Information Extraction Technologies in Chemistry
A Critical Review

Fraunhofer Institute Algorithms and Scientific Computing

Martin Hofmann-Apitius

The International Conference in

Trends for Scientific Information Professionals

Sitges (Barcelona), Spain. 21-24 October 2007

Structure of this Presentation

Text & Image Mining in Chemistry – a Critical Review

1. Representation of chemical information in scientific literature
2. Technologies for chemical named entity recognition
3. Technologies for information extraction from chemical structure depictions
4. Benchmarking activities in the biomedical arena & Call for a joint initiative for benchmarking of information extraction technology in the field of chemistry

Representation of Chemical Information in Scientific Literature

Fraunhofer Institute Algorithms and Scientific Computing

Sources of Chemical Information in Scientific Literature

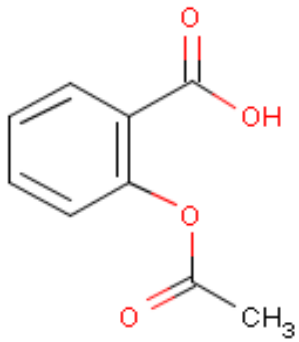
Scientific literature comprising chemical information is not limited to journal publications. Unstructured knowledge sources containing chemical information are:

- Journal articles
- Patents
- Books (incl. Chemical Handbooks)
- Doctoral Theses and Project Reports
- Package inserts / Chemical hazard documentation
- Websites

Identification and Representation in Chemistry

- **Trivial names (incl. brands):** Aspirin, Acetylsalicyclic acid, ….

- **Systematic nomenclatures:**
  - Mass formula: C9H8O4
  - SMILES: OC(=O)C1=C(C=CC=C1)OC(=O)C
  - InChI: 1/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)
  - IUPAC: pyrido[1",2":1',2']imidazo[4',5':5,6]pyrazino[2,3-*b*]phenazine



- **References to registration numbers** (e.g. CAS or Beilstein)

- **Structural formula:** universal language between chemists
  Chemical properties ~ atom composition + spatial arrangement

SCAI
Fraunhofer Institute
Algorithms and
Scientific Computing

Chemical Information Communication

- During the publication process

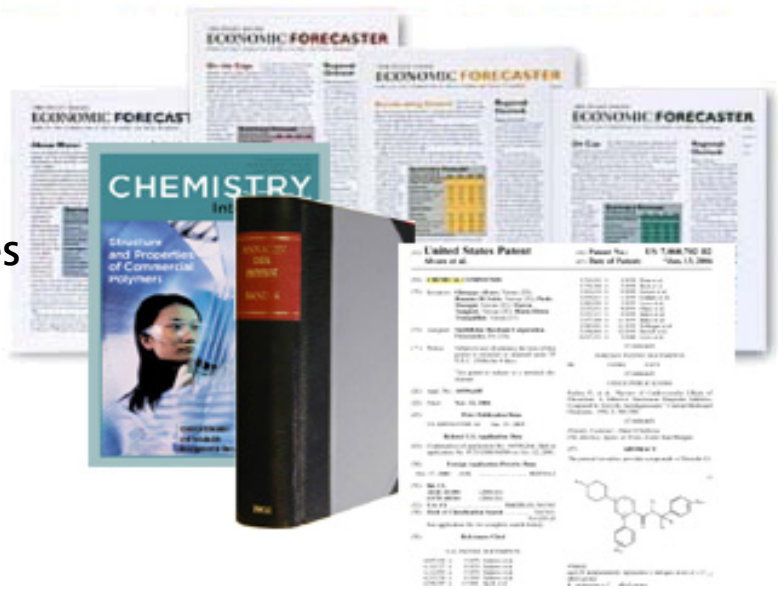
  chemical entities in text appear as

  trivial names, systematic and semi-systematic names

  moreover

  Molecule structures are published as images

⇒ the machine readable format is lost



SCAI
Fraunhofer Institute
Algorithms and
Scientific Computing

Technologies for Chemical Named Entity Recognition (NER)

Fraunhofer Institute Algorithms and Scientific Computing

Types of Chemical Named Entity Recognition (NER)

Three classes of automated Chemical NER can be distinguished:

- NER of chemical entities based on dictionaries
- NER of chemical entities based on rules (regular expressions)
- NER of chemical entities based on training sets (machine learning)

It is noteworthy that there exist also combinations of these approaches, e.g. using dictionaries as a feature in machine learning approaches.

Reports on Chemical Named Entity Recognition (NER)

Chemical NER is a rather new, emerging field. Information extraction specialists such as TEMIS have teamed up with cheminformatics specialists (in the case of TEMIS it is MDL) and cheminformatics specialists such as InfoChem have started their own information extraction approaches and teamed up with research teams like our team at SCAI. Other combinations are e.g. InforSense (a workflow environment provider) and Linguamatics; or the text mining activities of Accelrys (SciTegic⟐s Pipeline Pilot Technology).

Benchmarking is difficult as chemical knowledge is traditionally much more proprietary than biological knowledge and therefore there is no joint critical assessment of information extraction technologies in chemistry (unlike the BioCreative assessment of text mining tools in molecular biology)

Recent momentum comes from groups like the Cambridge-based research groups of Murray-Rust, Corbett, Teufel, and others, who push open source developments in chemical knowledge management and chemical NER.

Chemical NER in the Open Source Arena: OSCAR

OSCAR is an open source system for chemical NER developed by the groups of Corbett, Murray-Rust, Teufel and others in Cambridge.

OSCAR uses chemical dictionaries as well as rule-sets (regular expressions) including e.g. rules like “if it ends with -ase it is most likely an enzyme”.

OSCAR has been combined with a scientific document parser (SciXML format) and a recent paper by Batchelor and Corbett reports on first experiences with chemical NER, comparing NER of GO-terms with NER of chemical compounds in the same text (Proceedings of the Association for Computational Linguistics 2007; Demo & Poster Session).

OSCAR comprises a machine learning approach combined with dictionary features and rule-sets, that are used to define features.

What we Learned from OSCAR Publications and Preliminary Tests done at Fraunhofer SCAI ...

OSCAR is - like many open source tools - poorly documented, but source code is available

OSCAR authors provide annotation guidelines and - most important - they published on inter-annotator agreement ... Valuable information for the assessment of the quality of training corpora

The developers of OSCAR still see some issues with the algorithmic approaches taken:

- no stemming / lemmatization (detrimental to IUPAC expressions anyway)

- context-dependent disambiguation in chemistry --> example plural forms

- enumerations are a significant problem

Entity-Types for chemical NER and Performance of chemical NER using LingPipe (Technology used in OSCAR)

Type	Description	Example
CM	chemical compound	citric acid
RN	chemical reaction	1,3-dimethylation
CJ	chemical adjective	pyrazolic
ASE	enzyme	methylase
CPR	chemical prefix	1,3-

Table 1: Named entity types

Configuration	P	R	F
`TokenShape`	67.0%	52.9%	59.1%
$+c$	71.2%	62.3%	66.5%
$+t$	67.4%	52.5%	59.0%
$+c+t$	73.3%	62.5%	67.4%
`CharLm`	62.7%	63.4%	63.1%
$+l$	59.8%	68.8%	64.0%
$+t$	71.1%	70.0%	70.5%
$+l+t$	75.3%	73.5%	74.4%

Table 5: LingPipe performance using different configurations. c = custom token classifier, l = chemical name lists, t = custom tokeniser

Corbett, Batchelor and Teufel ***Annotation of Named Chemical Entities***
BioNLP 2007: Biological, translational, and clinical language processing, pages 57-64, Prague, June 2007

Chemical NER in *Patent* Literature

Chemical NER from patent literature is not trivial at all. Although e.g. Accelrys claims that their text analytics collection of the Pipeline Pilot is able to handle patents, we believe that current text mining technology still faces significant problems when applied to patents at production scale. Patents are complicated because:

- they are not necessarily written to be understandable

- many patents exist as PDF and have to be OCR-ed, which introduces errors

- patents contain a lot of chemical information in images and not in characters

- patents are full text documents with sometimes hundreds of page. Classical NLP tools are not able to work on such large documents - or, if the do, it takes ages.

Chemical Named Entity Recognition (NER) using CRFs

Conditional Random Fields (CRF) is a machine learning approach that is quite emerging in the text mining community. In the gene mentioning task of the BioCreative 2006 critical assessment, CRFs were applied by all top-ranking groups.

CRFs are based on graphical models for sequence-type information and text is nothing else but a sequence of characters and empty positions. CRFs are trained and work at the token level.

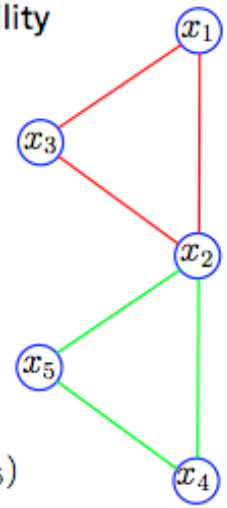
Undirected Graphical Models

- Decomposition (factorization) of a given probability distribution:

$$p(\vec{x}) = \frac{1}{Z} \prod_C \Psi_C(\vec{x}_C)$$

  - C: maximal cliques in the independency graph
  - Ψ_C: potential functions
  - $\vec{x}_C$: subset of $\vec{x}$ relevant for C
  - Z: normalization given by $Z = \sum_{\vec{x} \in \mathcal{X}} \prod_C \Psi_C(\vec{x}_C)$
- Concrete example:

$$p(\vec{x}) = \frac{1}{Z} \cdot \Psi_1(x_1, x_2, x_3) \cdot \Psi_2(x_2, x_4, x_5)$$



Appearance of IUPAC in Patents

- 2-[5-(p-tolylthiocarbonyl)pentyl]isoindoline-1,3-quinone
- 2-benzoyl-N-(2,4,6-triphenylpyridin-1-yl)benzamide
- 10-cycloheptyl-6-(4-propoxyphenyl)-5-thia-2,9,10-triazabicyclo[5.3.0]dec-11-ene-3,8-dione
- N-(2-methoxyphenyl)-4-[4-(trifluoromethyl)phenyl]carbonyl-1-thia-4,8-diazaspiro[4.5]decane-8-carboxamide
- 1-hexoxy-4-methyl-hexane
- 4-[(2,5-dichlorothiophen-3-yl)sulfonylamino]benzoic acid
- 2-[[4-benzyl-5-(2-bromophenyl)-1,2,4-triazol-3-yl]thio]-N-methyl-N-phenyl-acetamide
- allyl N-[2-[6-[[(3-cyanophenyl)-oxo-methyl]amino]benzothiazol-2-yl]thioethyl]carbamate

