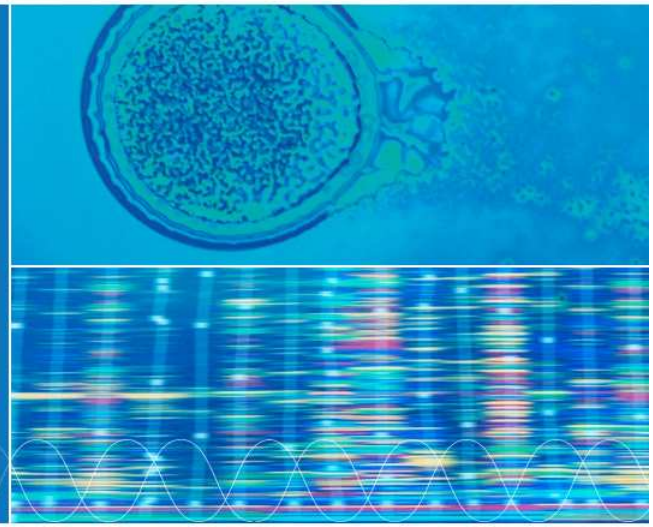
ARIADNE

Semantic Search in Biomedical Science Content

Applications for an NLP Engine

Martin Griffies| October 2010.



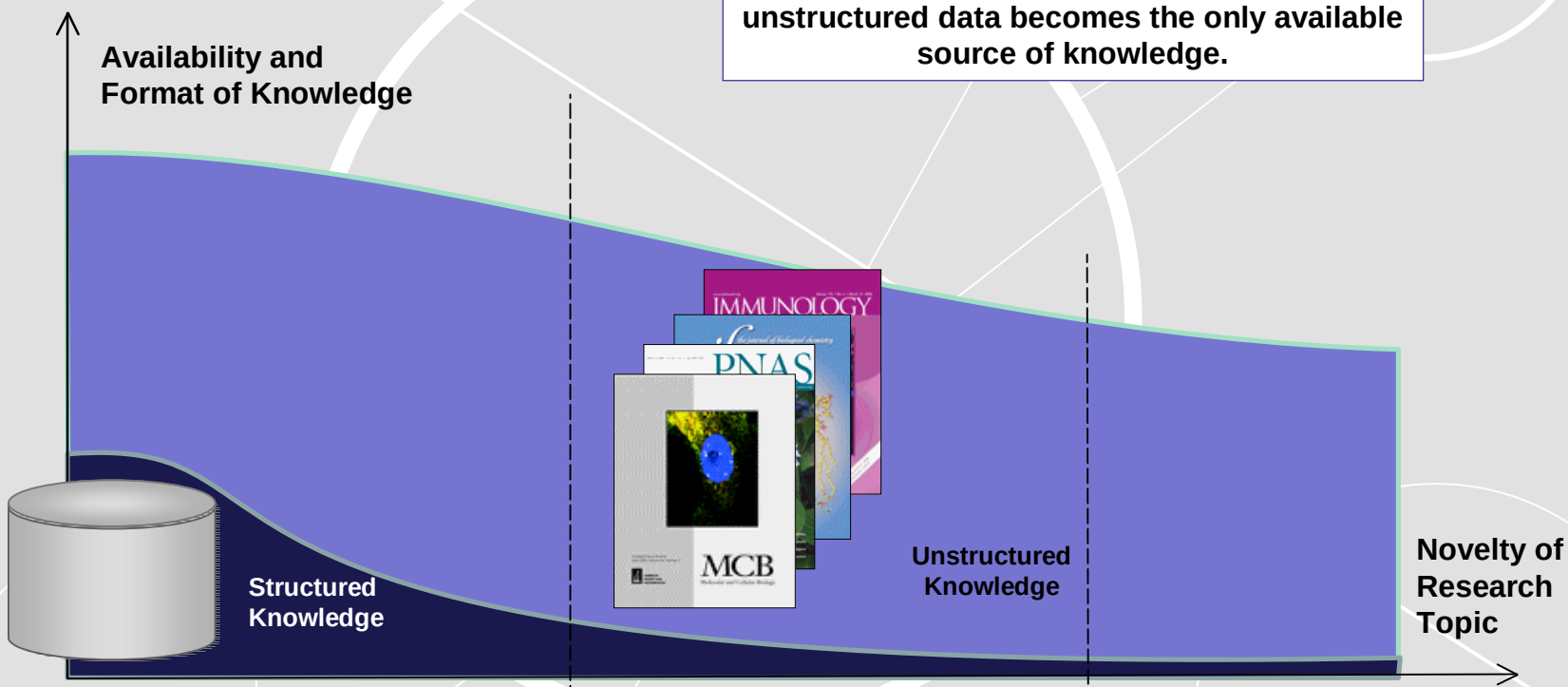
ARIADNE

We have a problem

Too much information, too little knowledge

- External data
  - Free : PubMed, ClinicalTrials.gov, Open Access journals, Espace, ..
  - Subscription journals, digests, secondary sources
  - WWW search results
  - News feeds, advertisements
  - Conferences
  - etc
- Internal data
  - Structured: databases, tables, curated data
  - Unstructured – reports, scanned documents.

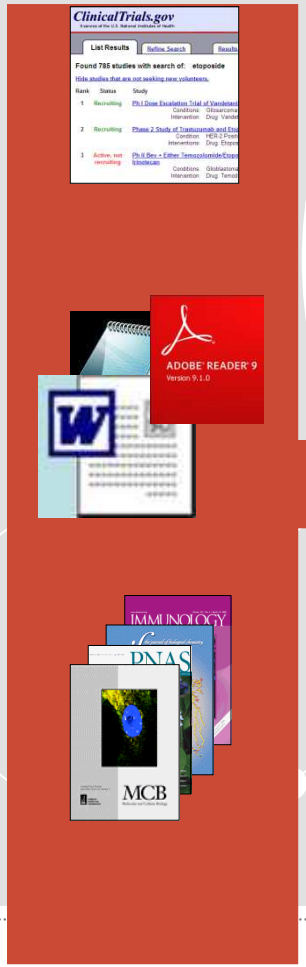
The Evolving Challenge



ARIADNE

The challenge of volume

More is not necessarily better

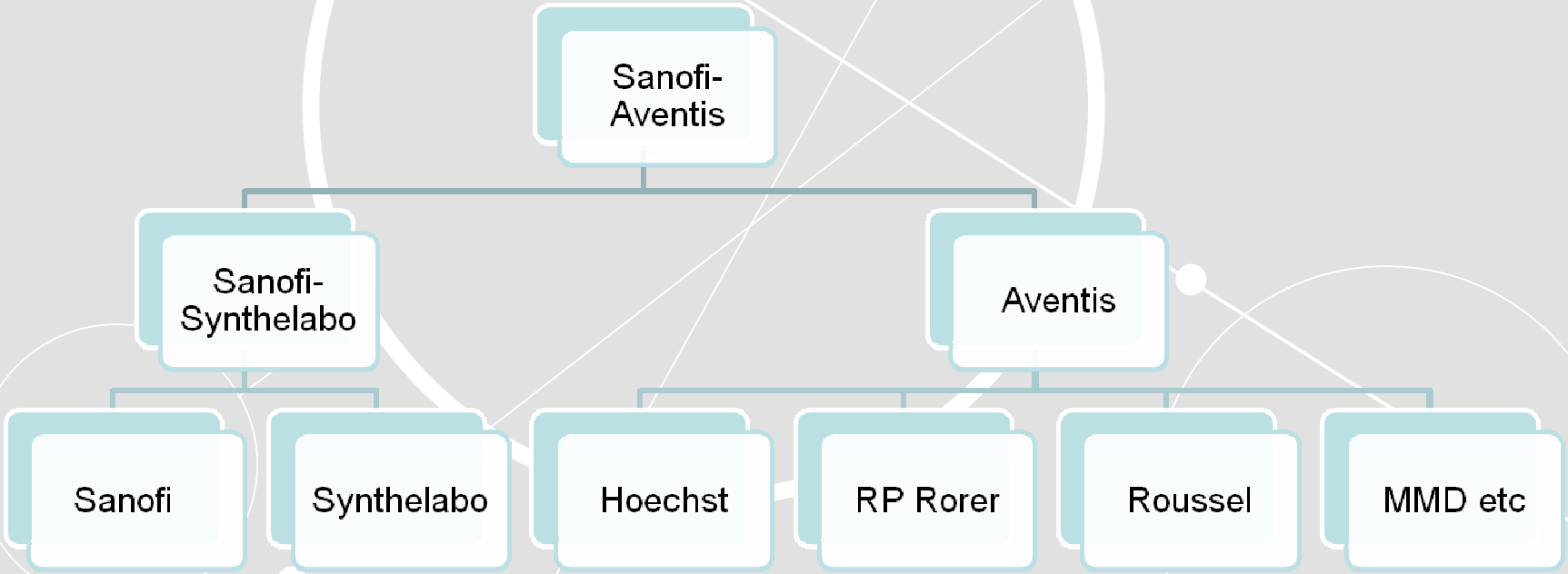


- 20 million PubMed abstracts
- 500.000 USPTO applications / year
- PubChem: 72 million records / 28 m structures.

