Linguamatics

# Semantic Insights using Agile NLP-based Text Mining

David Milward

ICIC
Sitges, Spain
October 18-21, 2009

Linguamatics

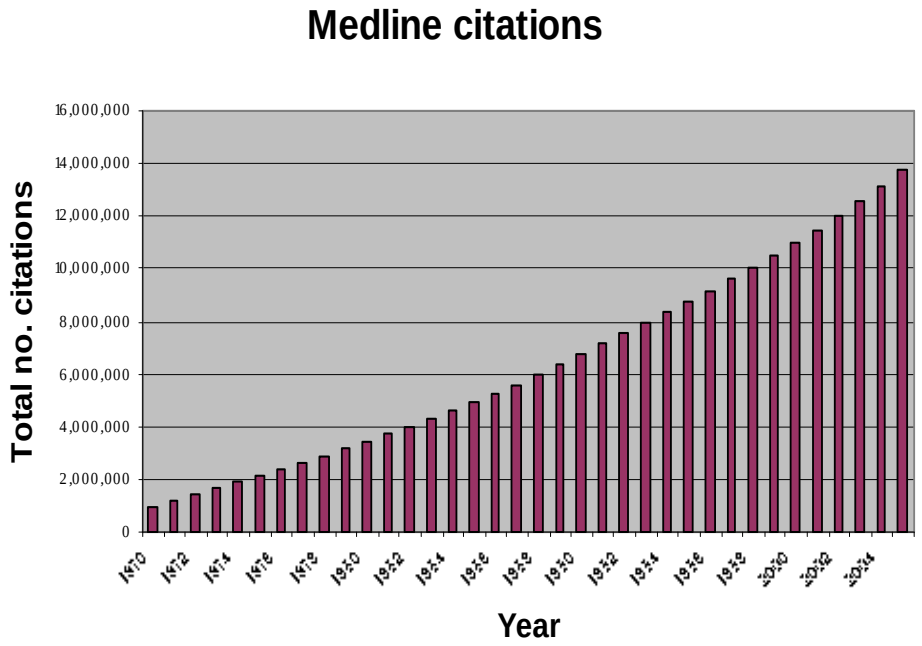
# Overview

- Agile Text Mining
- Extraction from large documents
  - Case study on extracting quantitative information from compound safety reports
- Semantic Insights via
  - Linking information extracted from different documents
  - Linking structured and unstructured knowledge
  - Linking chemical and biological knowledge

Linguamatics

# Knowledge Discovery Challenges

- Need to use available information to make better decisions
- Integrate knowledge from different sources
  - External
    - literature
    - news
    - web
  - Internal
    - experimental data
    - reports
    - databases
  - Automatically derive insight
  - Get to weak signals



Linguamatics

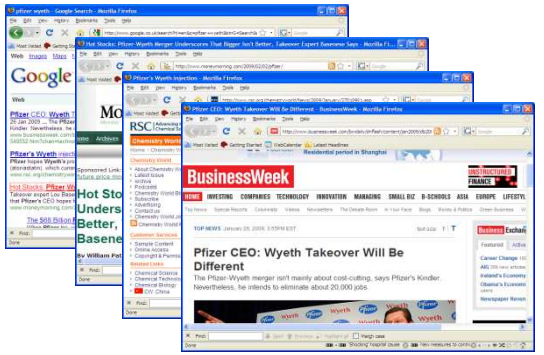
# Overview

- Agile Text Mining
- Extraction from large documents
  - Safety report case study
- Semantic Insights via
  - Linking information extracted from different documents
  - Linking structured and unstructured knowledge
  - Linking chemical and biological knowledge