Recognition of IUPAC in Patents: First Steps

- First idea: Artifical Corpus
- Replace gene names by randomly choosen IUPAC names in BioCreative Corpus
  - Comparison with alkaline phosphatases and 5-nucleotidase
  - Comparison with 7-chloro-2-(1-methylaminoethyl)-3-(4-methylphenyl)quinazolin-4-one and 1-cyclopropyl-1-(4-pyridylmethyl)-3-tosyl-urea
- Corpus with 15000 sentences with correct IUPAC names

SCAI
Fraunhofer Institute Algorithms and Scientific Computing

Lesson learned: Artificial Corpus =/= Real Patent Situation

- First idea: Artifical Corpus
- Replace gene names by randomly choosen IUPAC names in BioCreative Corpus
  - Comparison with alkaline phosphatases and 5-nucleotidase
  - Comparison with 7-chloro-2-(1-methylaminoethyl)-3-(4-methylphenyl)quinazolin-4-one and 1-cyclopropyl-1-(4-pyridylmethyl)-3-tosyl-urea
- Corpus with 15000 sentences with correct IUPAC names
- Evaluation with bootstrapping:
  precision: 97.65%, recall 97.1% F-score 97.37% ☺
- Annotation of a small corpus for real evaluation:
  precision: 58.82%, recall 35.24%, F-score 44.08% ☹

SCAI
Fraunhofer Institute Algorithms and Scientific Computing

BioMedical Objects can link from text to database entries

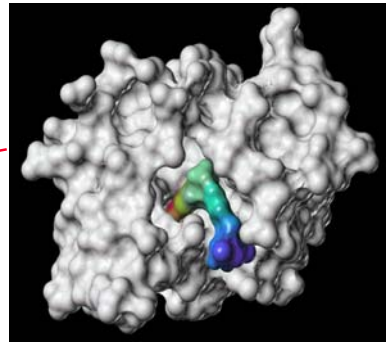
Abstract

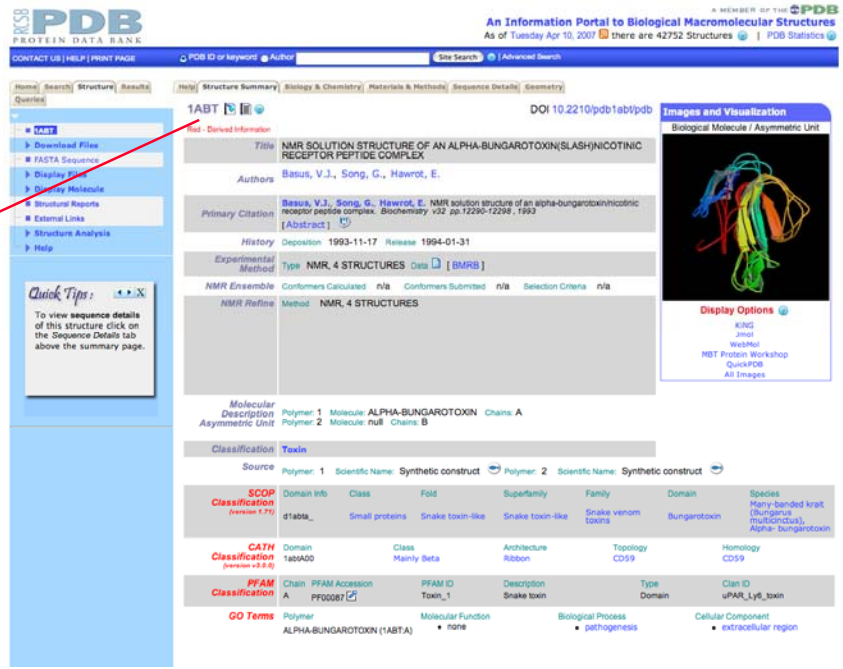
Background: Mast cell-derived prostaglandin D_2 (PGD_2), may contribute to eosinophilic inflammation and mucus production in allergic asthma. Chemoattractant receptor homologous molecule expressed on TH_2 cells (CRTH2), a high affinity receptor for prostaglandin D_2, mediates trafficking of TH_2-cells, 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.





SCAI
Fraunhofer Institute Algorithms and Scientific Computing

Chemical Objects can link from text to virtual experiments ...

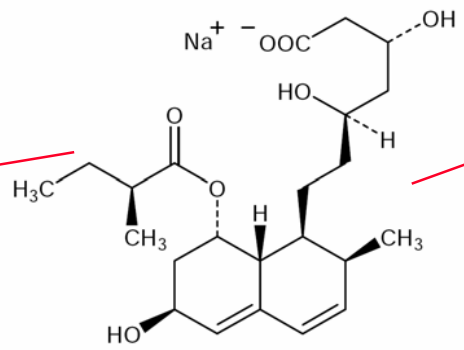
Abstract

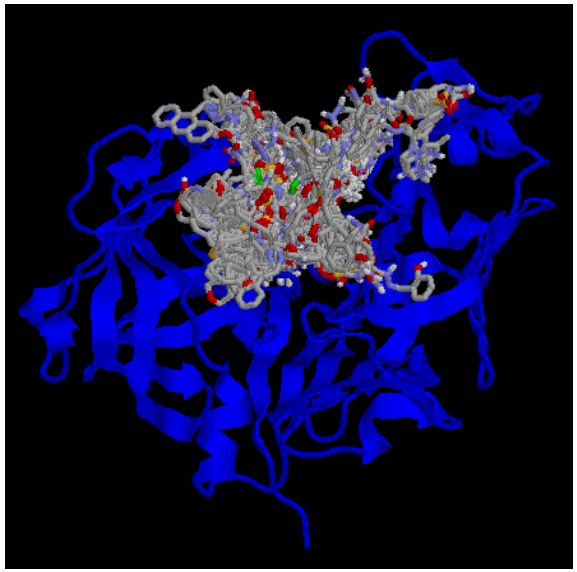
Background: Mast cell-derived prostaglandin D_2 (PGD_2), may contribute to eosinophilic inflammation and mucus production in allergic asthma. Chemoattractant receptor homologous molecule expressed on TH_2 cells (CRTH2), a high affinity receptor for prostaglandin D_2, mediates trafficking of TH_2-cells, 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.





DOCKING:
Simulated ligand binding

SCAI
Fraunhofer Institute
Algorithms and
Scientific Computing

Name 2 Structure

Conversion of names to structures offers a couple of options that can help to verify molecule information.

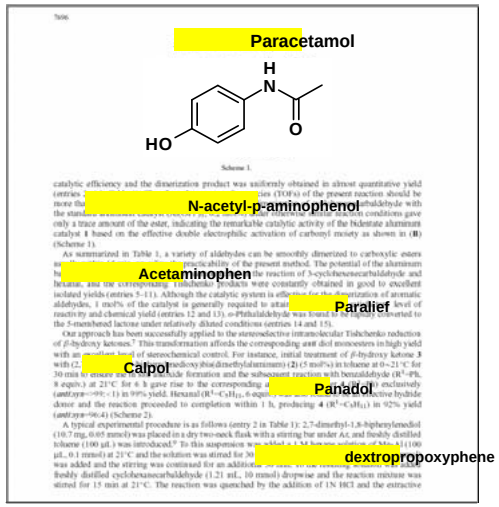
Names (trivial names as well as IUPAC, InChi or CAS designators) can be converted to structures using one of the about 4 tools available on the market.

Once the name is converted to a structure, all the nice gimmics that work on structures can be done ... Alignments, fragment searches, similarity searches, identity etc.

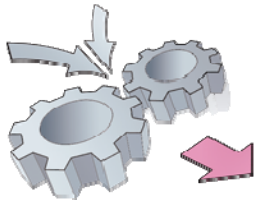
Now: how good are the current N2S tools ? Our colleagues at InfoChem did a preliminary study ...

Chemical Named Entity Recognition

Document



Annotation



Automatic recognition of chemical names

List of chemical names

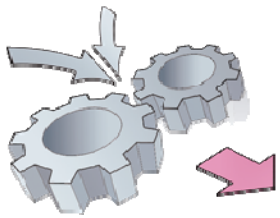
- Paracetamol
- Acetaminophen
- N-acetyl-p-aminophenol
- Calpol®
- Panadol®
- ….

- InfoChem is using the annotation software from IBM (*Chemical Annotator V.2.01*)
- It recognizes and extracts/highlights chemical names from text
- It uses various dictionaries of names and fragments

Name to Structure Conversion Tool (N2S)

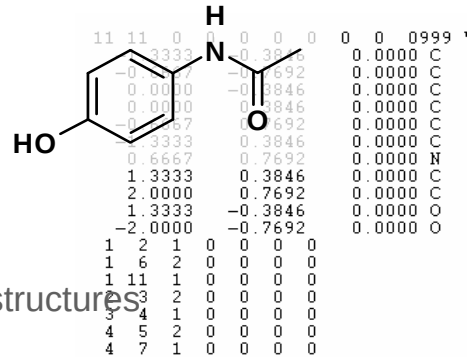
List of chemical names

- Paracetamol
- Acetaminophen
- N-acetyl-p-aminophenol
- Calpol®
- Panadol®
- ….