ARIADNE

The challenge of Origin

Time, location, language: Integration.

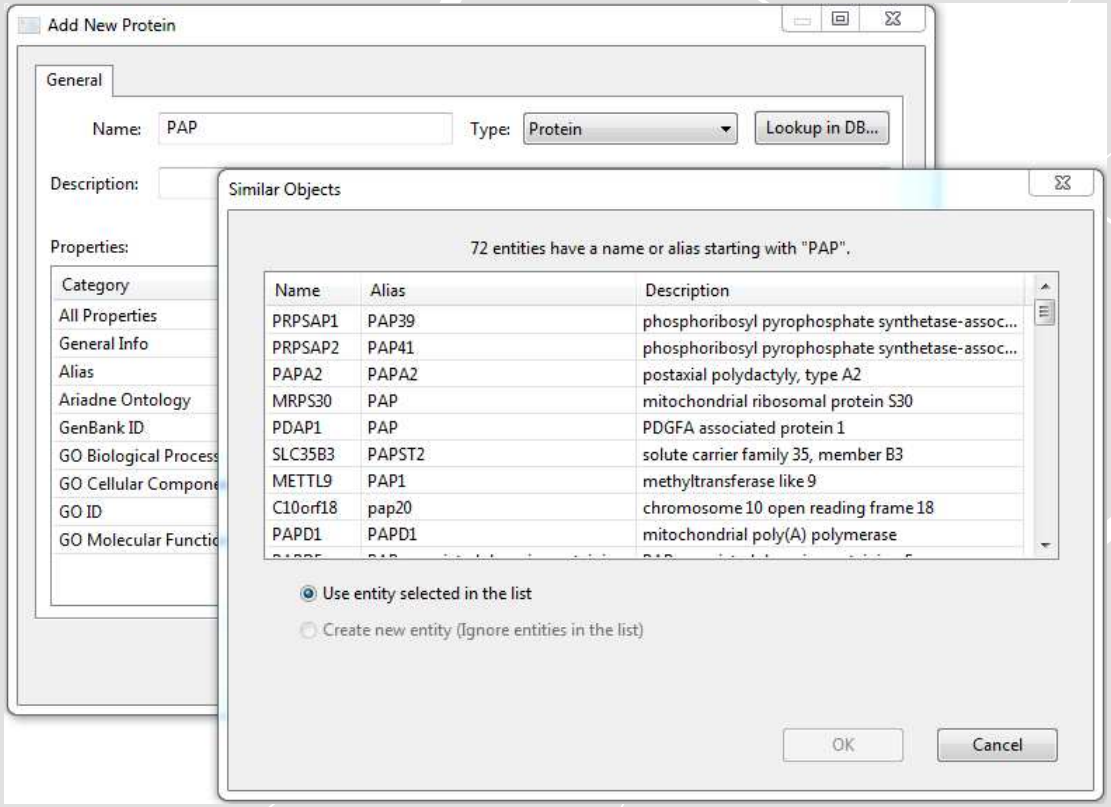


ARIADNE

The challenge of Meaning

The need for Semantics

- 1700 synonyms for APAP
- 70+ entities named “PAP”



ARIADNE

The Search for meaning

90% of Searches are Keyword-based

- Retrieves documents, not facts.
- But facts / assertions / relationships are (usually) what users want.
- RELEVANCE is needed

NLP Processing

11940574:7 Because **Axin2** has been shown to$_{Main}$ associate with and inhibit **beta-catenin** abundance and function,$_{Main}$ we hypothesized that$_{Main_That}$ **Axin 2**, which is affecting proliferation of MEF cell line$_{NPmod}$ can work in a negative feedback pathway$_{Main}$, regulating **Wnt** signaling and thus controlling apoptotic process$_{VPmod}$.

Syntactic normalization:

1. Verbal conjunction
2. VP conjunctions
3. NP conjunctions
4. NP modifiers
5. VP modifiers
6. Some syntactic movements
7. NP prepositional structure
8. Relational clauses
9. Nominalizations

Decomposed primitive sentences (triplets):

Axin2	**associate**	**beta-catenin abundance**
Axin2	**associate**	**beta-catenin function**
Axin2	**inhibit**	**beta-catenin abundance**
Axin2	**inhibit**	**beta-catenin function**
Axin2	**affect**	**MEF cell line proliferation**
Axin2	**work**	**negative feedback pathway**
Axin2	**regulate**	**Wnt signaling**
Axin2	**control**	**apoptotic process**

ARIADNE

From Text to Relations

"Axin binds beta-catenin and inhibits GSK-3beta activity in hepatocytes."

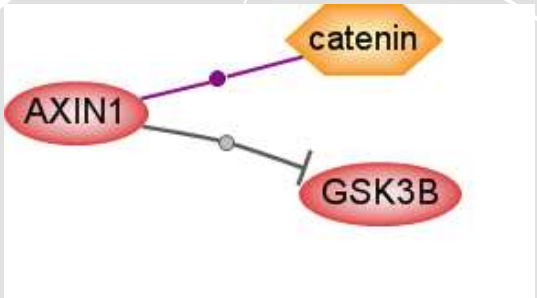
NLP: short statements (key facts in the form of triplets)

