Intelligent Systems for Knowledge Discovery in Biomedical Field

Tanja Urbancic
University of Nova Gorica, Nova Gorica, Slovenia and Jozef Stefan Institute, Ljubljana, Slovenia

In cooperation with:
Ingrid Petric
Bojan Cestnik
Marta Macedoni-Luksic

Summer School on Intelligent Systems, University of Cyprus, July 2-6, 2007

Motivation

Better understanding of diseases
Risk detection and early diagnosis
Better treatment of patients

For a health system:
- **more efficient, less expensive**

For a patient:
- **better healing**
- **better quality of life**

Some problems...

- for some diseases and disorders, causes and risk factors are not known
- some diseases have very diverse symptoms and/or courses of development
- several diseases share similar symptoms
- some cases are very rare
- some investigations are very expensive or invasive
- some cures are very expensive or have severe side-effects
- …

Towards solutions...

How medicine copes with these problems:

- epidemiologic studies
- studies related to genetics
- development of new diagnostic methods
- development of new cures and drugs
- ...

Knowledge technologies can help by

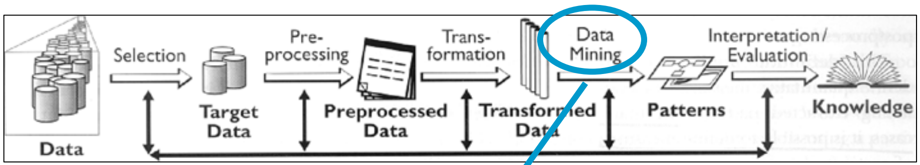
- development of tools for accessing the data, data analysis, knowledge discovery and decision support

Knowledge Discovery in Databases (KDD)
(Fayyad et al., 1996)

KDD - a process consisting of the following steps:

- **understanding the domain**
- **forming the data set and cleaning the data**
- **extracting regularities hidden in the data (formulating knowledge in the form of patterns or models)**
- **postprocessing of discovered knowledge**
- **exploiting the results**

Knowledge Discovery in Databases (KDD): process of identifying valid, novel, potentially useful and understandable patterns in data



- **Data Mining (DM):**
  - **a way of doing data analysis, aimed at finding patterns, revealing hidden regularities and relationships in the data**

Data mining in medicine

- Large quantities of data are collected
- Most often
  - predictive DM used for classification models (for diagnosis, prognosis, treatment planning)
  - data (represented in tables) collected from measurements or acquired by experts
  - overview of methods, examples and an exahaustive list of references e.g. in (Lavrac and Zupan, 2005)
- Data in different forms
  - e.g. images, texts

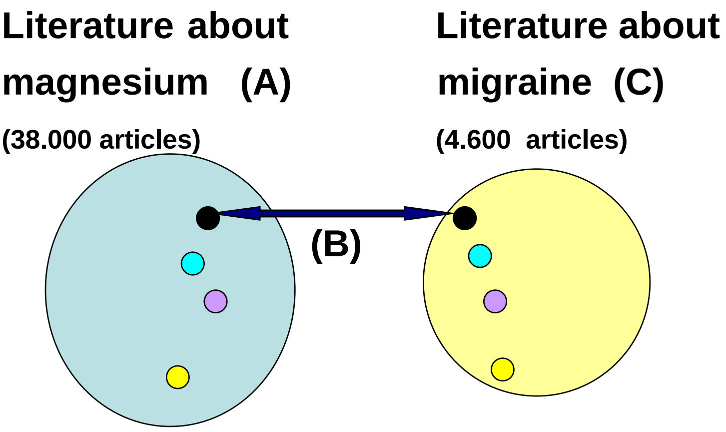
Text mining in biomedicine

(overview and examples in Cohen and Hersh, 2005)

- Extracting interesting information from biomedical knowledge represented in digital text forms
- Used also for
  - Relationship extraction
    (to recognize occurences of a pre-specified type of relationship, e.g. between genes and proteins)
  - Hypothesis generation
    (to uncover implicit relationships, worthy further investigation, e.g. potential new uses of drugs)

Swanson's example

(Swanson, 1990)



- 65 articles on migraine (BC) and
- 63 articles on magnesium (AB)

- Analysis resulted in 11 pairs of implicitly connected arguments

- Each of the 11 pairs consistent with, and suggestive of, the hypothesis that magnesium deficiency may be a causal factor in migrain.

