate the burden on the expert to sort out the meaningful explanations.

Some articles in MEDLINE are represented with titles, some with titles and abstracts, but only few of them have full text content available. Any of these articles' parts can be used as an input for RaJoLink. The choice of document parts that are taken into account when preparing the input for RaJoLink (e.g., titles, abstracts or full texts), has an important impact on the guidance of the entire text mining process. In this step, using abstracts rather than entire texts of articles is recommended. This recommendation can be argued by one of our previous studies [28], which pointed out that using only abstracts produced better results than using whole texts or only titles. Besides, the terminology used in abstracts has a stronger importance when compared to a randomly chosen part of the article's text corpus of the same size. In our supposition, the linking terms $b_1, b_2, \ldots b_r$ that are searched for by word intersections also have an even stronger connecting role between the two literatures this way.

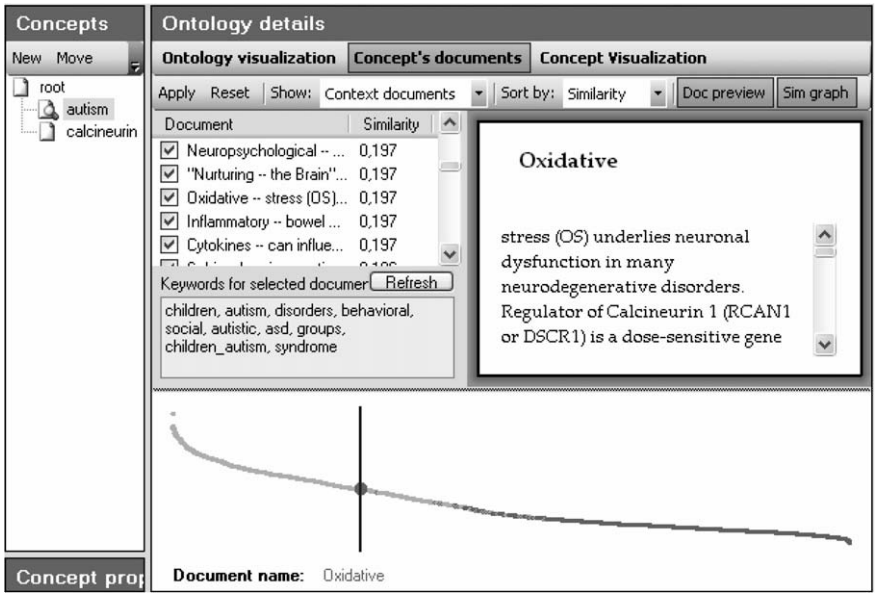

Fig. 5. OntoGen's similarity graph of a set of autism and calcineurin articles' abstracts. Two main article topics (*autism* and *calcineurin*) are listed on the left side of the window. As the autism topic is selected, the list of abstracts, which are in the relationship with this selected topic, is presented in the central part of the OntoGen's window. The distinctive calcineurin article (*Oxidative*) is visualized among the autism context documents.

Table 1
Schema of the RaJoLink method

Step	Input	Action	Tool, technique	Expert's involvement	Output
Ra	Set of records about *c*	1.1 Extraction of texts	Digital document archives		
		1.2 Data collection preprocessing	Word processing software		
		1.3 Identification of rare terms *r*	Word frequency statistics		
		1.4 Terms filtering	Content filtering	Indication of interesting rare terms	Rare terms $r_1, r_2, \ldots r_p$
Jo	Sets of records about $r_1, r_2, \ldots r_p$	2.1 Extraction of texts	Digital document archives		
		2.2 Data collections preprocessing	Word processing software		
		2.3 Search for joint terms	Word frequency statistics	Selection of a significant joint term	Joint term *a*
Link	Joint set of records about *a* and records about *c*	3.1 Extraction of texts	Digital document archives		
		3.2 Data collection preprocessing	Word processing software		
		3.3 Identification of content related A and *C* records	Text analysis		
		3.4 Search for linking terms *b*	Word intersection	Selection of meaningful linking terms	Linking terms $b_1, b_2, \ldots b_r$