Automatic conversion of chemical names to structures

Connection Table



- Goal: conversion of chemical names to structures (connection tables, e.g. MOLfile)
- Tools in test-phase: CambridgeSoft (*Name=Struct*, processed by IBM), ACD/Labs (*Name to Structure*), OpenEye (*Lexichem*)
- InfoChem (IC*N2S*) in development

Evaluation of N2S Tools

- First results

CambridgeSoft *Name=Struct*	ACDLabs *Name to Structure*	OpenEye *Lexichem*
Converts(1): 6.5 M	Converts(1): 3.5 M	Converts(1): 2.5 M
Philosophy: liberal, convert as much as possible	Philosophy: Rather strict, results more reliable	Philosophy: Rather strict
Test File: 2,164	Test File: 2,164	N/a
Conversion: 2,164 (100%)	Conversion: 1,142 (53%)	N/a
Correct: 1,486 (69%)	Correct: 985 (46%)	N/a
Wrong: 678 (31%)	Wrong: 157 (14%)	N/a

(1) Based on 6.5 M names abstracted from US patents 1976-2006 which could be converted with CambridgeSoft *Name=Struct*

Evaluation of N2S Tools Conclusions

- Challenges
  - “Real world names” instead of correct nomenclature
  - Ambiguous names (e.g. xylene: ortho/meta/para) => Mixture or error?
  - Recall/precision trade-off (correct conversion vs. high number of structures)
  - Deficiency of dictionaries (errors, missing fragments, trivial names, etc.)
  - Formally incorrect names (incorrect syntax, missing locants, typing errors etc.)
  - Chemical name mixed with text (e.g. methoxy-substituted benzene)

Technologies for Information Extraction from Chemical Structure Depictions

SCAI
Fraunhofer Institute
Algorithms and
Scientific Computing

Computer Systems are not Chemistry-aware

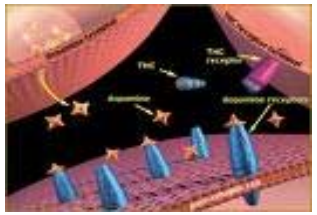
The universal language of chemistry is the graphical depiction of the chemical molecule structure. But even though chemical structure depictions can be readily interpreted by chemists, computers regard chemical structure depictions as a bunch of pixels. Consequently, structure depictions are information-rich, but cannot be used by computational approaches.

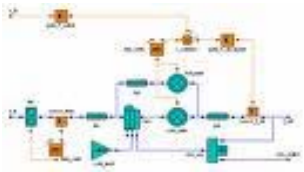
Computers are not chemistry-aware .... See example Google Image Search

Searching for Structural Information in Images

Google™ THC Search Images

Results 1 - 20 of about 113,000 for THC. (0.40 seconds)





New! Want to improve Google Image Search? Try Google Image Labeler.

SCAI
Fraunhofer Institute
Algorithms and
Scientific Computing

Computer-based Interpretation of Chemical Structure Depictions

4 different systems published so far:

- **Kékulé** (Joe R. McDaniel, Jason R. Balmuth Kekule: OCR-optical chemical (structure) recognition Journal of Chemical Information and Computer Sciences Date: 1992 Volume: 32 Issue: 4p. 373 - 378)
- **CLiDE** (P. Ibison, F. Kam, R.W. Simpson, C. Tonnelier, T. Venczel and A.P.Johnson: Chemical Structure Recognition and Generic Text Interpretation in the CLiDE project Proceedings on Online Information 92, 1992, London, England)
- **OSRA** (open source chemical OCR; see http://cactus.nci.nih.gov/osra/ started 2005 ?)
- **chemoCR** (see http://www.scai.fraunhofer.de/chemocr_references.html started 2004)

Reconstruction of Synthesis Pathways from Patents: an Example

@neuLink

Interaction Networks | Find Candidate Genes | MicroArray | Data Mining | Disease Models

Chromosome View shows the location of the found entity in a Chromosome browser.

Use Document Base Patents

X 10K Get Entities >

Enter full text search phrase

- Anatomy
- Disease
- DrugNames
- ProteinGene
- IUPAC

Documents | Entity | All

Show Patent

All | Anatomy | Disease | DrugNames | ProteinGene | IUPAC

Substituted pyridinyl-2-(diaza-bicyclo-alkyl)-pyrimidinone derivatives

Europäisches Patentamt
European Patent Office
Office européen des brevets

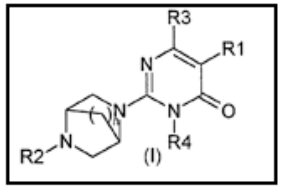
Date: 2004-09-08
File Identifier: EP03290571A1
EP Identifier: 1454908

Statistics

	Anatomy	Disease	Drug	Protein/Gene	IUPAC
Absolute	27	125	6	11	155
Top rank		'Neurodegenerative Diseases' Alzheimer disease		GSK3B	'2-(diaza-bicyclo -alkyl)-pyrimidone'

35 could be converted to structure using a single name to structure tool

The invention relates to a 2-(diaza-bicyclo-alkyl)-pyrimidone derivative represented by formula (I) or a salt thereof:



wherein:

- R1 represents a hydrogen atom, a C_{1-6} alkyl group or a halogen atom;

Biological Background Information

The invention relates also to a medicament comprising the said derivative or a salt thereof as an active ingredient which is used for preventive and/or therapeutic treatment of a **neurodegenerative disease** caused by abnormal activity of **GSK3β**, such as **Alzheimer disease**.

Technical Field

The present invention relates to compounds that are useful as an active ingredient of a medicament for preventive and/or therapeutic treatment of **neurodegenerative diseases** caused by abnormal activity of **GSK3β**.

Background Art

GSK3β (**glycogen synthase kinase 3β**) is a proline directed serine, threonine **kinase** that plays an important role in the control of metabolism, differentiation and survival. It was initially identified as an enzyme able to phosphorylate and hence inhibit glycogen synthase. It was later recognized that **GSK3β** was identical to **tau protein kinase 1** (**TPK1**), an enzyme that phosphorylates tau protein in epitopes that are also found to be hyperphosphorylated in **Alzheimer's disease** and in several taupathies. Interestingly, protein **kinase** B (**AKT**) **phosphorylation** of **GSK3β** results in a loss of its **kinase activity**, and it has been hypothesized that this inhibition may mediate some of the effects of neurotrophic factors. Moreover, **phosphorylation** by **GSK3β** of **β-catenin**, a protein involved in cell survival, results in its degradation by an ubiquitinilation dependent **proteasome pathway**.

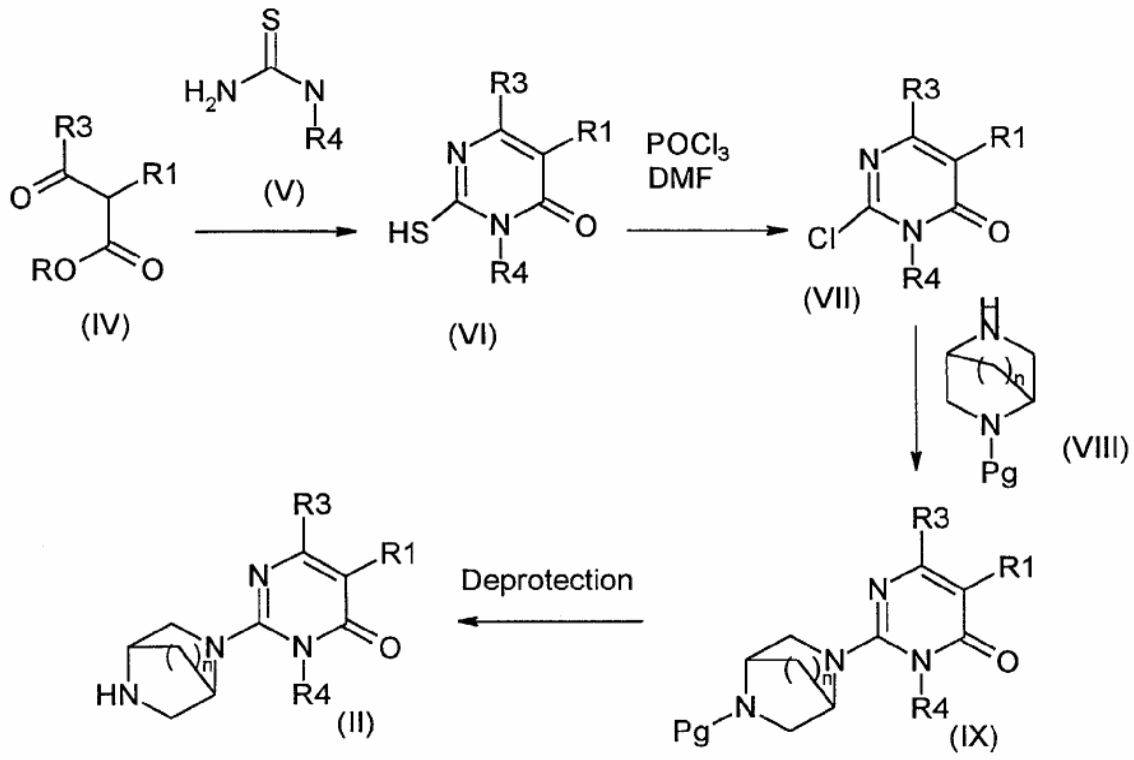
Thus, it appears that inhibition of **GSK3β** activity may result in neurotrophic activity. Indeed there is evidence that lithium, an uncompetitive inhibitor of **GSK3β**, enhances neuritogenesis in some models and also increases neuronal survival, through the induction of survival factors such as **Bcl-2** and the inhibition of the expression of proapoptotic factors such as **P53** and **Bax**.

Recent studies have demonstrated that β-amyloid increases the **GSK3β** activity and tau protein **phosphorylation**. Moreover, this **hyperphosphorylation** as well as the neurotoxic effects of β-amyloid are blocked by lithium chloride and by a **GSK3β** antisense mRNA. These observations strongly suggest that **GSK3β** may be the link between the two major **pathological processes** in **Alzheimer's disease**: abnormal **APP** (**Amyloid Precursor Protein**) processing and tau protein **hyperphosphorylation**.

Although tau **hyperphosphorylation** results in a destabilization of the neuronal **cytoskeleton**, the pathological consequences of abnormal **GSK3β** activity are, most likely, not only due to a pathological **phosphorylation** of tau protein because, as mentioned above, an excessive activity of this **kinase** may affect survival through the modulation of the expression of apoptotic and antiapoptotic factors. Moreover, it has been shown that β-amyloid-induced increase in **GSK3β** activity results in the **phosphorylation** and, hence the inhibition of **pyruvate dehydrogenase**, a pivotal enzyme in energy production and **acetylcholine synthesis**.

Altogether these experimental observations indicate that **GSK3β** may find application in the treatment of the neuropathological consequences and the cognitive and attention deficits associated with **Alzheimer's disease**, as well as other acute and chronic **neurodegenerative diseases**. These include, in a nonlimiting manner, **Parkinson's disease**, **tauopathies** (e.g. frontotemporoparietal dementia, corticobasal degeneration, **Pick's disease**, **progressive supranuclear palsy**) and other dementia including vascular dementia; acute stroke and others traumatic **injuries**; **cerebrovascular accidents** (e.g. age related **macular degeneration**); brain and **spinal cord trauma**; **peripheral neuropathies**; retinopathies and glaucoma.

