Innovative usage of unstructured information sources: From text- and data-mining to model-driven decision-support

Martin Hofmann-Apitius
Fraunhofer Institute for Algorithms and Scientific Computing SCAI, Sankt Augustin, Germany

“Innovative usage of unstructured information sources: From text- and data-mining to model-driven decision-support”

Prof. Dr. Martin Hofmann-Apitius
Head of the Department of Bioinformatics
Fraunhofer Institute for Algorithms and Scientific Computing

36th Annual IATUL Conference 2015
July 6 - 9, Hannover, Germany
Where I come from: Fraunhofer Society

- Founded 1949
- Europe’s largest applied research organisation
- 60 Research Institutes (7 Institutes in the US)
- > 23,000 Employees
- Annual Budget > 2 Billion Euro
- Financial model: 2/4 industry collaborations
  1/4 public funding
  1/4 institutional funding

*Joseph von Fraunhofer (1787 – 1826)
Scientist, Inventor and Entrepreneur
The Fraunhofer Institute Center Schloss Birlinghoven

- Largest research centre for informatics and applied mathematics in Germany
- Around 700 employees, thereof 500 scientists, approx. 200 students and trainees
- University links:
  - Bonn
  - Aachen
Expertise at the Department of Bioinformatics at SCAI

Fraunhofer SCAI Department of Bioinformatics currently comprises:

- 10 scientists
- 3 scientific software developers
- 7 PhD students
- ~ 5 Master students
- ~ 10 student workers

- predominantly computer scientists & biologists
- some PhD students via University of Bonn (Bonn-Aachen International Center for Information Technology)
Fraunhofer SCAI Department of Bioinformatics R&D activities:

1. Information extraction in the **life sciences**:
   - Recognition of named entities and relationships in text
   - Large-scale, automated Information Extraction

2. Integrative biology; disease modelling
   - Focus on neurodegenerative diseases

3. eScience, Grid- & Cloud- Computing / HPC (Cluster)
   - Focus on scaling of information extraction workflows
Imagine …

- **Cancer Patient, final stage, metastatic pancreas carcinoma**
  - Surgery, chemotherapy without success
  - Distant metastasis in bone marrow, lung and liver
  - Remaining life span: 2 – 4 weeks

- **Last chance: sequencing the cancer genome (< 10k€)**
  - Getting insight into mutations underlying cancer dysregulation
  - Understanding of mechanisms triggering uncontrolled growth
  - Identification of (experimental) compounds that inhibit tumour growth

- **This is not fiction – this is reality**
The challenge ...

- Cancer genome sequencing delivers vast amount of information
  - Tens to hundreds of mutations
  - Functional relevance of a significant number of mutations unclear
  - Contribution of mutation to tumour growth and metastasis?

- How do we assess the biological impact of genetic variation information?
  - Putting genetic variation information into a functional context
  - Reasoning over genetic variation information and inference of consequences
  - From inference to personalised recommendation .... within 2 weeks of time

- Let us see where we stand ...
Identification and normalisation of the relevant Life Science terminology is key for information retrieval, information extraction and inferring of knowledge.
Named Entity Recognition Technologies @SCAI

**ProMiner**

Recognition and Normalization of Named Entities in Scientific Text

- **Dictionary based approaches**
  - with normalisation and embedding in hierarchies
  - Genes/Proteins, Disease, Drugs, Cells...

- **Regular expression**
  - (partly to normalise)
  - Chrom. Locations, rs numbers...

- **Machine learning based approaches**
  - (CRF based)
  - SNP, IUPAC, epigenetic modifications

- **Combined entities**
  - Combined normalisation
  - SNP, Histonmodification
Different input output formats

MEDLINE abstracts

ASCII text

XML text, e.g. patents

Defined format

HTML text e.g. Journal articles, Web pages

PDF text e.g. Journal articles

1. Association of breast cancer resistance protein/ABCG2 phenotypes and novel promoter and intron 1 single nucleotide polymorphisms.