Argument 1 (magnesium literature)

- Mg is a natural calcium channel blocker.
- Stress and Type A behavior can lead to body loss of Mg
- Magnesium has anti-inflammatory properties.
- . . .

Argument 2 (migraine literature)

- Calcium channel blockers can prevent migraine attacks.
- Stress and Type A behavior are associated with migraine.
- Migraine may involve sterile inflammation of the cerebral blood vessels.
- . . .

Swanson's ABC model

Literature on agent A that causes phenomenon B

+

Literature on agent B that influences phenomenon C

Hypothesis: agent A may influence phenomenon C

- **ARROWSMITH (Smalheiser and Swanson, 1998)**
  - **connecting fish oil and Raynaud's syndrome**
  - **anticipating adverse drug reactions**
  - **identifying mechanisms by which bioactive compounds modulate cellular or organismal responses**
  - **identifying potential animal models for human disorders**

Work that followed

- **LitLinker (Pratt and Yetisgen-Yildiz, 2003)**
  - **potential causal links between biomedical terms**
- **DAD (Weeber et al., 2003)**
  - **new potential uses of the drug thalidomine**
- **BITOLA (Hristovski et al., 2005)**
  - **identifiction of disease candidate genes**
- **RaJoLink (Urbancic et al., 2007)**
  - **identifying potential relations that might contribute to better understanding of autism**

Generation of a hypothesis
A may influence C

For a given C, how do we find A?

Swanson:

Search proceeds via some intermediate literature (B) toward an unknown destination A. … Success depends entirely on the knowledge and ingenuity of the searcher.

Wheeler:

Our whole problem is to make the mistakes as fast as possible.

Hypothesis generation: a case study

(Urbancic, Petric, Cestnik and Macedoni-Luksic, 2007)

- **Motivation:**
  - To provide more systematical support in looking for A for the Swanson's ABC model
  - To contribute to the understanding of autism

Problem domain: Autism

- **Pervasive developmental disorders**
- **Anbormal development of cognitive, communication and social interaction skills**
- **Heterogeneity of disturbance**
  (ASD – autism spectrum disorders)

- **Very important:**
  Early diagnosis and treatment (3Y -> 1Y and less)

- **One of the problems:**
  Lack of studies about risk factors (Zerhouni 2004)