Axin	binds	**beta-catenin**
Axin	inhibits	**GSK-3beta** activity

Extracting relations

Axin - beta-catenin, relation: Binding

Axin -> GSK-3beta, relation: Regulation, effect: Negative

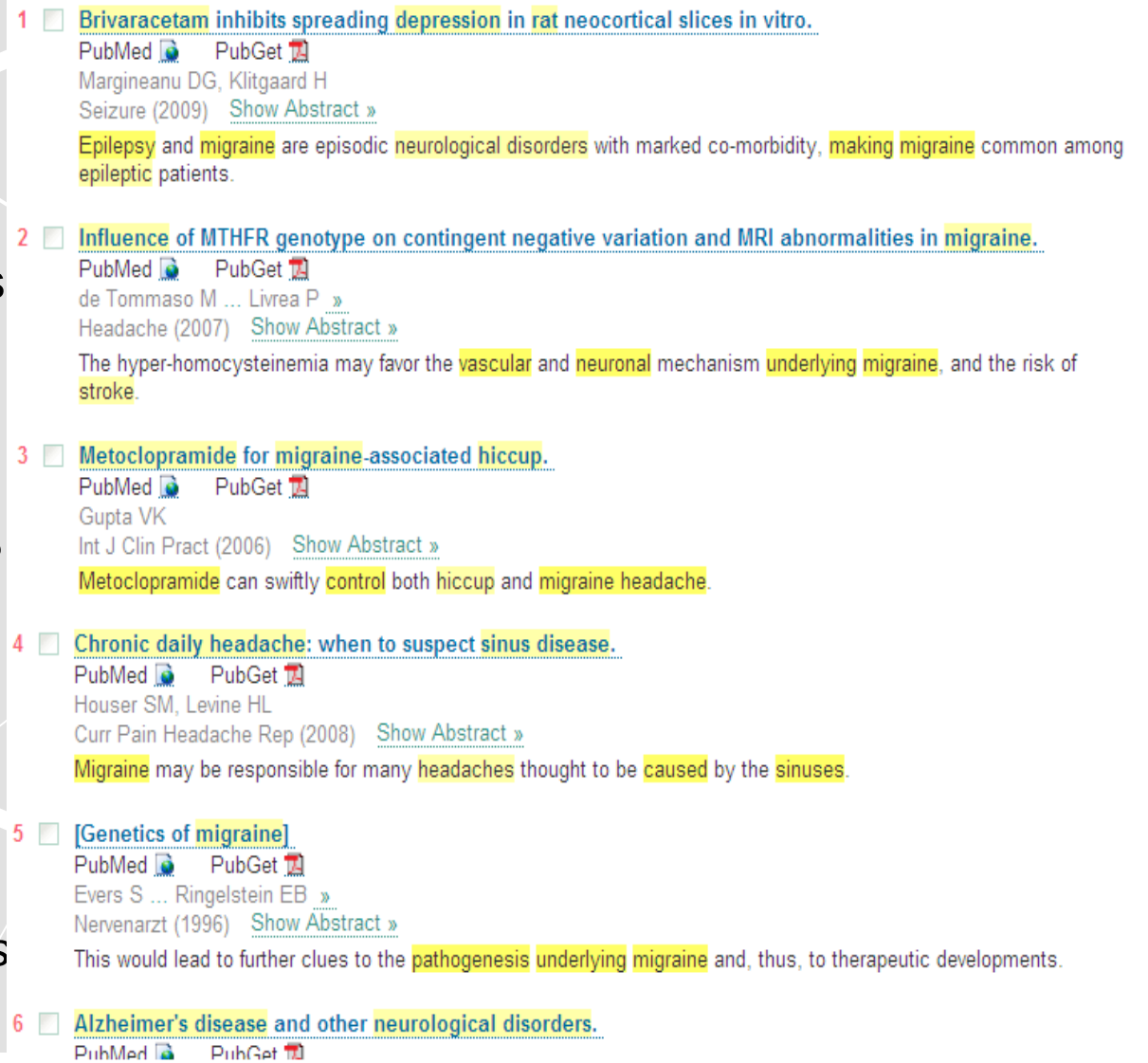


ARIADNE

Semantic Search System

- Finds the right documents
- Summarizes / precis findings
- Can also answer natural language questions:
  - "what causes migraine?

Can be tailored for specific domains by adding *ontologies*



ARIADNE

Applicable Technology

MedScan

Google

Internal Documents

Open Access

Dictionaries

IE Patterns

Grammar

Entity recognition

Pattern matching

ConceptScan

Relationship Extraction

Rules

Semantic processor

Cartridges

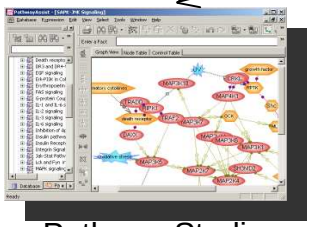
Mammalians

CI / etc

Toxicology

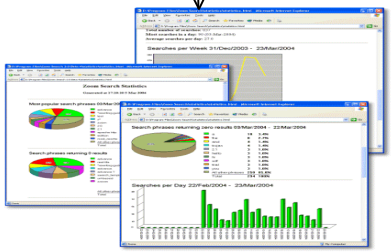
RNEF XML

Semantic document indexing/search



Pathway Studio Visualisation

Custom Databases

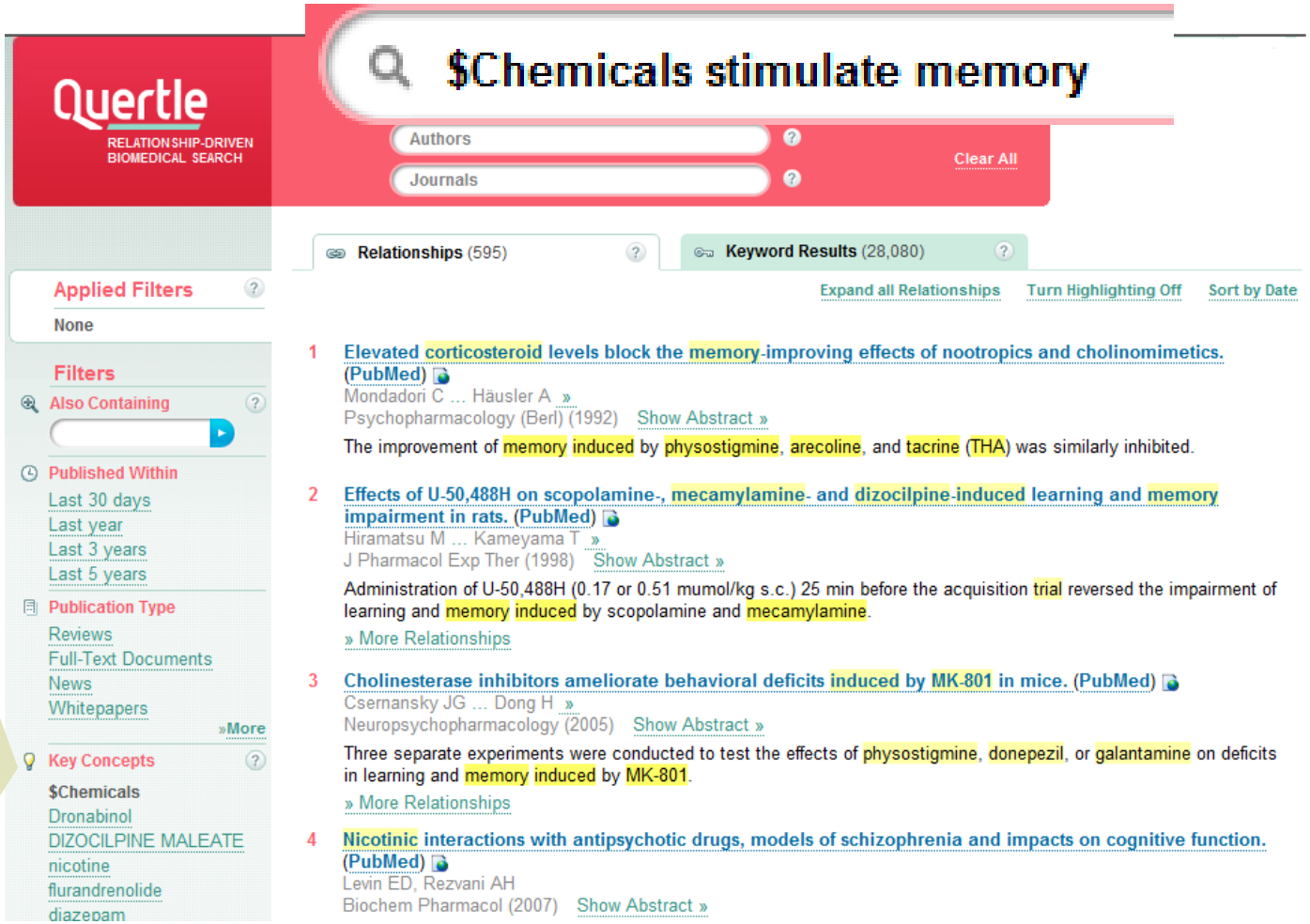


Knowledge and Analysis Reports

ARIADNE

Quertle - a Search Engine

- Dronabinol
- DIZOCILPINE MALEATE
- nicotine
- flurandrenolide
- diazepam
- Prodiamine
- DIZOCILPINE
- Trimethylstannane
- Benzodiazepine
- ECMA
- Angel dust
- okadaic acid
- DNA
- estrogen
- cycloheximide
- lorazepam
- Crospovidone
- streptozocin
- NMDA
- 8-OH-Dpat
- Trichloroethene
- Pentylenetetrazol
- baclofen
- 3-Quinuclidinyl benzilate
- Ibotenic acid
- kainic acid
- carbachol
- mecamylamine



