
Information Extraction from Patents: Combining Text- and Image-Mining



Fraunhofer Institute
Algorithms and
Scientific Computing

Martin Hofmann-Apitius

Bonn-Aachen International Centre for Information Technology (B-IT)
September 25, 2007

Status Report: Major Achievements at SCAI in 2006 / 07

ProMiner – Dictionary - based Named Entity Recognition

1. Successful participation in BioCreative 2006; rank 3 (F-score **0,799**) out of 21 participating systems in task gene protein name recognition and normalization [rank 2: F-score **0,804**; rank 1: F-score **0,810**]
2. Extension of ProMiner functionalities beyond biological entities: integration of chemical and medical dictionaries; extension towards integration of ontology-derived dictionaries (GO; ANEURIST disease-ontology)
3. ProMiner installed as UIMA service in testbed installation

Status Report: Major Achievements at SCAI in 2006 / 07

SCAI – Machine Learning - based Named Entity Recognition

1. Successful participation in BioCreative 2006; rank 4 (F-score **0,8633**) out of 21 participating systems in *gene mention* task [rank 1: F-score **0,8721**; rank 2: F-score **0,8683**; rank 3: F-score **0,8657**]
2. Extension of ProMiner functionalities towards entity-types that cannot easily be represented in dictionaries (e.g. certain SNP-information; IUPAC names)
3. IUPAC and SNP identification with conditional random field approach available as ProMiner add-ons.

Status Report: Major Achievements at SCAI in 2006 / 07

@neuLink

Development of a Persistence-Layer for Extracted Information in EU-Projekt
@neurIST (IST-027703)

1. Oracle-based system storing the results of ProMiner runs and integrating reference databases (UniProt; EntrezGene; dbSNP; PubChem) with textual objects
2. Used in EU-Project @neurIST for aggregation of disease-related information, including SNP-Information on the genes identified by ProMiner (in collaboration with the IMIM Institute in Barcelona)
3. Tagging of gene & protein names, drug trivial names, IUPAC names, disease terminology, chromosomal position information, SNP information and OMIM references in the entire MEDLINE

Seite 4



rupture



- + Genes / Proteins
- Chromosomal Location
- STS Marker
- mSNP
- Normalized SNP
- Drug Names
- OMIM Reference
- Disease Terminology
- Particular in TextMining Context
 - Aneurysm by Etiology
 - Sign
 - Associated Gene or Protein
 - Aneurysm Therapy
 - Drug Therapy
 - Supportive Care
 - Minimal Invasive Therapy
 - Surgical Therapy
 - + Grading System
 - Epidemiologic Quality
 - Intracranial Aneurysm Associ
 - Aneurysm by Shape
 - Intracranial Giant Aneurys
 - Intracranial Berry Aneurys

Entity

(when the query is ready)

as the Document Base.
using Full Text Search

the class + Grading System must be
ent
the class + Grading System must be
ent
presses of run Disease Terms and class
/stem in Entity View.

Help

- Reset Search
- Show (this) Information Screen
- Start Search
- Filter Results
- Expand / Collapse Tree Viewing
- Show these in results

Steps to pose a Query (Use Firefox)

1. Enter a Full Text Search into the grey field located below the icons on the top-left. (click on the grey field to access standard searches)
 - > Select 'kinase'
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.
 - > Click ONCE on Genes/Proteins to include it (green and plus)
3. Tick ONE Checkbox in the tree left to the item to choose this entity for display in the entity summary screen, then click





Drugs Associated with Aneurysm Rupture



Documents

Entity

The following entities relating to 'rupture' were found in 11414 documents.