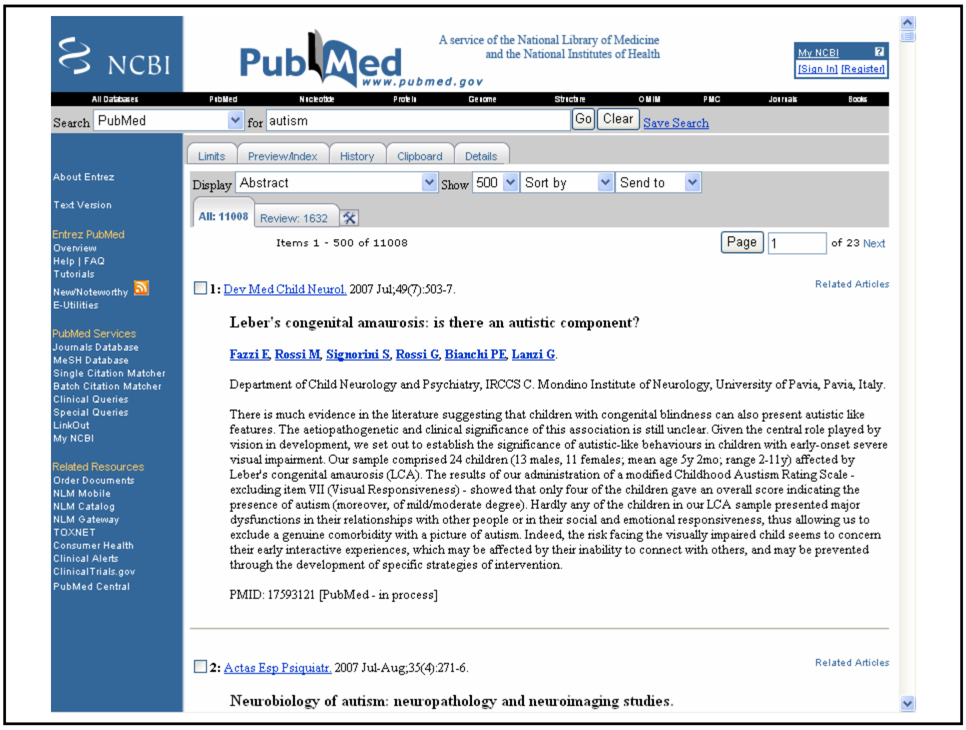
Data source: PubMed

- **US National Library of Medicine's bibliographic database**

- **more than 5.000 journals**
- **more than 15 M citations from mid-1950's to the present**
- **more than 1.500 complete references added daily**

- **10.821 documents with autis***
- **354 entire text in PubMed**
- **217 published in the last 10 years**


NCBI
PubMed www.pubmed.gov
A service of the National Library of Medicine and the National Institutes of Health
My NCBI
[Sign In] [Register]
All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books
Search PubMed for autism Go Clear Save Search
Limits Preview/Index History Clipboard Details
Display Summary Show 500 Sort by Send to
All: 11008 Review: 1632
Items 1 - 500 of 11008
Page 1 of 23 Next
About Entrez
Text Version
Entrez PubMed
Overview
Help | FAQ
Tutorials
New/Noteworthy
E-Utilities
PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
Special Queries
LinkOut
My NCBI
Related Resources
Order Documents
NLM Mobile
NLM Catalog
NLM Gateway
1: Fazzi E, Rossi M, Signorini S, Rossi G, Bianchi PE, Lanzi G.
Related Articles
Leber's congenital amaurosis: is there an autistic component?
Dev Med Child Neurol. 2007 Jul;49(7):503-7.
Abstract
PMID: 17593121 [PubMed - in process]
2: Paya B, Fuentes N.
Related Articles
Neurobiology of autism: neuropathology and neuroimaging studies.
Actas Esp Psiquiatr. 2007 Jul-Aug;35(4):271-6.
PMID: 17592791 [PubMed - in process]
3: Hayashi ML, Rao BS, Seo JS, Choi HS, Dolan BM, Choi SY, Chattarji S, Tonegawa S.
Related Articles
Inhibition of p21-activated kinase rescues symptoms of fragile X syndrome in mice.
Proc Natl Acad Sci U S A. 2007 Jun 25; [Epub ahead of print]
PMID: 17592139 [PubMed - as supplied by publisher]
4: Scheeren AM, Stauder JE.
Related Articles
Broader Autism Phenotype in Parents of Autistic Children: Reality or Myth?
J Autism Dev Disord. 2007 Jun 23; [Epub ahead of print]
PMID: 17588199 [PubMed - as supplied by publisher]


NCBI
PubMed www.pubmed.gov
A service of the National Library of Medicine and the National Institutes of Health
My NCBI
[Sign In] [Register]
Search PubMed for autism Go Clear Save Search
Limits Preview/Index History Clipboard Details
Display Abstract Show 500 Sort by Send to
All: 11008 Review: 1632
Items 1 - 500 of 11008
Page 1 of 23 Next
About Entrez
Text Version
Entrez PubMed
Overview
Help | FAQ
Tutorials
New/Noteworthy
E-Utilities
PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
Special Queries
LinkOut
My NCBI
Related Resources
Order Documents
NLM Mobile
NLM Catalog
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central
1: Dev Med Child Neurol. 2007 Jul;49(7):503-7.
Related Articles
Leber's congenital amaurosis: is there an autistic component?
Fazzi E, Rossi M, Signorini S, Rossi G, Bianchi PE, Lanzi G.
Department of Child Neurology and Psychiatry, IRCCS C. Mondino Institute of Neurology, University of Pavia, Pavia, Italy.
There is much evidence in the literature suggesting that children with congenital blindness can also present autistic like features. The aetiopathogenetic and clinical significance of this association is still unclear. Given the central role played by vision in development, we set out to establish the significance of autistic-like behaviours in children with early-onset severe visual impairment. Our sample comprised 24 children (13 males, 11 females; mean age 5y 2mo; range 2-11y) affected by Leber's congenital amaurosis (LCA). The results of our administration of a modified Childhood Autism Rating Scale - excluding item VII (Visual Responsiveness) - showed that only four of the children gave an overall score indicating the presence of autism (moreover, of mild/moderate degree). Hardly any of the children in our LCA sample presented major dysfunctions in their relationships with other people or in their social and emotional responsiveness, thus allowing us to exclude a genuine comorbidity with a picture of autism. Indeed, the risk facing the visually impaired child seems to concern their early interactive experiences, which may be affected by their inability to connect with others, and may be prevented through the development of specific strategies of intervention.
PMID: 17593121 [PubMed - in process]
2: Actas Esp Psiquiatr. 2007 Jul-Aug;35(4):271-6.
Related Articles
Neurobiology of autism: neuropathology and neuroimaging studies.

