Literature Mining: Towards Better Understanding of Autism

Tanja Urbančič[1,2], Ingrid Petrič[1], Bojan Cestnik[2,3], and Marta Macedoni-Lukšič[4]

[1] University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia
[2] Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[3] Temida, d.o.o., Dunajska 51, 1000 Ljubljana, Slovenia
[4] University Children's Hospital, University Medical Center, 1000 Ljubljana, Slovenia
tanja.urbancic@p-ng.si, ingrid.petric@p-ng.si,
bojan.cestnik@temida.si,
marta.macedoni-luksic@mf.uni-lj.si

Abstract. In this article we present a literature mining method RaJoLink that upgrades Swanson's ABC model approach to uncovering hidden relations from a set of articles in a given domain. When these relations are interesting from medical point of view and can be verified by medical experts, they represent new pieces of knowledge and can contribute to better understanding of diseases. In our study we analyzed biomedical literature about autism, which is a very complex and not yet sufficiently understood domain. On the basis of word frequency statistics several rare terms were identified with the aim of generating potentially new explanations for the impairments that are observed in the affected population. Calcineurin was discovered as a joint term in the intersection of their corresponding literature. Similarly, NF-kappaB was recognized as a joint term. Pairs of documents that point to potential relations between the identified joint terms and autism were also automatically detected. Expert evaluation confirmed the relevance of these relations.

Keywords: literature mining, knowledge discovery, biomedical literature, autism.

1 Introduction

The amount and the speed of growth of scientific information available online have strongly influenced the way of work in the research community which calls for new methods and tools to support it. Biomedical field is a very good example, with MEDLINE database, the primary component of PubMed (the United States National Library of Medicine's bibliographic database), which covers more than 5.000 journals published in more than 80 countries, contains more than 15 million citations from the mid-1950's to the present, and increases for more than 1.500 complete references daily [15]. Knowledge technologies, especially knowledge discovery based on data mining and text mining, offer new possibilities by their ability to uncover hidden relationships in data [7]. Several examples of the applications in the biomedical field are included into a presentation of European data mining projects given in [20]. When

R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.): AIME 2007, LNAI 4594, pp. 217–226, 2007.

a set of articles serves as a source of data, the process is typically called literature mining.

An early and very illustrative example of literature mining goes back into 1990, when Swanson presented his ABC model for discovering complementary structures in disjoint journal articles, leading to new hypotheses about diseases [19]. In his work he investigates whether an agent A influences a phenomenon C. To do this, he looked for interconnecting Bs, such that A causes phenomenon B (as reported in an article in literature about A) and the same B influences C (as reported in another article in literature about C). If articles about A (called also A literature) and articles about C (C literature) have few or no published papers in common, such discovered connections can turn out to be previously unknown. If they are also interesting from a medical point of view and can be verified by medical experts, they represent new pieces of knowledge and contribute to better understanding of diseases. This is particularly important in the case of complex pathological conditions, not yet sufficiently understood. To facilitate the discovery of hypotheses by linking findings across literature, Swanson and his colleagues designed a set of interactive software that is available on a web-based system called Arrowsmith [18].

Swanson's work inspired several researchers that continued his line of research. Pratt and Yetisgen-Yildiz [14] designed LitLinker that uses data mining techniques to identify correlations among concepts and then uses those correlations for discovery of potential causal links between biomedical terms. Weeber et al. [22] experimented with Swanson's idea of searching the literature for generating new potential therapeutic uses of the drug thalidomide with the use of a concept-based discovery support system DAD on the scientific literature. Another example of discovering new relations from bibliographic database according to Swanson's model is identification of disease candidate genes by an interactive discovery support system for biomedicine Bitola [9].