The presented linking approach suggests the novel way to improve the evidence gathering phase when analyzing individual *a* terms in their potential connection with the term *c*. In fact, even Srinivasan and colleagues, who declared to have developed the algorithms that require the least amount of manual work in comparison with other studies [15], still need significant time and human effort for collecting evidence relevant to the hypothesized connections. In the comparable RaJoLink's approach, a subject expert should be involved only in the conclusive actions of the step Link to accelerate the choice of significant linking terms. The presented procedural elements of the steps *Ra*, *Jo*, and *Link* are summarized in Table 1.

4. Application of the RaJoLink method to the literature on autism

We have examined the method on the literature on autism. Autism is a complex neurodevelopmental disorder with the estimated prevalence of 5.8 per 1000 children [29]. It belongs to a group of pervasive developmental disorders that the fourth revised edition of Diagnostic and Statistical Manual of Mental Disorders categorizes as a group of symptoms of neurological development, associated with early brain mechanisms [30]. It is mainly manifested as impairment in social relatedness, communication and as repetitive routines and restricted interests [31].

An interaction of multiple risk factors is considered to contribute to the autistic disorders. However, the aetiologies of autism are still largely unknown. Distinctive neuropathological, genetic and environmental studies are of central interest to autism research. In our literature-based study of autism we have mainly focused our attention on biochemical substrates and neurological mechanisms.

Many mechanisms that might be lying behind neurological disorders of autistic patients have been examined and important advances have been recently made in understanding neural systems that process various types of information. Imaging and other examinations of the autistic brains have shown several brain irregularities, among them the altered neuroanatomy [32], as well as abnormal cellular neurochemistry [33]. Within this context, molecular changes in brain development and neurotransmission, morphological distinctions of particular neurons, and regional brain volume abnormalities have frequently been reported [34].

Various neurological conditions in autism reflect also in the heterogeneity of the possible neuropathological causes of this disorder. Therefore, we have started our autism research with an assumption that neurological substances, processes and transformations play a central role in the pathology of autism.

5. Experiment and results

We retrieved the documents for application of the RaJoLink method to the autism literature using the PubMed search engine. PubMed provides access to the U.S. National Library of Medicine's premiere bibliographic database MEDLINE, as well as to some additional sources of bibliographic information on diseases, public health, pharmacy, pharmacology and other biomedical topics. PubMed automatically retrieves and displays citations on the entered search terms. A single citation may include abstract, full text in PubMed Central or links to full text available elsewhere. We performed a PubMed search of articles on autism for the last 10 years and thus collected 214 documents with their full text available in the PubMed Central. We computed the preprocessing of documents by converting the HTML and PDF papers to text, and by deleting graphics, paragraph marks, and manual line breaks from the full text versions, so that each document occupied one record in the input file. Obtaining and handling of full texts of literature requires extra time in terms of locating and converting them into a plain text format. Unlike titles and abstracts, which are available in HTML, the full texts from the early issues of some journals are provided only as PDF files. Consequently, their handling requires extra processing time. Also, the full-length articles can be only accessed from journals that are available for free or via a particular subscription. However, the full-length biomedical articles contain an abundance of data and if user could capture important information from them, it is worth spending additional time on obtaining and processing such text.

After analyzing the literature on autism, we generated its statistics of terms and a compiled list of rare terms. About 2000 terms were fully automatically detected in the step *Ra*. Each of them appeared only in one of the documents from our set of 214 autism documents, which means that only terms with $n(T)$ equal to 1 were selected as rare. We disregarded the terms that are not subject-oriented, such as the words: *atomic*, *bundle*, *checkout*, and *dipper*. In our experiment we considered only the *D12* second-level category from the 2008 MeSH tree structure, i.e. *Amino Acids, Peptides, and Proteins* [18]. Thus we chose meaningful rare terms, as in our experimental case the words: *lactoylglutathione*, *synaptophysin*, and *calcium channels*.