Linguamatics

# Search, Information Extraction and Agile Text Mining

- Document search
  - Provides the most relevant documents for a query
- Information Extraction
  - Finds specific assertions using Natural Language Processing
- Agile Text Mining
  - Provides answers to ad-hoc questions




| company | activity | company |
|---|---|---|
| Sanofi | bid | Aventis |
| Roche | partner | Antisoma |

| Chemical | Dosage | | Doc |
|---|---|---|---|
| Cyclosporine | 5 mg/kg/day | 10 | 1552052 |
| | 2.5 mg/kg/day | 7 | 9568413 |
| | 3 mg/kg/day | 5 | 16280301 |
| | 2.5 mg/kg | 4 | 10468801 |

Linguamatics

# "Which, What, Who?"

- Go directly to answers, e.g. find all the genes associated with a specific disease
- "Which gene" searches for >10,000s genes and all their synonyms
- Provides highlighted evidence and link to the document:
  - Read a sentence or a single document to convince you that the gene is relevant (or not)

| Entrez Genes | | Doc |
|---|---|---|
| ERBB2 | 470 | 15870086 |
| BRCA1 | 248 | 15564800 |
| BRCA2 | 160 | 15986445 |
| PGR | 129 | 15272277 |
| EGFR | 101 | 16280056 |
| VEGFA | 76 | 15897560 |
| TP53 | 73 | 15583825 |
| INS | 68 | 15805581 |
| CCND1 | 51 | 16140974 |
| AKT1 | 50 | 16619501 |

Linguamatics

# "Which, What, Who?"

- Find which gene mutations are mentioned specifically related to the disease, e.g. breast cancer

| Entrez Genes | Mutation | | Doc |
|---|---|---|---|
| ERBB2 | | 469 | 15870086 |
| | Ile655Val | 3 | 15970791 |
| | Ile654Val | 1 | 15550452 |
| BRCA1 | | 248 | 15564800 |
| | 4153delA | 1 | 15980987 |
| | C61G | 1 | 15980987 |
| BRCA2 | | 160 | 15986445 |
| | C5972T | 1 | 16280055 |
| | 999del5 | 1 | 16418514 |

Linguamatics

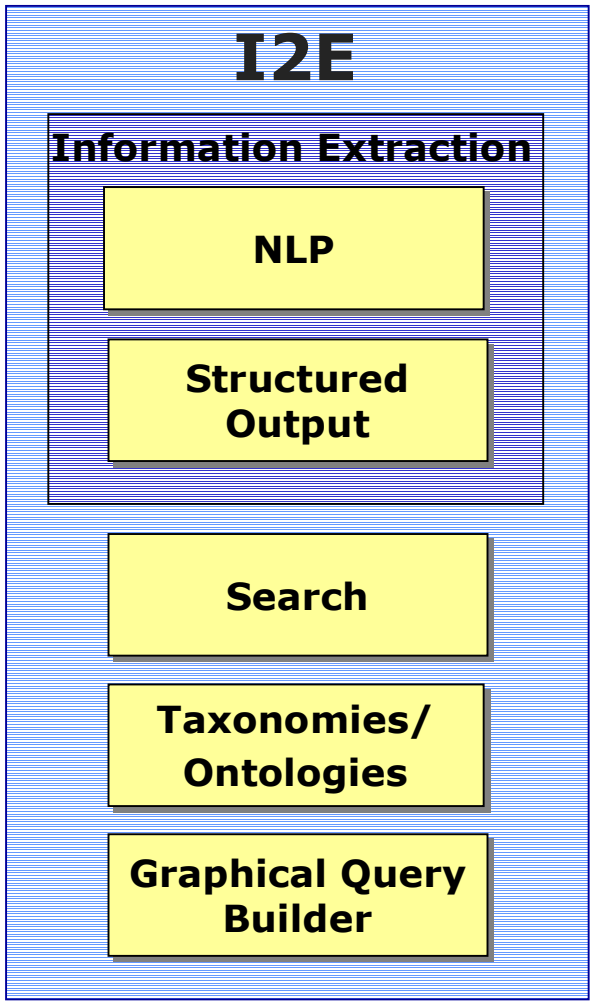
# "Tell me about X"