In Swanson's approach, either a specific agent A or a more general A category has to be determined in advance, so a target relationship that will be checked as a hypothesis has to be set before the process. In our work described in this article, we wanted to broaden the usability of the approach by suggesting how candidates for A can also be determined in a semi-automatic way. Our approach is based on identification of rare terms that could provide new explanations for the observed phenomena. The system automatically produces intermediate results, but is driven by human choices which rely on background knowledge of the domain. Expert's explicit involvement in the process enables for more focused and faster obtainment of results that experts find interesting and meaningful.

For our testing domain we chose autism. Autism belongs to a group of pervasive developmental disorders that are characterized by early delay and abnormal development of cognitive, communication and social interaction skills of a person. In most cases, these disorders have unclear origin. According to the Autism Society of America, autism is now considered to be an epidemic. The increase in the rate of autism revealed by epidemiological studies and government reports implicates the importance of external or environmental factors that may be changing. There's also an enormous increase of information in the field of autism research, which too often seems a fragmented tapestry stitched from differing analytical threads and theoretical

patterns. This is the reason why studies seeking for factors that can help to put pieces together into a single, coherent object are so important.

In the next section we give a brief overview of data and tools that were used in our experiments. In section 3 we present a method RaJoLink which for a given phenomenon C generates candidate agents A that could contribute to better understanding of C and gives pairs of articles connecting C and A via linking terms B as a basis for expert evaluation of discovered relations. Results in the autism domain are also presented. In Section 4 we apply the same method on a domain, restricted as suggested by a medical expert. Finally, we summarize the most important findings.

2 Brief Overview of Experimental Set-Up: Data and Tools

In our study, articles about autism in the PubMed database served as a source of data. Among 10.821 documents (found till August 21, 2006) that contained derived forms of autis*, the expression root for autism, there were 354 articles with their entire text published in the PubMed Central database. Due to a noticeable shift in the focus of investigations about autism, we further restricted this set of articles to those published in the last ten years, which resulted in a set of 214 articles used as an initial source of data in our study. Later in the experiment, other subsets of PubMed articles were selected as a source of data, as described in sections 3 and 4.

In our work we used OntoGen [8], the interactive tool for semi-automatic construction of ontologies, accessible with detailed description at http://ontogen.ijs.si/index.html. OntoGen is based on machine learning and text mining techniques that automatically extract topics covered within the input documents and thus support the user of the system to organize those documents into a topic ontology. This primary functionality of OntoGen helped us to obtain an overview of the fundamental concepts of autism domain knowledge, but was in our case not crucial for the sake of knowledge discovery. We rather used OntoGen's other functionalities, such as generation of word frequency statistics and determination of document similarity on the basis of bag-of-words representation and cosine similarity.

3 Method RaJoLink

The proposed method for uncovering relations that could contribute to understanding of a phenomenon C (in our case, autism) consists of the following steps:

1. Identification of *n* interesting rare terms C_R_1, C_R_2, …C_R_n in literature about C.
2. Search for a joint term A in the literatures about C_R_1, C_R_2, …C_R_n.
3. Search for linking terms B_1, B_2, …., B_m such that for each B_i there exists a pair of articles, one from literature A and one from literature C, both mentioning Bi.
4. Expert evaluation in which it is checked whether obtained pairs of articles contribute to better understanding of C.

We call the method RaJoLink after its key elements: *Ra*re terms, *Jo*int terms and *Link*ing terms. In the following sections we describe these steps in more detail and illustrate them by the autism example.

3.1 Identification of Interesting Rare Terms in the Domain Literature

From the processed text file of autism articles we obtained also a *.txt.stat file with statistical incidence of terms as they appeared in the input documents collection. Initially we took the autism.txt.stat file that OntoGen formed while constructing autism ontologies and allocated our attention to the rarest records. This simple retrieval technique enables quick identification of interesting items.