www.quertle.com

ARIADNE

Cartridges & Ontologies

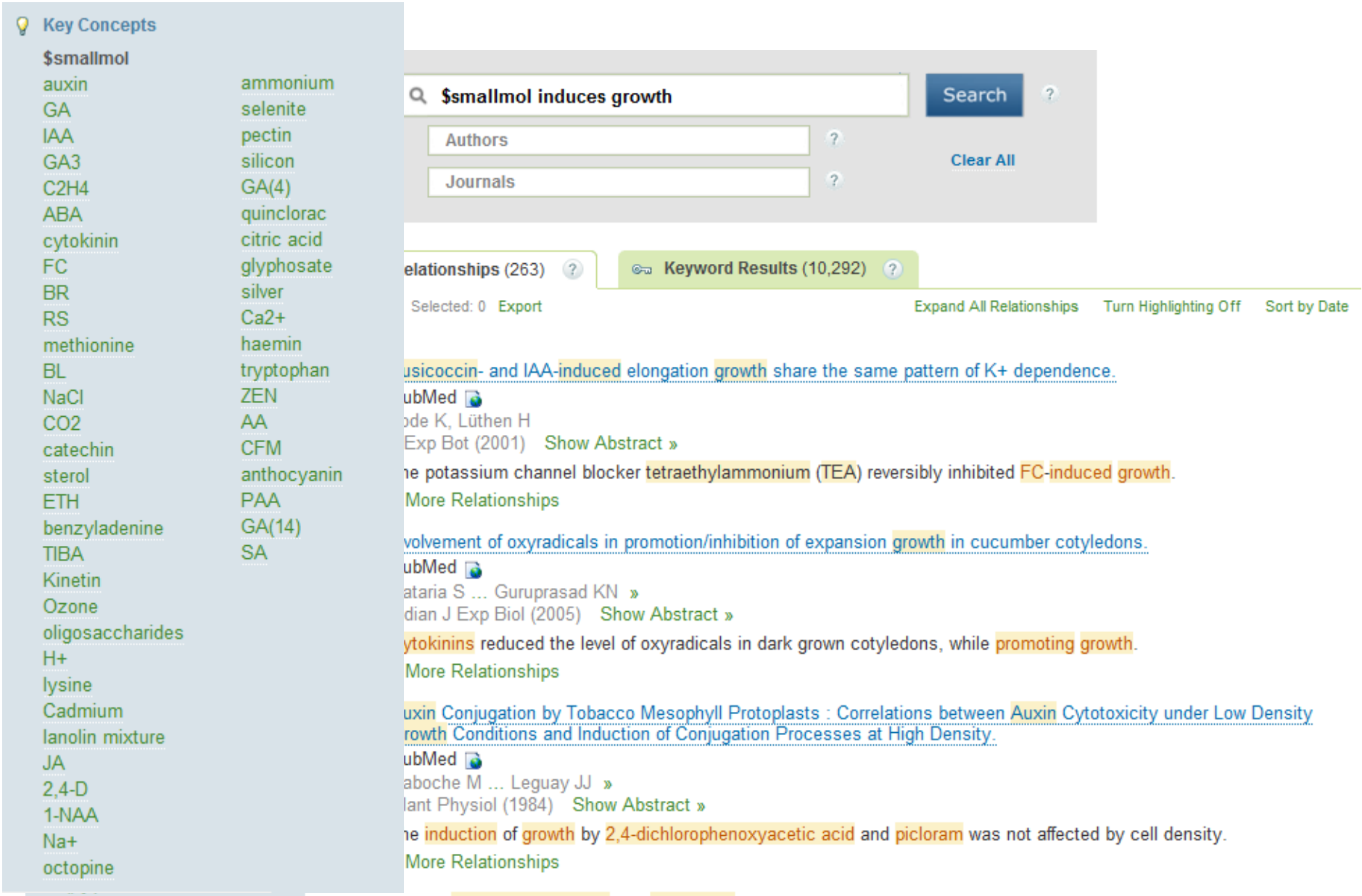
Flexibility & Domain Coverage

- Single engine, multiple domains
  - In-house dictionaries: canonical / preferred terms
  - Standard terminologies:
    - MeSH,
    - GO,
    - SNOMED,
    - MedDRA etc

- Clinical, AE records,
- Toxicology, Plants, Proteins,
- Competitive Intelligence

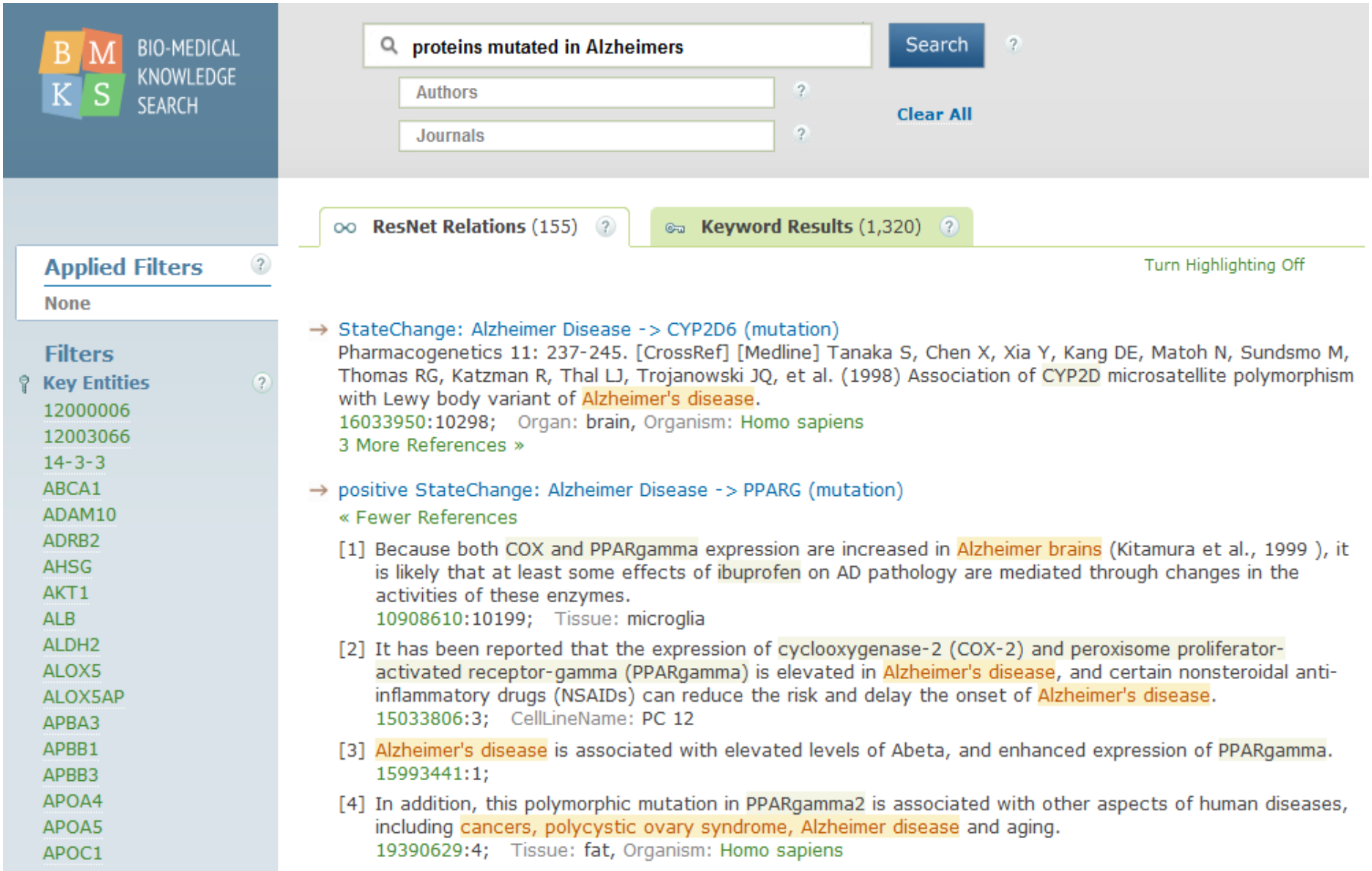
ARIADNE

BMKS App – Semantic Search (Plants)