- Search would provide documents most about X
- Here, profiling X by summarising information from millions of docs

| Pharmacologic Substance | Relation | Entity | | Doc | | Hit | qID |
|---|---|---|---|---|---|---|---|
| Cyclosporine | treats | Psoriasis | 81 | 1401311 | 3 | Cyclosporine therapy for psoriasis: a cell cycle-derived dosing schedule. | 7 |
| | dosage | 5 mg/kg/day | 15 | 1552052 | 3 | Alterations in renal function in psoriasis patients treated with cyclosporine, 5 mg/kg/day. | 13 |
| | causes | nephrotoxic | 2 | 11210296 | 1 | Methotrexate treatment can lead to bone marrow suppression and hepatotoxicity, and cyclosporine can cause nephrotoxicity. | 2 |
| | inhibit | Growth factor | 2 | 8884530 | 1 | FK506 and cyclosporin A inhibit growth factor-stimulated human keratinocyte proliferation by blocking cells in the G0/G1 phases of the cell cycle. | 1 |
| | affect | PGP | 1 | 15788683 | 1 | CONCLUSIONS: Cyclosporin A modulates Pgp, MRP-1, BCRP, and LRP, and this broad-spectrum activity may contribute to its clinical efficacy. | 1 |
| | IC50 | 3.0 ng/ml | 1 | 9585801 | 1 | We classified these patients into two groups on the basis of their PBMC sensitivity to cyclosporine with use of the median cyclosporine IC50 (3.0 ng/ml) of these patients as the cutoff point. | 12 |

# Interactive Information Extraction (I2E)

- **Natural Language Processing (NLP):** using linguistics to interpret the meaning of unstructured text sources.
- **Structured Output:** presenting extracted information with drill-down to supporting evidence.
- **Search Engine Approach:** returning results in real time.
- **Domain Knowledge Plug-in:** ability to provide semantic search capabilities with domain knowledge such as thesauri and taxonomies
- **Graphical User Interface**: users can define, share, and adapt queries



Linguamatics

# Overview

- Agile Text Mining
- Extraction from large documents
  - Safety report case study
- Semantic Insights via
  - Linking information extracted from different documents
  - Linking structured and unstructured knowledge
  - Linking chemical and biological knowledge

Linguamatics

# Extracting Data from Safety Reports

Mean $C_{max}$ and $AUC_{0-24h}$ in males at 3 and 15 mg/kg were 21 and 154 ng/mL and 169 and 1350 ng•h/mL, respectively, for free base and 37 and 490 ng/mL and 170 and 2790 ng•h/mL, respectively, for HCl salt.

- Extracting toxicity data for structure-activity modeling
- Reports in PDF format, up to 1000 pages, often as scanned electronic images
- Complex sentence construction
- Necessary data may be within text or within tables
- Inconsistent identifiers, e.g. for compounds, both in format and over time

Ref.: Nigel Greene, Pfizer Inc.; David Milward, Linguamatics; Richard Williams, Lhasa Ltd. (2009) Unlocking Toxicity Data for Structure-Activity Modeling By Semi-Automated Extraction from Study Reports, Bio-IT World 2009

# Semi-Automated Approach

| Exposure Scenario | Dose Used in Risk Assessment, UF | FQPA SF* and Endpoint for Risk Assessment | Study and Toxicological Effects |
|---|---|---|---|
| Short- and Intermediate-Term Incidental Oral (1 to 30 days and 1 month to 6 months) | NOAEL= 15 mg/kg/day | LOC for MOE = 300 (Residential) | Co-critical studies: subchronic oral (rat); subchronic neurotoxicity (rat) developmental toxicity (rat); LOAEL = 50 mg/kg/day based on reductions in body weight, body weight gain and food consumption. |