In the step *Jo*, we collected the PubMed abstracts of articles about lactoylglutathione, synaptophysin and calcium channels literature, respectively. We searched for the terms that the three textual files have in common and thus singled out joint terms for the literatures on the three rare terms. From several joint terms that were found automatically, *calcineurin*, a protein phosphatase that is widely present in mammalian brain [35], was chosen for further investigation.

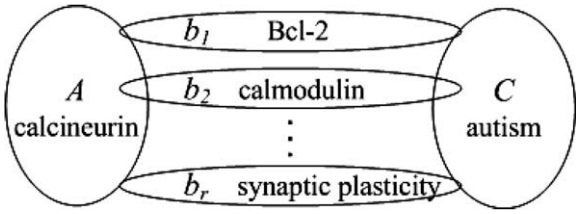

Fig. 6. Venn diagram of arguments (b_i) found as connection between the scientific literature on autism (C) and the scientific literature on calcineurin (A).

We began the final step of our method by retrieving abstracts of articles on autism as well as articles on calcineurin from PubMed database. In the combined set of literature on autism and literature on calcineurin, we were looking for exceptions within each of the two subgroups of literature, i.e. the calineurin articles among autism major groups of articles, and vice versa. In fact, according to the semantic similarity measure some articles on calcineurin were found in the subgroup of articles on autism. Similarly, there were some articles on autism in the subgroup of articles on calcineurin. Such exceptions led us to terms *Bcl-2*, *calmodulin*, *synaptic plasticity*, and 10 other linking terms between the literature on autism and the literature on calcineurin. Results are demonstrated according to Swanson's ABC model in a Venn diagram (Fig. 6).

Note that the reapplication of the RaJoLink method on a restricted set of records mentioning both, autism and fragile X, resulted in some other findings relevant for autism research. More concrete, as presented in [9], NF-kappaB was identified as a joint term with potential role in autism.

6. Evaluation of experimental results

6.1. Contribution to understanding of autism

The qualitative evaluation of the generated hypothesis that calcineurin is connected with autism was performed by a medical expert. She confirmed the results, which drew attention to interesting connections between two well developed, but not sufficiently connected fields [9]. In particular, she justified this statement by the recent calcineurin studies, indicating that calcineurin participates in intracellular signalling pathways that regulate synaptic plasticity and neuronal activities [36]. In addition, she commented that an impaired synaptic plasticity is thought to be also a consequence of the lack of FMR1 protein in fragile X syndrome, which is one of the identified causes of autism [37].

To the present, no direct evidence of calcineurin role in the autism phenomena has been reported on the Internet. Therefore, investigations of the calcineurin role in autism would be of great interest. Similarly, also the relation between autism and NF-kappaB, which was confirmed as relevant for the autism research in [9], would be interesting for further examinations.

6.2. Required human effort

Due to an enormous quantity of articles available on-line, literature-based knowledge discovery has become a very time consuming and laborious task. The RaJoLink method is intended to support experts on their way towards new discoveries. It covers both, open discovery and closed discovery processes. In both processes it tends to reduce the required human effort.

Swanson stated that in open discovery processes success depends entirely on the knowledge and ingenuity of the searcher [5]. The aim of RaJoLink is to reduce the search space, thus making the task easier for the searcher. At the same time, by focusing on rare terms the system identifies the candidates that are most likely to lead towards meaningful unpublished relations. This way, the system automatically produces intermediate results. Search space is further reduced by human choices of rare terms and joint terms. In addition, human involvement in these steps assures that search process concentrates on those parts of the search space that are interesting and meaningful for a subject expert. Based on these strategies, RaJoLink is designed to make the expert's involvement easy and efficient.

The search for the linking terms b in closed discovery process is combinatorially complex [5,27]. Here, RaJoLink gives priority to the terms b from two records, one from the set of records A and one from the set of records C, so that records in which b appears are close with respect to the similarity measure. The final choice of linking terms is, again, provided by a subject expert. Expert's explicit involvement in the process enables more focused and faster obtainment of results which are meaningful and interesting for further investigation. In the key steps of the process, RaJoLink supports the expert by listing term candidates sorted by estimated potential for the knowledge discovery.