ARIADNE

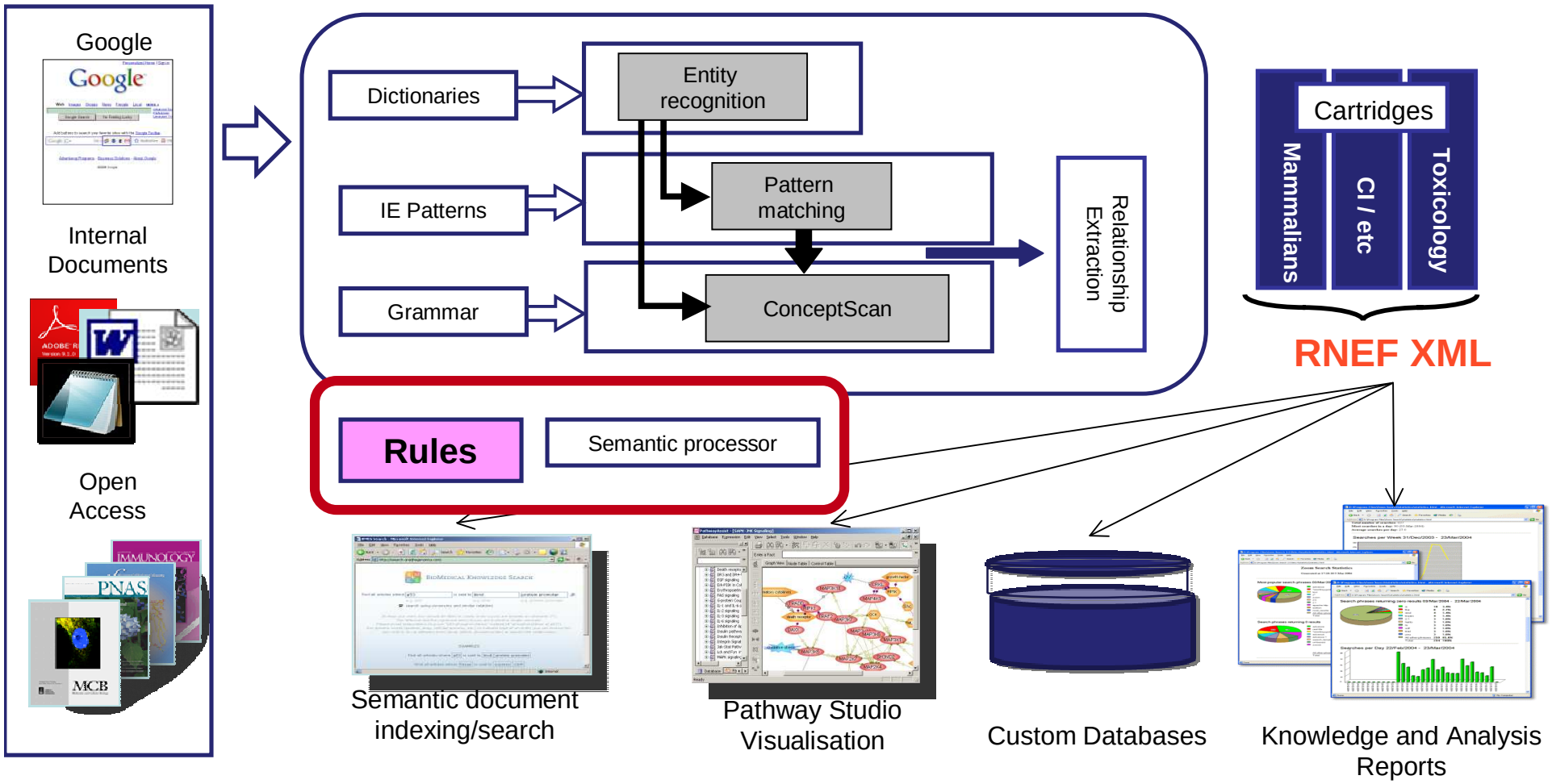
BMKS Application – Translational Medicine



Applicable Technology

ARIADNE

MedScan

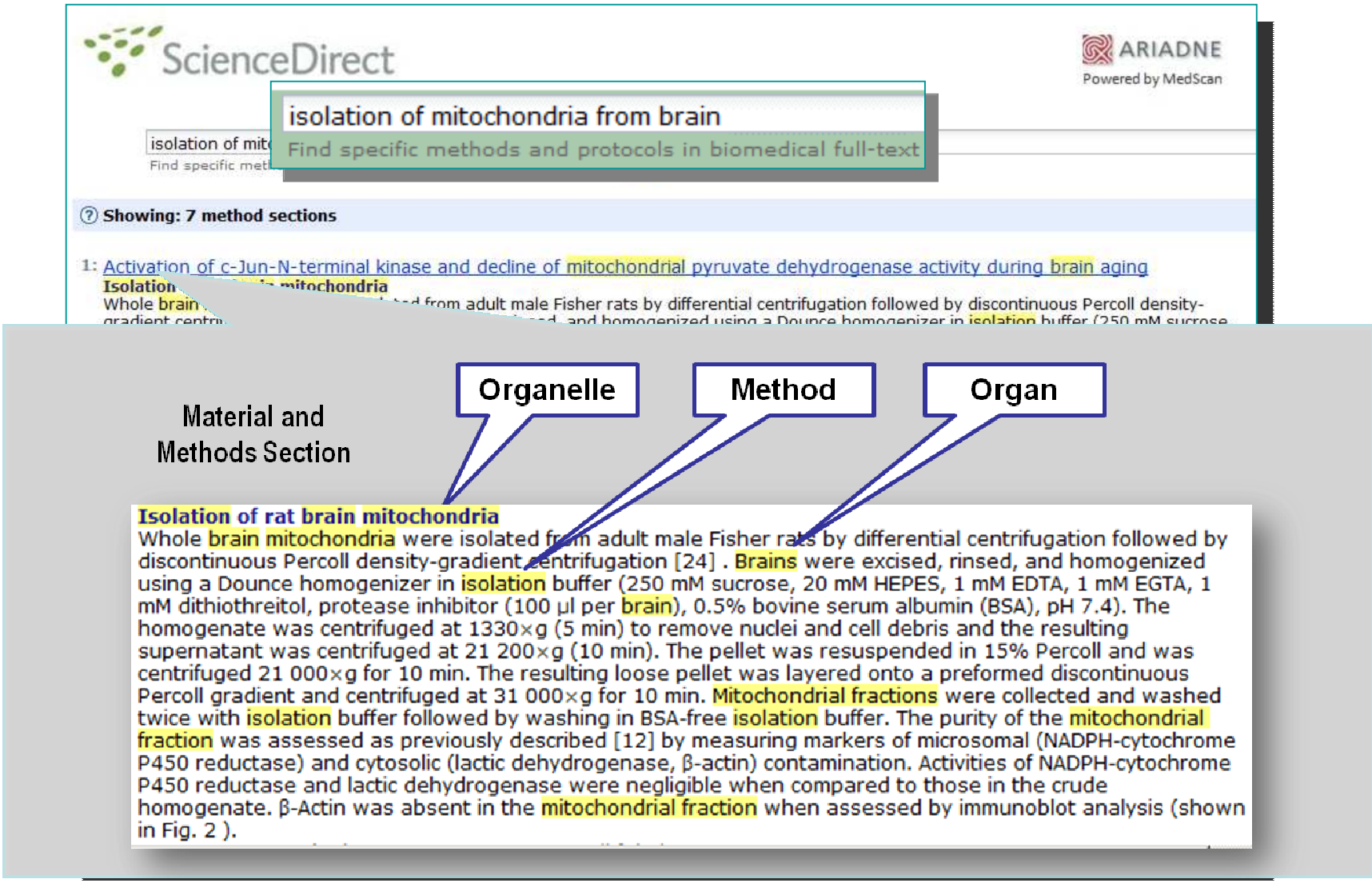


Examples of MedScan customization projects performed for customers:

1) Immunology: facts about cell types
2) Biomarkers: automatic classification of biomarkers
3) ClinicalTrials.gov: import of XML formatted clinical trial reports
4) Phosphosites: extraction of protein phosphorylation sites
5) Mutations: extraction of mutations linked to diseases

ARIADNE

Use of Rules: Materials & Methods



ARIADNE

Cell, Volume 139, Issue 6, 1084-1095, 11 December 2009

doi:10.1016/j.cell.2009.11.015

Article

The Structural Basis for mRNA Recognition and Cleavage by the Ribosome-Dependent Endonuclease RelE

Cajetan Neubauer[1, 4], Yong-Gui Gao[1, 4], Kasper R. Andersen[2, 4], Christine M. Dunham[1, 5], Ann C. Kelley[1], Jendrik Hentschel[1], Kenn Gerdes[3], V. Ramakrishnan[1], and Ditlev E. Brodersen[2]

[1] MRC Laboratory of Molecular Biology, Cambridge CB2 0QH, UK

[2] Department of Molecular Biology, Aarhus University, DK-8000 Aarhus C, Denmark

[3] Institute for Cell and Molecular Biosciences, The Medical School, University of Newcastle, Newcastle NE2 4HH, UK

Corresponding author

Corresponding author

[4] These authors contributed equally to the work

[5] Present address: Department of Biochemistry, Emory University School of Medicine, Atlanta, GA 30322, USA

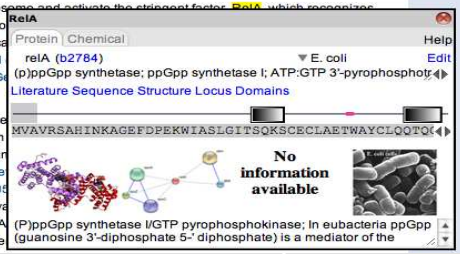
Summary

Translational control is widely used to adjust gene expression levels. During the stringent response in bacteria, mRNA is degraded on the ribosome by the ribosome-dependent endonuclease, RelE. The molecular basis for recognition of the ribosome and mRNA by RelE and the mechanism of cleavage are unknown. Here, we present crystal structures of *E. coli* RelE in isolation (2.5 Å) and bound to programmed *Thermus thermophilus* 70S ribosomes before (3.3 Å) and after (3.6 Å) cleavage. RelE occupies the A site and causes cleavage of mRNA after the second nucleotide of the codon by reorienting and activating the mRNA for 2′-OH-induced hydrolysis. Stacking of A site codon bases with conserved residues in RelE and 16S rRNA explains the requirement for the ribosome in catalysis and the subtle sequence specificity of the reaction. These structures provide detailed insight into the translational regulation on the bacterial ribosome by mRNA cleavage.