Export options: CSV | Excel | XML

- +rupture
- Genes / Proteins
- Chromosomal Location
- STS Marker
- mSNP
- + Drug Names**
- OMIM Reference
- Disease Terminology
 - Particular in TextMining Context
 - Risk Factor

Entity in Doc View	Relative Entropy	Entity Count	Links
Oxytocin	0.2344	441	D
Phenobarbitone	0.0614	213	D
Aspirin	0.0546	223	D
Collagenase	0.0467	128	D
Dinoprostone	0.0441	188	D
Ampicillin	0.0428	118	D
Betamethasone	0.0328	65	D
Fluorescein	0.0318	110	D
Sargramostim	0.0237	196	D
Choriogonadotropin alfa	0.0234	61	D
Streptokinase	0.0203	57	D
Misoprostol	0.0183	50	D
Deferoxamine	0.0169	52	D
Albendazole	0.0167	38	D
Methotrexate	0.0161	102	D
Enthecavirin	0.0142	30	D



Documents

Entity

+ "breast cancer" OR + "Breast can

The following entities relating to ' "breast cancer" OR "Breast cancer"' were found in 27853 documents.

- + Gene/Protein
-
-
- mSNP
- Normalized SNP
- Drug Names
- OMIM Reference
- Disease Terminology
- Particular in TextMining Context
- Risk Factor



Export options: CSV | Excel | XML

Entity in Doc View	Relative Entropy	Entity Count	
ESR1	0.3409	4652	
ERBB2	0.1641	2445	
BRCA1	0.097	1329	
BRCA2	0.0591	754	
CYP19A1	0.0322	606	
ABCG2	0.0321	410	
PGR	0.027	304	
TP53	0.0262	1332	
CSF3	0.0213	625	
MKI67	0.0213	472	
VEGF	0.0169	670	
SHBG	0.0168	386	
MUC1	0.0161	334	



@neuLink „Document View“

Documents

Entity

Result for "breast cancer" OR "Breast cancer", NER run 'FilteredMedAnalHomo_sapiens.syn.prt' for entity **ABCG2**, Page 0 > with 100 documents per page, totals to 310 and took 5329 ms.

All | Chromosomal Locations Drug Names Protein/Gene STS Marker Omim Reference Disease Terms rsSNP Normalized SNP

1. Cyclosporin A, tacrolimus and sirolimus are potent inhibitors of the human breast cancer resistance protein (ABCG2) and reverse resistance to mitoxantrone and topotecan.

PubMed 16404634 **Authors:** Anshul Gupta, Yang Dai, R Robert Vethanayagam, Mary F Hebert, Kenneth E Thummel, Jashvant D Unadkat, Douglas D Ross, Qingcheng Mao, **Date:** 2006-Sep-

PURPOSE: Several studies have demonstrated significant interactions between immunosuppressants (e.g., cyclosporin A) and chemotherapeutic drugs that are BCRP substrates (e.g., irinotecan), resulting in increased bioavailability and reduced clearance of these agents. One possible mechanism underlying this observation is that the immunosuppressants modulate the pharmacokinetics of these drugs by inhibiting BCRP. Therefore, the aim of this study was to determine whether the immunosuppressants cyclosporin A, tacrolimus and sirolimus are inhibitors and/or substrates of BCRP. **METHODS:** First, the effect of the immunosuppressants on BCRP efflux activity in BCRP-expressing HEK cells was measured by flow cytometry. **RESULTS:** Cyclosporin A, tacrolimus and sirolimus significantly inhibited BCRP-mediated efflux of pheophorbide A, mitoxantrone and BODIPY-prazosin. The EC(50) values of cyclosporin A, tacrolimus and sirolimus for inhibition of BCRP-mediated pheophorbide A efflux were 4.3 +/- 1.9 microM, 3.6 +/- 1.8 microM and 1.9 +/- 0.4 microM, respectively. Cyclosporin A, tacrolimus and sirolimus also effectively reversed resistance of HEK cells to topotecan and mitoxantrone conferred by BCRP. When direct efflux of cyclosporin A, tacrolimus and sirolimus was measured, these compounds were found not to be transported by BCRP. Consistent with this finding, BCRP did not confer resistance to the immunosuppressants in HEK cells. **CONCLUSION:** These results indicate that cyclosporin A, tacrolimus and sirolimus are effective inhibitors but not substrates of BCRP. These findings could explain the altered pharmacokinetics of BCRP substrate drugs when co-administered with the immunosuppressants and suggest that pharmacokinetic modulation by the immunosuppressants may improve the therapeutic outcome of these drugs.