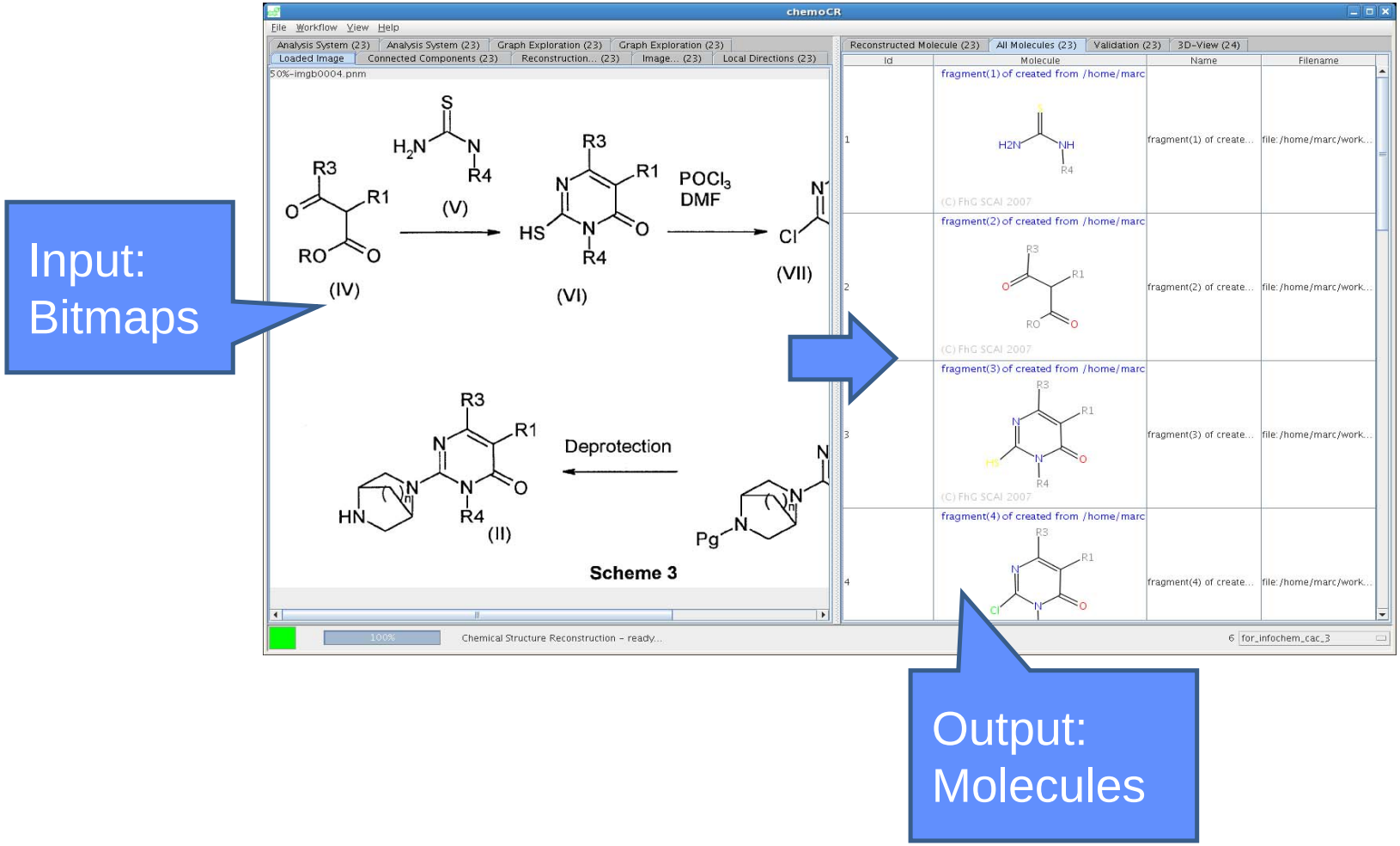
A Synthesis Reaction Schema



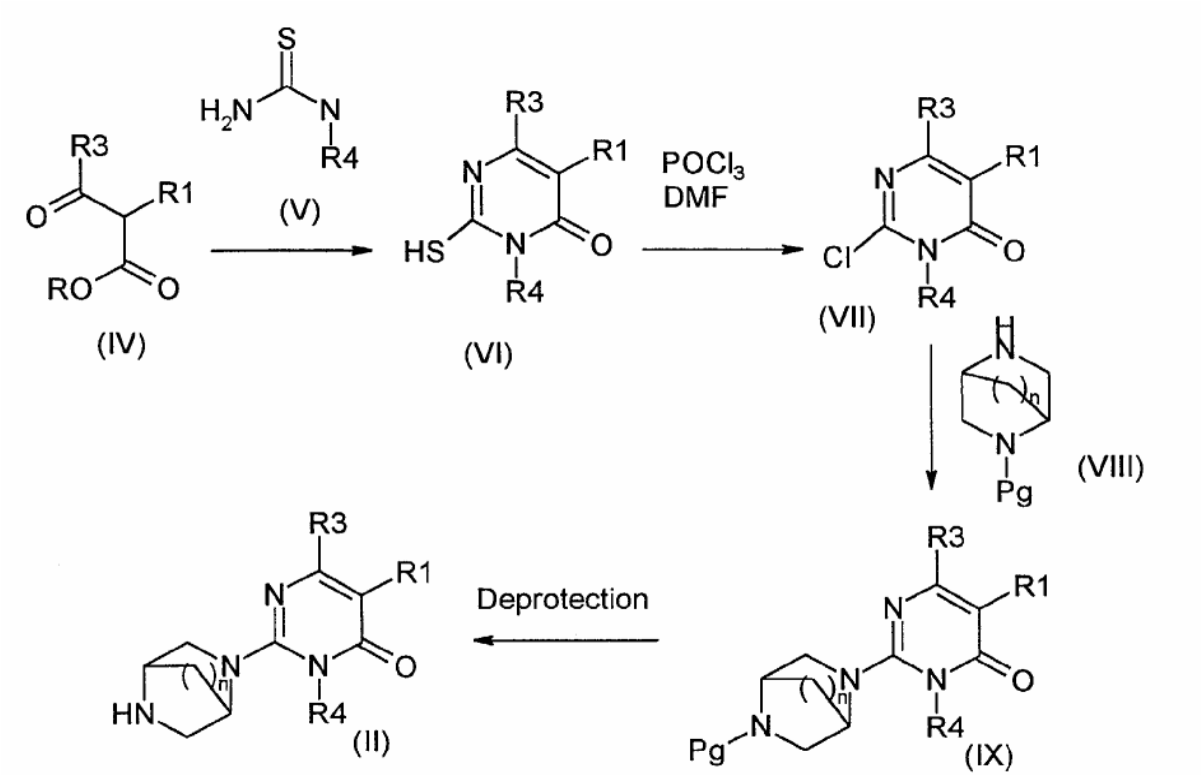
Scheme 3

Fraunhofer Institute Algorithms and Scientific Computing

ChemoCR GUI


Input:
Bitmaps
chemoCR
File Workflow View Help
Analysis System (23)
Analysis System (23)
Graph Exploration (23)
Graph Exploration (23)
Loaded Image
Connected Components (23)
Reconstruction... (23)
Image... (23)
Local Directions (23)
50%-imgb0004.pnm
(IV)
(V)
(VI)
(VII)
POCl3
DMF
Deprotection
(II)
Scheme 3
Reconstructed Molecule (23)
All Molecules (23)
Validation (23)
3D-View (24)
Id
Molecule
Name
Filename
fragment(1) of created from /home/marc
fragment(2) of created from /home/marc
fragment(3) of created from /home/marc
fragment(4) of created from /home/marc
(C) FhG SCAI 2007
Chemical Structure Reconstruction - ready...
Output:
Molecules