Getting acquainted with autism…

Which are the main topics in recent research of autism?

Which topics attract most attention?

How could the domain be structured?

Dataset of articles from PubMed Central

↓

Building ontology with OntoGen

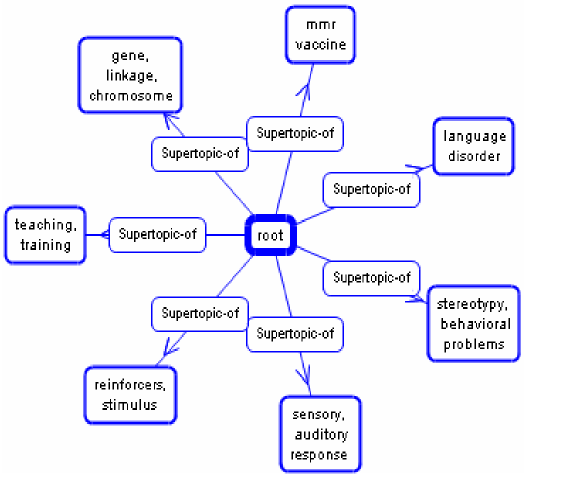
OntoGen v1.0

(Fortuna, Grobelnik, Mladenic, 2006)

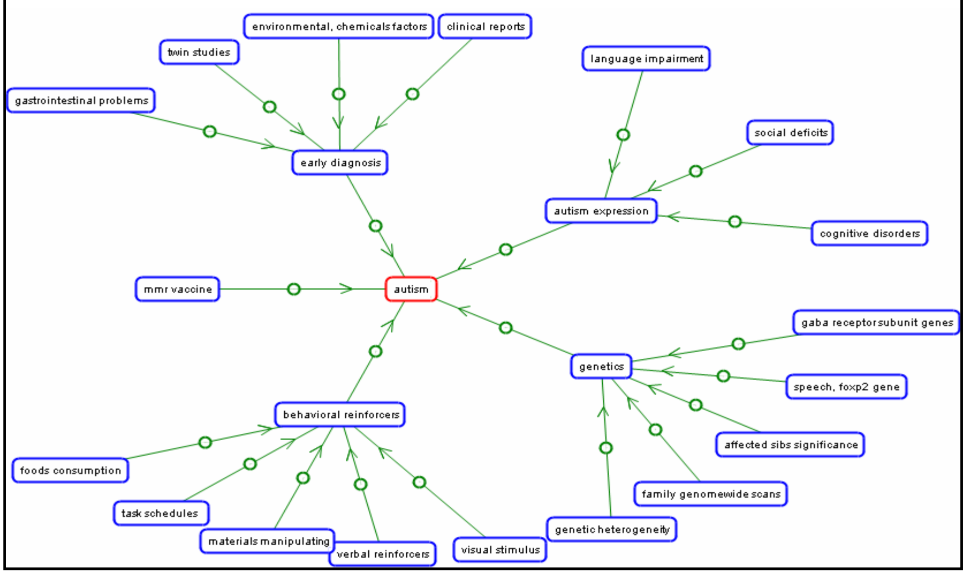
- **Designed for construction of topic ontologies**
- **Clustering algorithms used for topic suggestion**
- **Keyword extractions methods help the user to name the concept**
- **Interactive user interface**

Autism ontology example

Concepts of autism ontology with 7 subgroups, built on 214 abstracts from the PubMed Central database