Focusing on interesting terms that are very rarely mentioned in articles about autism is in our view more promising for new discoveries than exploring terms that are more frequent. With the proposed strategy we individuated a list of rare words that were, in regard of our autism background knowledge, viewed as promising in the sense that they could provide potential explanations of autistic disorders. From the 2192 rarest terms (terms that appeared once) all terms that were not domain specific were left out. From the remaining rare terms, 3 of them were chosen for further investigation: *lactoylglutathione, synaptophysin* and *calcium_channels*. Note that at the point of selecting interesting rare terms an expert's opinion might be very valuable.

To confirm the rarity of the chosen terms in the autism context, we searched the PubMed database for documents that contain the specific rare term together with the term autism. In fact, the term *lactoylglutathione* appeared only once together with the term autism, *synaptophysin* twice, and *calcium_channels* seven times.

3.2 Search for Joint Terms in the Literature About Rare Terms

Rare terms identified in the previous step served as a starting point for our deeper investigations of some pathological mechanisms that may lead to the autistic-like manifestations. Therefore we decided to search for the biomedical literature about *lactoylglutathione*, about *synaptophysin* and about *calcium_channels* that is publicly accessible in the PubMed database. As a result, we got three different sets of biomedical articles that were converted into three separate text files: the first one containing abstracts of *lactoylglutathione* articles, the second one with abstracts of *synaptophysin* articles and the third text file with abstracts of *calcium_channels* articles from PubMed. Further search in these domains enabled us to determine joint terms that appeared in all of them.

To find joint terms, we again used OntoGen, which besides constructing ontologies on the three input files of abstracts, created also three *.txt.stat files. Each of the three *.txt.stat files contained the statistical incidence of terms as they appeared in the processed documents from each of the input datasets. From the statistical files taken together, we identified joint terms that appeared in all of them. In other words, these joint terms appear in the lactoylglutathione literature, in the synaptophysin literature as well as in the calcium_channels literature. From several joint terms that were found automatically, *calcineurin* was chosen for further investigation.

The reasoning behind this step is the following: If there are some rare terms that appear in the autism literature and they all have a joint term in their intersection, it is worthwhile checking if this joint term has some connections to autism. If the autism literature and the joint term literature have few or no published papers in common, such an uncovered connection might contribute to better understanding of autism.

3.3 Search for Linking Terms and the Corresponding Pairs of Articles

For detecting some pairs of calcineurin-autism articles that would help us to build prominent relationships between the domain of calcineurin and the domain of autism, we used a set of calcineurin articles abstracts and a set of autism articles abstracts, which we extracted from PubMed. We united the two sets of abstracts in a single dataset and thus generated a database of calcineurin+autism domain. We used this combined dataset as input for OntoGen and built a new ontology on it in order to obtain the information about the similarity between documents from the input dataset.