From Picture to Reaction: Pre-Processing

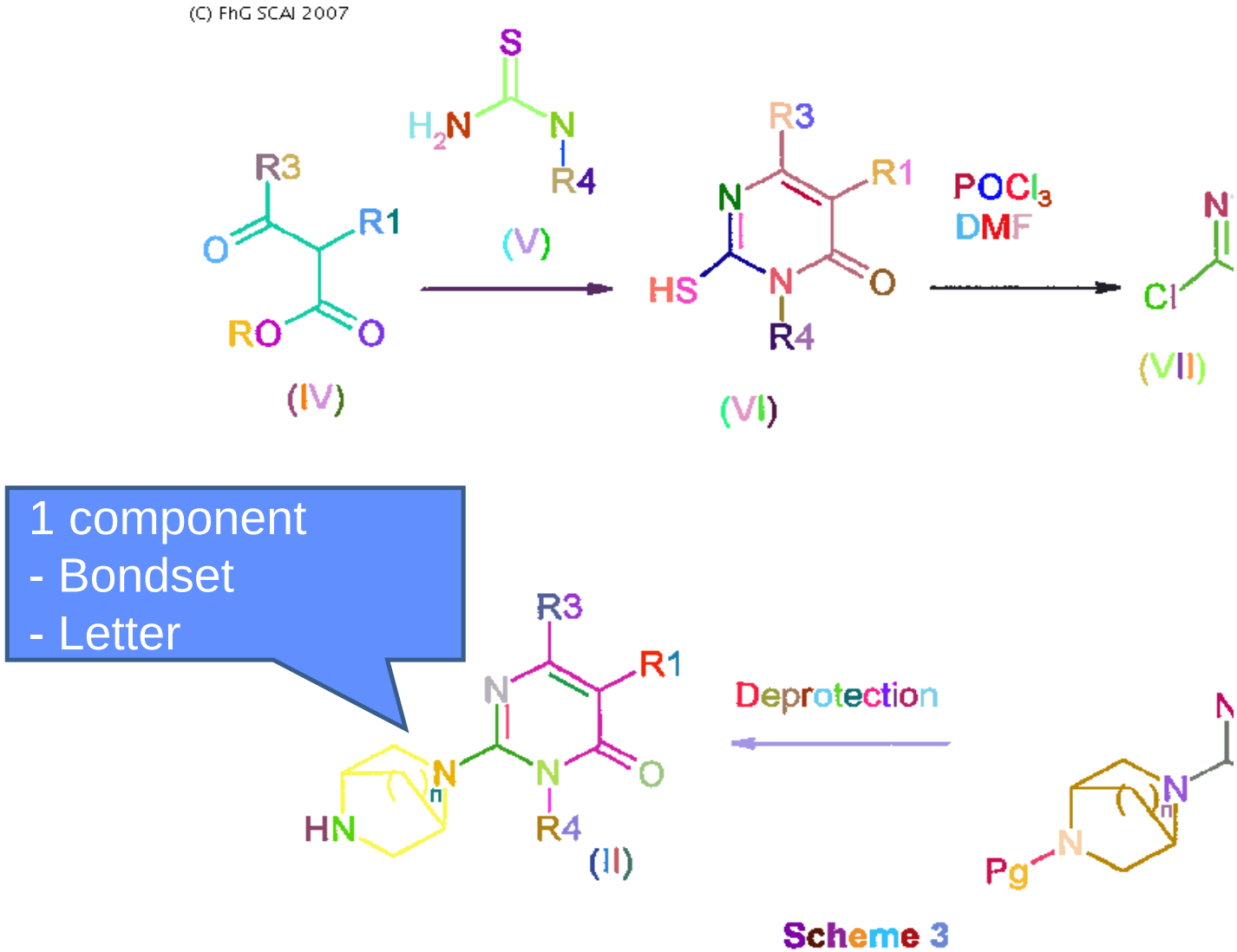


Scheme 3

0: Picture

BMP after scaling
& binarization

From Picture to Reaction: Identification of Connected Components

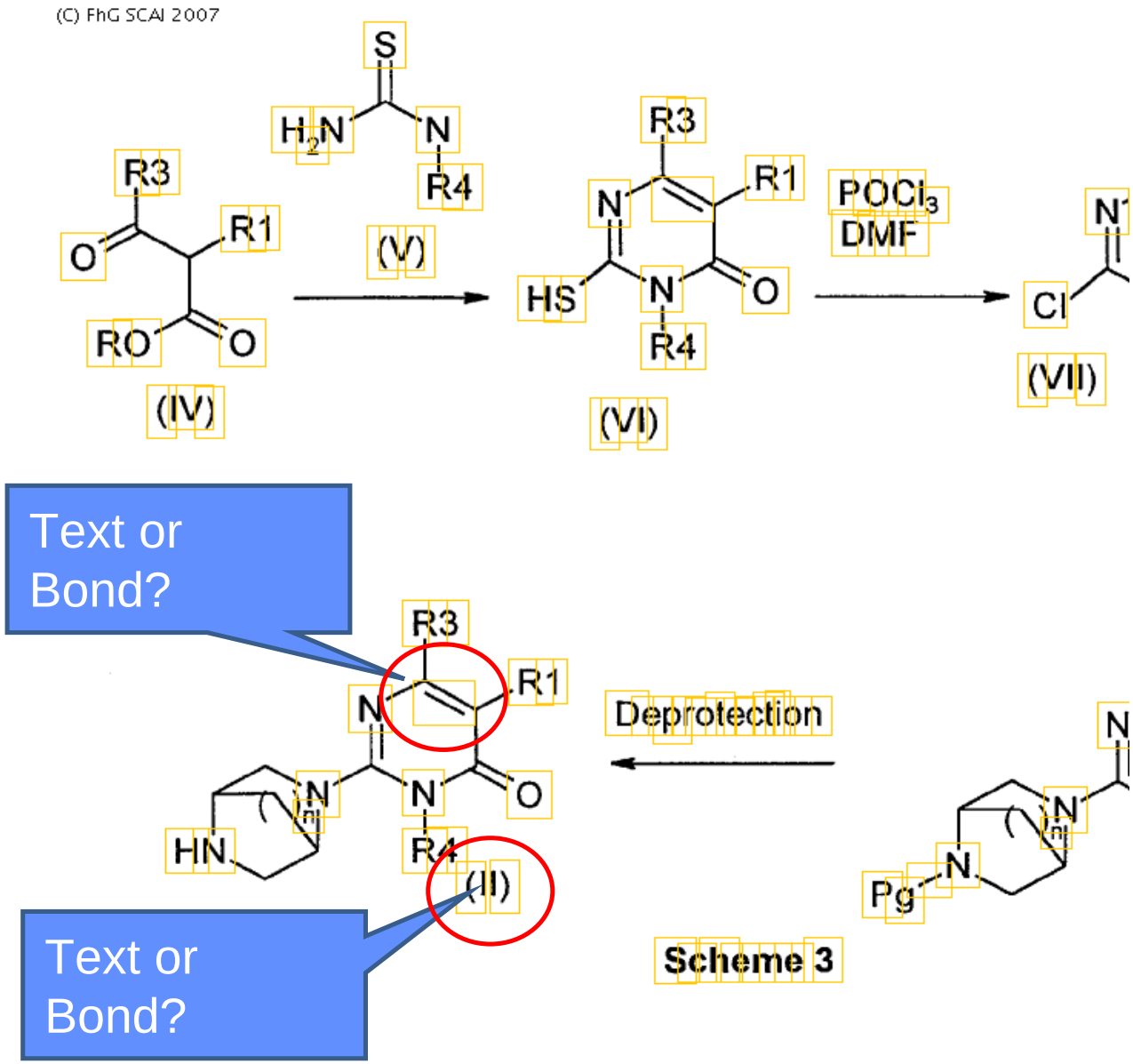


Scheme 3

0: Picture

1: Connected Components

From Picture to Reaction: Tagging of Characters

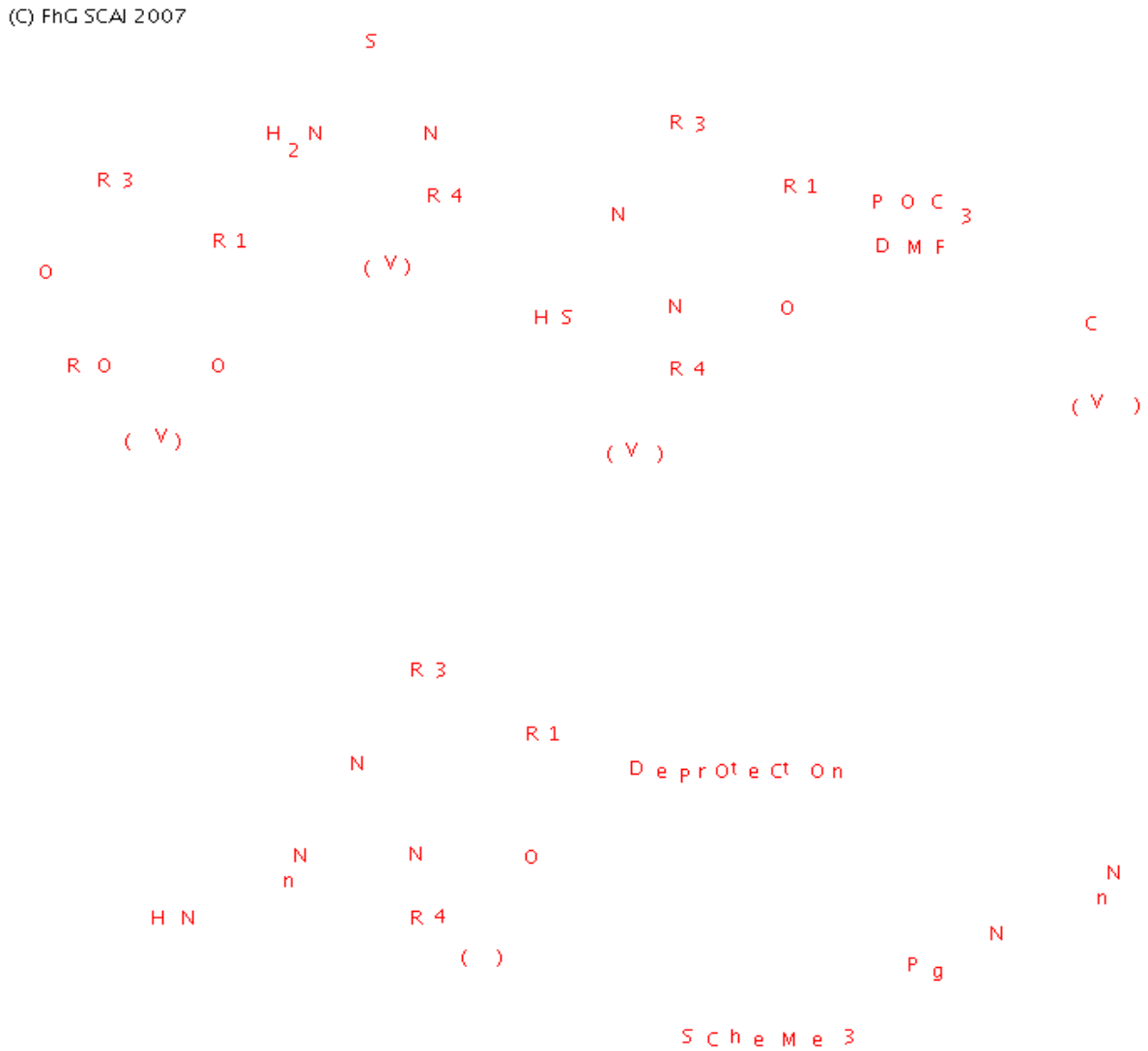


0: Picture

1: Connected Components

2: Tag Text

From Picture to Reaction: OCR of Identified Characters



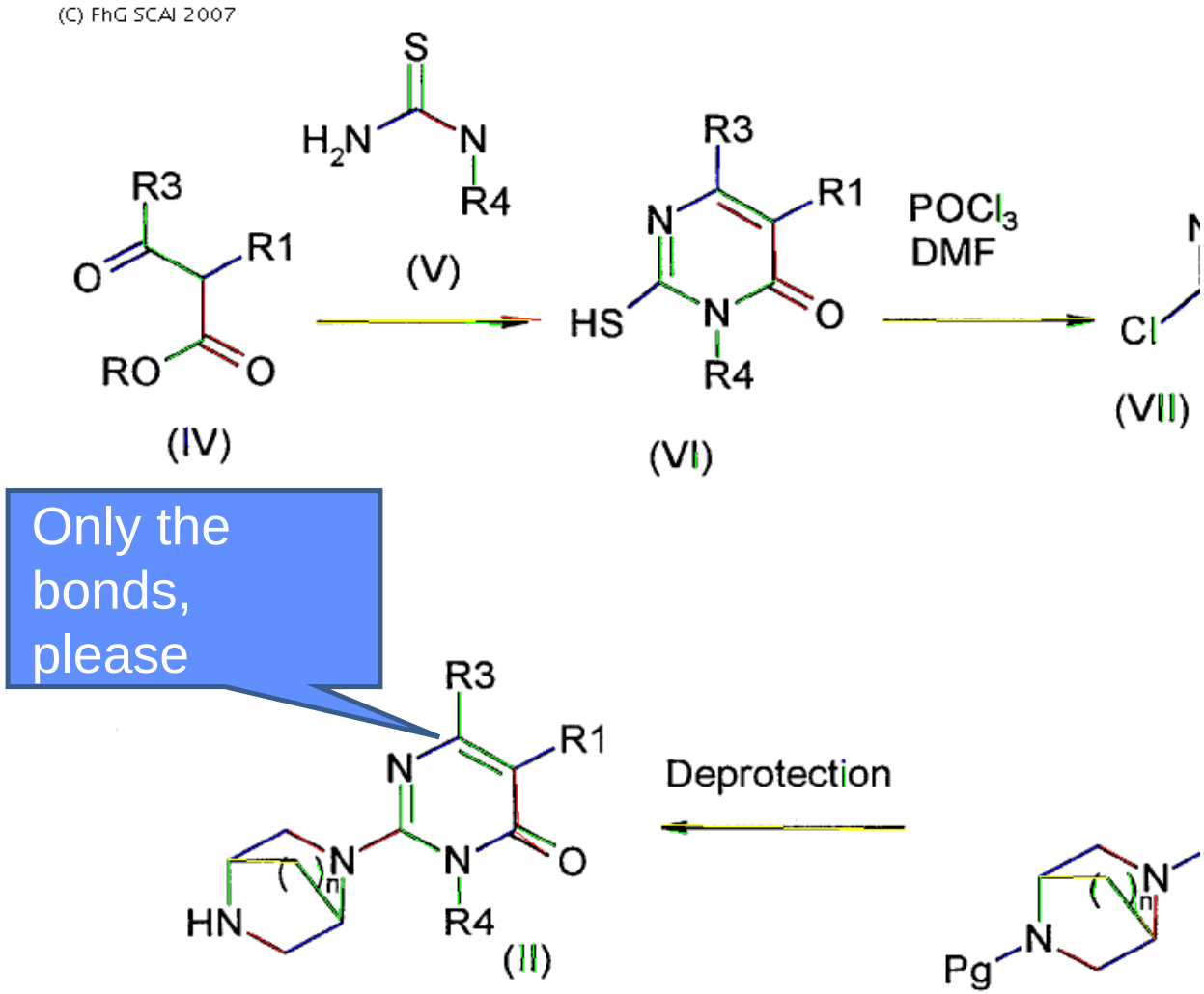
0: Picture

1: Connected Components

2: Tag Text

3: OCR

From Picture to Reaction: Identification of Lines (Bonds)