| Exposure Scenario | Dose Used in Risk Assessment, UF | FQPA SF* and Endpoint for Risk Assessment | Study and Toxicological Effects |
|---|---|---|---|
| Short- and Intermediate-Term Incidental Oral (1 to 30 days and 1 month to 6 months) | NOAEL= 15 mg/kg/day | LOC for MOE = 300 (Residential) | Co-critical studies: subchronic oral (rat); subchronic neurotoxicity (rat) developmental toxicity (rat); LOAEL = 50 mg/kg/day based on reductions in body weight, body weight gain and food consumption. |

| | Query | Doc | Chemical | LOAEL NOAEL | Dosage | Change | Observation |
|---|---|---|---|---|---|---|---|
| ☐ | levels_of_adverse_events (Run: 2008-12-01 14:22:11) | 00001.htm | (E)-acetamiprid | LOAEL | 17.5 mg/kg/day | reduce | body weight gain |
| ☐ | levels_of_adverse_events (Run: 2008-12-01 14:22:11) | 00001.htm | (E)-acetamiprid | LOAEL | 17.5 mg/kg/day | reduce | reduced body weight |
| ☐ | levels_of_adverse_events (Run: 2008-12-01 14:22:11) | 00001.htm | (E)-acetamiprid | LOAEL | 50 mg/kg/day | reduce | body weight |
| ☑ | levels_of_adverse_events (Run: 2008-12-01 14:22:11) | 00001.htm | (E)-acetamiprid | LOAEL | 50 mg/kg/day | reduce | body weight gain |
| ☐ | levels_of_adverse_events (Run: 2008-12-01 14:22:11) | 00001.htm | (E)-acetamiprid | LOAEL | 50 mg/kg/day | reduce | food consumption |
| ☐ | levels_of_adverse_events (Run: 2008-12-01 14:22:11) | 00001.htm | (E)-acetamiprid | LOAEL | 51.0 mg/kg/day | reduce | litter size |
| ☐ | levels_of_adverse_events (Run: 2008-12-01 14:22:11) | 00001.htm | (E)-acetamiprid | LOAEL | 51.0 mg/kg/day | reduce | pup weights in both generations |
| ☐ | levels_of_adverse_events (Run: 2008-12-01 14:22:11) | 00001.htm | (E)-acetamiprid | LOAEL | 51.0 mg/kg/day | reduce | reductions in litter size |
| ☐ | levels_of_adverse_events (Run: 2008-12-01 14:22:11) | 00001.htm | (E)-acetamiprid | LOAEL | 51.0 mg/kg/day | reduce | viability |
| ☐ | levels_of_adverse_events (Run: 2008-12-01 14:22:11) | 00008.htm | Bifenazate | LOAEL | 400 mg/kg/day | reduce | decreased body weight |
| ☐ | levels_of_adverse_events (Run: 2008-12-01 14:22:11) | 00016.htm | (E)-Clothianidin | LOAEL | 31.2 mg/kg/day | reduce | decreased mean body weight gain |
| ☐ | levels_of_adverse_events (Run: 2008-12-01 14:22:11) | 00016.htm | (E)-Clothianidin | LOAEL | 75 mg/kg/day | increase | an increased litter incidence of a missing lobe |

**OCR/PDF conversion**

**Example PDF from EPA**

**I2E Text Mining**

**Results Curation**

Linguamatics

# Extracting Quantitative Information

- Clinical chemistry, e.g. serum chemistry, hematology, urinalysis
- Clinical signs, observations
- Complex, precise patterns used to extract from "respectively" constructions, e.g. for exposure

