Direct Use of Information Extraction from Scientific Text for Modeling and Simulation in the Life Sciences

SCAI

Fraunhofer Institute Algorithms and Scientific Computing

Martin Hofmann-Apitius

Department of Bioinformatics
Fraunhofer Institute for Algorithms and Scientific Computing (SCAI)

Fraunhofer-Campus Schloss Birlinghoven

Institutes

- Algorithms and Scientific Computing **SCAI**
- Intelligent Analysis and Information Systems **IAIS**
- Applied Information Technology **FIT**

600 Scientists, 200 Students

Linked to Universities Bonn, Aachen and Cologne

Fraunhofer Institute Algorithms and Scientific Computing

b-it

universitätbonn

Direct Use of Information Extraction from Scientific Text for Modeling and Simulation in the Life Sciences

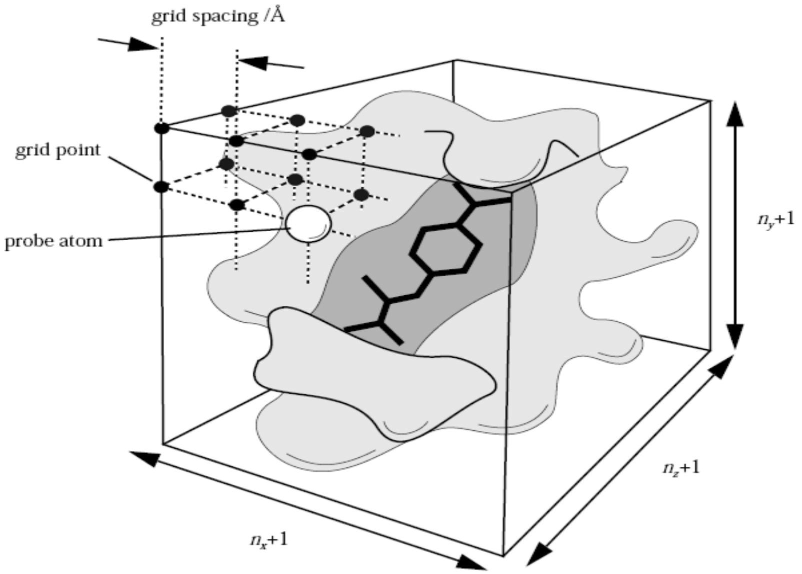
Paradigm Changes in the Life Sciences

In the Life Sciences, we currently observe significant paradigm changes

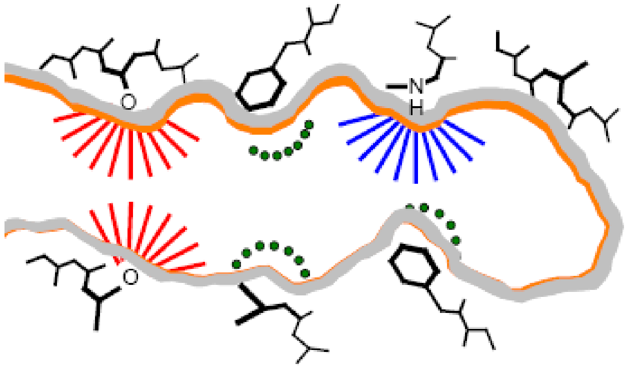
- the “omics” paradigm has lead to a flood of data and a flood of publications
- a single researcher cannot keep track with all the relevant (and related) literature any more
- everything is connected; genetics, molecular biology, biochemistry, pharmaceutical chemistry and organic chemistry are “networked”
- Biology and Chemistry and Medicine are more and more turning into quantitative sciences, described by mathematical models and with the option of using simulation (*in silico* experimentation)

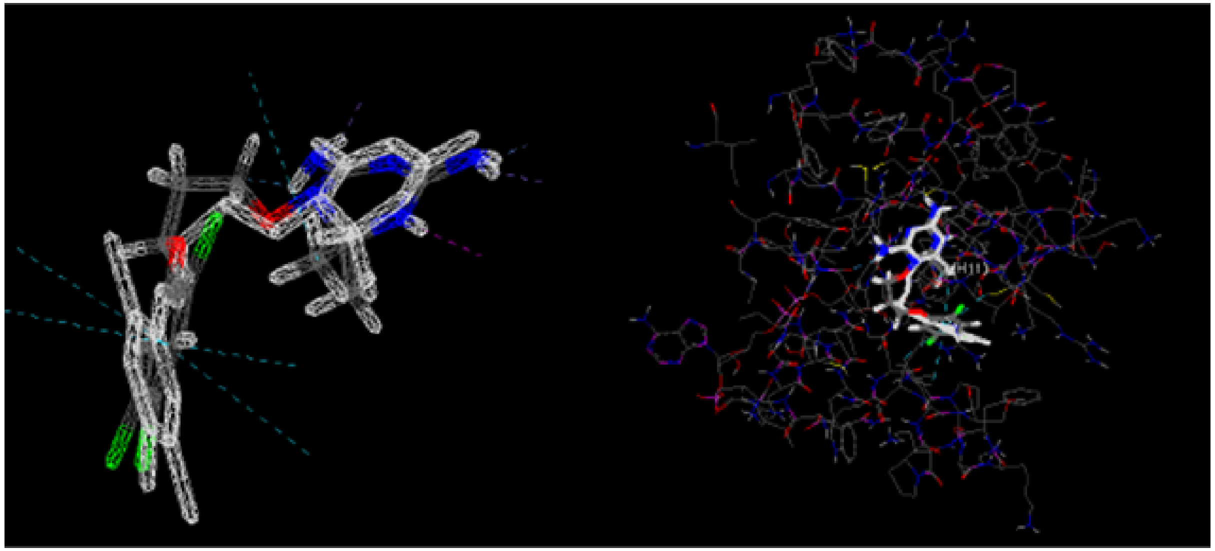
Virtual Screening as an Example

AutoDock - affinity grid maps for each atom type



FlexX – different types of interactions (interaction points)





Large-Scale in silico Experimentation: eSciences

Large-scale *in silico* experimentation & eSciences: the WISDOM project

- Large-scale virtual screening for novel drugs against Malaria (*plasmodium falciparum*)
- International collaboration based on the EGEE grid computing infrastructure (with thousands of CPUs connected worldwide)
- Millions of protein – drug interaction simulations; equivalent to more than 80 years of permanent computing on a single CPU
- However, WISDOM was based on a rather “physical” scenario: a virtual representation of a chemical compound is positioned into the binding geometry of a protein, of which the 3D-structure is known.

Simulation and Knowledge – Driven Approaches

First principle – based sciences

- Physics
- Engineering
- Physico-Chemistry

→Based on mathematical models

→Simulation approaches can be easily applied

Descriptive, empirical sciences

- Biology
- Pharmaceutical Chemistry
- Medicine

→Based on knowledge represented in the literature

→Very complex, difficult to simulate

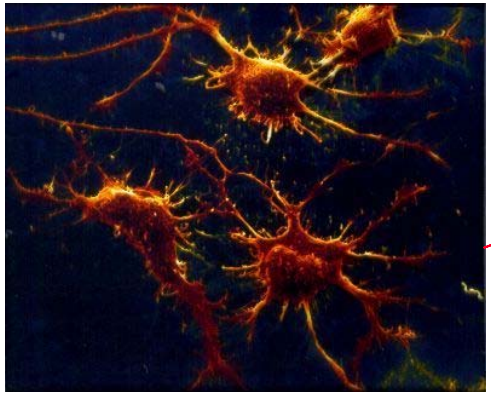
Making Use of The Wealth of Knowledge that is Out There

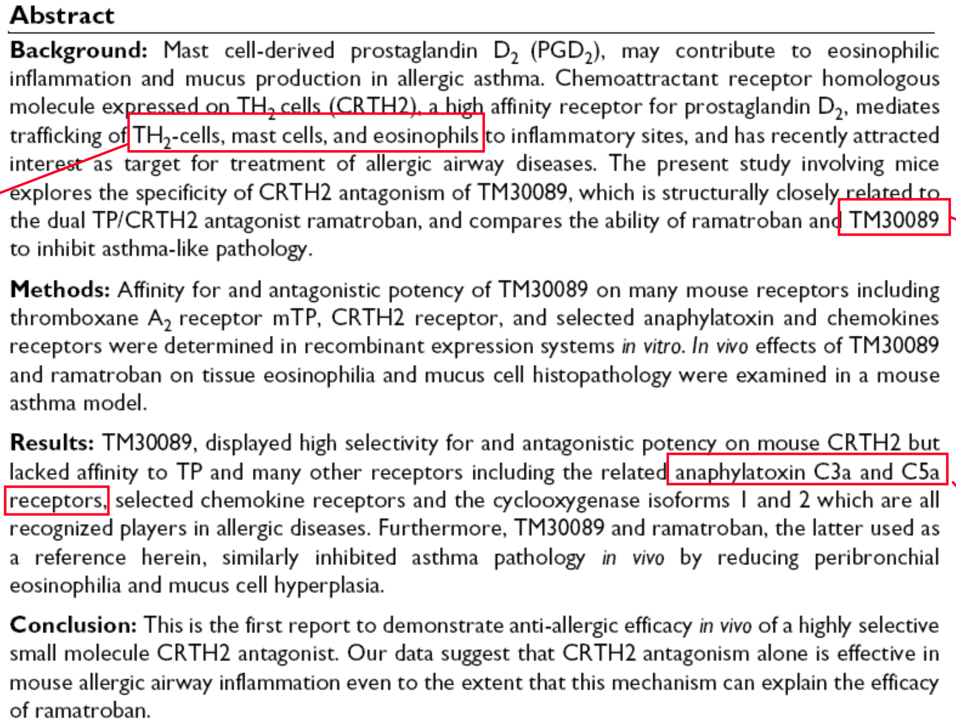
Literature as a Main Source for Knowledge in the Life Sciences

- Biology and Medicine are still to a large extend *empirical* sciences
- Complex: very high number of entities and relationships
- Lots of data on genes and proteins in databases
- However, biodatabases do only comprise data and not necessarily knowledge (data + models)
- Expressiveness of natural language in text is much higher; therefore scientific text is a much better source for biomedical knowledge

→ How do we get access to the knowledge and how can we model it?

Biological, Medical and Chemical Objects in Text





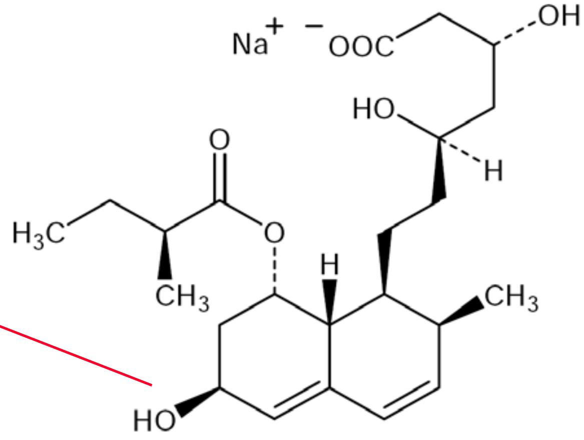
Abstract

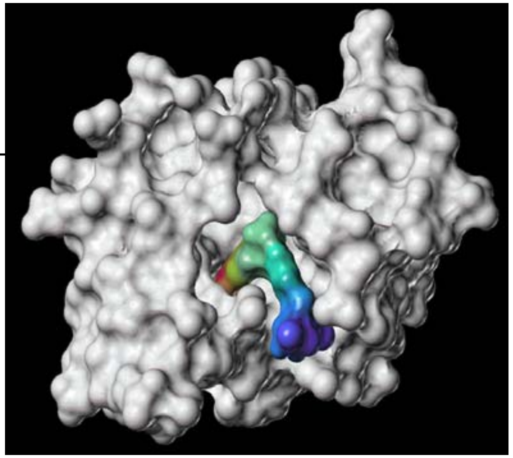
Background: Mast cell-derived prostaglandin D$_2$ (PGD$_2$), may contribute to eosinophilic inflammation and mucus production in allergic asthma. Chemoattractant receptor homologous molecule expressed on TH$_2$ cells (CRTH2), a high affinity receptor for prostaglandin D$_2$, mediates trafficking of TH$_2$-cells, mast cells, and eosinophils to inflammatory sites, and has recently attracted interest as target for treatment of allergic airway diseases. The present study involving mice explores the specificity of CRTH2 antagonism of TM30089, which is structurally closely related to the dual TP/CRTH2 antagonist ramatroban, and compares the ability of ramatroban and TM30089 to inhibit asthma-like pathology.

Methods: Affinity for and antagonistic potency of TM30089 on many mouse receptors including thromboxane A$_2$ receptor mTP, CRTH2 receptor, and selected anaphylatoxin and chemokines receptors were determined in recombinant expression systems *in vitro*. *In vivo* effects of TM30089 and ramatroban on tissue eosinophilia and mucus cell histopathology were examined in a mouse asthma model.

Results: TM30089, displayed high selectivity for and antagonistic potency on mouse CRTH2 but lacked affinity to TP and many other receptors including the related anaphylatoxin C3a and C5a receptors, selected chemokine receptors and the cyclooxygenase isoforms 1 and 2 which are all recognized players in allergic diseases. Furthermore, TM30089 and ramatroban, the latter used as a reference herein, similarly inhibited asthma pathology *in vivo* by reducing peribronchial eosinophilia and mucus cell hyperplasia.

Conclusion: This is the first report to demonstrate anti-allergic efficacy *in vivo* of a highly selective small molecule CRTH2 antagonist. Our data suggest that CRTH2 antagonism alone is effective in mouse allergic airway inflammation even to the extent that this mechanism can explain the efficacy of ramatroban.





Technologies for Information Extraction from Literature

During the last five years, substantial progress has been made in the area of automated text analysis. In particular in life science informatics there is a strong community developing new methods and tools for the automated recognition and extraction of information from scientific literature.