2. Breast cancer resistance protein (BCRP/ABCG2) is expressed by progenitor cells/reactive ductules and hepatocytes and its expression pattern is influenced by disease etiology and species type: possible functional consequences.

PubMed 16709727 **Authors:** Sara Vander Borgh, Louis Libbrecht, Aezam Katoonizadeh, Jos van Pelt, David Cassiman, Frederik Nevens, Alfons Van Lommel, Bryon E Petersen, Johan Fevery, Peter L Jansen, Tania A Roskams, **Date:** 2006-Sep-

Status Report: Major Achievements at SCAI in 2006 / 07

ChemoCR – reconstruction of chemical information from structure depictions

1. Partnership between InfoChem and SCAI on further development of ChemoCR
2. First large scale test runs provide basis for classification of types of depictions
3. First steps towards multi-modal information extraction: using ProMiner and ChemoCR simultaneously on patent literature

Seite 9

Beyond Analysis of MEDLINE: Information Extraction from Patents

Current challenges in text mining (at least at SCAI):

1. Information extraction from large corpora of full text documents
2. Combination of text analysis and analysis of chemical structure depictions
3. Normalization (mapping of text objects to database reference entries) issues: ProMiner normalized so far on UniProt and EntrezGene; mapping of SNPs to reference databases and mapping of chemical names and structures to chemical reference structures are completely new challenges

Seite 10

Named Entity Recognition in Full-Text Patents

@neuLink

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

Use Document Base: **Medline** | Documents | Entity | All

X 10K | Get Entities >

Enter full text search phrase

Show Patent

All | Anatomy | Disease | DrugNames | ProteinGene | IUPAC

Valproic acid and derivatives thereof as histone deacetylase inhibitors

Date: 2002-01-09
File Identifier: EP00114088A1
EP Identifier: 1170008

Statistics

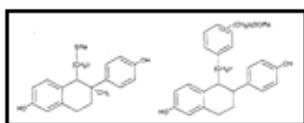
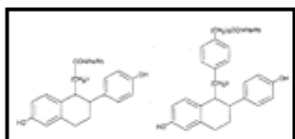
	Anatomy	Disease	Drug	Protein/Gene	IUPAC
Absolute	12	37	19	59	0
Top rank		'breast cancer' ' <u>teratocarcinoma</u> '	' <u>Valproic Acid</u> '	'PPARD' 'NCOR1' 'TFAP2A'	

Chemical Entities Related to Target Estrogen Receptor

Said compound also has the agonistic effect. It is therefore required to develop an anti-estrogenic compound which is substantially or completely free of agonistic effect and which can effectively block the **estrogen receptor**.

In addition, it has been known that 7 \pm -substituted derivatives of estradiol, for example, 7 \pm -(CH₂)₁₀CONMeBu derivatives, are steroidal anti-estrogenic agents without agonistic effect (see, EP-A 0138504, **USP 4,659,516**). Further, an estradiol derivative having a 7 \pm -(CH₂)₉SOC₅H₆F₅ substituent has also been disclosed (see, Wakeling et al., **Cancer Res.**, 1991, 51, 3867).