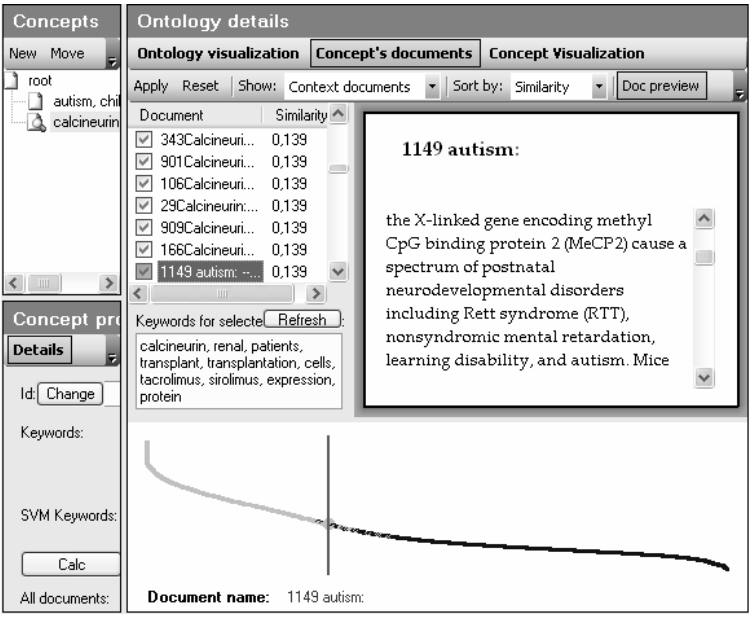

Fig. 1. OntoGen's representation of the set of autism and calcineurin articles according to their similarities. Two main topics (*autism and calcineurin*) are listed on the left side of the OntoGen's window. As the calcineurin topic is selected, the list of documents that are in the relationship with it is presented in the central part of the window. An outlying autism article (*1149 autism*) can be viewed inside the calcineurin context documents due to its similarity with the neighboring documents.

From the OntoGen's similarity graph (Figure 1) we could quickly notice, which documents are semantically strongly related to each of our research domains, autism and calcineurin, respectively, because they were clearly positioned on the two opposite sides of the similarity curve. However, regarding our goal of looking for relations between our two domains of research, the most prominent examples from the input dataset should be positioned on those graph sides, where the autism articles lay near the calcineurin articles. Therefore we focused our attention on the groups of the calcineurin-autism articles that were positioned in the vicinities according to their similarity. In this way we obtained pairs of calcineurin-autism articles containing

terms with similar meanings. We used such terms as a hypothetical conjunct of calcineurin and autism domain. As the candidate hypotheses for calcineurin and autism relationship we found thirteen pairs of PubMed articles that, when put together, could connect the two categories, autism and calcineurin, respectively. Three of such pairs of articles are listed in Table 1.

This step of the method was inspired by the Swanson's ABC approach. Note that instead of determining agent A in advance, in our approach it is generated as a joint term in the second step of the method.

Table 1. Hypotheses for autism and calcineurin relationship

Autism literature	Calcineurin literature
Fatemi et al. [6] reported a reduction of *Bcl-2* (a regulatory protein for control of programmed brain cell death) levels in autistic cerebellum.	Erin et al. [5] observed that calcineurin occurred as a complex with *Bcl-2* in various regions of rat and mouse brain.
Belmonte et al. [3] reviewed neuropathological studies of cerebral cortex in autism indicating abnormal *synaptic* and columnar structure and neuronal migration defects.	Chen et al. [4] reported about the decrease in protein ubiquitination in synaptosomes and in nonneuronal cells that may play role in the regulation of *synaptic* function by a calcineurin antagonist FK506.
Huber et al. [10] showed evidences about an important functional role of fragile X protein, an identified cause of autism, in regulating activity-dependent *synaptic plasticity* in the brain.	Winder and Sweatt [23] described the critical role of protein phosphatase 1, protein phosphatase 2A and calcineurin in the activity-dependent alterations of *synaptic plasticity.*

3.4 Expert's Comment

The medical expert in our team confirmed that the generated relations draw attention to interesting connections between two well developed, but not sufficiently connected fields. In particular, she justified this statement by the following comment: Calcineurin is a calcium/calmodulin-dependent protein phosphatase [17]. Recent studies indicate that it participates in intracellular signaling pathways, which regulate synaptic plasticity and neuronal activities [16]. 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 [11].

4 Re-application of RaJoLink on a Restricted Domain

The four steps described in Section 4 resulted in uncovered relations that according to the expert evaluation could present a contribution towards better understanding of autism.

In addition, it should be mentioned that a full table with identified pairs of related articles proved to be very useful in our dialog with the domain expert since it guided the discussion very efficiently towards new ideas for further investigations. More concretely, the suggestion was to have a closer look at the significance of the fragile

X protein loss in autism as reported by Huber et al. [10]. This evaluation significantly helped us in reducing the hypothesis space. It encouraged us to further mine the data on autism in its particular relation to the fragile X. We did it by reapplying RaJoLink method, this time on a restricted set of articles that dealt with both, autism and fragile X, as follows.