Named Entity Recognition and Normalisation

1. Association of breast cancer resistance protein/ABCG2 phenotypes and novel promoter and intron 1 single nucleotide polymorphisms.

The hypothesis was tested that sequence diversity in breast cancer resistance protein (BCRP)‘s cis-regulatory region is a significant determinant of BCRP expression. The BCRP promoter and intron 1 were resequenced in lymphoblast DNA from the polymorphism discovery resource (PDR) 44 subset. BCRP single nucleotide polymorphisms (SNPs) were genotyped in donor human livers, intestines, and lymphoblasts quantitatively phenotyped for BCRP mRNA expression. Carriers of the -15822C>T SNP had lower BCRP expression in multiple tissues. The intron 1 SNP 16702C>T was associated with high expression in livers; 1143G>A was associated with low expression in intestines; 12283T>C was associated with higher expression in the PDR44 and White livers. The -15894C>T promoter SNP was significantly associated with higher BCRP expression in multiple tissues. Patients with the -15894C>T genotype had substantially higher clearance of p.o. imatinib. We next determined whether BCRP expression was related to polymorphic alternative splicing or alternative promoter use. Liver polyclonally expressed an alternatively spliced mRNA [splice variant (SV) 1] skipping exon 2. Although SV1 livers did not uniformly carry the exon 2 G34A allele, 90% of G34A livers expressed SV1 (versus 4% of 34GG livers). BCRP mRNA was significantly lower among Hispanic livers with the G34A variant genotype and may be due, in part, to polymorphic exon 2 splicing. Analysis of allele expression imbalance (AEI) showed that PDR44 samples with AEI had lower BCRP mRNA expression; however, no linked cis-polymorphisms were identified. BCRP used multiple promoters, and livers differentially using alternative exon 1b had lower BCRP. In conclusion, BCRP expression in lymphoblasts, liver, and intestine is associated with novel promoter and intron 1 SNPs.

Neoplasms by Site
Breast Neoplasms (D001943)
Breast Neoplasms, Male Carcinoma, Ductal, Breast Phyllodes Tumor

DB00619: imatinib
refSNP ID: rs2231137
ABCG2-Position: 18897
Avg Het: 0.203+/−0.246
SCAIView functionalities

- **Document View** - Displays all the documents retrieved based on the search query. Entity classes can be selected that you want to highlight. By default, Documents are displayed according the date (newest on the top).

- **Entity View** - Displays named entities under the column entities and are linked to corresponding abstracts

- **Export** - PMID and Entity tables can be exported to text files or excel sheets
Motivation
UIMA workflow and UI applications

- UIMA (Unstructured Information Management Architecture) is a software architecture for deploying and developing unstructured information management application.
- Originally developed by IBM, now open source.
- **Unstructured information application** may be defined as a software system designed to analyses large volume of unstructured information in order to discover, organise, and deliver knowledge to the end user.
- Thus this architecture provides analytical platform by converting unstructured text to structured information.
UIMA based analysis at SCAI

- BEL like Statement Extraction
- Co-occurrence and Tri-occurrence based relationship extraction
- Machine Learning based relationship extraction
- Topic Modelling
- Term Frequency based Analysis
- ...
Capturing Knowledge on Causes and Effects: OpenBEL

Subject: The abundance of molecules designated by the name “corticosteroid” in the CHEBI namespace.

Predicate: decreases

Object: The biological process designated by the name “Tissue Damage” in the NCI namespace.
OpenBEL: Capturing of Knowledge and “encoding” of data

Phosphorylation of glycogen synthase kinase 3beta at Threonine, 668 increases the degradation of Amyloid precursor protein.
BEL forms Graphs

- Subject-predicate-object “triples”
- Object of one triple can be subject of another
- Putting them together makes arbitrarily large knowledge graphs
- Reasoning over causal relationships becomes a graph traversal
The World´s largest Computable Model for Alzheimer´s Disease

Project Goals of the Work of Gurnoor Singh

- Implement a solution integrating document retrieval via SCAIVew and analytical tools which extracts information/knowledge based on a UIMA workflow

- A generic solution which works well on any analysis workflow (‘wrapper’)

- A well distributed, flexible, and efficient solution for multitasking

- Show application by, performing an exemplary analysis which measures the difference of information gain between abstract representation and full text representation of Biomedical journals
Exemplary Application

- Biological Research Question: “Is there a difference in Information Gain between abstract and full text of a document?”
- Information Gain can be measured as number of unique BEL like statements.
- Install daemon for analysing BEL Like Statement.