| | | | | | | |
|---|---|---|---|---|---|---|
| AUC | 2630 | ng-h/mL | 5 | mg/kg | Day 1 | Mean AUC 0-24h values were 2630, 98800, and 785000 ng-h/mL on Day 1 and 3190, 116000, and 1340000 ng-h/mL on Day 30 at 5, 50, and 250 mg/kg, respectively. |
| AUC | 98800 | ng-h/mL | 50 | mg/kg | Day 1 | Mean AUC 0-24h values were 2630, 98800, and 785000 ng-h/mL on Day 1 and 3190, 116000, and 1340000 ng-h/mL on Day 30 at 5, 50, and 250 mg/kg, respectively. |
| AUC | 785000 | ng-h/mL | 250 | mg/kg | Day 1 | Mean AUC 0-24h values were 2630, 98800, and 785000 ng-h/mL on Day 1 and 3190, 116000, and 1340000 ng-h/mL on Day 30 at 5, 50, and 250 mg/kg, respectively. |

Linguamatics

# Overview

- Agile Text Mining
- Extraction from large documents
  - Safety report case study
- Semantic Insights via
  - Linking information extracted from different documents
  - Linking structured and unstructured knowledge
  - Linking chemical and biological knowledge

Linguamatics

# "How does A relate to B"

## Potential Mechanism of Action of Compound on Disease

**<Cyclosporine>-<relation>-<Gene>-<relation>-<Psoriasis>**

| Cyclosporine | Compound-Gene | Entrez Genes | Gene-Disease | Psoriasis | | Doc | | Hit | | Doc | | Hit |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cyclosporine | inhibit | Interferon | affect | Psoriasis | 1 | 1789988 | 1 | In addition, ciclosporin blocked the interferon-gamma-induced increase in epidermal 12(S)-HETE binding. | ▸ 9 | 9856816 | 1 | ... helper 1-type cytokines such as interferon-gamma in psoriasis. |
| Cyclosporine | inhibit | IL8 | affect | Psoriasis | ▸ 2 | 9588080 | 1 | It was found out that CsA inhibits IL-8 production by stimulated THP-1 monocyte cell ... | ▸ 8 | 11378328 | ▸ 2 | Interleukin-8-positive neutrophils in psoriasis. |
| Cyclosporine | affect | CALM3 | affect | Psoriasis | 1 | 2277142 | 1 | Cyclosporine binds to calmodulin with low affinity, and ... | ▸ 7 | 1879887 | 1 | Epidermal calmodulin levels in psoriasis before & after therapy. |
| Cyclosporine | inhibit | Growth factor | affect | Psoriasis | ▸ 2 | 8884530 | 1 | FK506 and cyclosporin A inhibit growth factor-stimulated human keratinocyte proliferation by blocking cells in the ... | ▸ 6 | 14962110 | 1 | Single-nucleotide polymorphisms of vascular endothelial growth factor in psoriasis of early onset. |

Linguamatics

# "Which B relate to A and C?"

## Potential Biomarkers found in Serum/Plasma

**<Disease>-<relation>-<Gene>-<location>-<Serum/Plasma>**

| Class1 | Relation | Class2 | Tissue | | Doc | | Hit | | Doc | | Hit |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Breast Cancer | associate | ERBB2 | plasma | 35 | 15756435 | 6 | Detection of Her2/neu, c-MYC and ZNF217 gene amplification during breast cancer progression using fluorescence in situ hybridization. | 1 | 19372565 | 1 | ... protein in tumor tissue and the HER-2 extracellular domain in plasma were used to show interdiction ... |
| Breast Cancer | associate | TNF | plasma | 3 | 15999154 | 3 | However, the effect of TNFA and TNFB gene polymorphisms on the expression of steroid receptors in breast cancer cells is not well documented. | 11 | 7542532 | 2 | ... in a significant rise in plasma TNFalpha levels(0.05 +/- 0.05 ng ... |
| Breast Cancer | associate | IL6 | serum | 3 | 16115908 | 1 | The IL6 polymorphism was significantly associated with breast cancer. | 11 | 19099662 | 2 | ... in pigs: TNF-alpha, IL-6 expressions in serum and rates of MODS after ... |
| Breast Cancer | associate | INS | plasma | 5 | 16393696 | 2 | Insulin-like growth factor II mediates resveratrol stimulatory effect on cathepsin D in breast cancer cells. | 10 | 19095054 | 2 | ... and a 38.5% increase in plasma insulin at 60 min, compared ... |