With the goal to discover unsuspected associations between pieces of knowledge about autism and fragile X, we retrieved articles from PubMed that contain information about autism and that at the same time talk about the fragile X. We found 41 articles with their entire text published in the PubMed Central, which served as our input file of data on autism and fragile X. As in the case of our data mining on pure autism articles, we used them for ontology construction with OntoGen, and next, we utilized the OntoGen's statistical *.txt.stat file for the identification of those terms that rarely appeared in the PubMed documents collected in our input dataset.

When searching the statistical file we concentrated our attention on listed terms that appeared only in one document from the input dataset. In this way we chose some of these rare terms for the following text mining on the smallest pieces of autism and fragile X knowledge. The following three were chosen based on background knowledge: *BDNF (brain-derived neurotrophic factor), bicuculline*, and *c-Fos*. By searching in PubMed articles that treat each of the three selected terms domains, we constructed three separate ontologies. Afterwards, we searched the OntoGen's *.txt.stat files to find some interesting words that the listed domains have in common as joint terms. We found several promising terms belonging to three of the domains. One of such terms, which we found in the BDNF*.txt.stat file, in the bicuculline*.txt.stat file, as well as in the c-Fos*.txt.stat file, was the term *NF-kappaB*. Figure 2 illustrates how resulting joint terms were obtained for autism domain and for autism+fragile_X domain.

For a given hypotheses of NF-kappaB and autism relationship we found pairs of PubMed articles that could connect the knowledge about the transcription factor NF-kappaB with the domain of autism. We present three of such pairs in Table 2.

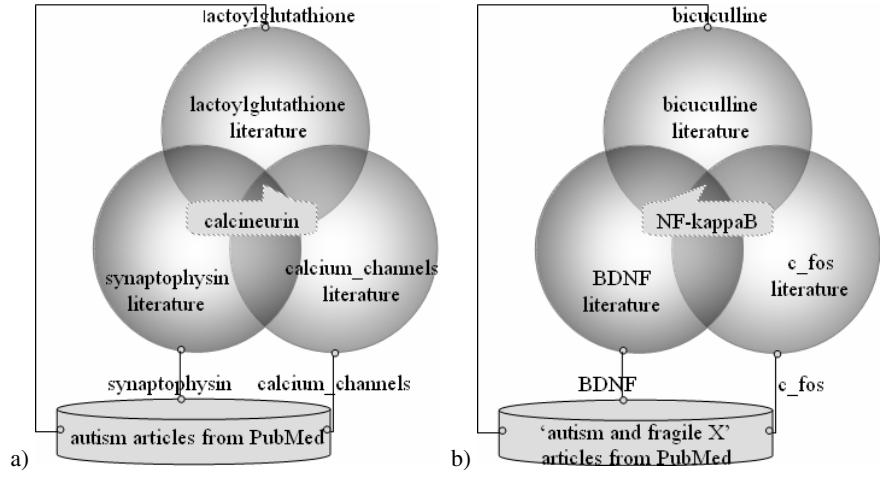

Fig. 2. Results obtained on (a) autism domain, and (b) autism+fragile_X domain

Table 2. Hypotheses for autism and NF-kappaB relationship

Autism literature	NF-kappaB literature
Araghi-Niknam and Fatemi [2] showed reduction of *Bcl-2*, an important marker of apoptosis, in frontal, parietal and cerebellar cortices of autistic individuals.	Mattson [12] reported in his review that activation of NF-kappaB in neurons can promote their survival by inducing the expression of genes encoding antiapoptotic proteins such as *Bcl-2* and the antioxidant enzyme Mn-superoxide dismutase.
Vargas et al. [21] reported altered *cytokine* expression profiles in brain tissues and cerebrospinal fluid of patients with autism.	Ahn and Aggarwal [1] reported that on activation NF-kappaB regulates the expression of almost 400 different genes, which include enzymes, *cytokine*s (such as TNF, IL-1, IL-6, IL-8, and chemokines), adhesion molecules, cell cycle regulatory molecules, viral proteins, and angiogenic factors.
Ming et al. [13] reported about the increased urinary excretion of an *oxidative stress* biomarker - 8-iso-PGF2alpha in autism.	Zou and Crews [24] reported about increase in NF-kappaB DNA binding following *oxidative stress* neurotoxicity.