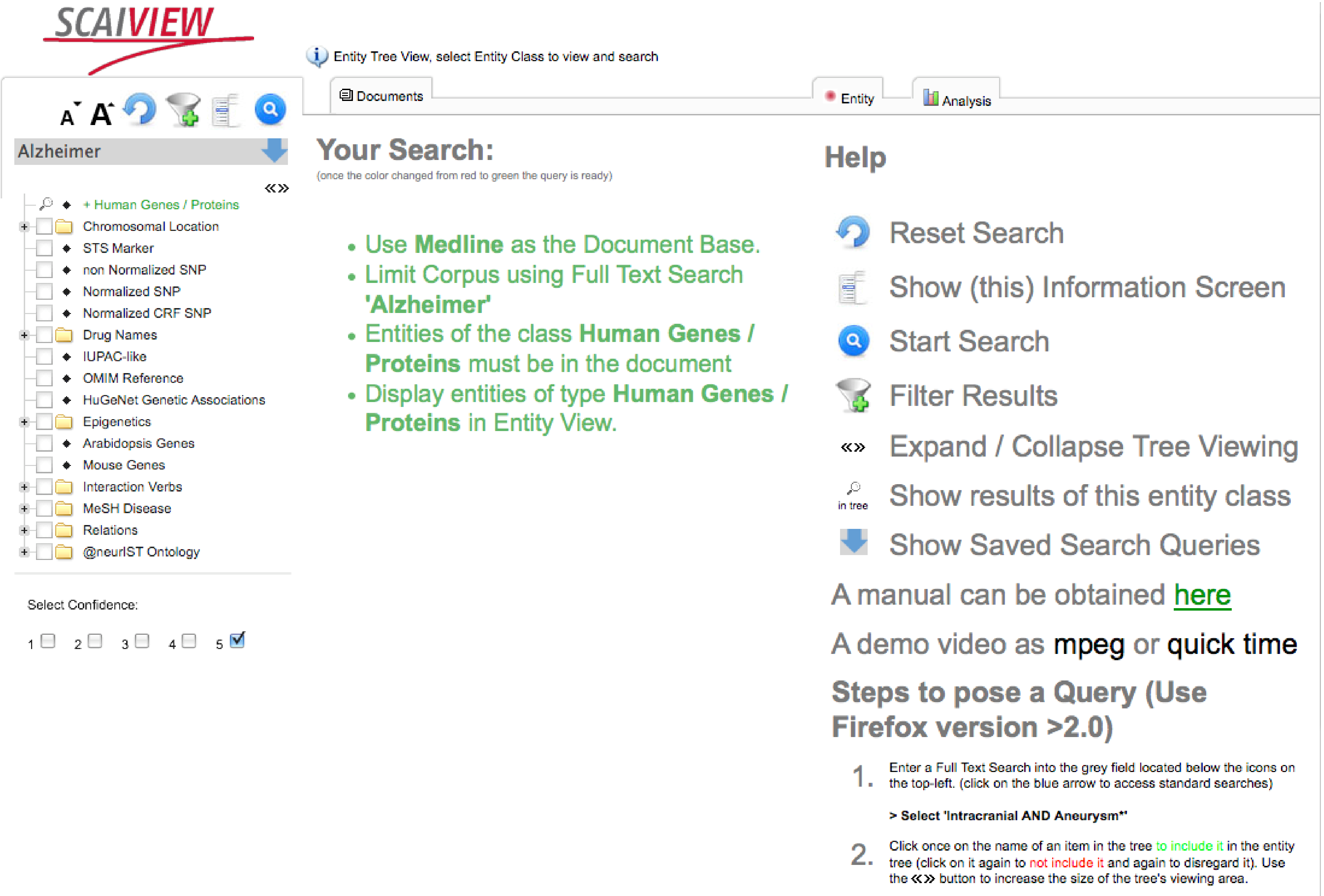
Our group at Fraunhofer SCAI has developed three tools that enable mining in literature:

-**ProMiner,** a solution for named entity recognition based on rules and dictionaries (a “reading machine” for biomedical text)

-**ChemoCR**, a software that identifies chemical structure depictions in full text (a “reading machine” for chemical structure depictions)

-**SCAIView**, a text mining environment that supports end-users

ProMiner & SCAIView



Your Search:

(once the color changed from red to green the query is ready)

- Use **Medline** as the Document Base.
- Limit Corpus using Full Text Search **'Alzheimer'**
- Entities of the class **Human Genes / Proteins** must be in the document
- Display entities of type **Human Genes / Proteins** in Entity View.

Help

Reset Search

Show (this) Information Screen

Start Search

Filter Results

«» Expand / Collapse Tree Viewing

in tree Show results of this entity class

Show Saved Search Queries

A manual can be obtained here

A demo video as mpeg or quick time

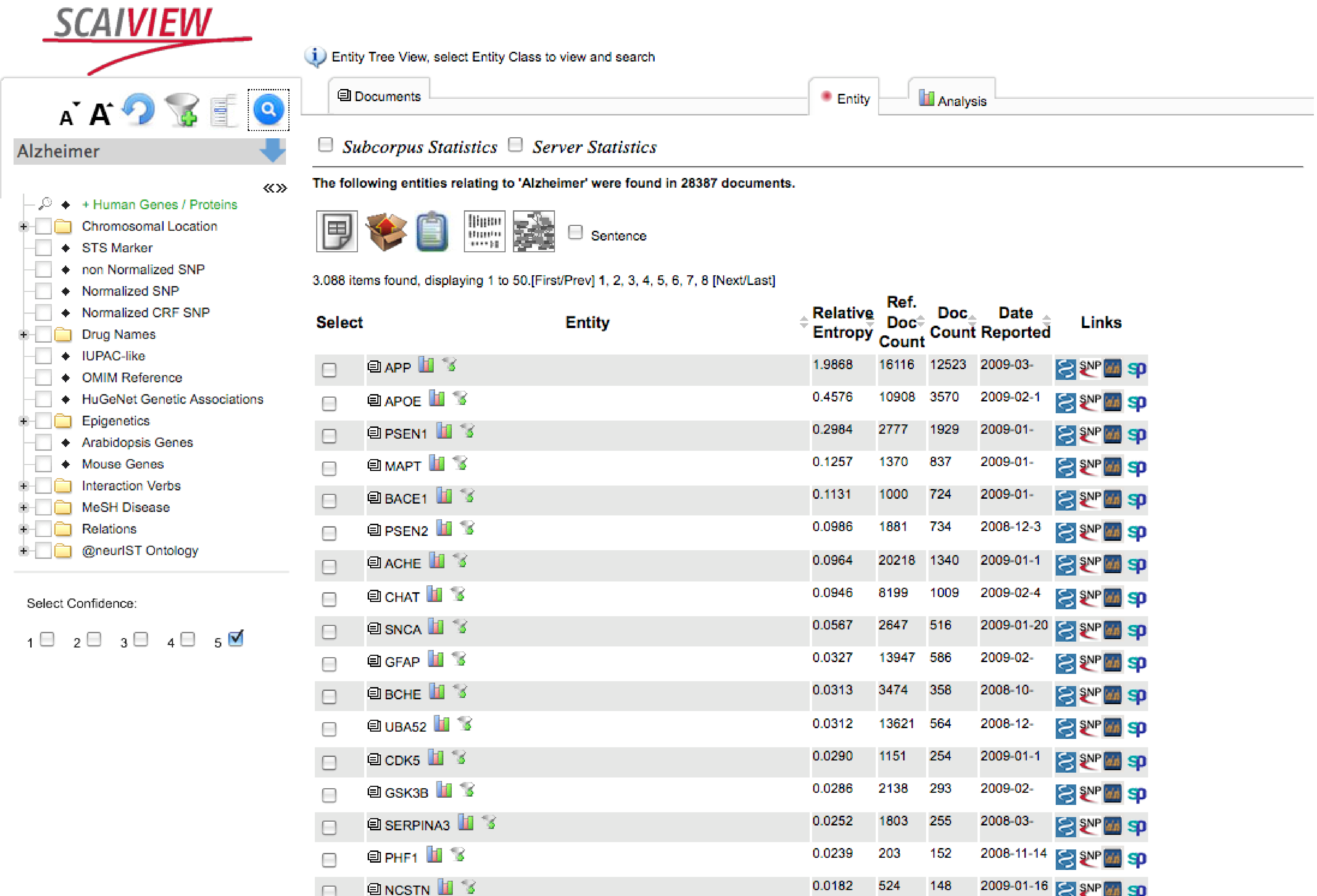
Steps to pose a Query (Use Firefox version >2.0)

1. Enter a Full Text Search into the grey field located below the icons on the top-left. (click on the blue arrow to access standard searches)

   > Select 'Intracranial AND Aneurysm*'

2. Click once on the name of an item in the tree to include it in the entity tree (click on it again to not include it and again to disregard it). Use the «» button to increase the size of the tree's viewing area.

ProMiner & SCAIView



Entity Tree View, select Entity Class to view and search

- + Human Genes / Proteins
- □ Chromosomal Location
- □ STS Marker
- □ non Normalized SNP
- □ Normalized SNP
- □ Normalized CRF SNP
- □ Drug Names
- □ IUPAC-like
- □ OMIM Reference
- □ HuGeNet Genetic Associations
- □ Epigenetics
- □ Arabidopsis Genes
- □ Mouse Genes
- □ Interaction Verbs
- □ MeSH Disease
- □ Relations
- □ @neurIST Ontology

Select Confidence:

1 □ 2 □ 3 □ 4 □ 5 ☑

□ *Subcorpus Statistics* □ *Server Statistics*

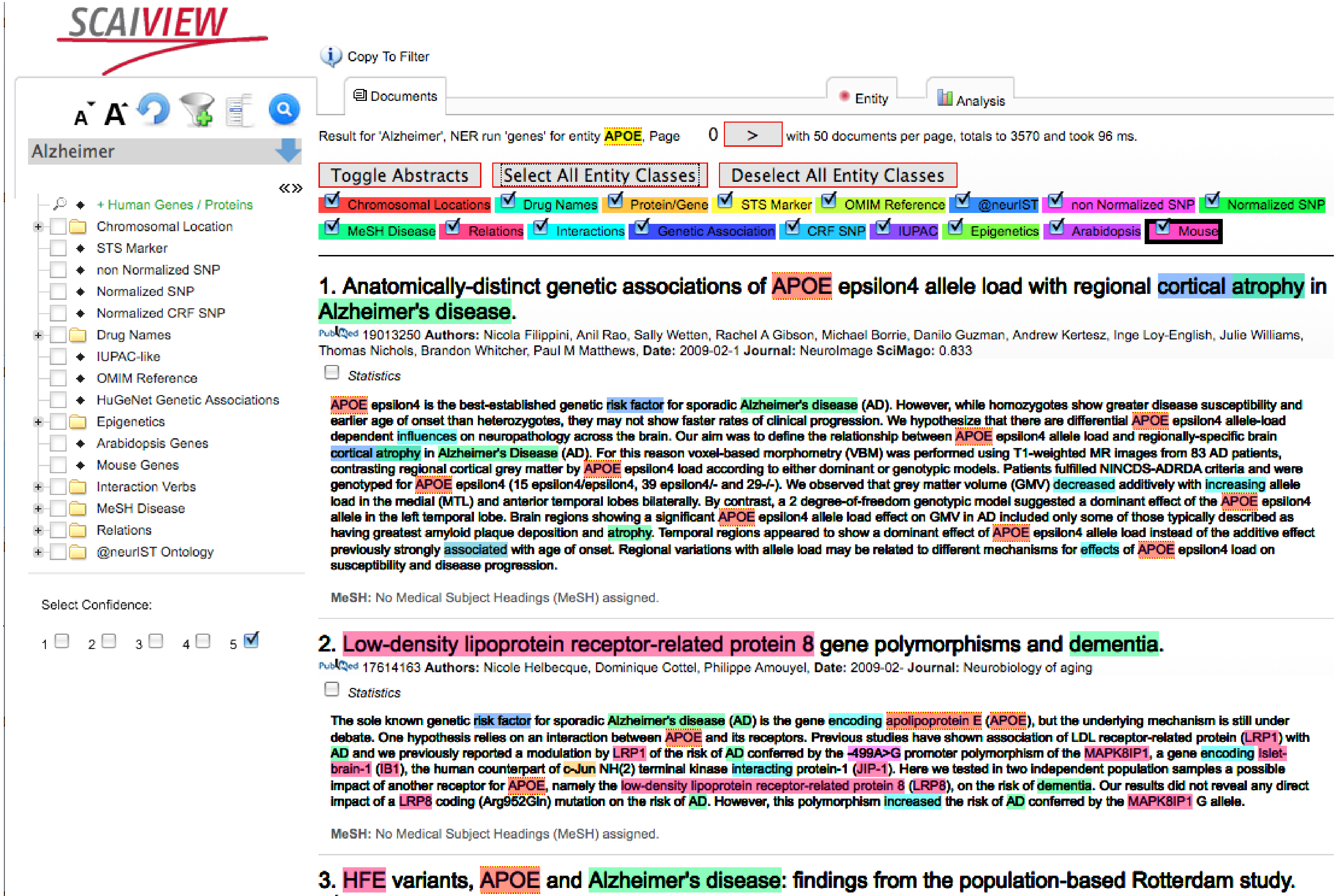
The following entities relating to 'Alzheimer' were found in 28387 documents.

□ Sentence

3.088 items found, displaying 1 to 50.[First/Prev] 1, 2, 3, 4, 5, 6, 7, 8 [Next/Last]

Select	Entity	Relative Entropy	Ref. Doc Count	Doc Count	Date Reported	Links
□	APP	1.9868	16116	12523	2009-03-	SNP sp
□	APOE	0.4576	10908	3570	2009-02-1	SNP sp
□	PSEN1	0.2984	2777	1929	2009-01-	SNP sp
□	MAPT	0.1257	1370	837	2009-01-	SNP sp
□	BACE1	0.1131	1000	724	2009-01-	SNP sp
□	PSEN2	0.0986	1881	734	2008-12-3	SNP sp
□	ACHE	0.0964	20218	1340	2009-01-1	SNP sp
□	CHAT	0.0946	8199	1009	2009-02-4	SNP sp
□	SNCA	0.0567	2647	516	2009-01-20	SNP sp
□	GFAP	0.0327	13947	586	2009-02-	SNP sp
□	BCHE	0.0313	3474	358	2008-10-	SNP sp
□	UBA52	0.0312	13621	564	2008-12-	SNP sp
□	CDK5	0.0290	1151	254	2009-01-1	SNP sp
□	GSK3B	0.0286	2138	293	2009-02-	SNP sp
□	SERPINA3	0.0252	1803	255	2008-03-	SNP sp
□	PHF1	0.0239	203	152	2008-11-14	SNP sp
□	NCSTN	0.0182	524	148	2009-01-16	SNP sp

ProMiner & SCAIView



Result for 'Alzheimer', NER run 'genes' for entity APOE, Page 0 > with 50 documents per page, totals to 3570 and took 96 ms.

1. Anatomically-distinct genetic associations of APOE epsilon4 allele load with regional cortical atrophy in Alzheimer's disease.