Another autism ontology example



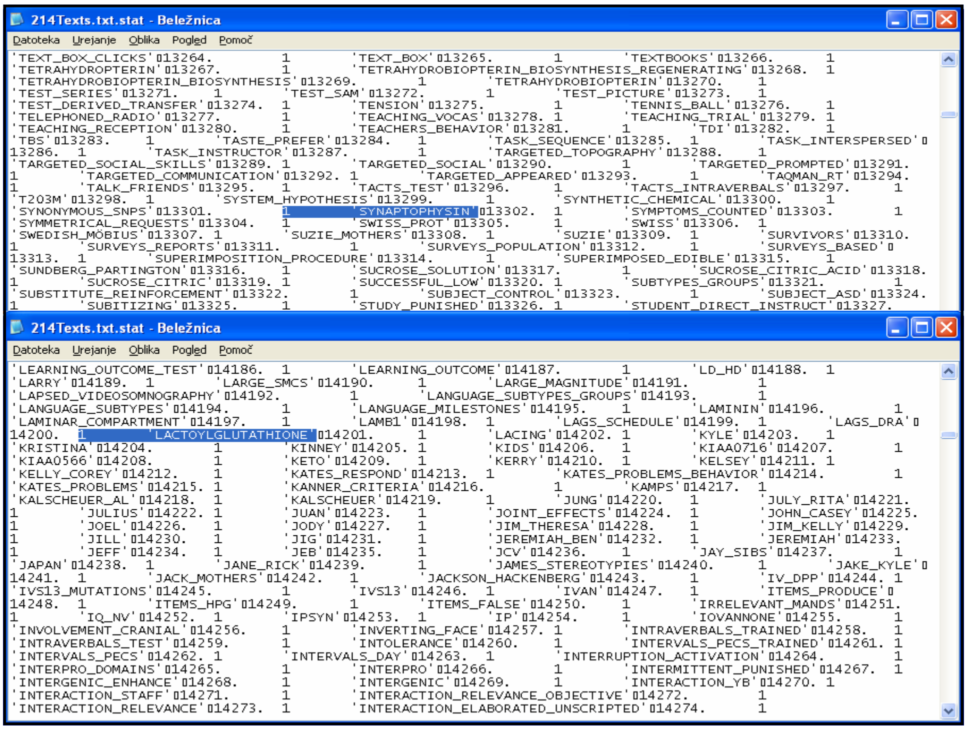
Looking for rare terms

Dataset of articles from PubMed Central

Ontology construction with OntoGen

***.txt.stat files**

Rare terms



For further investigation we choose:

- lactoylglutathione
- synaptophysin
- calcium channels

Why?

- icrease of polarity of glyoxalase I in autistic brain, glyoxalase system involves lactoylglutathione
- altered synaptic function in autism, synaptophysin is a protein localized to synaptic vesicles
- abnormal calcium signalling in some autistic children

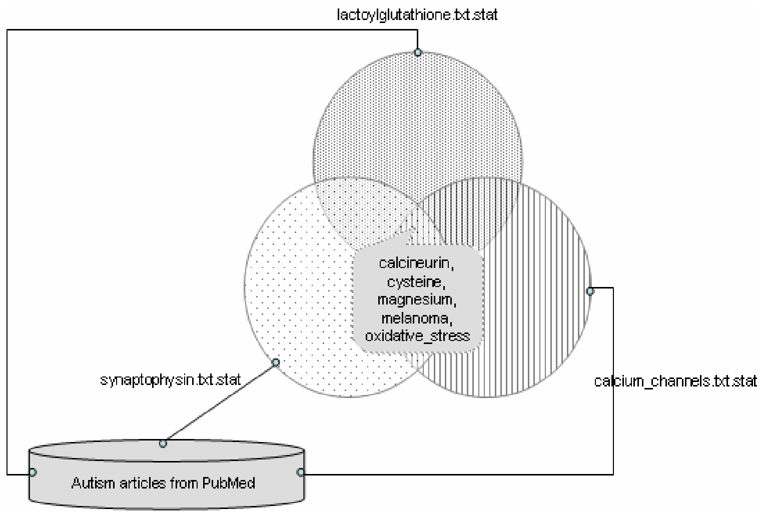
Do chosen rare terms have something in common?