7. Discussion

The huge sets of biomedical articles, which augment the space of possible hypotheses for problem solutions, represent a significant challenge in the field of biomedical discoveries based on literature. The vast amount of textual data might lead to an information overload. Extracting useful knowledge in the form of uncovered relations can therefore be very time consuming. To assist researchers in this process, we have proposed the method of combining word frequency statistics and semantic text analysis techniques following the Swanson's ABC model approach. When identifying plausible new relations across disjoint biomedical literatures, Swanson's investigation of the subject of interest prevailingly starts by already having a hypothesis [11], which is then tested in the closed discovery process. In the proposed RaJoLink method, we concentrate on open discovery, in which hypotheses are not known in advance. A fundamental difference from the previously proposed models of the open discovery approach and the unique contribution of the RaJoLink method therefore lies in the rarity principle that we apply to the open literature-based discovery. In fact, we use rare terms identified in the literature on c to guide the search for new hypotheses. We have applied various statistical and computational tools in innovative ways to speed-up and improve the knowledge discovery process.

The main novelty of our work is the development of a methodology for user-guided knowledge discovery using the rarity principle together with the notion of bisociation. Current literature-based approaches are limited as they depend heavily on simple, associative information search. Usually, association is computed using measures of similarity or co-occurrence. Association techniques have many advantages: they are effective in identifying related information and they are easy and efficient to implement. Consequently, they are a technique of choice for many applications. Even though, due to their 'hard-wired' underlying criteria of association or similarity, these methods fail to discover relevant information, which is not related in obvious associative ways. Especially information related across diverse contexts is hard to recognize with the conventional associative approach. In these cases the context-crossing connections, called bisociations, are often needed for creative, innovative discoveries.

Bisociative relationships can only be discovered on the basis of a sufficiently large and diverse underlying corpus of information. In our case this corpus are MEDLINE papers. The larger the corpus is, the more likely it is to contain bisociative relationships. The

RaJoLink approach, proposed in this paper, has the potential for bisociative relation discovery as it allows switching between contexts (papers from different areas) by exploring rare terms in the intersection between contexts.

As Swanson was interested in specific aspects of diseases research, namely the diet and food deficiency, our research approach has been driven by neurological concepts that in our case have served as a semantic filter at different stages of the experiment. Moreover, we have been able to additionally limit the search space in the discovery process by following only those terms, which rarely appeared in whole texts of the subject under investigation. Using whole texts of articles in this phase is therefore another difference between our proposed discovery process and Swanson's or Weeber's processes, since they used mostly titles and abstracts of articles.

In our training datasets we have managed to identify rare, joint and linking terms that have led us to discover the relations between calcineurin and autism. Although no direct assessment of calcineurin involvement in the phenomenon of autism has been published yet, we have been able to identify significant links between calcineurin and autism literature by combining the articles from two domains in a single set of literature and searching for semantic similarities among documents from such combined input set. However, we observe that it is difficult to isolate important domain concepts, such as neurological aspects, without proper background knowledge. Therefore, by interacting with a subject expert, the entire process of knowledge discovery can benefit in terms of speed and guidance towards meaningful solutions. Especially in the cases where no lexical classification is available for the researched domain, the subject expert can provide the necessary background knowledge for the identification of meaningful results.

8. Conclusions

In this article, an innovative method for literature-based discovery is presented. The method is named RaJoLink after its key procedural elements, which are: rare terms, joint terms and linking terms. The results of the initial experimental case studies suggest that the RaJoLink method can enhance an existing literature-based discovery. The main advantage of RaJoLink lies in the support of the open discovery processes by the innovative use of rare terms from the problem domain literature to guide the generation of new hypotheses. Therefore the crucial step of the method is in selecting rare terms. We managed to employ the rarity as a principle and a means to find new interesting pieces of knowledge that were previously available in the dispersed literature and could be linked together.