The expert's comment to these finding was as follows. It is thought that autism could result from an interaction between genetic and environmental factors with an oxidative stress and immunological disorders as potential mechanisms linking the two [3], [13]. Both of the mechanisms are related to NF-kappaB as the result of our analysis. The activation of the transcriptional factor NF-kappaB was shown to prevent neuronal apoptosis in various cell cultures and in vivo models [12]. Oxidative stress and elevation of intracellular calcium levels are particularly important inducers of NF-kappaB activation. In addition, various other genes are responsive to the activation of the NF-kappaB, including those for cytokines. In this way the NF-kappaB can be involved in the complex linkage between the immune system and autism [3], [21]. So, according to our analysis one possible point of convergence between "oxidative stress" and "immunological disorder" paradigm in autism is NF-kappaB.

5 Conclusions

We present a literature mining method for searching pairs of papers in disjoint literatures that could, when linked together, contribute to a better understanding of complex pathological conditions, such as autism. We focus on rare terms to generate potentially new explanations for the impairments that are observed in the affected population. With this goal we further review the main aspects of the chosen rare terms with the ontology construction on each of these starting point domains. It is on these latter aspects that we focus furthermore, as we attempt to investigate whether any of chosen rare terms relate to each other. In fact, our assumption is that such known relations might lead us to discovering implicit knowledge about autism in previously unrelated biomedical literature. With the calcineurin and NF-kappaB examples we finally illustrate the potential of literature mining to detect links between unrelated biomedical articles.

By detecting published evidence of some autism findings on one hand that coincide with specific calcineurin and NF-kappaB observations on the other hand, we present possible relationships between autism and calcineurin literature, as well as between autism and NF-kappaB literature. Further research about timing, environmental conditions, maturational differences in brain development, and other determinants of calcineurin and NF-kappaB involvement in autism spectrum disorders is needed for stronger evidence, but in any case, the method has proved its potential in supporting experts on their way towards new discoveries in biomedical field.

Acknowledgments. This work was partially supported by the Slovenian Research Agency program Knowledge Technologies (2004-2008). We thank Nada Lavrač for her suggestion to use OntoGen and Blaž Fortuna for his discussions about OntoGen's performance.

References

1. Ahn, K.S., Aggarwal, B.B.: Transcription Factor NF-{kappa}B: A Sensor for Smoke and Stress Signals. Annals of the New York Academy of Sciences 1056, 218–233 (2005)
2. Araghi-Niknam, M., Fatemi, S.H.: Levels of Bcl-2 and P53 are altered in superior frontal and cerebellar cortices of autistic subjects. Cellular and Molecular Neurobiology 23(6), 945–952 (2003)
3. Belmonte, M.K., Allen, G., Beckel-Mitchener, A., Boulanger, L.M., Carper, R.A., Webb, S.J.: Autism and abnormal development of brain connectivity. The Journal of Neuroscience 27(42), 9228–9231 (2004)
4. Chen, H., Polo, S., Di Fiore, P.P., De Camilli, P.V.: Rapid Ca2+-dependent decrease of protein ubiquitination at synapses. Proceedings of the National Academy of Sciences of the United States of America 100(25), 14908–14913 (2003)
5. Erin, N., Bronson, S.K., Billingsley, M.L.: Calcium-dependent interaction of calcineurin with Bcl-2 in neuronal tissue. Neuroscience 117(3), 541–555 (2003)
6. Fatemi, S.H., Stary, J.M., Halt, A.R., Realmuto, G.R.: Dysregulation of Reelin and Bcl-2 proteins in autistic cerebellum. Journal of Autism and Developmental Disorders 31(6), 529–535 (2001)
7. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: Knowledge Discovery and Data Mining: Towards a Unifying Framework. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, pp. 82–88 (1996)
8. Fortuna, B., Grobelnik, M., Mladenić, D.: Semi-automatic Data-driven Ontology Construction System. In: Proceedings of the 9th International multi-conference Information Society IS-2006, Ljubljana, Slovenia, pp. 223–226 (2006)
9. Hristovski, D., Peterlin, B., Mitchell, J.A., Humphrey, S.M.: Using literature-based discovery to identify disease candidate genes. International Journal of Medical Informatics 74, 289–298 (2005)
10. Huber, K.M., Gallagher, S.M., Warren, S.T., Bear, M.F.: Altered synaptic plasticity in a mouse model of fragile X mental retardation. Proceedings of the National Academy of Sciences of the United States of America 99(11), 7746–7750 (2002)
11. Irwin, S., Galvez, R., Weiler, I.J., Beckel-Mitchener, A., Greenough, W.: Brain structure and the functions of FMR1 protein. In: Hagerman, R.J., Hagerman, P., J. Fragile X syndrome. The Johns Hopkins University Press, Baltimore, pp. 191–205 (2002)