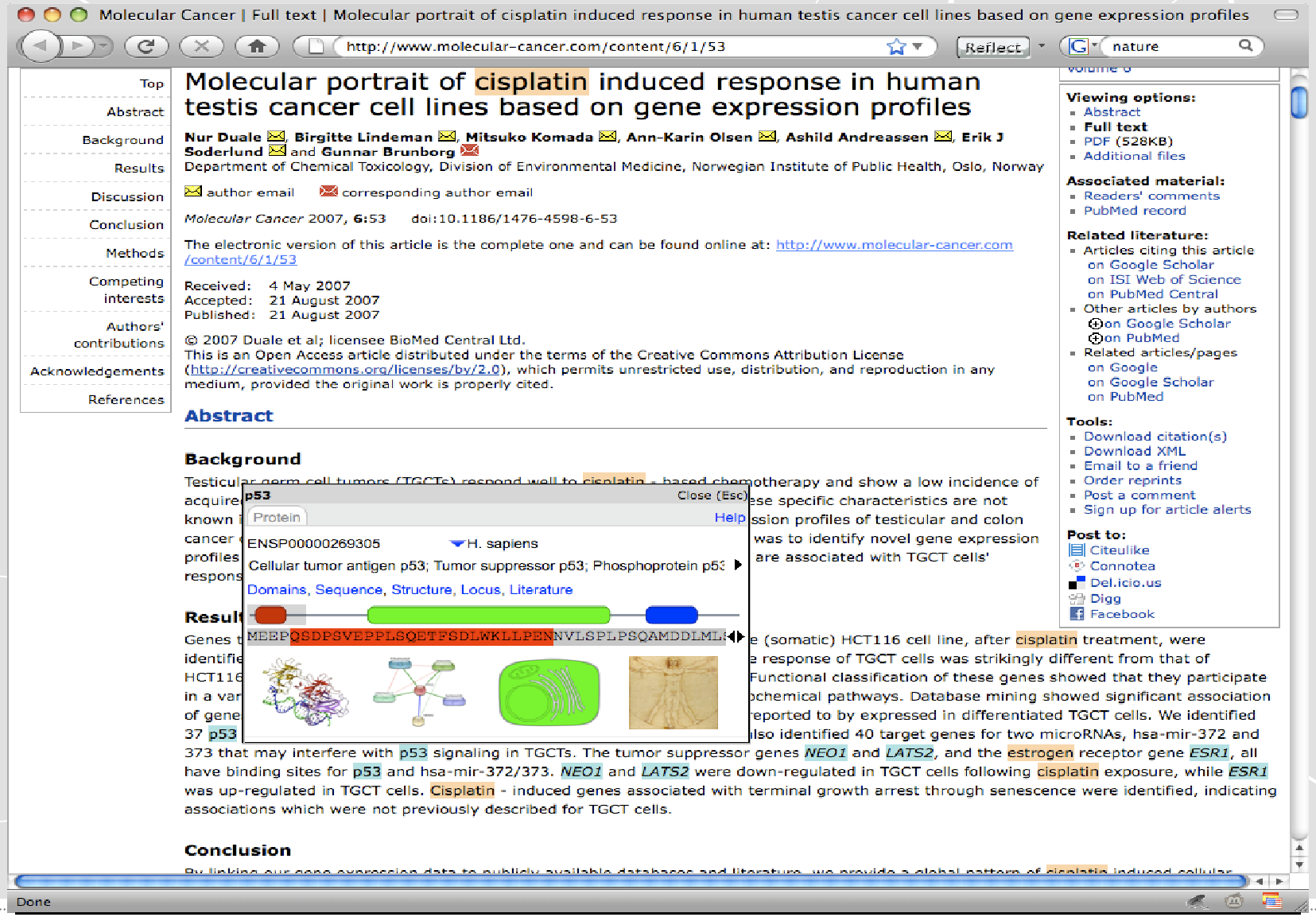
INTRODUCTION

Rapid adaptation to environmental stress is vital for free-living bacteria. During deprivation of nutrients, uncharged transfer RNAs (tRNAs) bind to the ribosome and activate the stringent factor, RelA, which recognizes stalled ribosomes and catalyzes synthesis of the signal nucleotide, (p)ppGpp. This alarmone regulates the *stringent response*, a far-reaching adaptation o (Potrykus and Cashel, 2008). The stringent response also leads to activation of RelE, an effective inhibitor of protein synthesis. Under normal physiologica RelB, and is inactive (Christensen and Gerdes, 2003,Christensen et al., 2001,Galvani et al., 2001,Gotfredsen and Gerdes, 1998,Li et al., 2009,Overgaard degradation of RelB by Lon protease, RelE is able to bind the ribosome and specifically cleave messenger RNA (mRNA) in the A site (Christensen and G 2009,Pedersen et al., 2003).

Such toxin-antitoxin pairs are very common in bacteria, and the RelE superfamily also encompasses HigB, YoeB, YafQ, and YhaV, associated with a varie Gerdes, 2006,Grady and Hayes, 2003,Prysak et al., 2009,Schmidt et al., 2007). The crystal structure of YoeB showed that its fold and catalytic mechanism (Kamada and Hanaoka, 2005). RelE is structurally similar to YoeB (Li et al., 2009,Takagi et al., 2005) but lacks the conserved catalytic histidine and glutan an intrinsic nuclease activity of the ribosome (Garza-Sanchez et al., 2008,Hayes and Sauer, 2003,Kamada and Hanaoka, 2005,Li et al., 2009,Sunohara e pausing ribosomes are recovered for rescue by tmRNA in the absence of stringent response factors (Hayes and Sauer, 2003,Kamada and Hanaoka, 200 most likely a result of the combined action of RelE-like endonucleases and exonucleases like RNase II (Garza-Sanchez et al., 2009). RelE-induced cleava site codon, although it is occasionally also seen after the third nucleotide and, upon peptide release, even in the E site (Pedersen et al., 2003). The mRNA efficiency, with the UAG and UGA stop codons and sense codons like UCG and CAG, among the most efficiently cleaved (Pedersen et al., 2003). Togethe

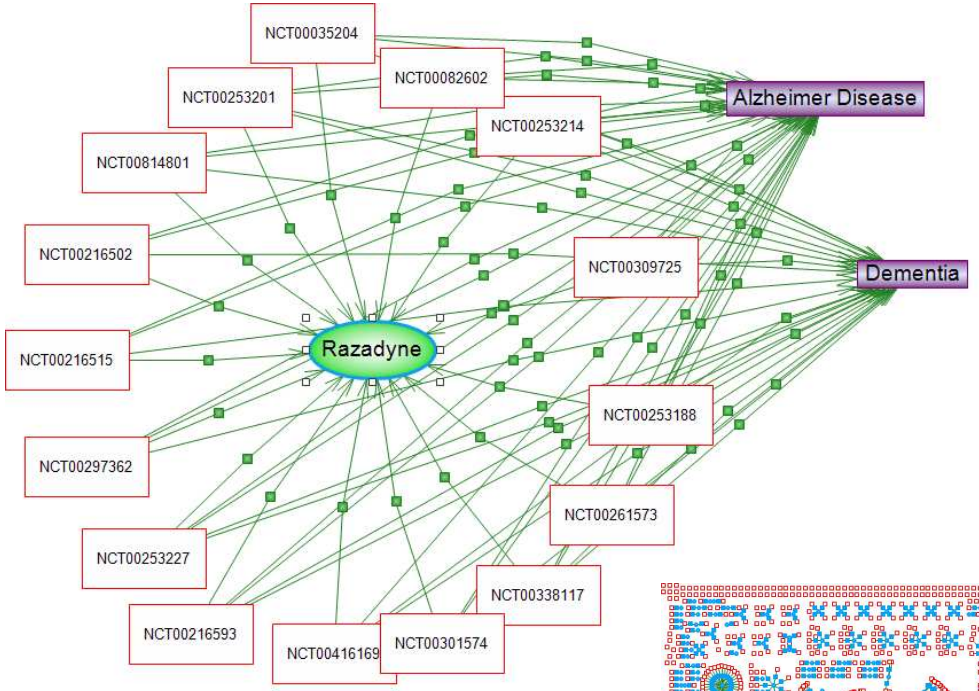


ARIADNE



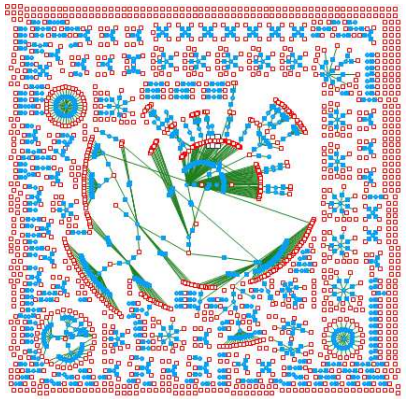
ARIADNE

Project – ClinicalTrials.gov



Data Components:

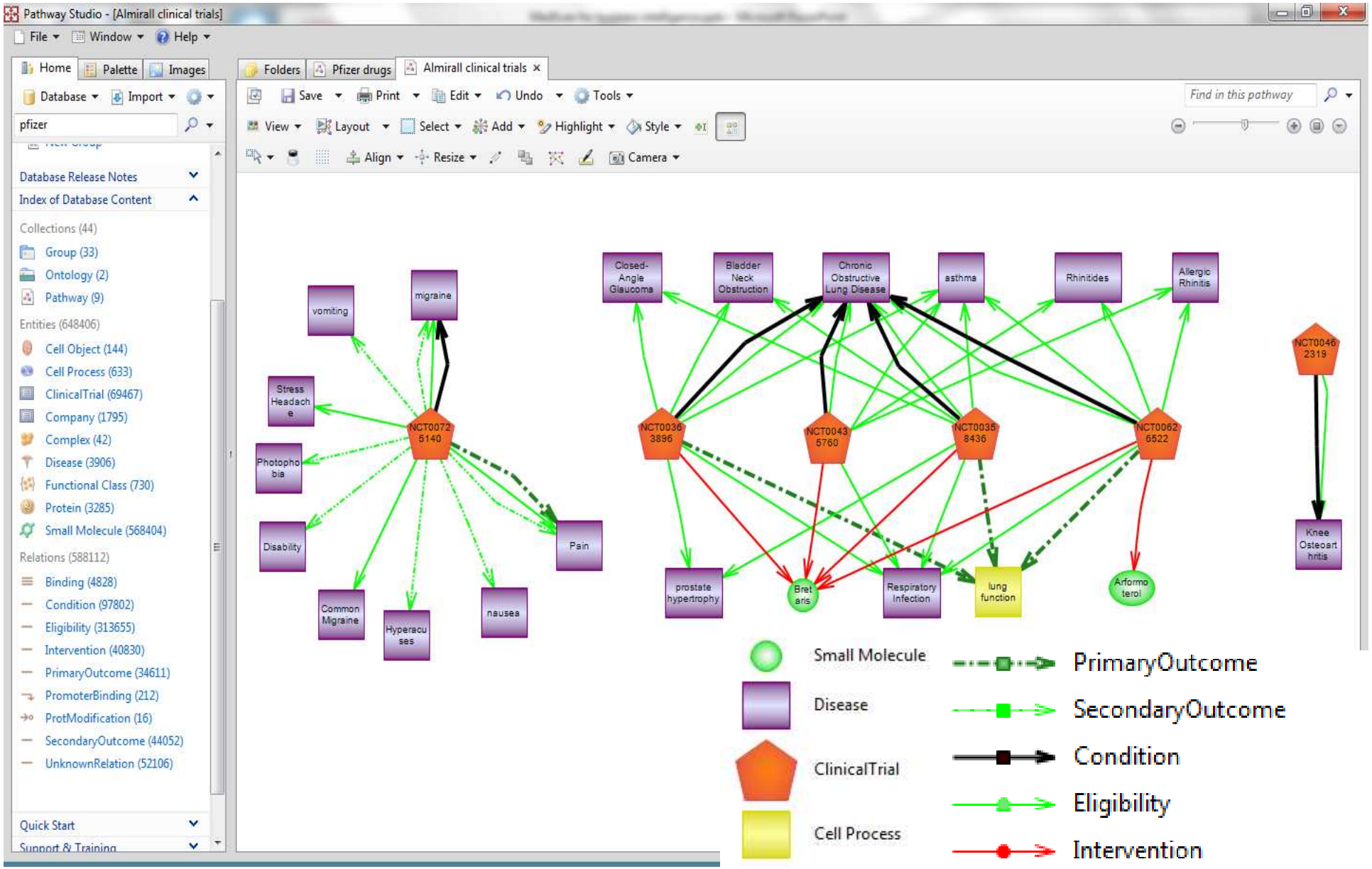
- Study
- Status
- Start
- Company
- Phase
- Primary Outcome



Use clinical trials data with Tox/Small molecule cartridge for abstract deduction.

ARIADNE

Information extracted by Clinical trial cartridge. Statistics from >85,000 XML files



ARIADNE

Extracted fields:

ClinicalTrials.gov
A service of the U.S. National Institutes of Health

Home Search Study Topics Glossary Search

Full Text View | Tabular View | No Study Results Posted | Related Studies

A Trial Assessing LAS34273 in Moderate to Severe Stable Chronic Obstructive Pulmonary Disease (COPD)

This study has been completed.

First Received: August 10, 2006 Last Updated: September 1, 2010 History of Changes

Sponsor:	Almirall, S.A.
Collaborator:	Forest Laboratories
Information provided by:	Almirall, S.A.
ClinicalTrials.gov Identifier:	NCT00363896

Purpose

To evaluate the efficacy and safety of LAS 34273 compared to placebo in patients with moderate to severe COPD during one year of treatment.

Condition	Intervention	Phase
Chronic Obstructive Pulmonary Disease	Drug: Aclidinium bromide	Phase III

Study Type: Interventional
Study Design: Allocation: Randomized
Control: Placebo Control
Intervention Model: Parallel Assignment
Masking: Double-Blind

Official Title: Clinical Trial Assessing Efficacy and Safety of LAS34273 in Moderate to Severe Stable Chronic Obstructive Pulmonary Disease (COPD) Patients

Resource links provided by NLM:

MedlinePlus related topics: COPD (Chronic Obstructive Pulmonary Disease)

Drug Information available for: Aclidinium bromide

U.S. FDA Resources

Further study [illegible] provided by Almirall, S.A.:

Primary Outcome Measures:
- Lung function

Secondary Outcome Measures:
- Exacerbations and Quality of Life

Study Start Date: August 2006
Study Completion Date: May 2008
Primary [illegible] Date: May 2008 (Final data collection date for primary outcome measure)

Eligibility

Ages Eligible for Study: 40 Years and older
Genders Eligible for Study: Both
Accepts Healthy Volunteers: No