Linguamatics

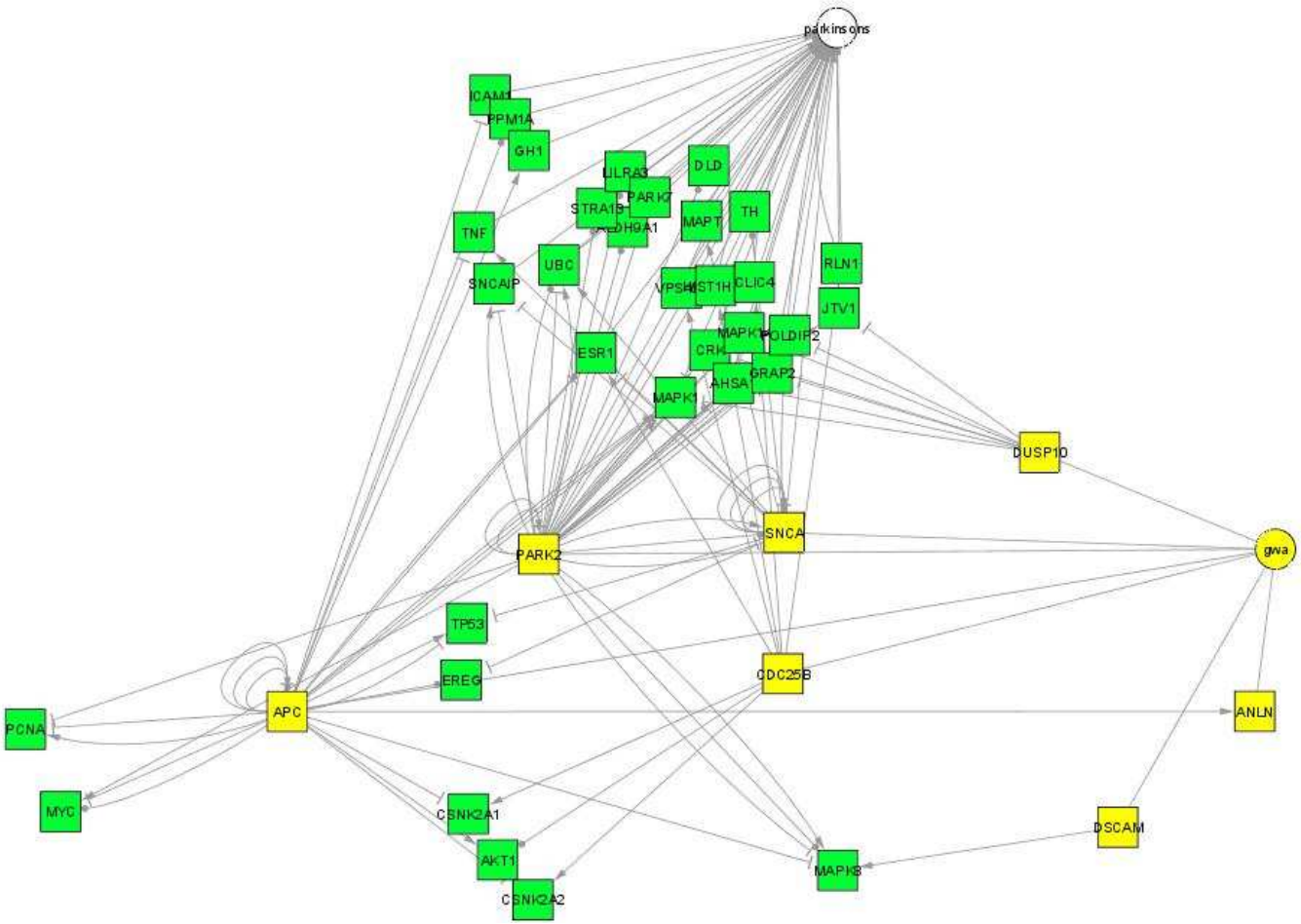
# Overview

- Agile Text Mining
- Extraction from large documents
  - Safety report case study
- Semantic Insights via
  - Linking information extracted from different documents
  - Linking structured and unstructured knowledge
  - Linking chemical and biological knowledge