To make RaJoLink an easy-to-use tool for efficient knowledge discovery in biomedicine, our future work will include implementation of different visualisations of results. To provide more guidance to the users, the role of parameters values (such as the threshold frequency of terms to be marked as rare, the number of selected rare terms, etc.) will be systematically investigated. Another part of our further study will be the comparative application of the complete RaJoLink method, as well as of its different steps, on titles, abstracts and full texts as distinct parts of scientific articles. Furthermore, as RaJoLink method is generally applicable, we will test it also on other problem domains.

Acknowledgment

This work was partially supported by the Slovenian Research Agency program Knowledge Technologies (2004–2008).

References

[1] PubMed. Overview [Online], 2007. Available from: URL:<http://www.ncbi.nlm.nih.gov/>.
[2] Fayyad U, Piatetsky-Shapiro G, Smyth P. Knowledge discovery and data mining: towards a unifying framework. In: Proceeding of the second international conference on knowledge discovery and data mining, Oregon: Portland; 1996. p. 82–8.
[3] Shortliffe EH. The adolescence of AI in medicine: will the field come of age in the '90s? Atif Intell Med 1993;5(2):93–106.
[4] Swanson DR. Undiscovered public knowledge. Libr Q 1986;56(2):103–18.
[5] Swanson DR. Medical literature as a potential source of new knowledge. Bull Med Libr Assoc 1990;78(1):29–37.
[6] Belmonte MK, Allen G, Beckel-Mitchener A, Boulanger LM, Carper RA, Webb SJ. Autism and abnormal development of brain connectivity. J Neurosci 2004;24(42):9228–31.
[7] Zerhouni EA for National Institutes of Health and National Institute of Mental Health. Congressional Appropriations Committee Report on the State of Autism Research. Department of Health and Human Service, Bethesda MD, 2004.
[8] Petrič I, Urbančič T, Cestnik B. Discovering hidden knowledge from biomedical literature. Informatica 2007;31(1):15–20.
[9] Urbančič T, Petrič I, Cestnik B, Macedoni-Lukšič M. Literature mining: towards better understanding of autism. In: Bellazzi R, Abu-Hanna A, Hunter J, editors. AIME 2007. Proceedings of the 11th conference on artificial intelligence in medicine in Europe. The Netherlands: Amsterdam; 2007. p. 217–26.
[10] Smalheiser NR, Swanson DR. Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. Comput Methods Programs Biomed 1998;57(3):149–53.
[11] Swanson DR, Smalheiser NR, Torvik VI. Ranking indirect connections in literature-based discovery: the role of medical subject headings (MeSH). J Am Soc Inf Sci Tech 2006;57(11):1427–39.
[12] Weeber M, Vos R, Klein H, de Jong-van den Berg LTW. Using concepts in literature-based discovery: simulating Swanson's Raynaud–fish oil and migraine–magnesium discoveries. J Am Soc Inf Sci Tech 2001;52(7):548–57.
[13] Weeber M. Drug discovery as an example of literature-based discovery. In: Džeroski S, Todorovski L, editors. Computational discovery, 2007. p. 290–306.
[14] Lindsay RK, Gordon MD. Literature-based discovery by lexical statistics. J Am Soc Inf Sci 1999;50(7):574–87.
[15] Srinivasan P, Libbus B, Sehgal AK. Mining MEDLINE: postulating a beneficial role for curcumin longa in retinal diseases. In: BioLINK 2004: linking biological literature, ontologies, and databases, Boston, MA. p. 33–40.
[16] Hristovski D, Peterlin B, Mitchell JA, Humphrey SM. Using literature-based discovery to identify disease candidate genes. Int J Med Inform 2005;74(2–4):289–98.