0: Picture

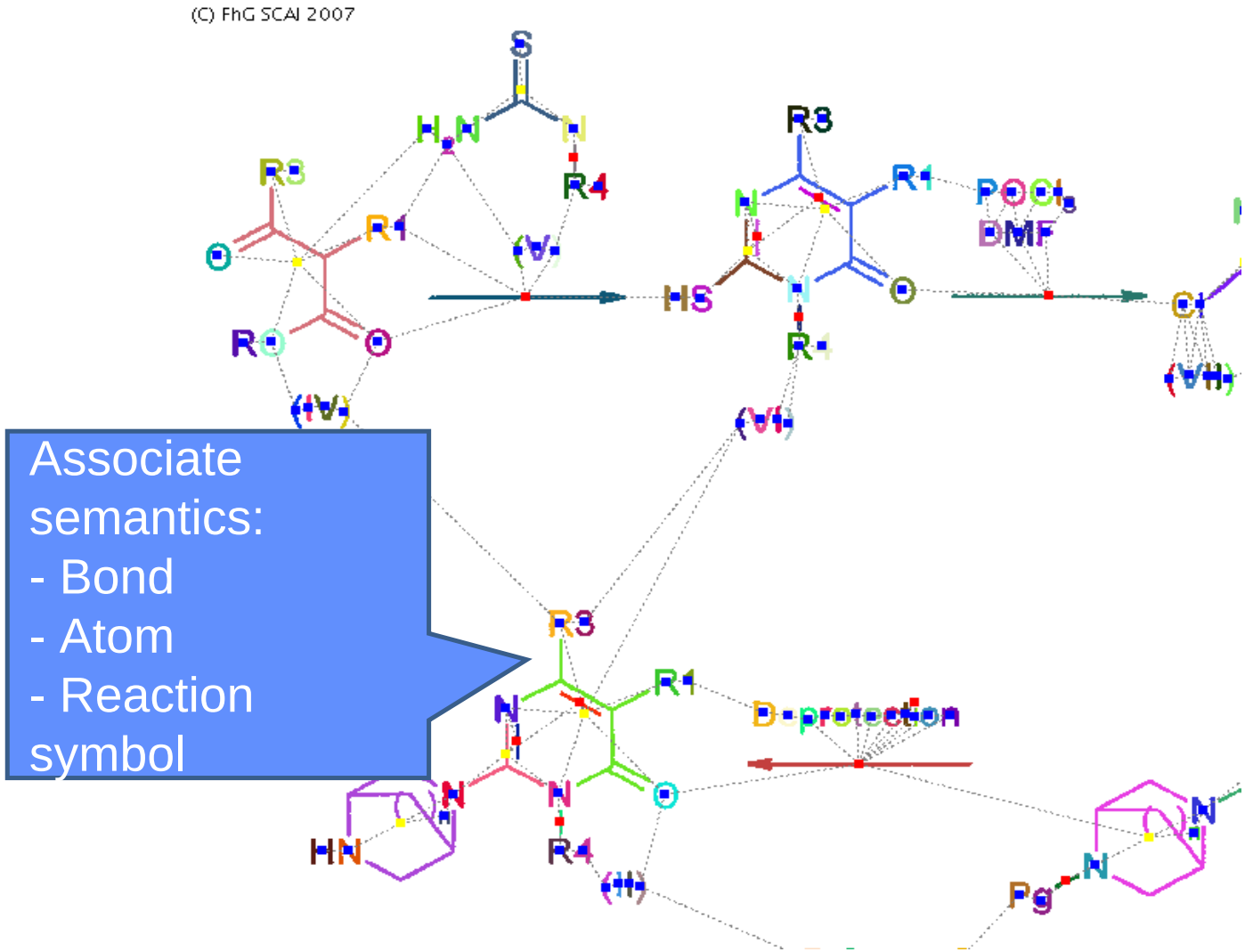
1: Connected Components

2: Tag Text

3: OCR

4: Vectorizer

From Picture to Reaction: Applying Chemical Knowledge



0: Picture

1: Connected Components

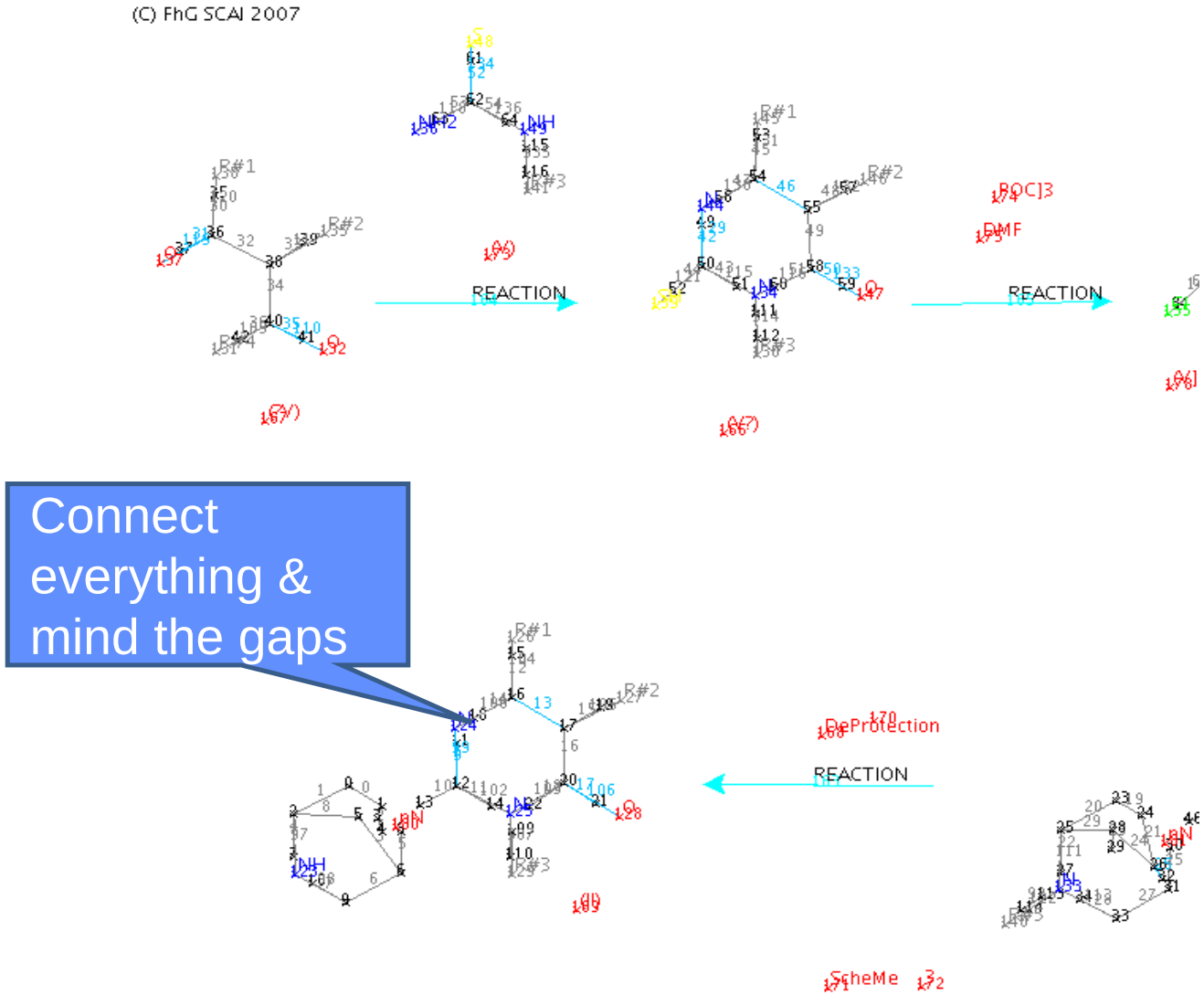
2: Tag Text

3: OCR

4: Vectorizer

5: Expert System

From Picture to Reaction: Interpretation as Chemical Graphs



0: Picture

1: Connected Components

2: Tag Text

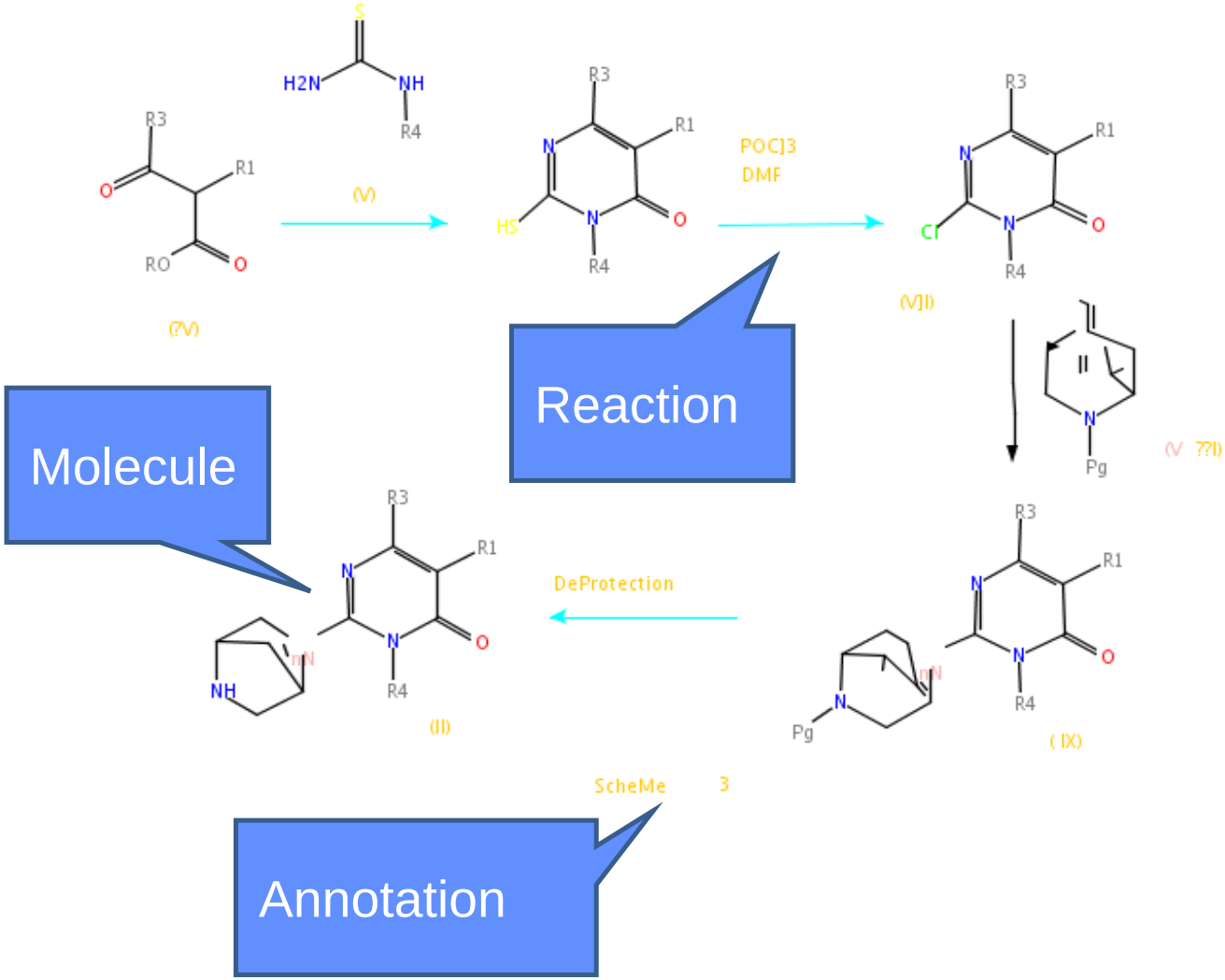
3: OCR

4: Vectorizer

5: Expert System

6: Chemical Graph

From Picture to Reaction: Representation Format



0: Picture

1: Connected Components

2: Tag Text

3: OCR

4: Vectorizer

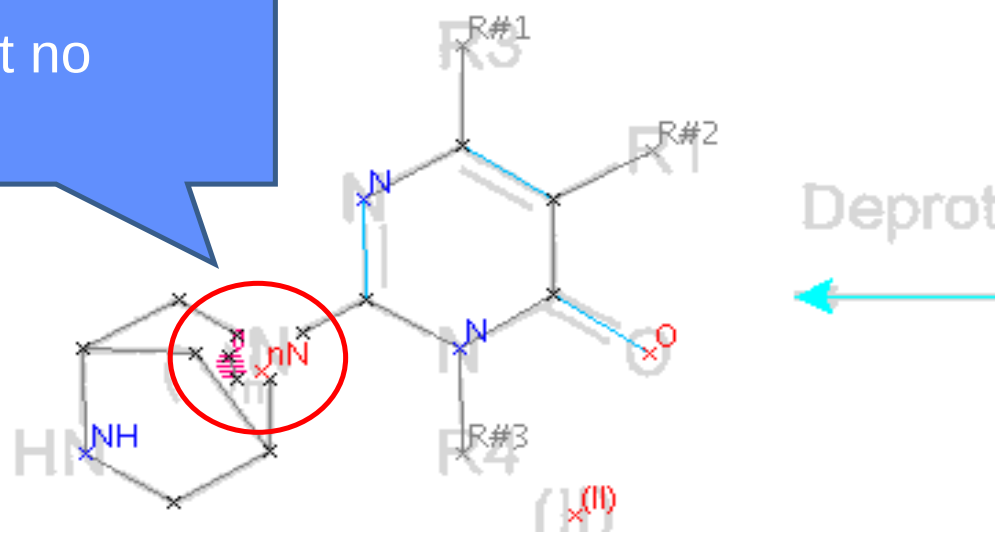
5: Expert System

6: Chemical Graph

7: Molecule

From Picture to Reaction: Error / Problem Detection

Sorry, but no Markush



List of errors

ID	Color	Object type	Object ID	Error type	Specification	Description
0		Caption	15	UNKNOWNATOM...	ERROR	nN is missing in l...
1		Caption	16	UNKNOWNATOM...	ERROR	nN is missing in l...