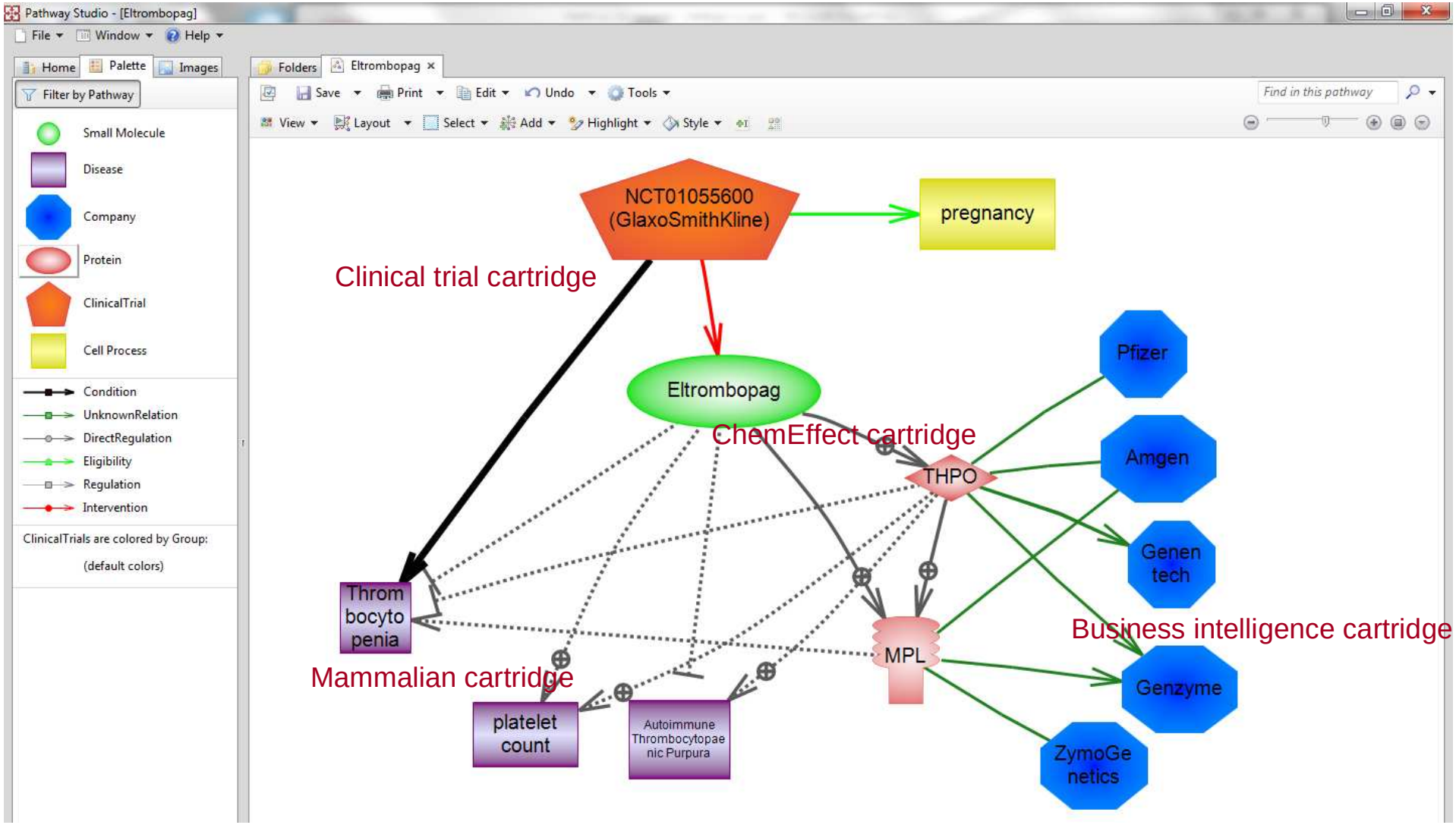
Criteria
Inclusion Criteria:
- Males and females aged ≥ 40 years with a clinical diagnosis of moderate to severe stable COPD

Exclusion Criteria:
- History or current diagnosis of asthma, recent respiratory tract infection or acute COPD exacerbation, life expectancy of less than 1 year, known symptomatic prostatic hypertrophy, bladder neck obstruction or narrow-angle glaucoma

ARIADNE

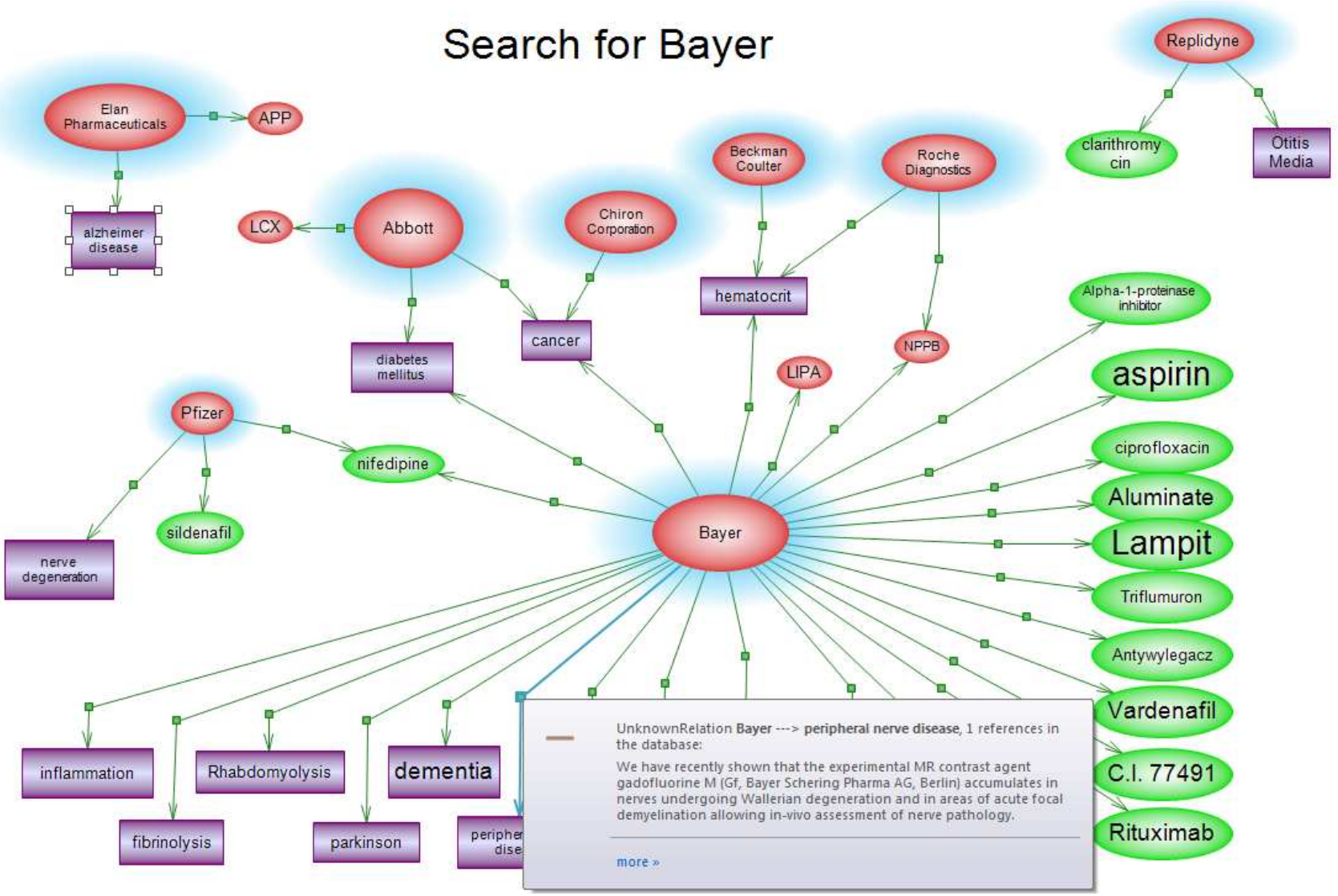
Combined cartridges:

What companies are developing TPO therapies?



ARIADNE

Project - Business Intelligence



ARIADNE

Solution requirements

- Highly Curated Dictionaries and Patterns
  - On-going manual effort aided with statistical analysis
- Advanced NLP Algorithms:
  - Context sensitive disambiguation
  - Pattern / Dictionary based entity detection
  - Parallel Entity Detection
  - ConceptScan® Technology
- Curator
  - Statistical error model for rules-based "error detection and correction" algorithm

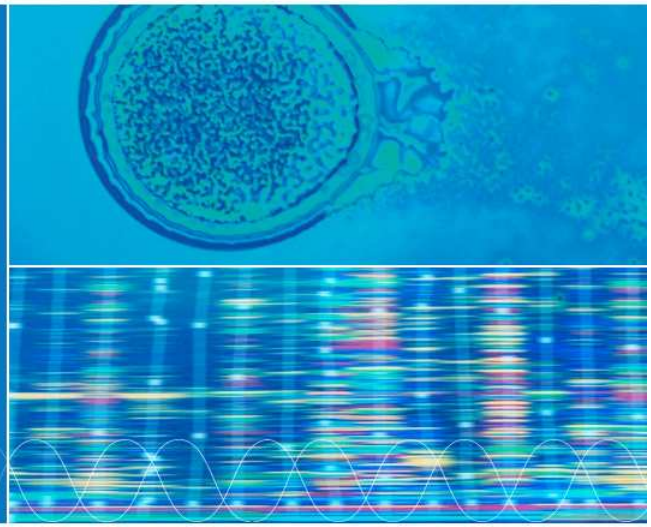
ARIADNE

ARIADNE

Thank You.
Questions?

Applications for an NLP Engine

mjg@ariadnegenomics.com



Acknowledgements:
Nikolai Daraselia, Anton Yuryev,(Ariadne)
Jud Dunham, (Elsevier)
Reinhard Schneider, EMBL.