Linguamatics

# Linking Candidate GWAS Genes with Diseases:
## Case Study with InforSense: Parkinson's Disease

**<GWAS>-<relation>-<Gene>* -<relation>-<Parkinson's>**




**Cytoscape Interaction Network**

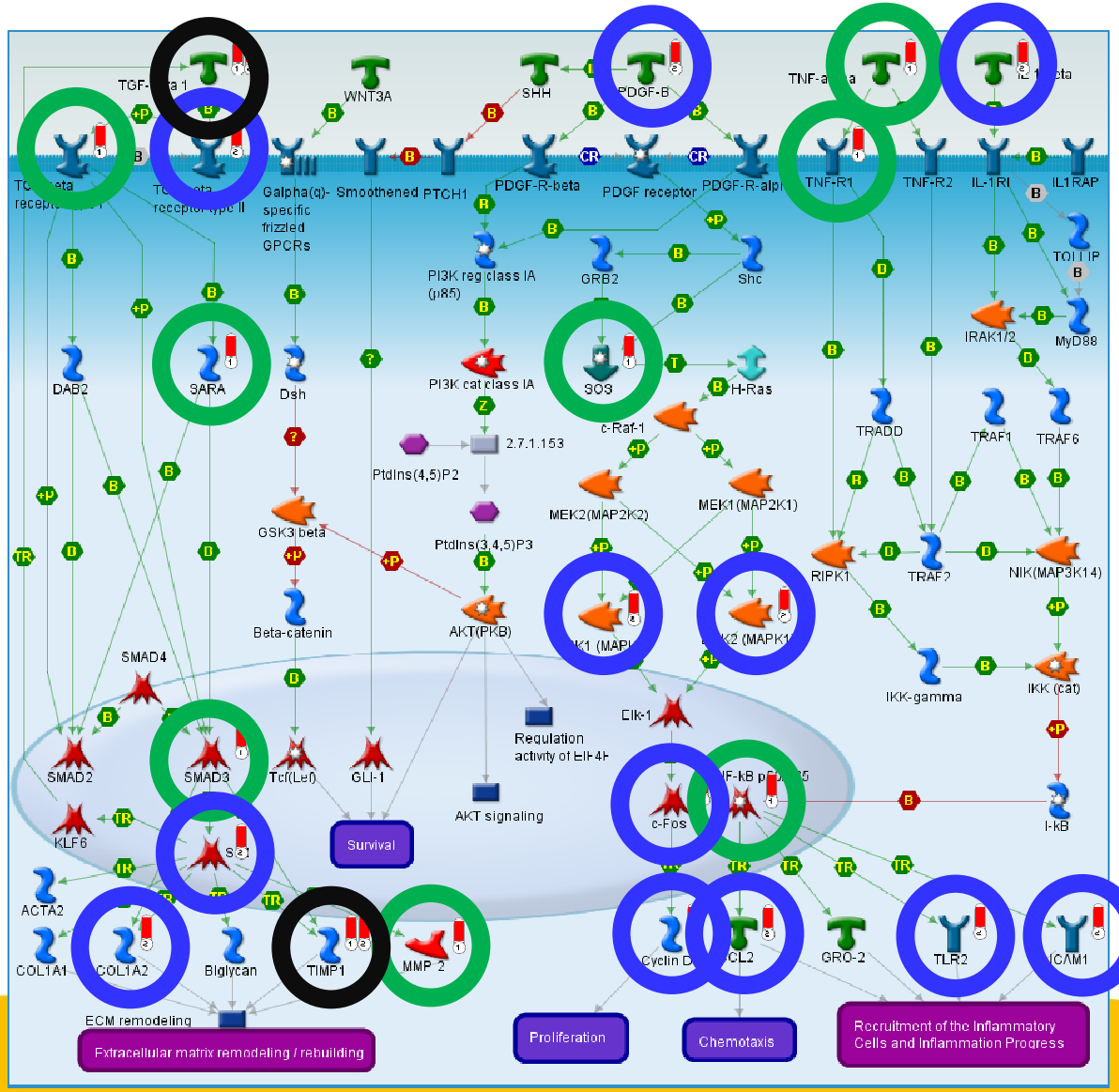
# Merging Structured and Unstructured:
## Case Study with GeneGo: Liver Fibrosis