PubMed 19013250 **Authors:** Nicola Filippini, Anil Rao, Sally Wetten, Rachel A Gibson, Michael Borrie, Danilo Guzman, Andrew Kertesz, Inge Loy-English, Julie Williams, Thomas Nichols, Brandon Whitcher, Paul M Matthews, **Date:** 2009-02-1 **Journal:** NeuroImage **SciMago:** 0.833

□ *Statistics*

APOE epsilon4 is the best-established genetic risk factor for sporadic Alzheimer's disease (AD). However, while homozygotes show greater disease susceptibility and earlier age of onset than heterozygotes, they may not show faster rates of clinical progression. We hypothesize that there are differential APOE epsilon4 allele-load dependent influences on neuropathology across the brain. Our aim was to define the relationship between APOE epsilon4 allele load and regionally-specific brain cortical atrophy in Alzheimer's Disease (AD). For this reason voxel-based morphometry (VBM) was performed using T1-weighted MR images from 83 AD patients, contrasting regional cortical grey matter by APOE epsilon4 load according to either dominant or genotypic models. Patients fulfilled NINCDS-ADRDA criteria and were genotyped for APOE epsilon4 (15 epsilon4/epsilon4, 39 epsilon4/- and 29-/-). We observed that grey matter volume (GMV) decreased additively with increasing allele load in the medial (MTL) and anterior temporal lobes bilaterally. By contrast, a 2 degree-of-freedom genotypic model suggested a dominant effect of the APOE epsilon4 allele in the left temporal lobe. Brain regions showing a significant APOE epsilon4 allele load effect on GMV in AD included only some of those typically described as having greatest amyloid plaque deposition and atrophy. Temporal regions appeared to show a dominant effect of APOE epsilon4 allele load instead of the additive effect previously strongly associated with age of onset. Regional variations with allele load may be related to different mechanisms for effects of APOE epsilon4 load on susceptibility and disease progression.

MeSH: No Medical Subject Headings (MeSH) assigned.

2. Low-density lipoprotein receptor-related protein 8 gene polymorphisms and dementia.

PubMed 17614163 **Authors:** Nicole Helbecque, Dominique Cottel, Philippe Amouyel, **Date:** 2009-02- **Journal:** Neurobiology of aging

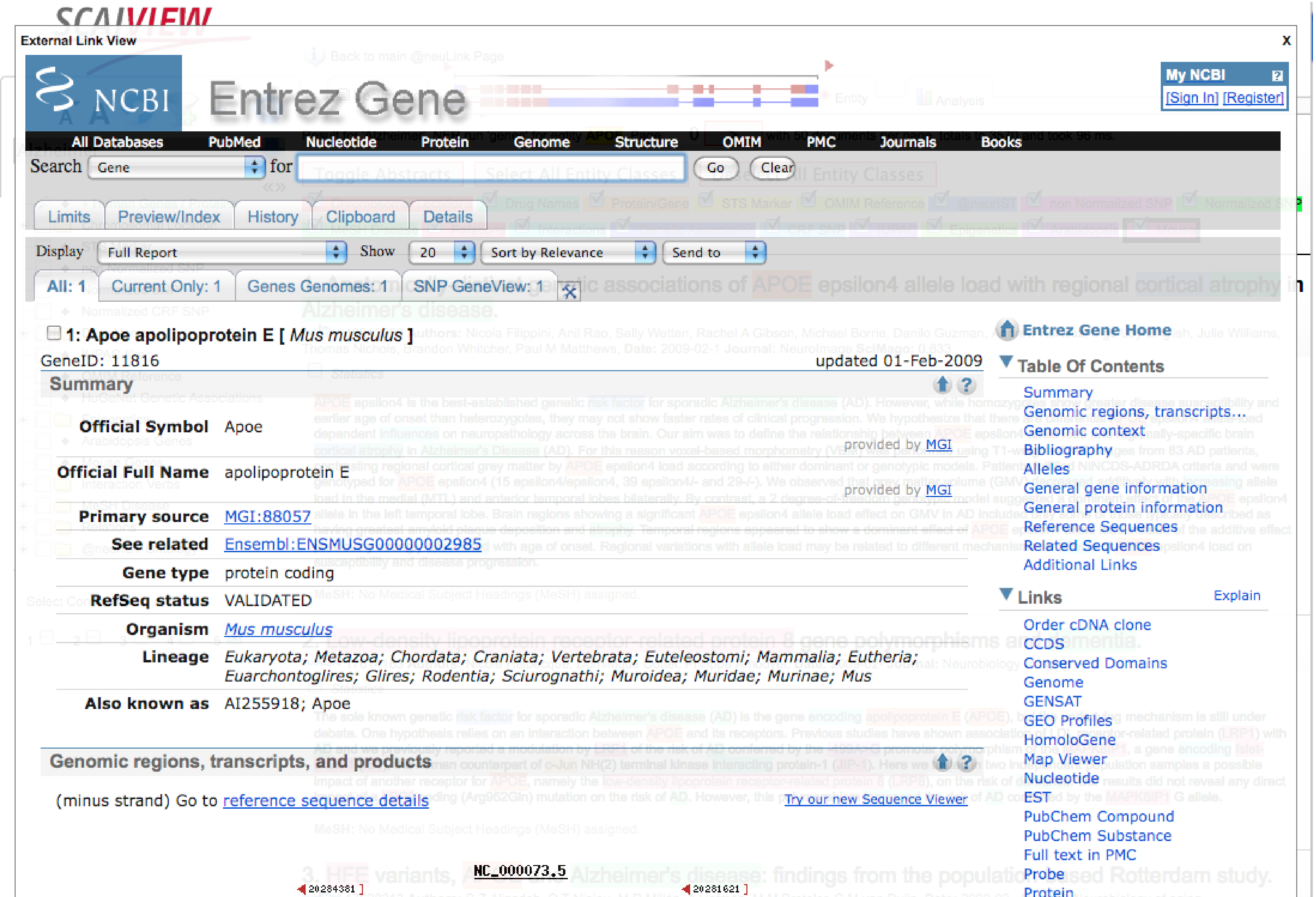
□ *Statistics*

The sole known genetic risk factor for sporadic Alzheimer's disease (AD) is the gene encoding apolipoprotein E (APOE), but the underlying mechanism is still under debate. One hypothesis relies on an interaction between APOE and its receptors. Previous studies have shown association of LDL receptor-related protein (LRP1) with AD and we previously reported a modulation by LRP1 of the risk of AD conferred by the -499A>G promoter polymorphism of the MAPK8IP1, a gene encoding Islet-brain-1 (IB1), the human counterpart of c-Jun NH(2) terminal kinase interacting protein-1 (JIP-1). Here we tested in two independent population samples a possible impact of another receptor for APOE, namely the low-density lipoprotein receptor-related protein 8 (LRP8), on the risk of dementia. Our results did not reveal any direct impact of a LRP8 coding (Arg952Gln) mutation on the risk of AD. However, this polymorphism increased the risk of AD conferred by the MAPK8IP1 G allele.

MeSH: No Medical Subject Headings (MeSH) assigned.

3. HFE variants, APOE and Alzheimer's disease: findings from the population-based Rotterdam study.

ProMiner & SCAIView



External Link View

Entrez Gene

My NCBI [Sign In] [Register]

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

Search Gene for Go Clear

Limits Preview/Index History Clipboard Details

Display Full Report Show 20 Sort by Relevance Send to

All: 1 Current Only: 1 Genes Genomes: 1 SNP GeneView: 1

□ 1: Apoe apolipoprotein E [*Mus musculus*]

GeneID: 11816 updated 01-Feb-2009

Summary

Official Symbol	Apoe provided by MGI
Official Full Name	apolipoprotein E provided by MGI
Primary source	MGI:88057
See related	Ensembl:ENSMUSG00000002985
Gene type	protein coding
RefSeq status	VALIDATED
Organism	*Mus musculus*
Lineage	*Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea; Muridae; Murinae; Mus*
Also known as	AI255918; Apoe

Genomic regions, transcripts, and products

(minus strand) Go to reference sequence details Try our new Sequence Viewer

NC_000073.5

20284381] 20281621]

Entrez Gene Home

Table Of Contents

Summary
Genomic regions, transcripts...
Genomic context
Bibliography
Alleles
General gene information
General protein information
Reference Sequences
Related Sequences
Additional Links

Links Explain

Order cDNA clone
CCDS
Conserved Domains
Genome
GENSAT
GEO Profiles
HomoloGene
Map Viewer
Nucleotide
EST
PubChem Compound
PubChem Substance
Full text in PMC
Probe
Protein

Making Use of The Wealth of Knowledge that is Out There:

Example 1:

Using Text-based Information for the Prediction of Pharmaceutical Activities of Drugs

(Master Thesis of Harsha Gurulingappa, B-IT and Fraunhofer SCAI)

Task: ATC Classification of yet Unclassified Drugs

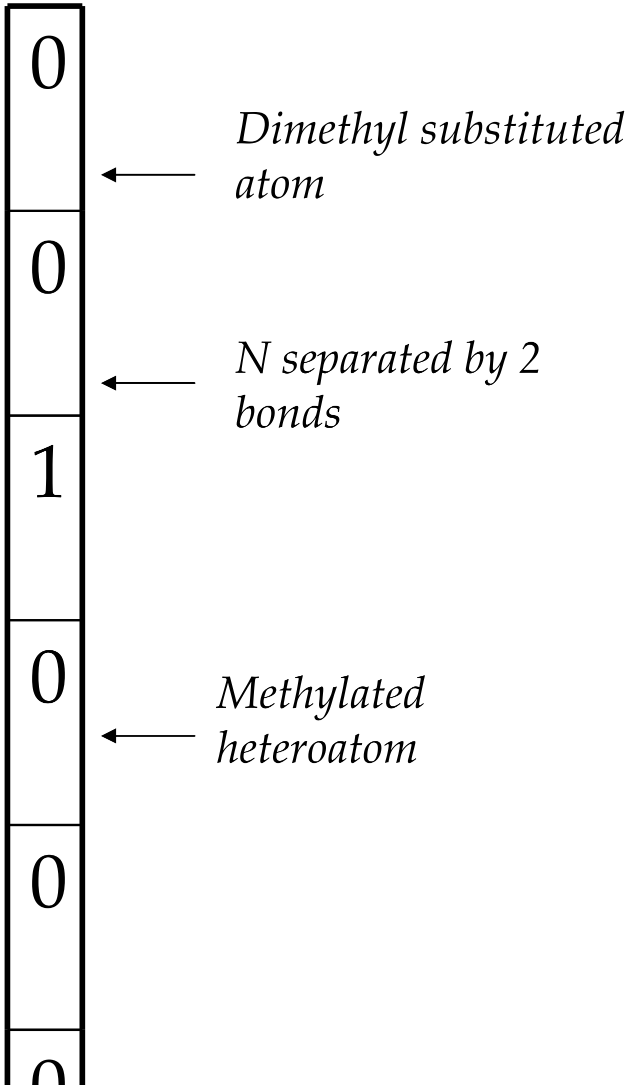
Goal of this study:

- developed of a method for predicting putative ATC classes for unclassified drug terms.
- develop a new paradigm for strategies aiming at identifying potential secondary applications for existing drugs.
- Use of textual features/evidences for characterization of drugs

→ Disadvantage: Highly dependent on Information Availability.

Fingerprints: Method for the Prediction of Chemical Properties

MACCS keys: Structural Keys



0

0

1

0

0

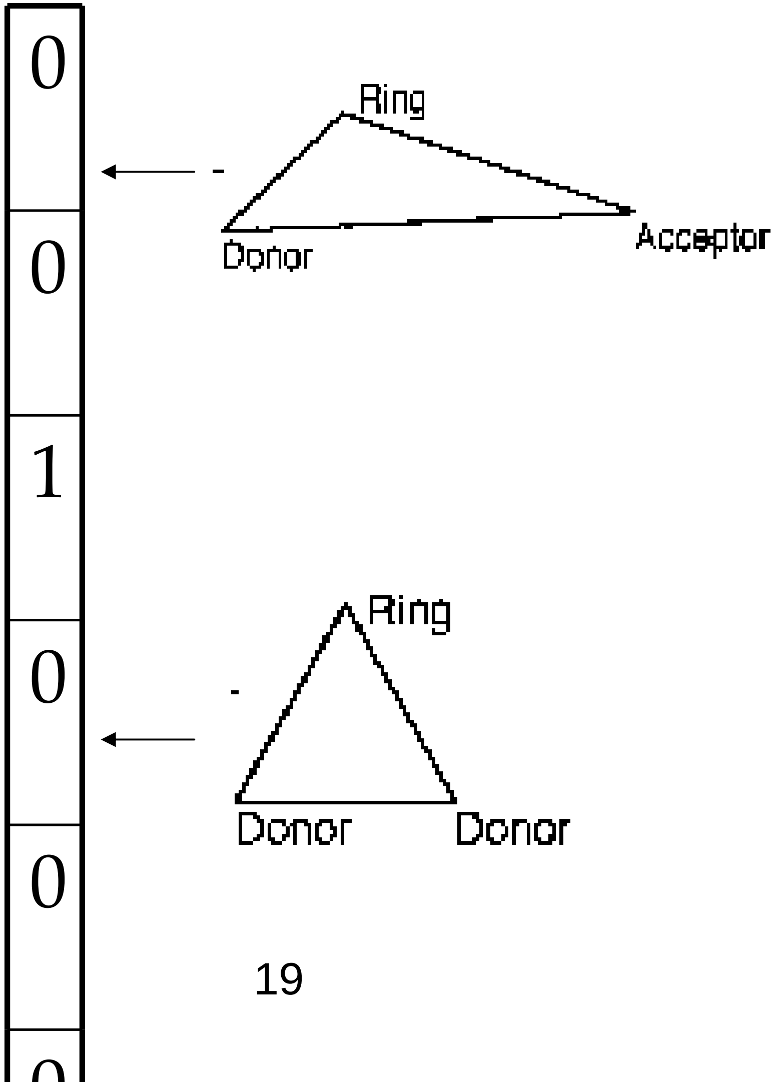
0

Dimethyl substituted atom

N separated by 2 bonds

Methylated heteroatom

TGT keys: 3 Point Pharmacophore based fingerprints



0

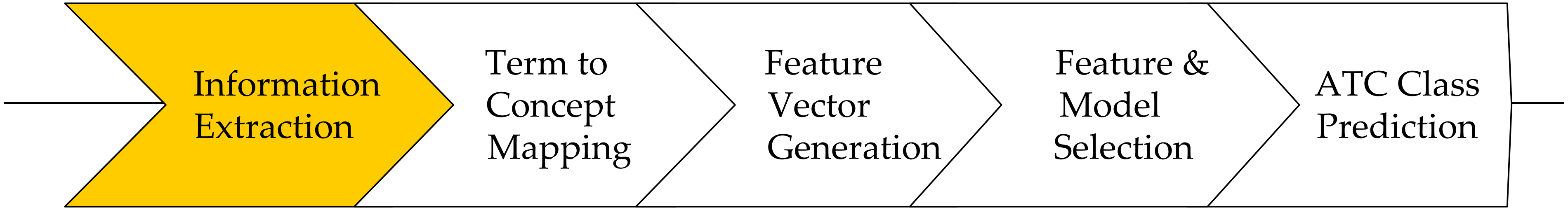
0

1

0

0

0



Information Extraction

Term to Concept Mapping

Feature Vector Generation

Feature & Model Selection

ATC Class Prediction

Timolol is a beta adrenoceptor blocker

A vasodilator like propatyl nitrate, can open the ...

Digitoxin, a cardiac stimulant is responsible for ...