12. Mattson, M.P.: NF-kappaB in the survival and plasticity of neurons. Neurochemical Research 30(6-7), 883–893 (2005)
13. Ming, X., Stein, T.P., Brimacombe, M., Johnson, W.G., Lambert, G.H., Wagner, G.C.: Increased excretion of a lipid peroxidation biomarker in autism. Prostaglandins, Leukotrienes, and Essential Fatty Acids 73(5), 379–384 (2005)
14. Pratt, W., Yetisgen-Yildiz, M.: LitLinker: Capturing Connections across the Biomedical Literature. In: Proceedings of the International Conference on Knowledge Capture (K-Cap'03), Florida, pp. 105–112 (2003)
15. PubMed: Overview (January 2007) http://www.ncbi.nlm.nih.gov/
16. Qiu, S., Korwek, K.M., Weeber, E.J.: A fresh look at an ancient receptor family: emerging roles for density lipoprotein receptors in synaptic plasticity and memory formation. Neurobiology of Learning and Memory 85(1), 16–29 (2006)
17. Rusnak, F., Mertz, P.: Calcineurin: Form and Function. Physiological Reviews 80(4), 1483–1521 (2000)
18. Smalheiser, N.R., Swanson, D.R.: Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. Computer Methods and Programs in Biomedicine 57, 149–153 (1998)
19. Swanson, D.R.: Medical literature as a potential source of new knowledge. Bulletin of the Medical Library Association 78(1), 29–37 (1990)
20. Van Someren, M., Urbančič, T.: Applications of machine learning: matching problems to tasks and methods. The Knowledge Engineering Review 20(4), 363–402 (2006)
21. Vargas, D.L., Nascimbene, C., Krishnan, C., Zimmerman, A.W., Pardo, C.A.: Neuroglial activation and neuroinflammation in the brain of patients with autism. Annals of Neurology 57(1), 67–81 (2005)
22. Weeber, M., Vos, R., Klein, H., De Jong-van den Berg, L.T., Aronson, A.R., Molema, G.: Generating Hypotheses by Discovering Implicit Associations in the Literature: A case Report of a Search for New Potential Therapeutic Uses for Thalidomide. Journal of the American Medical Informatics Association 10(3), 252–259 (2003)
23. Winder, D.G., Sweatt, J.D.: Roles of serine/threonine phosphatases in hippocampal synaptic plasticity. Nature reviews Neuroscience 2(7), 461–474 (2001)
24. Zou, J., Crews, F.: CREB and NF-kappaB Transcription Factors Regulate Sensitivity to Excitotoxic and Oxidative Stress Induced Neuronal Cell Death. Cellular and Molecular Neurobiology 26(4-6), 383–403 (2006)