2		Caption	17	UNKNOWNATOM...	ERROR	(V is missing in l...
3		ChemicalBond	2	VECTOR_ERROR	PROBLEMATIC	Lonely bond (bo...
4		ChemicalBond	66	VECTOR_ERROR	PROBLEMATIC	Lonely bond (bo...
5		ChemicalBond	74	VECTOR_ERROR	PROBLEMATIC	Lonely bond (bo...

Close

0: Picture

1: Connected Components

2: Tag Text

3: OCR

4: Vectorizer

5: Expert System

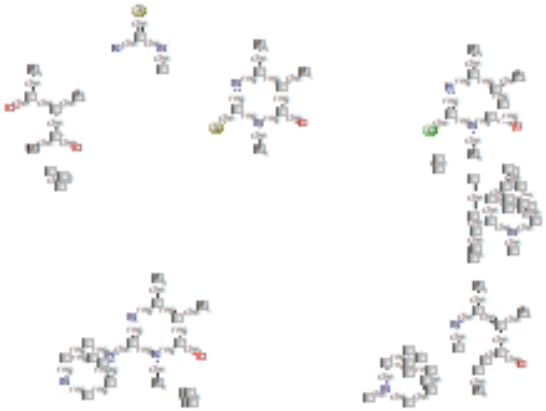
6: Chemical Graph

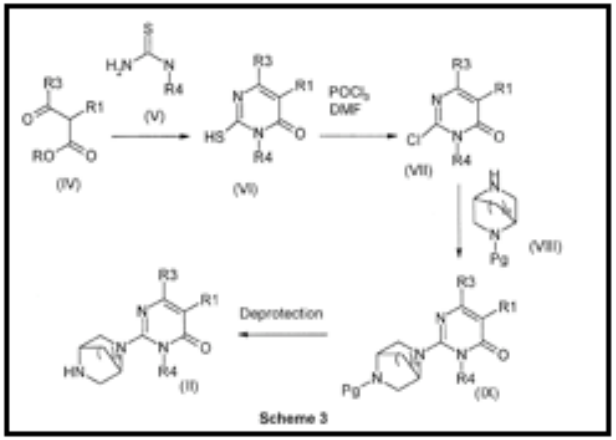
7: Molecule

8: Validation

Reaction Schema Reconstructed by ChemoCR: Embedding the Resulting SDF in Patent Document View

The compound of formula (II) may be prepared according to the method defined in scheme 3.





(In the above scheme the definition of R1, R3, R4 and n are the same as already described for compound of formula (I)).

Fraunhofer Institute Algorithms and Scientific Computing

Potential Knowledge Gain Through ChemoCR Analysis

One of the key questions associated with multi-modal patent mining is: do we gain from being able to simultaneously analyze text and chemical structure depictions?