- Corpus: Collection of PMID as defined by user query in SCAIView
BELIEF Workflow

1. TXT
2. Various NLP tools
3. SD etc.
4. NER with multiple dictionaries
5. ProMiner
6. RE preprocessing
7. Linear Classifier, TEES etc.
8. Relation extraction tools
9. BEL Writer
10. BEL

Unstructured Information Management Architecture
An Apache Project

© Fraunhofer SCAI
What text mining can deliver – NER & Normalisation

3. Simvastatin inhibits induction of matrix metalloproteinase-9 in rat alveolar macrophages

Simvastatin: A derivative of lovastatin and potent competitive inhibitor of 3-hydroxy-3-methylglutaryl coenzyme A reductase (hydroxymethylglutaryl COA reductase), which is the rate-limiting enzyme in cholesterol biosynthesis.

3) may play an important role in emphysematous change in chronic obstructive pulmonary disease (COPD), one of the major causes of disability worldwide. We previously reported that simvastatin, an inhibitor of HMG-CoA reductase, attenuates emphysematous lungs of rats exposed to cigarette smoke. However, it remained uncertain how cigarette smoke induced MMP-9 and how the induced MMP-9 expression in alveolar macrophages (AMs), a major source of MMP-9 in the lungs of COPD patients.

Simvastatin signaling for MMP-9 induction and the inhibitory mechanism of simvastatin on MMP-9 induction in AMs exposed to isolated rat AMs, CSE induced MMP-9 expression and phosphorylation of ERK and Akt. A chemical inhibitor of MEK1/2 or ERK or Akt, respectively, and also inhibited CSE-mediated MMP-9 induction. Simvastatin reduced CSE-mediated MMP-9 expression and phosphorylation of ERK and Akt. Farnesyl pyrophosphate (FPP) or geranylgeranyl pyrophosphate (GGPP) similar to farnesyl or GGPP transferase suppressed CSE-mediated MMP-9 induction. Simvastatin attenuated CSE-mediated activation of RAS and phosphorylation of ERK, Akt, p65, IkappaB, and nuclear AP-1 or NF-kappaB activity. Taken together, these results suggest that simvastatin may inhibit CSE-mediated MMP-9 induction, primarily by blocking prenylation of RAS in the signaling pathways, in which Raf-MEK-ERK, PI3K/Akt, AP-1, and IkappaB-NF-kappaB are involved.
What text mining can deliver – NER & Normalisation

- Recall and Precision rates are between 70% and 90% for biomedical NER

- **Recall**: How many of the existing names does the system detect
- **Precision**: How many of the detected names are correct

http://www.scaiview.com
BELIEF Workflow

Various NLP tools → SD etc. → ProMiner → NER with multiple dictionaries → RE preprocessing → Relation extraction tools (Linear Classifier, TEES etc.) → BEL Writer → BEL
## Current dictionaries included