Non-steroidal anti-estrogenic agents without agonistic effect have been first reported by Wakeling et al. in 1987 (see, A. Wakeling and Bowler, J. Endocrinol., 1987, 112, R7). Meanwhile, U.S. Patent No. 4,904,661 discloses phenol derivatives having anti-estrogenic activity. These phenol derivatives generally have a naphthalene scaffold and include, typically, the following compounds:



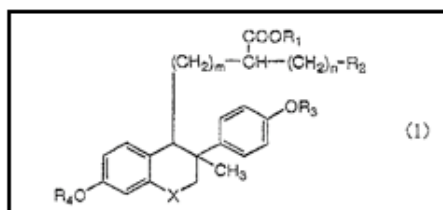
Some chroman and thiochroman derivatives have been reported as anti-estrogenic compounds having no agonistic effect (WO 98/25916). Although the existing anti-estrogenic compounds having no agonistic effect show a substantial therapeutic effect when administered via intravenous or subcutaneous injection, they show a highly reduced therapeutic effect when administered orally, probably due to their low bioavailability by oral route, etc. Therefore, for convenience's sake in the case of administration, it is desired to develop anti-estrogenic compounds which show a sufficient effect when administered orally and at the same time have no agonistic effect.

DISCLOSURE OF THE INVENTION

The object of the present invention is to provide chroman or thiochroman derivatives which have anti-estrogenic activity and are advantageous in pharmaceutical use.

The present inventors researched anti-estrogenic activity of compounds having various structures. As a result, we have found that chroman or thiochroman derivatives of general formula (1) could show a good anti-estrogenic activity in substantial absence of agonistic effect and that they provided a sufficiently high activity even when administered orally. The present invention has been accomplished on the basis of this finding.

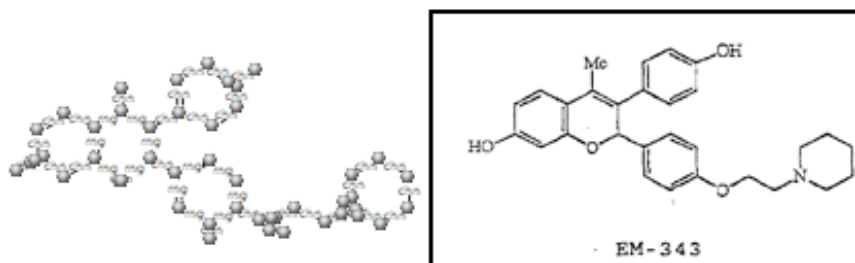
Namely, the present invention provides a compound having the following general formula (1):



Multimodal Tagging: Embedding ChemoCR Reconstruction

In treating **diseases** caused by abnormal **tissue** growth that is dependent upon a certain sexual steroidal hormone such as estrogen, it is highly important to significantly inhibit, more preferably completely eliminate, the effect induced by the hormone. For this purpose, it is desirable to reduce the level of hormone capable of acting on the steroidal hormone receptor site. For instance, anti-estrogenic agents are commonly administered for alternative or combination therapy to limit the production of estrogen to the amount less than required to activate the receptor site. However, such conventional technique for blocking estrogen production could not sufficiently inhibit the effect induced through the **estrogen receptor**. Practically, even when estrogen is completely absent, some of the receptors may be activated. It was therefore considered that estrogen antagonists could provide better therapeutic effect in comparison to the technique for blocking only the production of sexual steroidal hormone. Thus, numerous estrogen antagonists have been developed. For example, many patent publications including U.S. Patent Nos. 4,760,061, 4,732,912, 4,904,661, 5,395,842 and WO 96/22092 disclose various anti-estrogenic compounds. Sometimes, however, prior art antagonists may themselves act as agonists, and therefore activate rather than block the receptor. For example, **Tamoxifen** has been most widely used as an anti-estrogenic agent. However, this agent has a disadvantage that it exhibits estrogenic activity in some organs (see, M. Harper and A. Walpole, J. Reprod. Fertile., 1967, 13, 101).

As another non-steroidal anti-estrogenic compound, WO 93/10741 discloses a benzopyran derivative having an aminoethoxyphenyl substituent(s) (Endorecherche), the typical compound of which is EM-343 having the following structure:



Linkout to PubChem through InChI

TECHNICAL FIELD

External Link View

The present invention relates to chroman or thiochroman derivatives having anti-estrogenic activity.