Looking for joint terms

Word	Total	calcium_channels	lactoylglutathione	synaptophysin
'CYPRUS_INSTITUTE'	1	1		
'CYS'	2	1		1
'CYS_CYS'	1			1
'CYS_CYS_CYS'	1			1
'CYSTEINE'	3	1	1	1
'CYSTEINE_MOTIF'	1			1
'CYSTEINE_RESIDUE'	1	1		
'CYSTEINE_RESIDUES'	2		1	1
'CYSTEINE_RICH'	1	1		
'CYSTEINE_RICH_DOMAINS	1	1		
'CYSTEINE_STRING'	1			1
'CYSTEINE_STRING_PROTE	1			1
'CYSTEINYL'	1		1	
'CYSTEINYLGLYCINE'	1		1	

Record: 5654 of 30071

Joint terms connect sets of literature



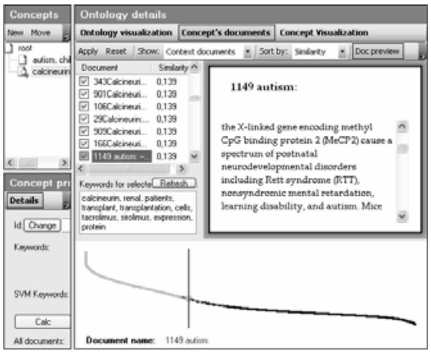
For further investigation we choose

- **clacineurin**

Calcineurin is calcium- and calmodulin-dependent serine/threonine protein phosphatase, which is widely present in mammalian tissues, with the highest levels found in brain.

To the present, no direct evidence of calcineurin role in autism has been reported on the internet.

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 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*) is inside the calcineurin context documents due to its similarity with the neighboring documents.



Argument 1 (calcineurin literature)	Argument 2 (autism literature)
• Erin et al. (2003) observed that calcineurin occured as a complex with Bcl-2 in various regions of rat and mouse brain.	• Fatemi et al. (2001) reported a reduction of Bcl-2 (a regulatory protein for control of programmed brain cell death) levels in autistic cerebellum.
• Cofanet al. (2005) published their article about effect of calcineurin inhibitors on low-density lipoprotein oxidation.	• Qiu et al. (2006) described the low-density lipoprotein receptors that regulate cholesterol transport, in neuropsychiatric disorders, such as autism.
• Zhabotinsky et al. (2006) described induction of long-term depression that depends on calcineurin.	• Bear et al. (2004) reported about loss of fragile X protein, an identified cause of autism that increased long-term depression in mouse hippocampus.
• . . .	• . . .

Towards the hypotheses… (1)

Which rare terms are promising for hypotheses generation?

Background knowledge is crucial.

Can we automatise selection of promising rare terms?

Partially, in some cases.
E.g., selecting terms from a neurobiological dictionary.

But at the moment,

Expert's involvement is crucial.

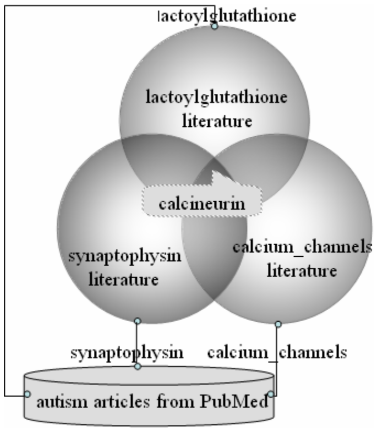
Towards the hypotheses… (2)

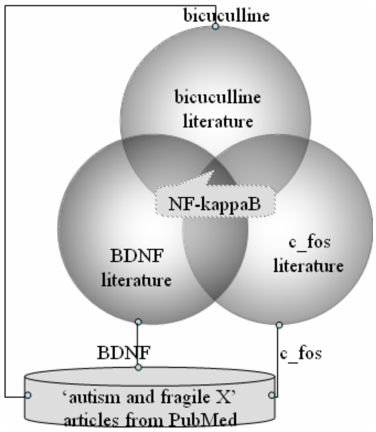
Do these pairs of documents point towards useful hypotheses?

Expert's evaluation is crucial.

"Continue with fragil X, this would really be interesting…"

Results obtained on autism domain, and on autism+fragile_X domain





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, *cytokines* (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.

Expert's evaluation (reported in Urbancic et al., 2007)

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.

Hypothesis A may influence C generation:

Being interested in C, how do we find A?