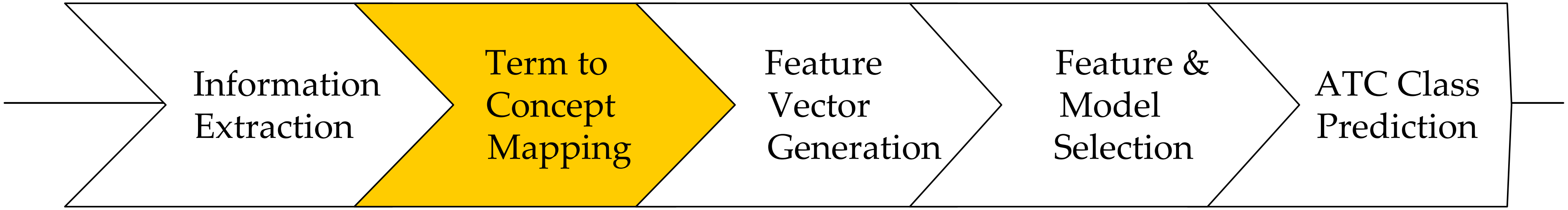
🟦 Drug Term

🟥 Property Term

⬛ Free Text

Hearst Patterns: Express direct relationship between drugs and their properties.

[Kolářik et al., 2007]



Information Extraction

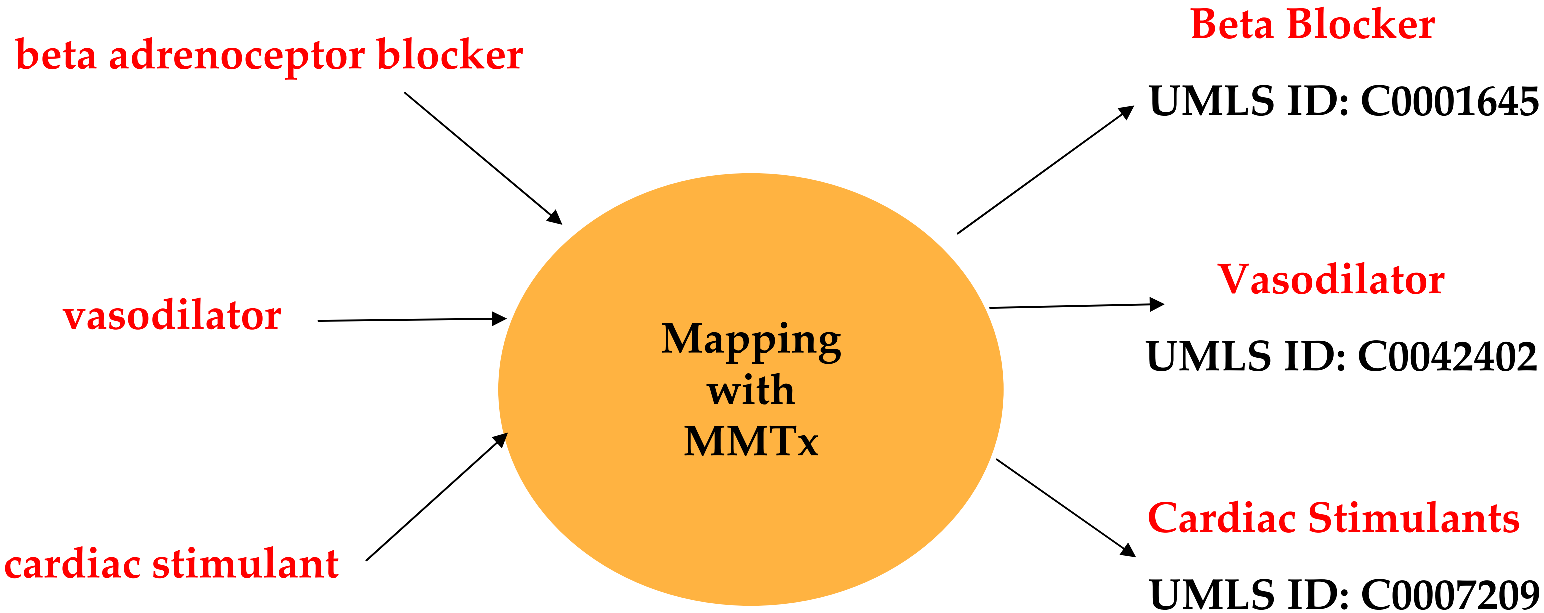
Term to Concept Mapping

Feature Vector Generation

Feature & Model Selection

ATC Class Prediction

Map extracted property terms to concepts in UMLS*



beta adrenoceptor blocker

vasodilator

cardiac stimulant

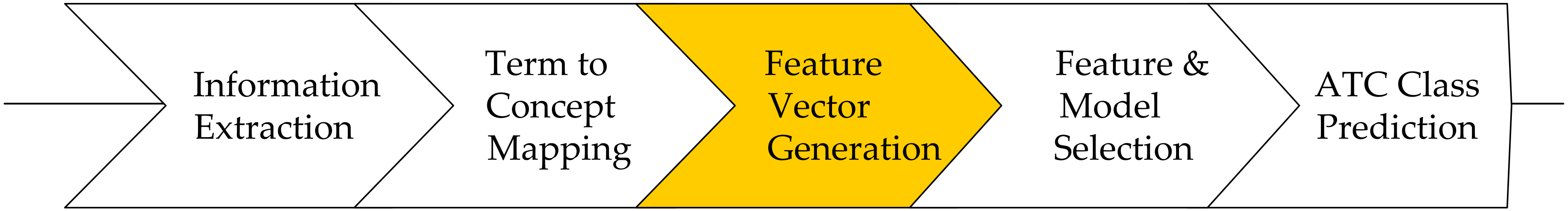
Mapping with MMTx

Beta Blocker
UMLS ID: C0001645

Vasodilator
UMLS ID: C0042402

Cardiac Stimulants
UMLS ID: C0007209

*UMLS: Unified Medical Language System



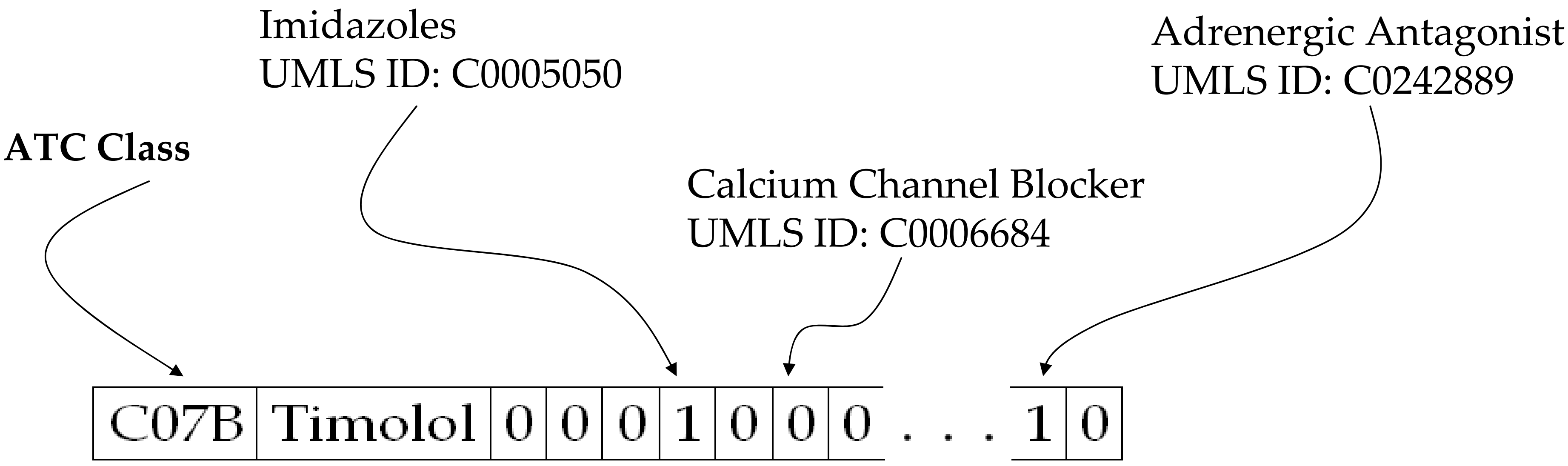
Information Extraction

Term to Concept Mapping

Feature Vector Generation

Feature & Model Selection

ATC Class Prediction



Imidazoles
UMLS ID: C0005050

Adrenergic Antagonist
UMLS ID: C0242889

ATC Class

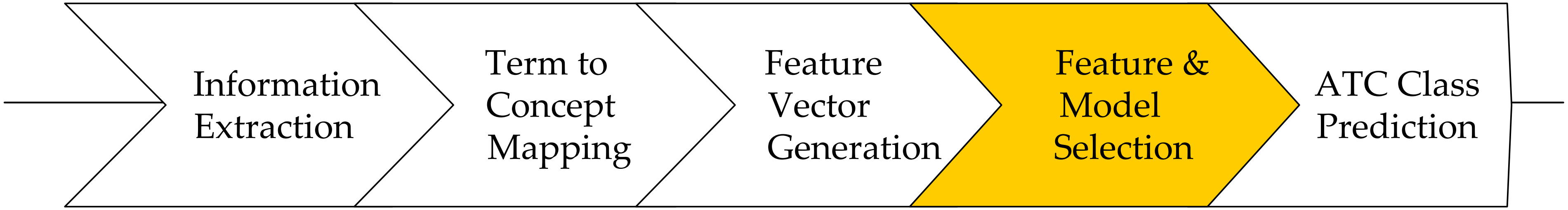
Calcium Channel Blocker
UMLS ID: C0006684

C07B	Timolol	0	0	0	1	0	0	0	. . .	1	0

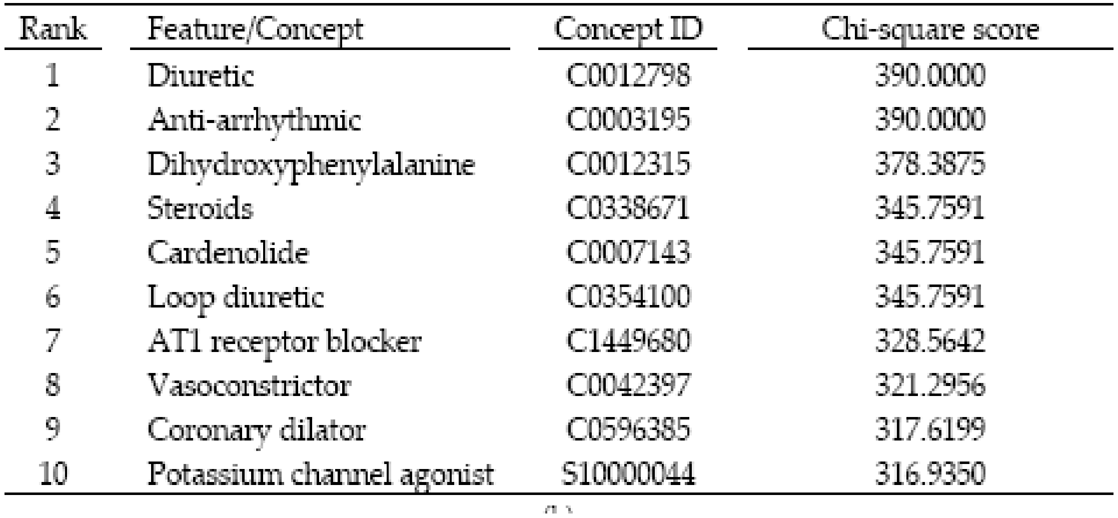
Binary Feature Vector: '0': Feature Absent & '1': Feature Present

C07B	Timolol	0	0	0	7	0	0	0	. . .	4	0

Weighted Feature Vector: '0': Feature Absent & '≥1': Corpus Frequency of the Feature

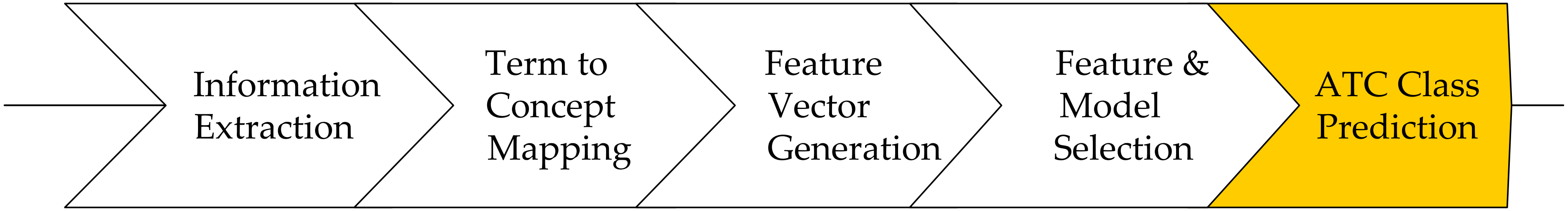


- Feature Selection
  - > Mutual Information
  - > Chi-square criterion



Rank	Feature/Concept	Concept ID	Chi-square score
1	Diuretic	C0012798	390.0000
2	Anti-arrhythmic	C0003195	390.0000
3	Dihydroxyphenylalanine	C0012315	378.3875
4	Steroids	C0338671	345.7591
5	Cardenolide	C0007143	345.7591
6	Loop diuretic	C0354100	345.7591
7	AT1 receptor blocker	C1449680	328.5642
8	Vasoconstrictor	C0042397	321.2956
9	Coronary dilator	C0596385	317.6199
10	Potassium channel agonist	S10000044	316.9350

- Models/Classifiers
  - > Naïve Bayes
  - > Nearest Neighbor
  - > Decision Tree
  - > Support Vector Machine



