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 exhaustive 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 occurrences 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)

Literature about magnesium (A) (38,000 articles)

Literature about migraine (C) (4,600 articles)

• 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 migraine.
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)
  – identification 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
• Abnormal 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
Leber's congenital amaurosis: is there an autonomic component?

Fernandez M, Sperduto D, Fernández B, Blanchard DG, Levent G


There is much evidence in the literature suggesting that children with congenital blindness can develop autonomic dysfunctions. The orthopedic and neural signs of this association are not precise. The oculomotor abnormalities, the nystagmus and the mydriasis are the most apparent, and the absence of the corneal reflex is often present. The presence of autonomic dysfunction in these children is determined, usually, by the involvement of the sympathetic nervous system. The aim of this study is to evaluate the presence of autonomic dysfunction in children with Leber's congenital amaurosis. The sample comprised 14 children (11 males, 3 females; mean age 7.2 years, range 3-13) affected by Leber's congenital amaurosis (LCA). The study of the autonomic evaluation of a standard Children's Ocular Functioning Test (COFT) showed that only five of the children gave an overall score indicating the presence of autonomic dysfunction (27.3%), with no apparent pattern.

PMID: 17599125 [PubMed - in process]
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?
• Increase 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

<table>
<thead>
<tr>
<th>Word</th>
<th>Total</th>
<th>calcium_channels</th>
<th>lactoylglutathione</th>
<th>synaptophysin</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYPRUS_INSTITUTE</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYS</td>
<td>2</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>CYS_CYS</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYS_CYS_CYS</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYSINE</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CYSINE_MOTIF</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYSINE_RESIDUE</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYSINE_RESIDUES</td>
<td>2</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>CYSINE_RICH</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYSINE_RICH_DOMAINS</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYSINE_STRING</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYSINE_STRING_PROTE</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYSTEINYL</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>CYSTEINYL_CYSTEINE</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Records: 14 | 4 | 5 | 3 | 3 of 3087
For further investigation we choose

- calcineurin

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.

![OntoGen representation](image)

### Argument 1
(calcineurin literature)

- Erin et al. (2003) observed that calcineurin occurred as a complex with Bcl-2 in various regions of rat and mouse brain.
- Cofanet al. (2005) published their article about effect of calcineurin inhibitors on low-density lipoprotein oxidation.
- Zhabotinsky et al. (2006) described induction of long-term depression that depends on calcineurin.

### Argument 2
(autism literature)

- Fatemi et al. (2001) reported a reduction of Bcl-2 (a regulatory protein for control of programmed brain cell death) levels in autistic cerebellum.
- Qiu et al. (2006) described the low-density lipoprotein receptors that regulate cholesterol transport, in neuropsychiatric disorders, such as autism.
- 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

<table>
<thead>
<tr>
<th>Autism literature</th>
<th>NF-kappaB literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Araghí-Niknam and Fatemi [2] showed reduction of Bcl-2, an important marker of apoptosis, in frontal, parietal and cerebellar cortices of autistic individuals.</td>
<td>Mattson [12] reported in his review that activation of NF-kappaB in neurons can promote their survival by inducing the expression of genes encoding ant apoptotic proteins such as Bcl-2 and the antioxidant enzyme Mn-superoxide dismutase.</td>
</tr>
<tr>
<td>Vargas et al. [21] reported altered cytokine expression profiles in brain tissues and cerebrospinal fluid of patients with autism.</td>
<td>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.</td>
</tr>
<tr>
<td>Ming et al. [13] reported about the increased urinary excretion of an oxidative stress biomarker - 8-iso-PGF2alpha in autism.</td>
<td>Zou and Crews [24] reported about increase in NF-kappaB DNA binding following oxidative stress neurotoxicity.</td>
</tr>
</tbody>
</table>
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)

<table>
<thead>
<tr>
<th>Step</th>
<th>Input</th>
<th>Action</th>
<th>Tool, technique</th>
<th>Human involvement</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ra</td>
<td>Set of articles about domain of interest (about phenomenon C)</td>
<td>1.1 Extraction of texts</td>
<td>Digital document archives</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.2 Data collection preprocessing</td>
<td>Word processing software</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.3 Identification of rare (R) terms</td>
<td>Word frequency statistics</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.4 Semantic filtering</td>
<td>Latent semantic indexing</td>
<td>Indication of interesting R terms</td>
<td>R terms C_{R_1}, C_{R_2},...C_{R_p}</td>
</tr>
</tbody>
</table>
## Steps of the RaJoLink method (2)

<table>
<thead>
<tr>
<th>Step</th>
<th>Input</th>
<th>Action</th>
<th>Tool, technique</th>
<th>Human involvement</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jo</td>
<td>Sets of articles about $C_{R_1}$, $C_{R_2}$,..., $C_{R_p}$</td>
<td>2.1 Extraction of texts</td>
<td>Digital document archives</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.2 Data collections preprocessing</td>
<td>Word processing software</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.3 Identification of each dataset’s concepts and subconcepts</td>
<td>Word frequency statistics, Clustering</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.4 Search for joint terms</td>
<td>Word frequency statistics</td>
<td>Selection of significant joint terms</td>
<td>Joint terms $A_1$, $A_2$,...,$A_q$ (agents $A$)</td>
</tr>
</tbody>
</table>

## Steps of the RaJoLink method (3)

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

Selected bibliography
(continued)


• OntoGen: http://ontogen.ijs.si/