<table>
<thead>
<tr>
<th>Entity class</th>
<th>Resources</th>
<th>BEL namespace</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Genes/Proteins</td>
<td>EntrezGene/Uniprot</td>
<td>HGNC</td>
</tr>
<tr>
<td>Mouse Genes/Proteins</td>
<td>EntrezGene/Uniprot</td>
<td>MGI</td>
</tr>
<tr>
<td>Rat Genes/Proteins</td>
<td>EntrezGene/Uniprot</td>
<td>RGD</td>
</tr>
<tr>
<td>Protein family names</td>
<td>OpenBEL</td>
<td>PFH</td>
</tr>
<tr>
<td>Protein complex names</td>
<td>OpenBEL</td>
<td>NCH</td>
</tr>
<tr>
<td>Protein complex names</td>
<td>Gene Ontology</td>
<td>GOCCTERM</td>
</tr>
<tr>
<td>Chemical names</td>
<td>OpenBEL</td>
<td>SCHEM</td>
</tr>
<tr>
<td>Chemical names</td>
<td>ChEBI</td>
<td>CHEBI</td>
</tr>
<tr>
<td>Chemical names</td>
<td>ChEMBL</td>
<td>SCHEM</td>
</tr>
<tr>
<td>Disease names</td>
<td>MeSH</td>
<td>MESHD</td>
</tr>
<tr>
<td>Anatomy names</td>
<td>MeSH</td>
<td>MESHA</td>
</tr>
</tbody>
</table>
Use case: relation between small molecules (mainly protein inhibitors) and their targets

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>Recall rate initial version</th>
<th>Recall rate application adapted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes/Protein: (HGNC)</td>
<td>80 %</td>
<td>93 %</td>
</tr>
<tr>
<td>Chemical compounds: ChEBI</td>
<td>15 %</td>
<td>66 %</td>
</tr>
<tr>
<td>Chemical compounds: SCHEM</td>
<td>30 %</td>
<td>75 %</td>
</tr>
<tr>
<td>Chemical compounds: ChEBI + SCHEM + ChEMBL</td>
<td>not determined</td>
<td>91 %</td>
</tr>
<tr>
<td>Selventa-human-complex</td>
<td>40 %</td>
<td>46 %</td>
</tr>
<tr>
<td>GO-Complex</td>
<td>not determined</td>
<td>64 %</td>
</tr>
<tr>
<td>Selventa-human-complex + Complex</td>
<td>not determined</td>
<td>82 %</td>
</tr>
<tr>
<td>GO-Function</td>
<td>22 %</td>
<td>not determined</td>
</tr>
<tr>
<td>Selventa-human-families</td>
<td>8 %</td>
<td>77 %</td>
</tr>
</tbody>
</table>
Recognition and normalization of terminology

- Normalisation is needed!
- Use external and internal (OpenBEL) resources for named entity recognition (Mapping!)
- Combine various resources
- Adapt terminology to use cases (OpenBel namespaces provide no synonyms)
- Offer curators the annotation of different concepts

For relation extraction high recall is a precondition!!!
BELIEF Workflow

Various NLP tools

NER with multiple dictionaries

Relation extraction tools

Linear Classifier, TEES etc.

BEL Writer

BEL
Relation Extraction

Two kinds of relationship extraction tools are available which are tested and compared on common benchmark sets:

- The BioNLP shared tasks deliver a very detailed annotation for relationship extraction similar to the information needed for BEL
Relation Extraction

Two kinds of relationship extraction tools are available which are tested and compared on common benchmark sets:

- The BioNLP shared tasks deliver a very detailed annotation for relationship extraction similar to the information needed for BEL

IL-7 also delayed the decrease in the levels of cMyc expression
Relation Extraction

Two kinds of relationship extraction tools are available which are tested and compared on common benchmark sets:

- The BioNLP shared tasks deliver a very detailed annotation for relationship extraction similar to the information needed for BEL

- Simpler binary classification:
  
  IL 7 also delayed the decrease in the levels of cMyc ....

  IL7 – cMyc Relation: Yes

Classifies if a relation between 2 entities is existing but gives no information about the direction or type
Technology – Performance

- NLP (Sentence Detection ~6% error) 94
- NLP (Tokenization ~8% error) 86
- NER (Different Classes ~15% error) 73
- RelationExtraction (Multi-step ~25% error) 54

Propagated error!
What text mining can deliver: Relation Extraction – Example from BioNLP shared Task

Recall ~30% and Precision ~50% for regulation events – binary classification has higher recall and precision rates

That seems not very promising but many relations might be redundant!

http://bishop.scai.fraunhofer.de/scaiview/
BELIEF Workflow

Various NLP tools → SD etc. → ProMiner → RE preprocessing → Linear Classifier, TEES etc. → Relation extraction tools → BEL Writer → BEL
We implemented a rule set translating BioNLP SharedTask to BEL