[17] Yetisgen-Yildiz M, Pratt W. Using statistical and knowledge-based approaches for literature-based discovery. J Biomed Inform 2006;39(6):600–11.
[18] Nelson SJ, Johnston D, Humphreys BL. Relationships in medical subject headings. In: Bean CA, Green R, editors. Relationships in the organization of knowledge. New York: Kluwer Academic Publishers; 2001. p. 171–84.
[19] Mednick SA. The associative basis of the creative process. Psychol Rev 1962;69(3):220–32.
[20] Koestler A. The act of creation. New York: MacMillan Company; 1964.
[21] Barnett V, Lewis T. Outliers in statistical data. New York: Wiley; 1994.
[22] Juršič M, Mozetič I, Lavrač N. Learning ripple down rules for efficient lemmatization. In: IS 2007. Proceedings of the 10th international multiconference information society. Slovenia: Ljubljana; 2007. p. 206–9.
[23] Sebastiani F. Machine learning in automated text categorization. ACM Comput Surv 2002;34(1):1–47.
[24] Grobelnik M, Mladenić D. Extracting human expertise from existing ontologies. In: EU-IST Project IST-2003-506826 SEKT, 2004.
[25] Fortuna B, Grobelnik M, Mladenić D. Semi-automatic data-driven ontology construction system. In: Bohanec M, Gams M, Rajkovič V, Urbančič T, Bernik M, Mladenić D, et al., editors. IS-2006. Proceedings of the 9th international multi-conference information society. Slovenia: Ljubljana; 2006. p. 223–6.
[26] Salton G, Buckley C. Term weighting approaches in automatic text retrieval. Info Process Manage 1988;24(5):513–23.
[27] Hearst MA. Untangling text data mining. In: Dale R, editor. Proceedings of the 37th annual meeting of the association for computational linguistics. San Francisco, CA: Morgan Kaufmann Publishers; 1999. p. 3–10.
[28] Petrič I, Urbančič T, Cestnik B. Comparison of ontologies built on titles, abstracts and entire texts of articles. In: Bohanec M, Gams M, Rajkovič V, Urbančič T, Bernik M, Mladenić D, et al., editors. IS-2006. Proceedings of the 9th international multi-conference information society. Slovenia: Ljubljana; 2006. p. 227–30.
[29] Hirtz D, Thurman DJ, Gwinn-Hardy K, Mohamed M, Chaudhuri AR, Zalutsky R. How common are the "common" neurologic disorders? Neurology 2007;68(5):326–37.
[30] American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4th ed. Text Revision, Washington, DC. 2000.
[31] Georgiades S, Szatmari P, Zwaigenbaum L, Duku E, Bryson S, Roberts W, et al. Structure of the autism symptom phenotype: a proposed multidimensional model. J Am Acad Child Adolesc Psychiatry 2007;46(2):188–96.
[32] Bauman ML, Kemper TL. Neuroanatomic observations of the brain in autism: a review and future directions. Int J Dev Neurosci 2005;23(2–3):183–7.

[33] DeVito TJ, Drost DJ, Neufeld RW, Rajakumar N, Pavlosky W, Williamson P, et al. Evidence for cortical dysfunction in autism: a proton magnetic resonance spectroscopic imaging study. Biol Psychiatry 2007;61(4):465–73.
[34] Bethea TC, Sikich L. Early pharmacological treatment of autism: a rationale for developmental treatment. Biol Psychiatry 2007;61(4):521–37.
[35] Rusnak F, Mertz P. Calcineurin: form and function. Physiol Rev 2000;80(4):1483–521.
[36] Qiu S, Korwek KM, Weeber EJ. A fresh look at an ancient receptor family: emerging roles for low density lipoprotein receptors in synaptic plasticity and memory formation. Neurobiol Learn Mem 2006;85(1):16–29.
[37] Irwin S, Galvez R, Weiler IJ, Beckel-Mitchener A, Greenough W. Brain structure and the functions of FMR1 protein. In: Hagerman RJ, Hagerman PJ, editors. Fragile X syndrome. Baltimore: The Johns Hopkins University Press; 2002. p. 191–205.