Information Extraction

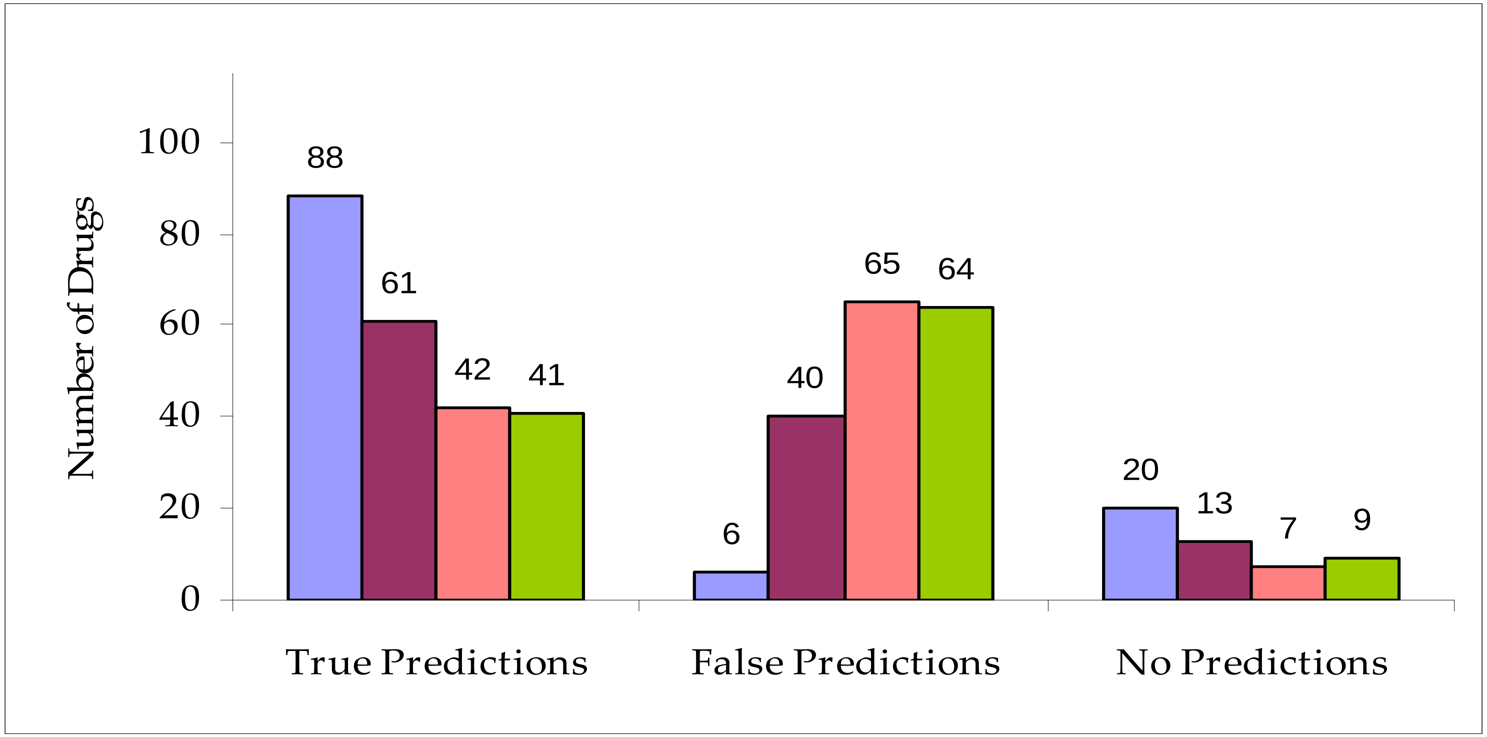
Term to Concept Mapping

Feature Vector Generation

Feature & Model Selection

ATC Class Prediction

Concept Based Vs Structure Based Approaches
Test Set = 114 Drugs



Concepts: 77.2% 🟦

SuperPred: 53.5% 🟪

MACCS Keys: 36.8% 🟥

TGT Keys: 35.8% 🟩

Making Use of The Wealth of Knowledge that is Out There:

Example 2:

Using Text-based Information for the Identification of Genes likely to mediate Susceptibility to Breast Cancer

(Master Thesis of Erfan Younesi, B-IT and Fraunhofer SCAI)

Task: Predicting Networks of Genes / Proteins that are Linked to the Clinical Progression and Outcome of the Disease

Goal of this study:

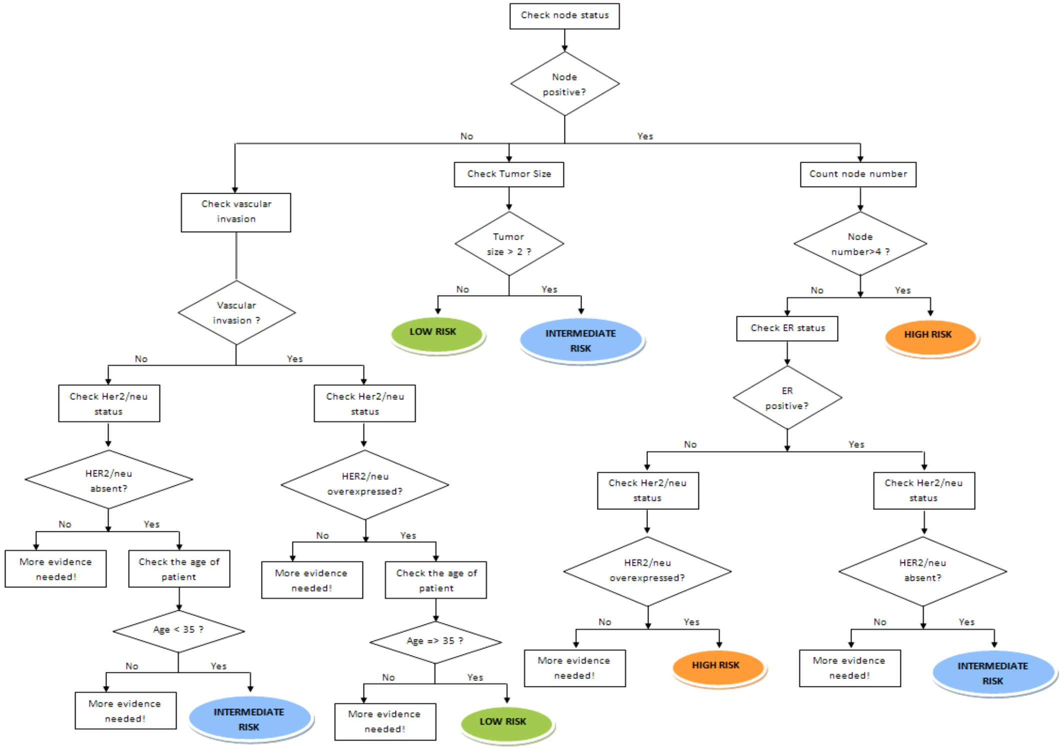
- Identification of networks of proteins functionally linked to tumorigenesis of breast cancer
- Identification of combinations of nodes in a network that can serve as markers for susceptibility to clinical treatment

→ Vision: using text mining to extract evidences for best clinical practice

Decision on Clinical Treatment Strategies for Breast Cancer

Current situation:

- Decisions made up on very few factors (e.g. Lymph Node status; ER+/-)
- Cooperativity of genes and proteins not taken into account



Check node status

Node positive?

No

Check vascular invasion

Vascular invasion ?

No

Check Her2/neu status

HER2/neu absent?

No

More evidence needed!

Yes

Check the age of patient

Age < 35 ?

No

More evidence needed!

Yes

INTERMEDIATE RISK

Yes

Check Her2/neu status

HER2/neu overexpressed?

No

More evidence needed!

Yes

Check the age of patient

Age => 35 ?

No

More evidence needed!

Yes

LOW RISK

Check Tumor Size

Tumor size > 2 ?

No

LOW RISK

Yes

INTERMEDIATE RISK

Yes

Count node number

Node number>4 ?

No

Check ER status

ER positive?

No

Check Her2/neu status

HER2/neu overexpressed?

No

More evidence needed!

Yes

HIGH RISK

Yes

Check Her2/neu status

HER2/neu absent?

No

More evidence needed!

Yes

INTERMEDIATE RISK

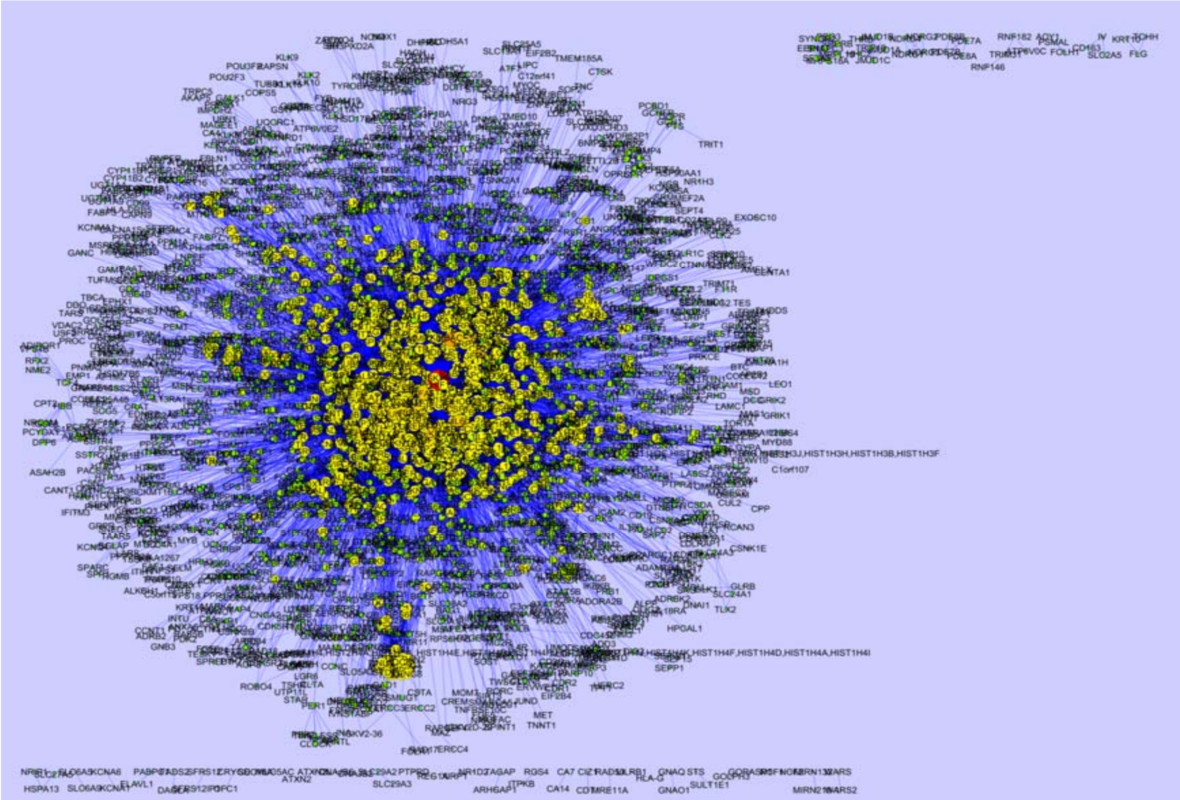
Yes

HIGH RISK

Definition of a Network of Interacting Proteins in Breast Cancer

Approach:

- Definition of a network of molecular entities strongly associated with breast cancer
- Network based on simple co-occurences in text

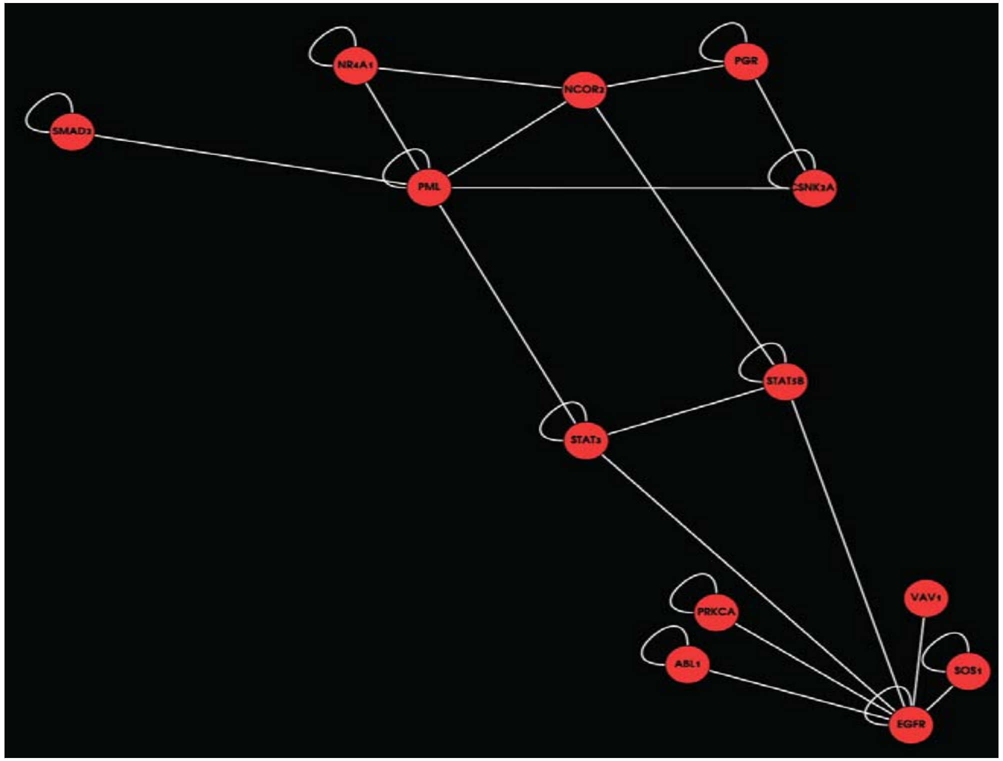


SCAI
Fraunhofer Institute
Algorithms and
Scientific Computing

Reduction of Complexity: Definition of a Minimum Network

Approach:

- Selection of subgraph based on
  - Network topology
  - Functional characterization
  - association with clinical outcome



Identification of Novel Breast Cancer Susceptibility Associations

23 novel associations between the minimum gene set associated with breast cancer susceptibility and clinical outcome could be identified

Associations	Cocitation frequencies	Novel association (not exists in PIANA)	Evidence of general relation between two genes from Literature (PMID)	Shared GO process	Shared KEGG pathway
AKT1 - EGFR	93	Y	14981538 -17686159 -18351692-16774943 -16419029-16546981-16288304-15800944	Nitric oxide anabolism, protein amino acid phosphorylation	MAPK signaling pathway, Focal adhesion, Colorectal cancer, Pancreatic cancer, Glioma
TP53 - EGFR	71	Y	18311481	Cell cycle, Response to stress, regulation of cell proliferation,	MAPK signaling pathway, Colorectal cancer, Pancreatic cancer, Glioma
PGR - EGFR	23	Y	1616857-1911227	regulation of epithelial cell proliferation	--
STAT3 - AKT1	16	Y	10853013-16288304-16728588	--	Jak-STAT signaling pathway, Adipocytokine signaling pathway, Pancreatic cancer
TNF - EGFR	10	Y	9829842-11221831	regulation of protein amino acid phosphorylation, cell-cell adhesion, regulation of cell proliferation	MAPK signaling pathway
CDH1 - PGR	6	Y	16512896	--	--
VDR - EGFR	4	Y	16087726-17377416	skeletal development	--

Making Use of The Wealth of Knowledge that is Out There:

Example 3:

A Look into the Future: A Computational Grand Challenge in the Area of Patent Mining

(ongoing collaboration between Fraunhofer SCAI and FZ Jülich)

Task: Annotation of All Chemical Structure Depictions in All Pharmaceutical Patents from EPO

Goal of this study:

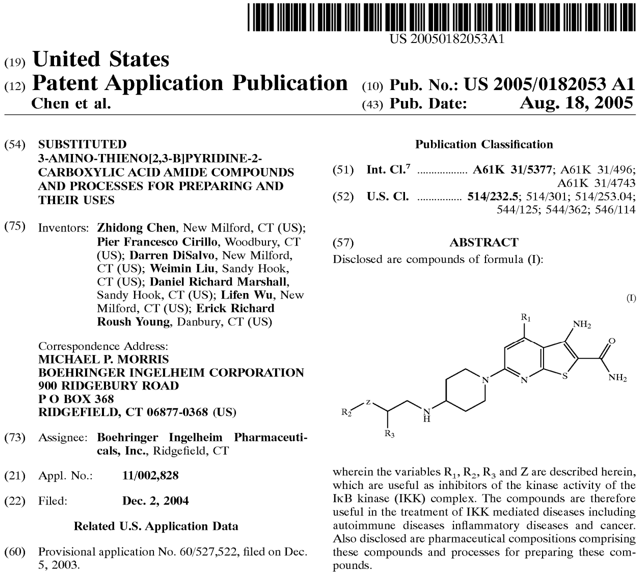
- Feasibility study for large-scale annotation of patents
- Grand computing challenge in the area of knowledge computing
- Demonstration of enhancement of retrieval in the area of chemistry by intelligent software (ChemoCR – chemical structure reconstruction)

→ Vision: using image mining to mine chemical IP at large scale

Reconstruction of Chemical Information

Step1: PDF Conversion

or



US 20050182053A1

(19) **United States**

(12) **Patent Application Publication**
Chen et al.