If there are some rare terms that appear in the C literature and they all have a joint term A in the intersection of their literature, it is worthwhile checking if this joint term has some connections to C via linking terms (B).

If C literature and A literature have few or no published papers in common, such (up to now uncovered) connection might contribute to better understanding of C.

RaJoLink method

Domain.txt.stat files

↓

Rare terms

↓ selection

RareTerm1.txt.stat & RareTerm2.txt.stat & …& RareTermN.txt.stat

↓

Joint terms

↓ selection

JointTerm = A, Domain = C

↓ Swanson's ABC model

Linking term B such that A -> B and B -> C

↓ evaluation

A may influence C

Steps of the RaJoLink method (1)

Step	Input	Action	Tool, technique	Human involvement	Output
Ra	Set of articles about domain of interest (about phenomenon C)	1.1 Extraction of texts	Digital document archives		
		1.2 Data collection preprocessing	Word processing software		
		1.3 Identification of rare (R) terms	Word frequency statistics		
		1.4 Semantic filtering	Latent semantic indexing	Indication of interesting R terms	R terms C_R_1, C_R_2,…C_R_p

Steps of the RaJoLink method (2)

Step	Input	Action	Tool, technique	Human involvement	Output
Jo	Sets of articles about C_R_1, C_R_2,…, C_R_p	2.1 Extraction of texts	Digital document archives		
		2.2 Data collections preprocessing	Word processing software		
		2.3 Identification of each dataset's concepts and subconcepts	Word frequency statistics, Clustering		
		2.4 Search for joint terms	Word frequency statistics	Selection of significant joint terms	Joint terms A_1, A_2,…A_q (agents A)

Steps of the RaJoLink method (3)

Step	Input	Action	Tool, technique	Human involv.	Output
Link	Joint set of articles about A_i and articles about C	3.1 Extraction of texts	Digital document archives		
		3.2 Data collection preprocessing	Word processing software		
		3.3 Identification of semantically related A_i and C documents	Semantic text analysis		
		3.4 Search for linking terms (agents B)	Word intersection	Selection of meaningful terms B_i	Linking terms B_1, B_2,… B_r

Case-study conclusions

- **Ontology construction is useful for systematical exploration of sets of articles and for getting insight into a new domain.**
- **It is worthwhile to explore rare terms for generation of hypotheses (RaJoLink method).**
- **Expert's involvement is crucial for speeding up the process (selections) and for evaluations of candidate hypotheses.**
- **Expert evaluation confirmed the relevance of discovered relations in the autism domain.**

Selected bibliography

- **U.M. Fayyad, G. Piatetski-Shapiro, P. Smith: The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM,* 39(11), 27-41, 1996**
- **N. Lavrac, B. Zupan: Data Mining in Medicine. In Maimon and Rokach (eds.) *Data mining and knowledge discovery handbook.* New York: Springer, pp. 1107-1137, 2005**
- **M. van Sommeren, T. Urbancic: Applications of machine learning: matching problems to tasks and methods, *The Knowledge Engineering Review*, Vol. 20(4), 363-402, 2006**
- **A.M. Cohen, W.R. Hersh: A Survey of Current Work in Biomedical Text Mining. Briefings in Bioinformatics, 6(1), pp. 57-71, 2005.**
- **D.R. Swanson, Medical literature as a potential source of new knowledge. *Bulletin of the Medical Library Association*, 78(1), 29-37, 1990**

Selected bibliography
(continued)

- **T. Urbancic, I. Petric, B. Cestnik, M. Macedoni-Luksic: Literature Mining: Towards Better Understanding of Autism. In: *Artificial Intelligence in Medicine LNAI 4594* (R. Belazzi, A. Abu-Hanna, J. Hunter eds.), Springer, 215-224, 2007**
- **I. Petric, T. Urbancic, B. Cestnik: Discovering Hidden Knowledge from Biomedical Literature, *Informatica* 31, 15-20, 2007**
- **B. Fortuna, M. Grobelnik, D. Mladenic: Semi-automatic Data-driven Ontology Construction System. *Proceedings of the 9th international multi-conference Information Society*, Ljubljana, Slovenia, 223-226, 2006**
- **OntoGen: http://ontogen.ijs.si/**
- **PubMed: Overview at http://www.ncbi.nlm.nih.gov/**