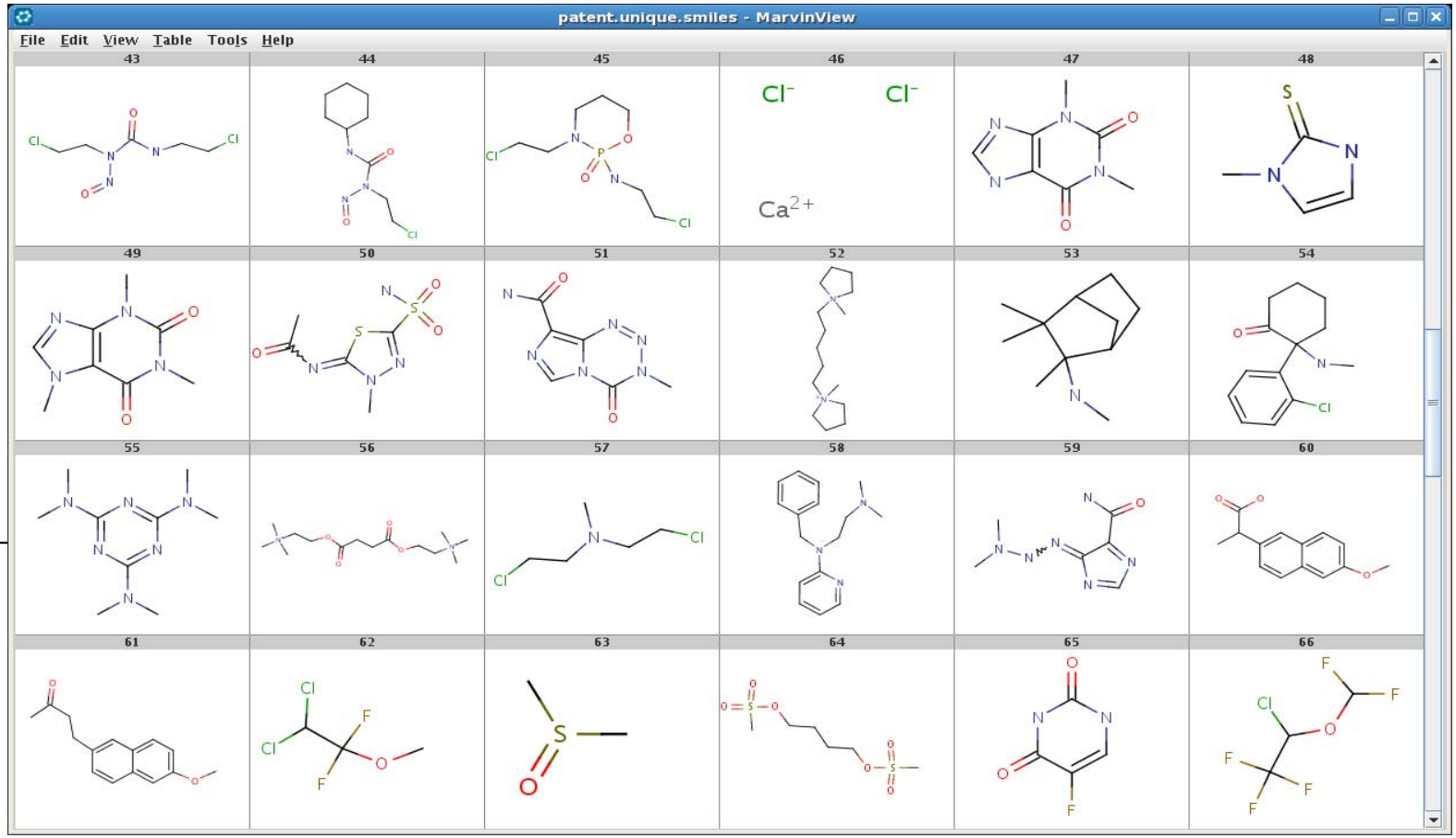
What is the “gain of knowledge” if we combine text analysis and image analysis?

Molecules can be Found in Text

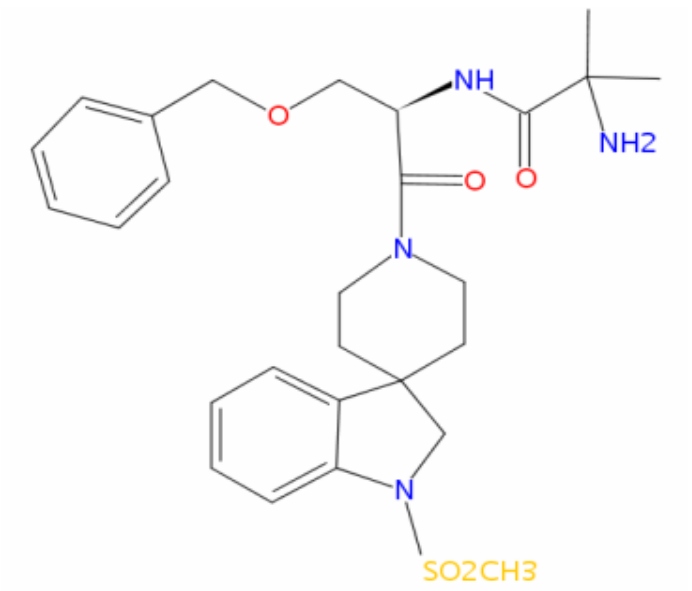
560 different molecules (fragments) identified in text

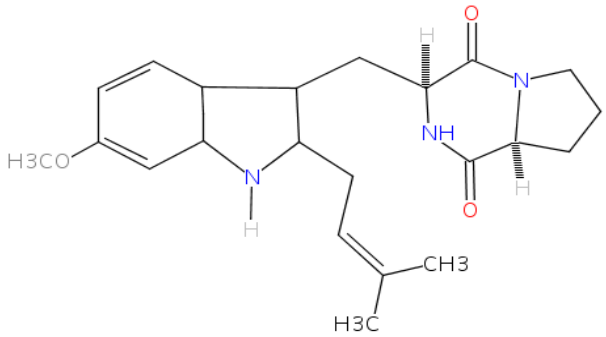
Mapping to PubChem via dictionary (name to InChI)

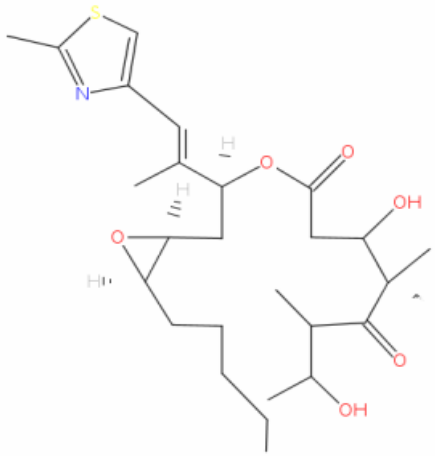
Mostly known structures

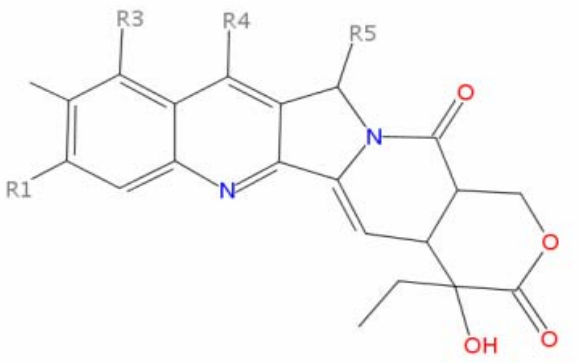


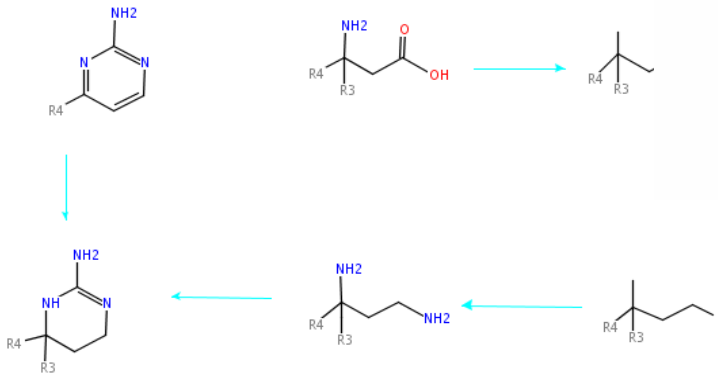
Structure Depictions Frequently Contain *Novel* Structures











Benchmarking activities in the biomedical arena & Call for a joint initiative for benchmarking of information extraction technology in the field of chemistry

SCAI
Fraunhofer Institute
Algorithms and
Scientific Computing

Lessons to be learned from the Biomedical Community

In the biomedical community benchmarking activities such as **CASP** (Critical Assessment of Techniques for Protein Structure Prediction) or **BioCreative** (Critical Assessment of Text Mining in Biology) have helped to assess the quality of current technology developments.

However, the development of the testing scenarios and the organisation of such critical assessments is a time consuming task (ask Lynette Hirschman).

Need for Benchmarking / Evaluations in Chemistry

Taken from: Corbett, Batchelor and Teufel ***Annotation of Named Chemical Entities*** BioNLP 2007: Biological, translational, and clinical language processing, pages 57-64, Prague, June 2007

A few chemical named entity recognition (Corbett and Murray-Rust, 2006; Townsend et al., 2005; Vasserman, 2004; Kemp and Lynch, 1998; Sun et al., 2007) or classification (Wilbur et al., 1999) systems have been published. A plugin for the GATE system[3] will also recognise a limited range of chemical entities. Other named entity recognition or classification systems (Narayanaswamy et al., 2003; Torii et al., 2004; Torii and Vijay-Shanker, 2002; Spasic and Ananiadou, 2004) sometimes include chemicals as well as genes, proteins and other biological entities. However, due to differences in corpora and the scope of the task, it is difficult to compare them. There has been no chemical equivalent of the JNLPBA (Kim et al., 2004) or BioCreAtIvE (Yeh et al., 2005) evaluations. Therefore, a corpus and a task definition are required.

A Consideration ...

BioCreative will be organised by Lynette Hirschman and her colleagues (Rolf Appweiler; Alfonso Valencia and others) again in 2008. At the 5th Symposium on Text Mining in the Life Sciences in September 2007 in Bonn, Lynette Hirschman gave a very impressive talk on the results of the critical assessment of text mining in biomedicine BioCreative 2004 and 2006.

In her talk, she called for suggestions from the community for new scenarios that could be worked on in the course of BioCreative 2008.

.... and a Suggestion

I suggest that we ask Lynette to open a parallel track in BioCreative, which focuses on information extraction in chemistry. Such ChemCreative should be aligned with BioCreative, but it needs significant input from the chemistry community.

Fraunhofer SCAI offers full support for establishing ChemCreative and we offer to team up with public (academic) professionals in the chemical information management arena to develop the appropriate scenarios and evaluation criteria.

Finally: *did you ever come to the point where you felt it would be just too good to have access to publicly available, well annotated corpora of chemical literature ?*

SCAI
Fraunhofer Institute
Algorithms and
Scientific Computing

Acknowledgement

In alphabetical order:

Holger Dach
Juliane Fluck
Christoph Friedrich
Tobias Gattermayer
Tobias Goecke
Carina Haupt
Roman Klinger
Corinna Kolarik
Peter Kral
Theo Mevissen
Bernd Müller
Chia-Hao Ou
A. Weihermüller
Marc Zimmermann