(10) **Pub. No.: US 2005/0182053 A1**

(43) **Pub. Date: Aug. 18, 2005**

(54) **SUBSTITUTED 3-AMINO-THIENO[2,3-B]PYRIDINE-2-CARBOXYLIC ACID AMIDE COMPOUNDS AND PROCESSES FOR PREPARING AND THEIR USES**

(75) Inventors: **Zhidong Chen**, New Milford, CT (US); **Pier Francesco Cirillo**, Woodbury, CT (US); **Darren DiSalvo**, New Milford, CT (US); **Weimin Liu**, Sandy Hook, CT (US); **Daniel Richard Marshall**, Sandy Hook, CT (US); **Lifen Wu**, New Milford, CT (US); **Erick Richard Roush Young**, Danbury, CT (US)

Correspondence Address:
MICHAEL P. MORRIS
BOEHRINGER INGELHEIM CORPORATION
900 RIDGEBURY ROAD
P O BOX 368
RIDGEFIELD, CT 06877-0368 (US)

(73) Assignee: **Boehringer Ingelheim Pharmaceuticals, Inc.**, Ridgefield, CT

(21) Appl. No.: **11/002,828**

(22) Filed: **Dec. 2, 2004**

Related U.S. Application Data

(60) Provisional application No. 60/527,522, filed on Dec. 5, 2003.

Publication Classification

(51) **Int. Cl.**^{7} **A61K 31/5377**; A61K 31/496; A61K 31/4743

(52) **U.S. Cl.** **514/232.5**; 514/301; 514/253.04; 544/125; 544/362; 546/114

(57) **ABSTRACT**

Disclosed are compounds of formula (I):

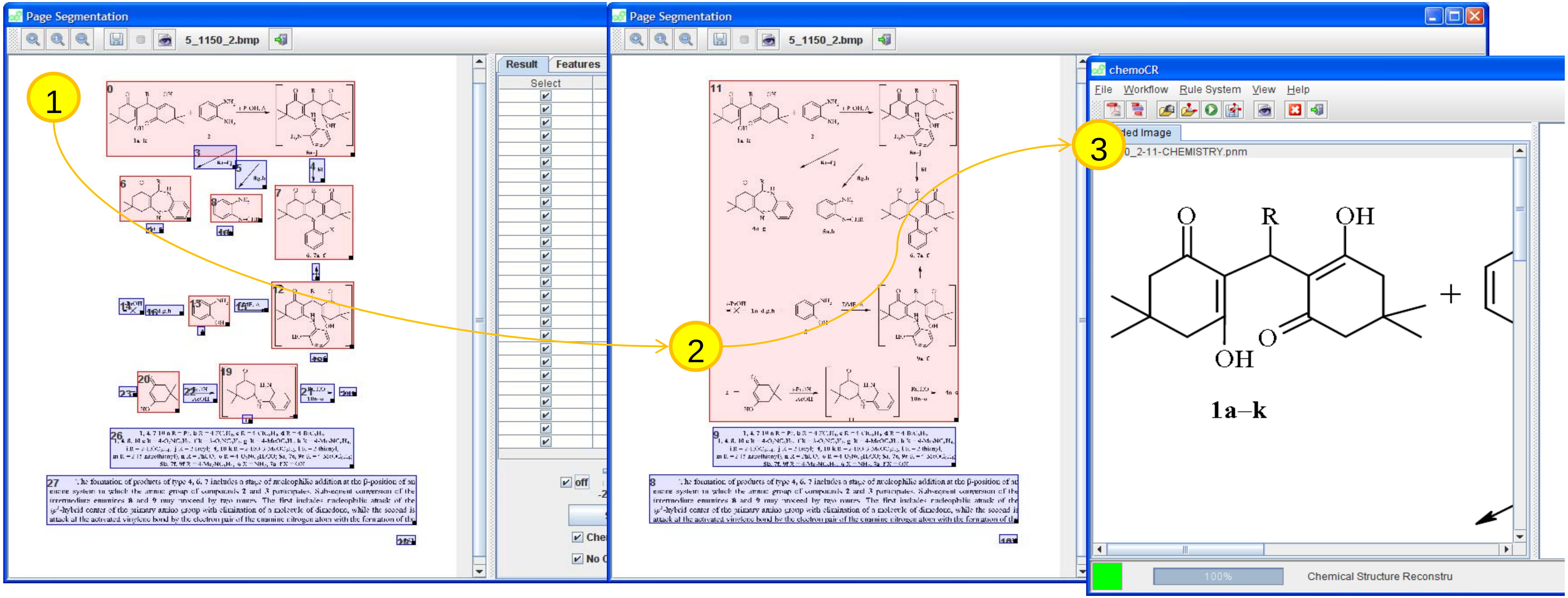
(I)

wherein the variables R_1, R_2, R_3 and Z are described herein, which are useful as inhibitors of the kinase activity of the IκB kinase (IKK) complex. The compounds are therefore useful in the treatment of IKK mediated diseases including autoimmune diseases inflammatory diseases and cancer. Also disclosed are pharmaceutical compositions comprising these compounds and processes for preparing these compounds.

Normalization of image: 250 DPI, grey scale

Reconstruction of Chemical Information

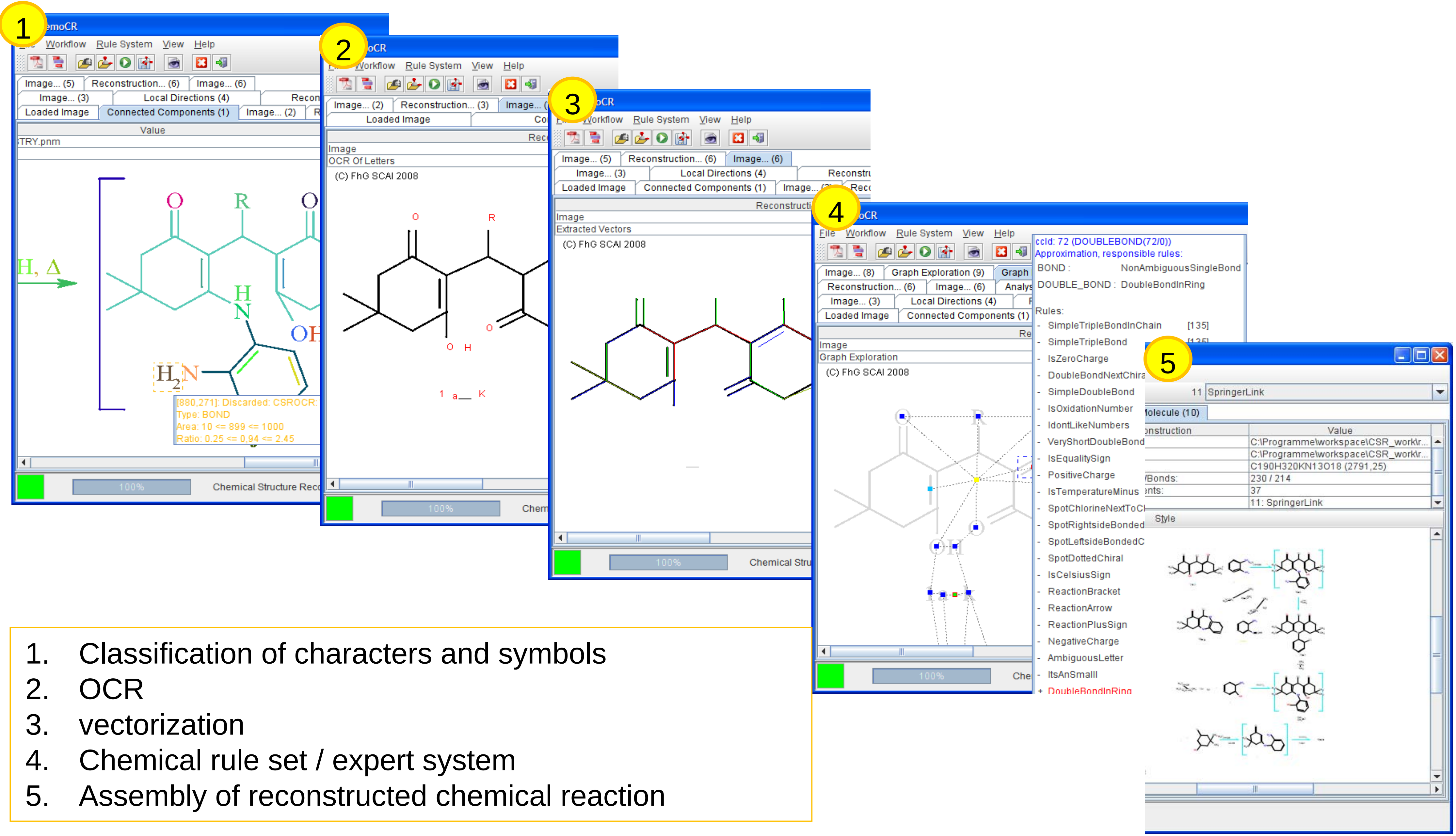
Step 2: Page Segmentation



1. Classification of interesting regions
2. Grouping of chemical reaction schemata
3. Transfer of chemical reaction schemata segment to reconstruction module

Reconstruction of Chemical Information

Step 3: Reconstruction of Chemical Information



[880,271]: Discarded: CSROCR
Type: BOND
Area: 10 <= 899 <= 1000
Ratio: 0.25 <= 0.94 <= 2.45

ccId: 72 (DOUBLEBOND(72/0))
Approximation, responsible rules:
BOND : NonAmbiguousSingleBond
DOUBLE_BOND : DoubleBondInRing

Rules:
- SimpleTripleBondInChain [135]
- SimpleTripleBond [135]
- IsZeroCharge
- DoubleBondNextChira
- SimpleDoubleBond
- IsOxidationNumber
- IdontLikeNumbers
- VeryShortDoubleBond
- IsEqualitySign
- PositiveCharge
- IsTemperatureMinus
- SpotChlorineNextToCl
- SpotRightsideBonded
- SpotLeftsideBondedC
- SpotDottedChiral
- IsCelsiusSign
- ReactionBracket
- ReactionArrow
- ReactionPlusSign
- NegativeCharge
- AmbiguousLetter
- ItsAnSmallI
+ DoubleBondInRing

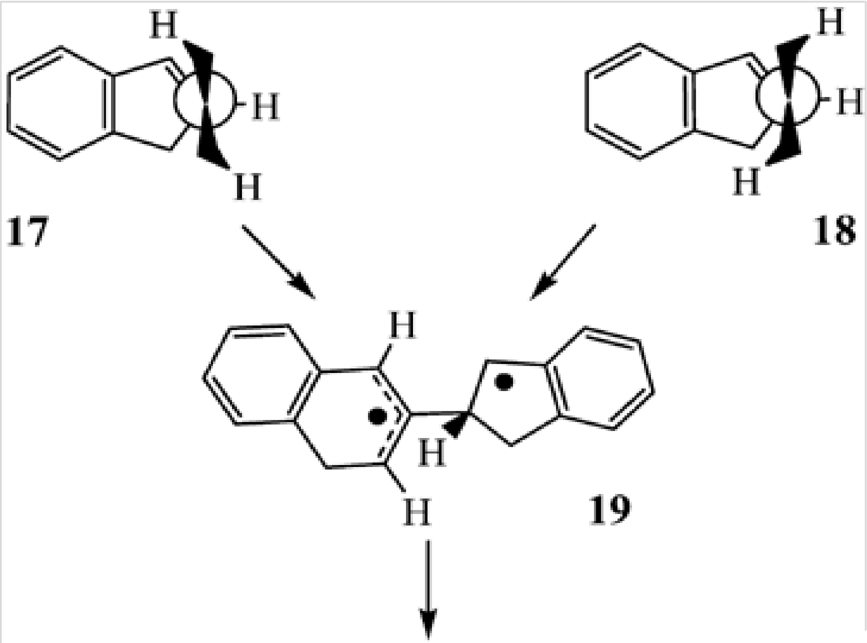
onstruction	Value
	C:\Programme\workspace\CSR_work\r...
	C:\Programme\workspace\CSR_work\r...
	C190H320KN13O18 (2791,25)
/Bonds:	230 / 214
ents:	37
	11: SpringerLink

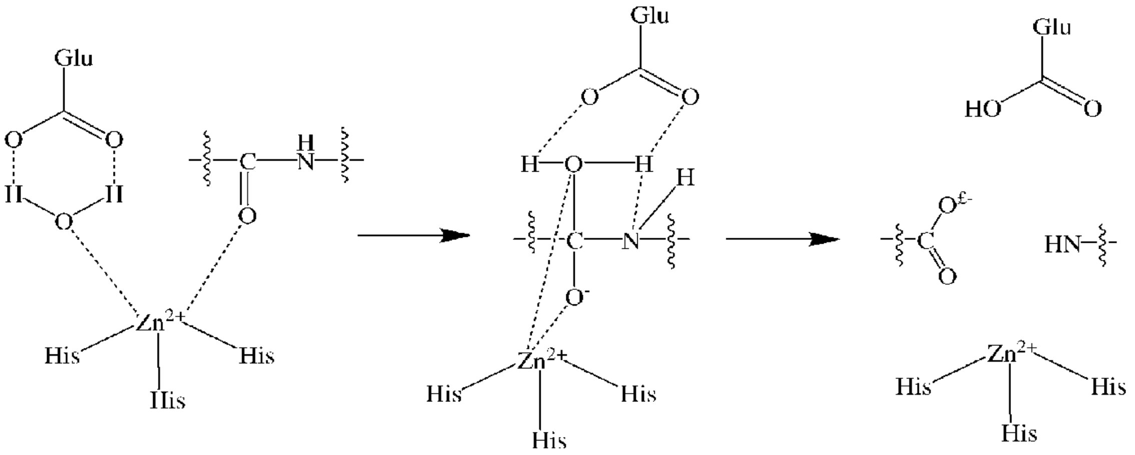
1. Classification of characters and symbols
2. OCR
3. vectorization
4. Chemical rule set / expert system
5. Assembly of reconstructed chemical reaction