Abundances: p(),..., pmod()
Modifications: deg()
Catabolism: tloc()
Location: complex()

BEL networks derived from qualitative translations of BioNLP Shared Task annotations.
The Association for Computational Linguistics (ACL) Sofia 2013
IL-7 also delayed the decrease in the levels of cMyc expression.
Relation extraction example result

Automatic extension to full statements in workflow:

Fixme -|r(HGNC:MYC)

p(HGNC:IL7) -| Fixme -|r(HGNC:MYC)
Automatic extension to full statements in workflow:

\[
\text{Fixme} \quad -| \quad r(\text{HGNC:MYC}) \\
\text{p(HGNC:IL7)} \quad -| \quad \text{Fixme} \quad -| \quad r(\text{HGNC:MYC}) \\
\text{p(HGNC:IL7)} \quad -- \quad r(\text{HGNC:MYC})
\]
What text mining currently does not deliver: Interpretations

- Generated Statement:
  - p(FIXME) -| p(HGNC:STAT4) -| p(FIXME) -> p(HGNC:IL10)

- Manual Statement:
  - p(HGNC:STAT4) -> p(HGNC:IL10)
Semi automatic BEL Knowledge Extraction Pipeline

Application
BeliefDashboard

Projects management and multiple document upload

- Create, delete and list functionality
BELIEF Dashboard Curation Interface

Detected concepts:
- Highlighted text entity in mouse over
- Provide a fast overview of all entities in the evidence

BEL statement
Context annotation
Edit/Delete/Export
BEL Editor – Update document information

Document information

Pubmed Id: 11350768

PMID: 11350768

Title: Nicotine infusion alters leptin and uncoupling protein 1 mRNA expression in adipose tissues of rats.

Journal: American journal of physiology. Endocrinology and metabolism; Vol. 280; Iss. 6

Authors: K Arai, K Kim, K Kaneko, M Iketeri, A Otagiri, N Yamauchi, T Shibasaki

Published: Jun/2001
Back to Gurnoor Singh: Experimental Setup

- Query disease under study: Alzheimer's Disease (AD)
- A total 10 jobs selecting top 25, 50, 100, 500, and 1000 text documents from both the SCAIView systems were exported to BEL processing daemon.
- BEL documents retrieved via SCAIview were analyzed using KAM navigator and Cytoscape
- Biological networks were further narrowed down to Protein-Protein Interactions networks
Information Graphs

![Graph showing the number of information graphs against the number of text documents, with two lines representing Abstract and Full-Text.]
Graph Topology

500- Abstract
Documents: 500
Nodes: 215
Edges: 509

500- Full Text
Documents: 500
Nodes: 1490
Edges: 5741

Manually Curated
Documents: 500
Nodes: 553
Edges: 3525
Plotting APP (yellow) and its First Neighbors

500- Abstract
Nodes: 58

500- Full Text
Nodes: 230

Manually Curated
Nodes: 121
Venn Diagrams

All triples

APP triples

500 Abstracts 500 Full-texts
0 110 1040
0 105 244
304
Manually curated

500 Abstracts 500 Full-texts
0 39 138
0 19 34
68
Manually curated
Text Mining & Decision Support

- Our cancer patient may die, because:
  - Nobody can read all the papers that contain relevant information in only two weeks
  - The publishing industry does not permit machines to do the job

- How will we reason over genetic variation information in a functional context?
  - See Naz et al., Briefings in Bioinformatics, in press

- Time for a “GRAND CHALLENGE”
  - Let us work together to organise a Grand Challenge that demonstrates how our cancer patient could be saved if automated text mining would be supported by the publishing industry
Innovative Text Mining and Decision Support

• We can use unstructured text like any database

• We can extract useful and interesting facts, such as triples that represent causal relationships in biomedicine

• We can use these semi-automated information extraction processes to generate a knowledge base in languages such as BEL

• In the future, such knowledge bases will enable decision support in life saving, time-critical scenarios