Identifying disease pathways with more confidence by combining information from GeneGo and I2E

Blue: 12 GG genes

Green: 8 LM genes

Black: 2 common in both sets



Linguamatics

# Overview

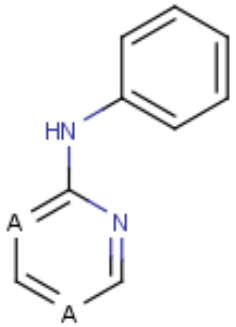
- Agile Text Mining
- Extraction from large documents
  - Safety report case study
- Semantic Insights via
  - Linking information extracted from different documents
  - Linking structured and unstructured knowledge
  - Linking chemical and biological knowledge

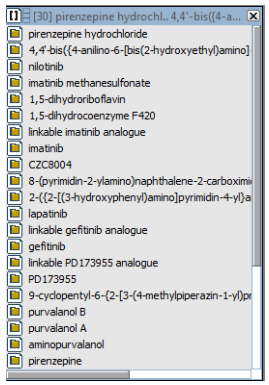
Linguamatics

# Linking Chemical and Biological Knowledge:
## Integrated Substructure Search from ChemAxon with I2E Text Mining

**<Structure>-<is-substructure>-<Chemical> -<relation>-<Target>**

**N(C1=CC=CC=C1)C1=*C=*C=N1**




| Class1 | Relation | Class2 | | Doc | | Hit |
|---|---|---|---|---|---|---|
| imatinib | inhibit | ABL1 | 2 | 16303243 | 1 | Imatinib, an inhibitor of BCR-ABL tyrosine kinase, also inhibits BCRP-mediated drug transport. |
| imatinib methanesulfonate | inhibit | ABL1 | 3 | 15803362 | 1 | BACKGROUND: Imatinib mesylate is a potent inhibitor of Abl, KIT, and PDGFR tyrosine kinases. |
| gefitinib | inhibit | EGFR | 2 | 15692759 | 2 | The clinical benefit and safety of the EGFR tyrosine kinase inhibitor gefitinib ('Iressa')1 was evaluated in this Phase II, multicentre study of patients with taxane and anthracycline pretreated, metastatic breast cancer. |
| lapatinib | inhibit | ERBB2 | 3 | 16452223 | 2 | Alternatively, inhibition of ErbB2 signaling using lapatinib (GW572016), a reversible small -molecule inhibitor of ErbB1/ErbB2 tyrosine kinases, at pharmacologically relevant concentrations, leads to marked inhibition of survivin protein with subsequent apoptosis. |

Linguamatics

# Semantic Insights using Agile Text Mining

- Derive structure from unstructured and semi-structured text
- Text treated as a database
  - With dynamic relationships
- Query results are structured, allowing
  - Further analysis
  - Visualisation
  - Gaps to be filled in structured knowledge
- Derive new insights by combining information from multiple
  - Documents
  - Data types
  - Disciplines