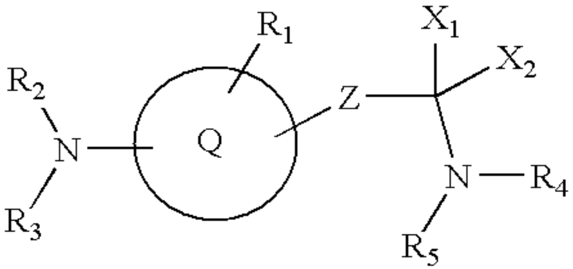
The Challenges

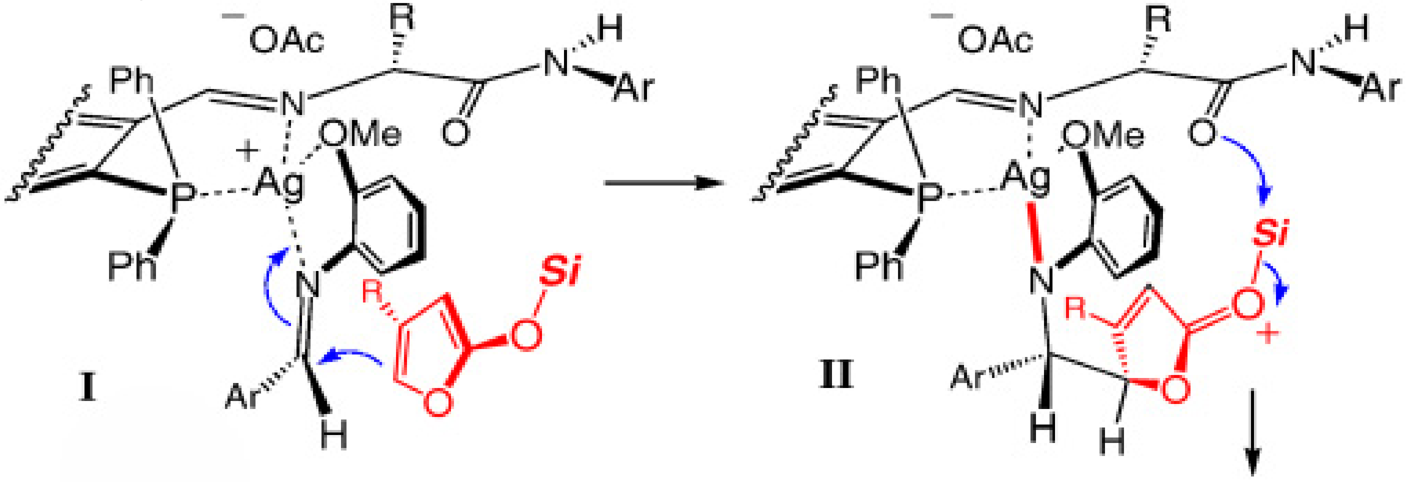
The following images give an idea how our current way of communicating chemical knowledge makes the life of computer scientists interesting

Semantik unklar

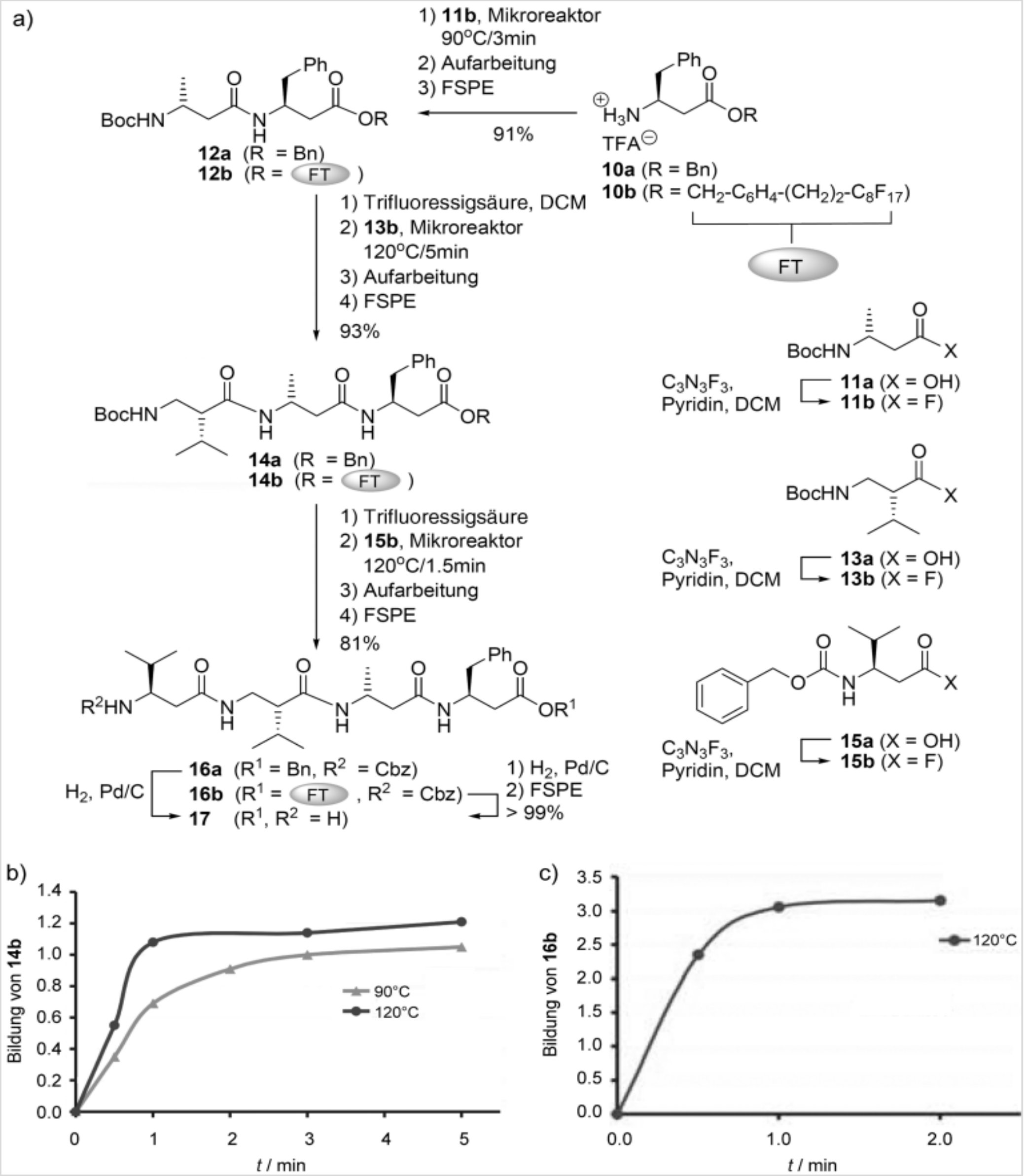


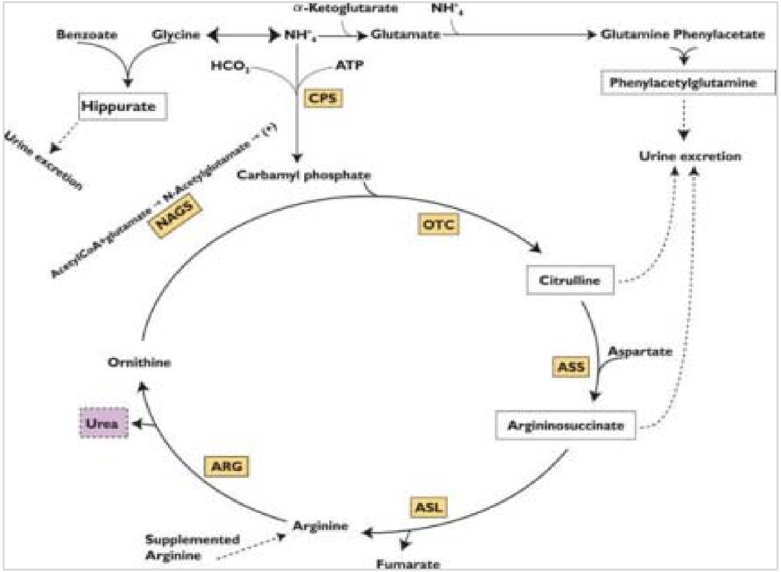


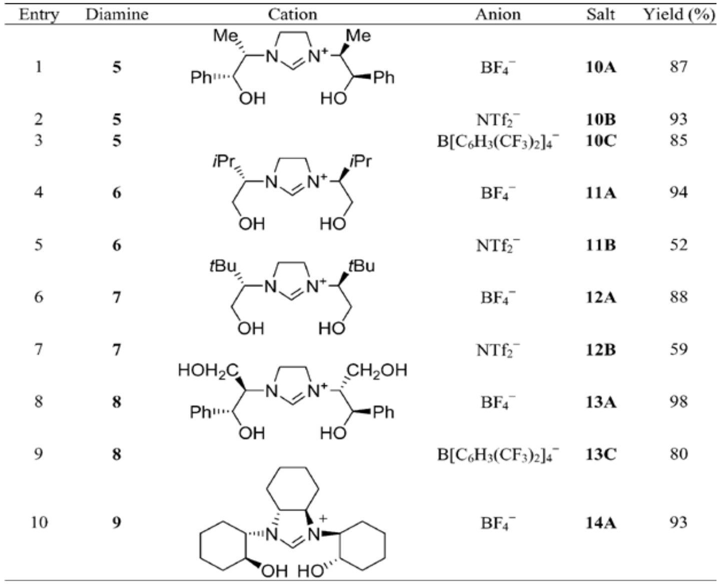




Komplex Zusammengesetzt



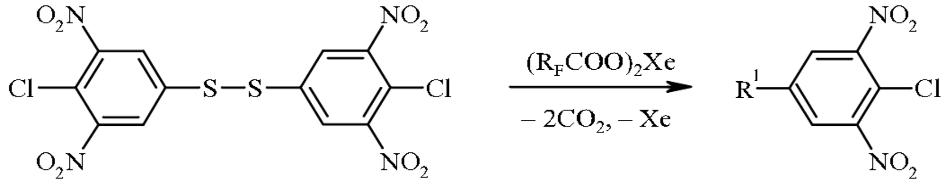




Entry	Diamine	Cation	Anion	Salt	Yield (%)
1	5		BF_4^-	**10A**	87
2	5		NTf_2^-	**10B**	93
3	5		$B[C_6H_3(CF_3)_2]_4^-$	**10C**	85
4	6		BF_4^-	**11A**	94
5	6		NTf_2^-	**11B**	52
6	7		BF_4^-	**12A**	88
7	7		NTf_2^-	**12B**	59
8	8		BF_4^-	**13A**	98
9	8		$B[C_6H_3(CF_3)_2]_4^-$	**13C**	80
10	9		BF_4^-	**14A**	93

Komplex Zusammengesetzt

$$2\ R_FCOOH + XeF_2 \longrightarrow (R_FCOO)_2Xe + 2\ HF$$



1 **2a,b**

2 a $R^1 = SCF_3$, **b** $R^1 = SC_2F_5$

$$RSSR + (R_FCOO)_2Xe \longrightarrow RSSR^{+\bullet} + (R_FCOO)_2Xe^{-\bullet}$$

$$(R_FCOO)_2Xe^{-\bullet} \longrightarrow R_FCOO^{\bullet} + R_FCOO^{-} + Xe$$

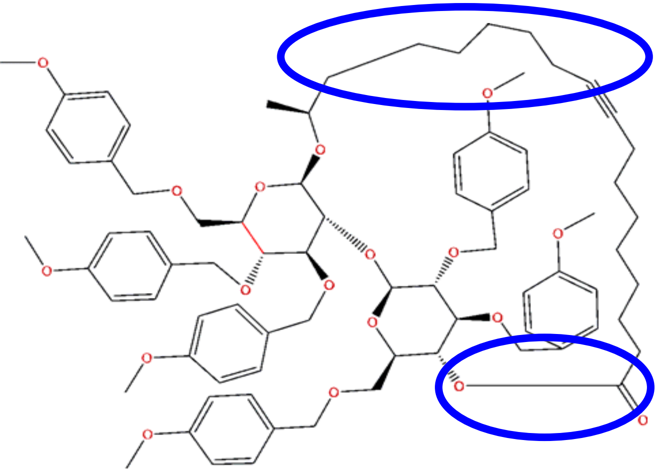
$$R_FCOO^{\bullet} \longrightarrow R_F^{\bullet} + CO_2$$

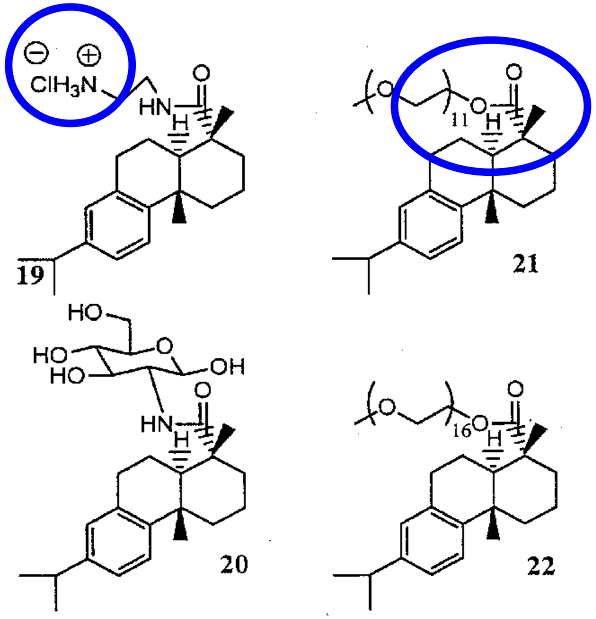
$$RSSR^{+\bullet} \longrightarrow RS^{\bullet} + RS^{+}$$

$$RS^{\bullet} + R_F^{\bullet} \longrightarrow RSR_F$$

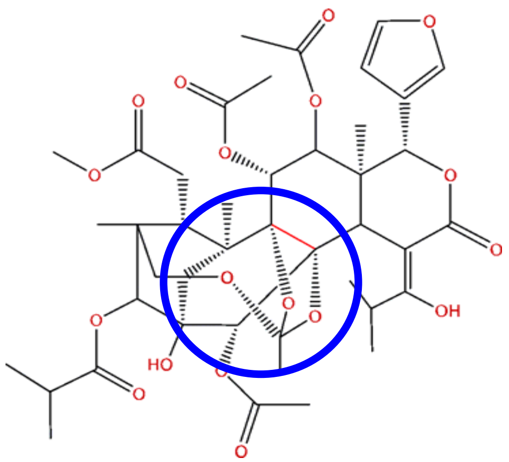
R = 1-chloro-2,6-dinitrophenyl; $R_F = C_nF_{2n+1}$ (when $n = 1$–2)

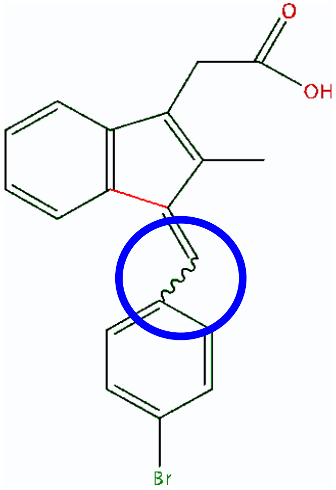
Zeichnerisch schwierig

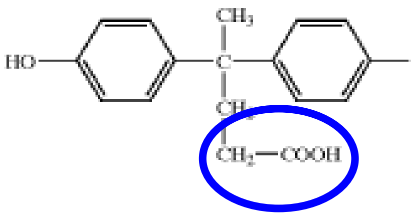


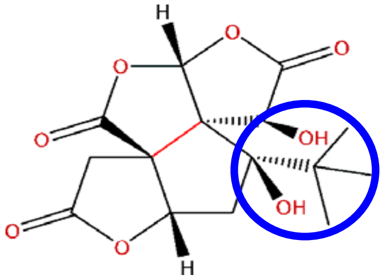


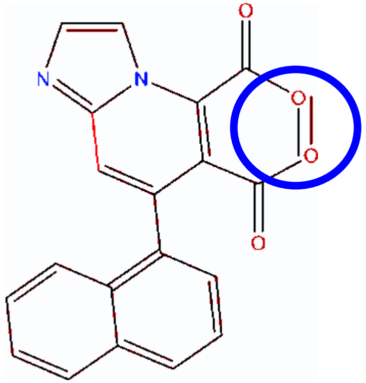
SCHEME 3. Dehydroabietic acid-based surfactants.

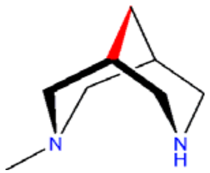




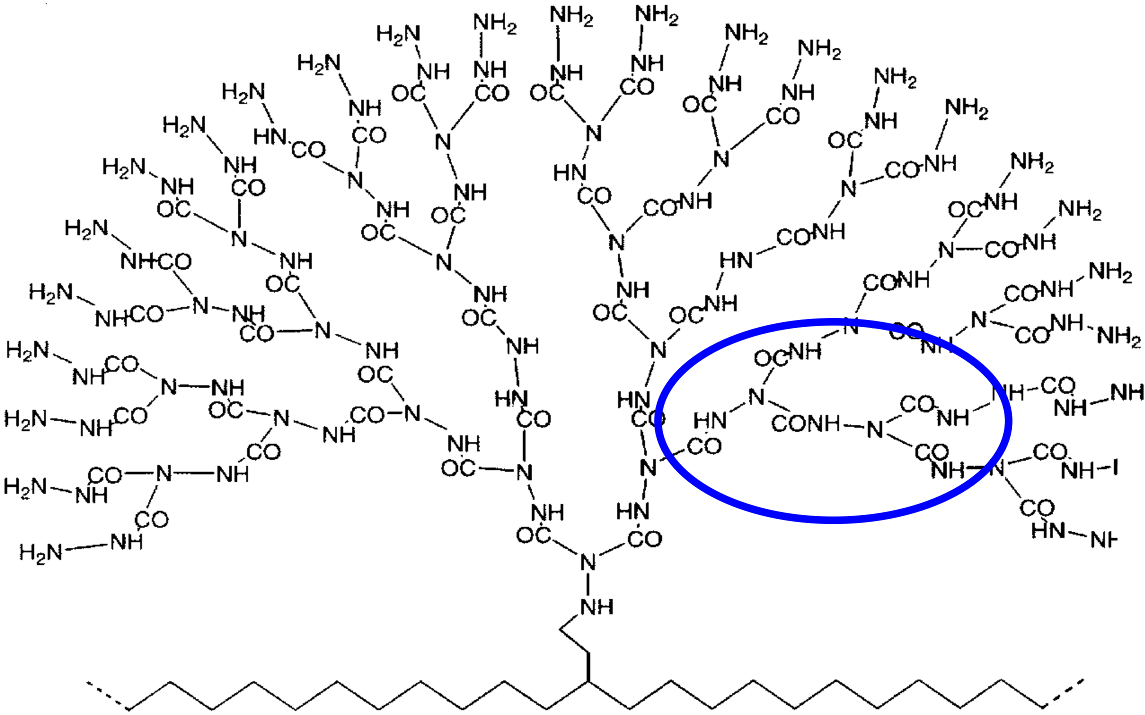


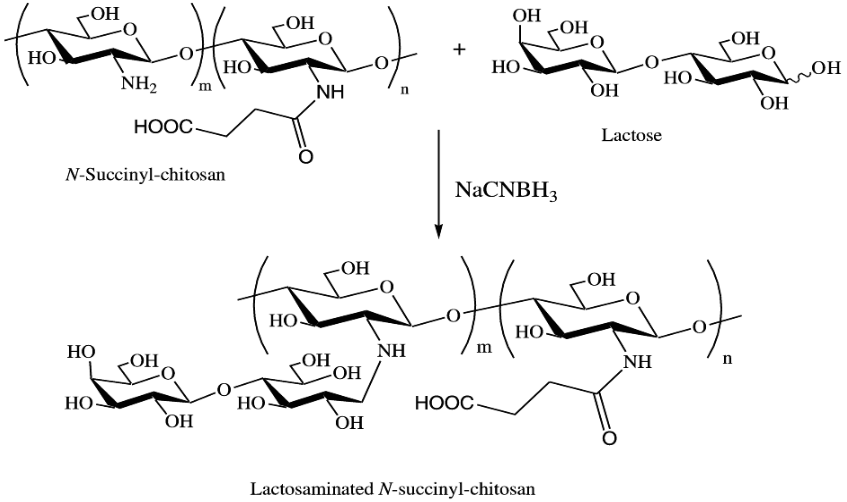


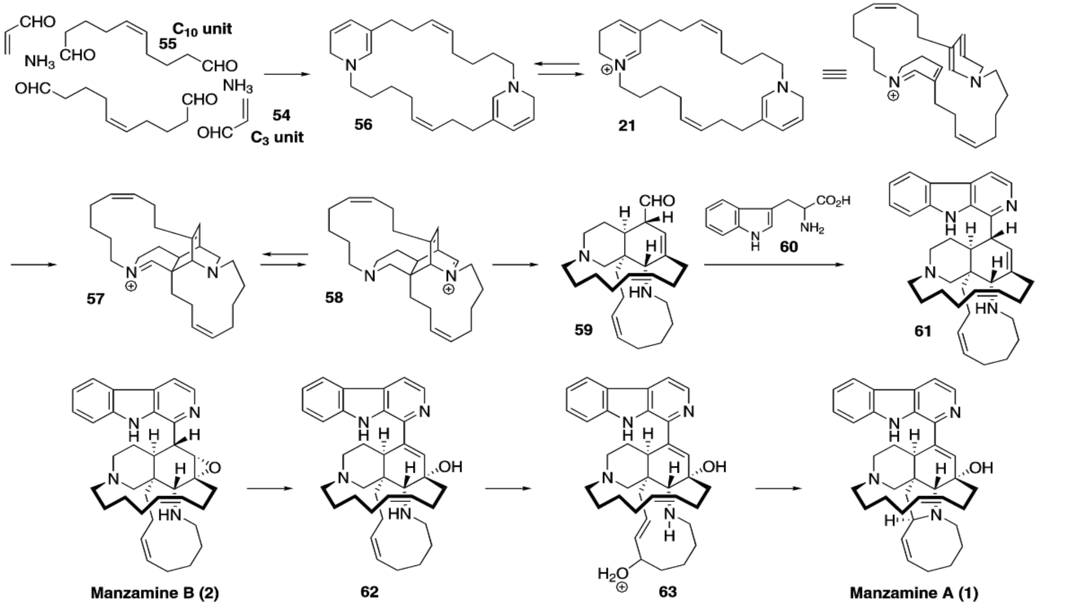




Zeichnerisch schwierig

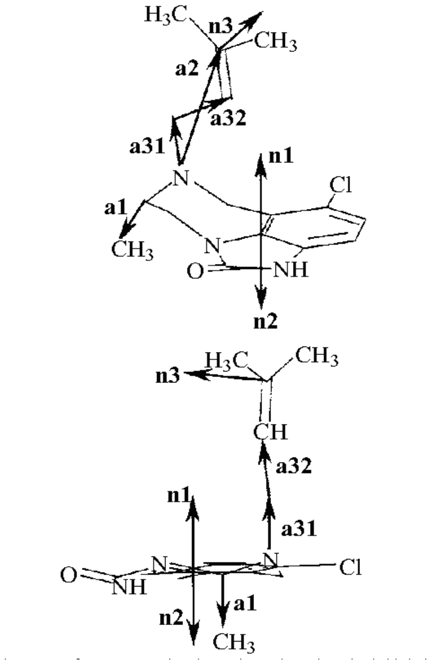


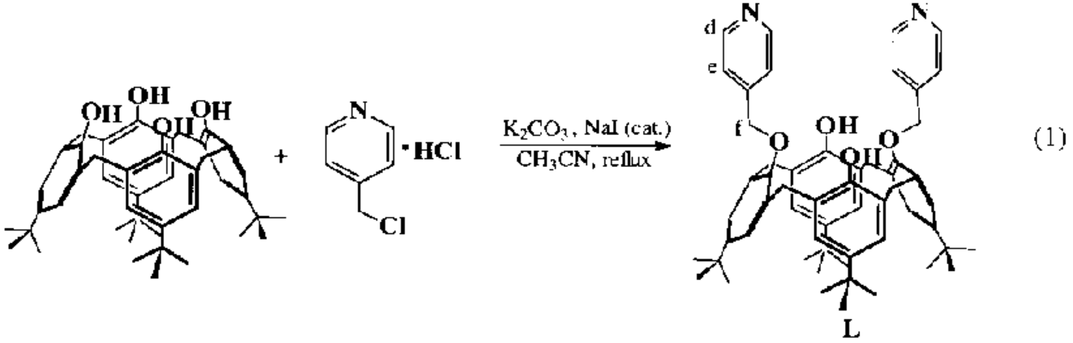


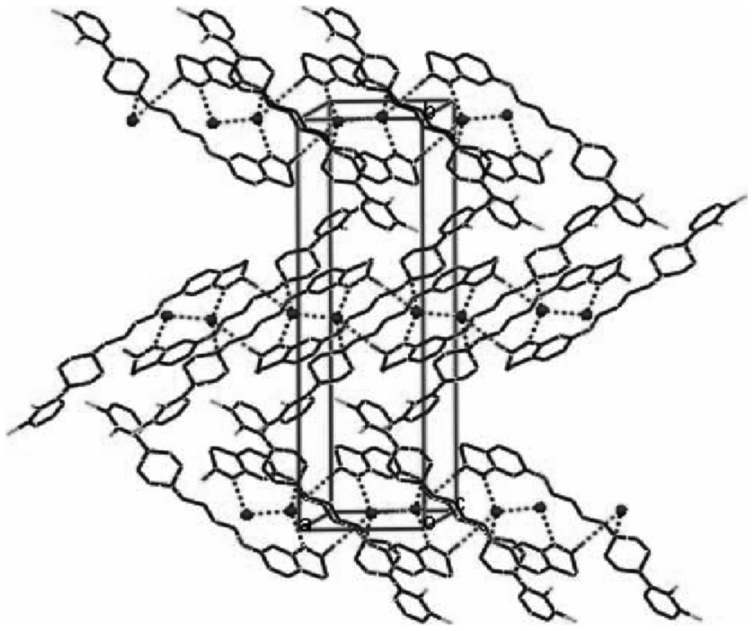


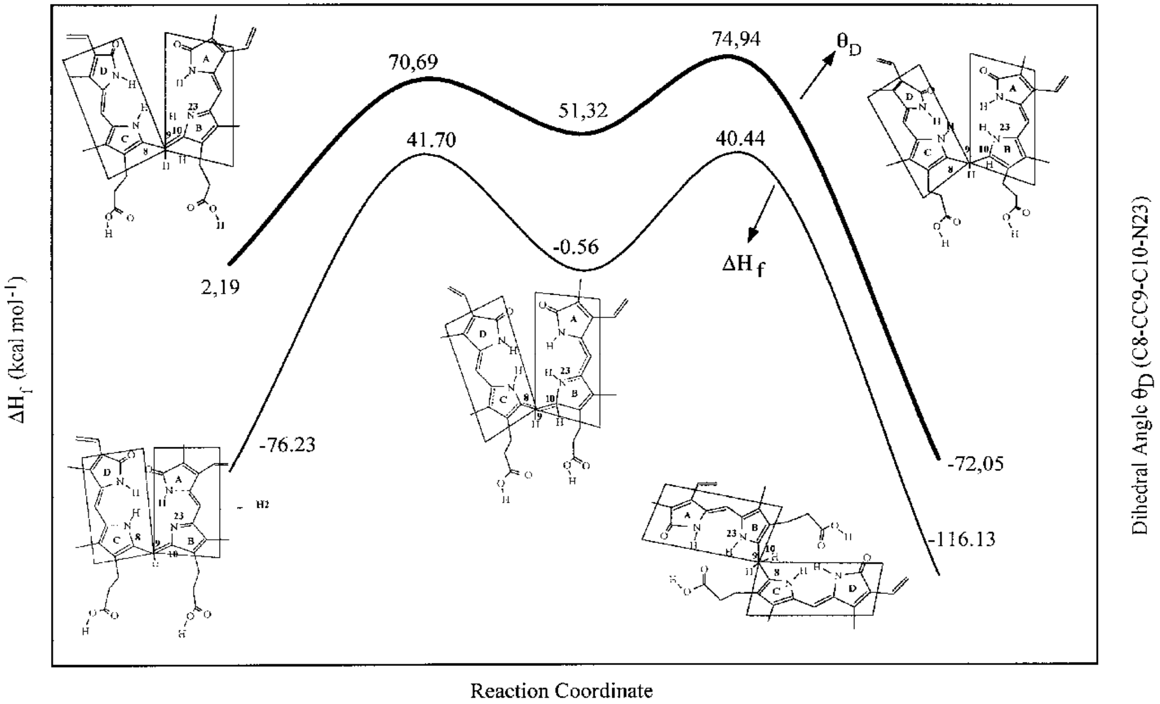
Scheme 5 Biosynthesis of manzamine A proposed by Baldwin and Whitehead

Völliger Wahnsinn









Acknowledgement

Olga Domanova

Juliane Fluck

Christoph Friedrich

Tobias Gattermayer

Harsha Gurulingappa

Carina Haupt

Vinod Kasam

Roman Klinger

Corinna Kolarik

Theo Mevissen

Bernd Müller

Thorsten Prinz

Paurush Praveen

Erfan Younesi

Antje Wolf

Marc Zimmermann