BACKGROUND ART

In treating diseases caused by abnormal effects induced by a hormone, upon a certain sexual steroidal hormone such as estrogen, it is highly important to significantly inhibit, more than the effect induced by the hormone, the effect of this hormone. For this purpose, it is desirable to reduce the level of hormone capable of acting on the steroidal hormone receptor site. However, such conventional technique for blocking estrogen production could not sufficiently inhibit the effect induced through the estrogen receptor. Practically, even when estrogen is completely absent, some of the receptors may be activated. It was therefore considered that estrogen antagonists could provide better therapeutic effect in comparison with estrogen antagonists have been developed. For example, many patent publications including U.S. Patent Nos. 4,760,000 and WO 93/22052 disclose steroid anti-estrogenic compounds. Sometimes, however, prior art antagonists may themselves act as agonists, and the steroid anti-estrogenic compound has been most widely used as an anti-estrogenic agent. However, this agent has a disadvantage that it exhibits estrogenic activity in some organs (see, M. Harper and A. Waipole, J. Reprod. Fert., 1967, 13, 101).

Information on biological activities of small molecules

PubChem

HOME SEARCH SITE MAP PubMed Entrez Structure GenBank PubChem Help

Compound Text Search GO

Structure Search Basic Advanced

Search Clear Save Query

Search Input: SMILES/SMARTS, InChI, CID or Formula Structure File Saved Query

InChI=1/C26H29NO/c1-4-25(21-11-7-5-8-12-21)26(22-13-14-15-16-17-18-19-20)23-24

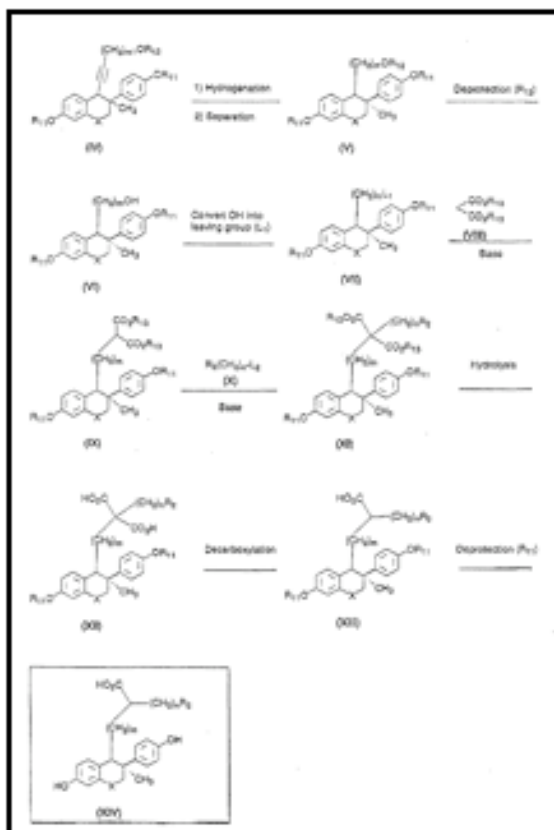
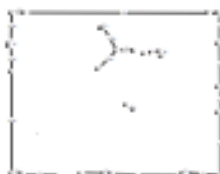
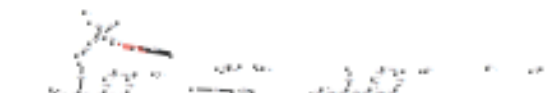
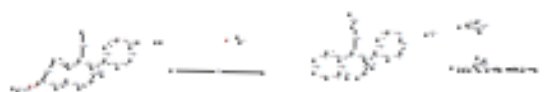
Sketch

Search Type:

Search using: Similar Compounds, score >= 90%

Embedded Reconstruction of Reaction Schemata

The compound of general formula (1) can be prepared according to any one of the following Reaction Schemes 1 to 10 (Processes 1 to 10).



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?

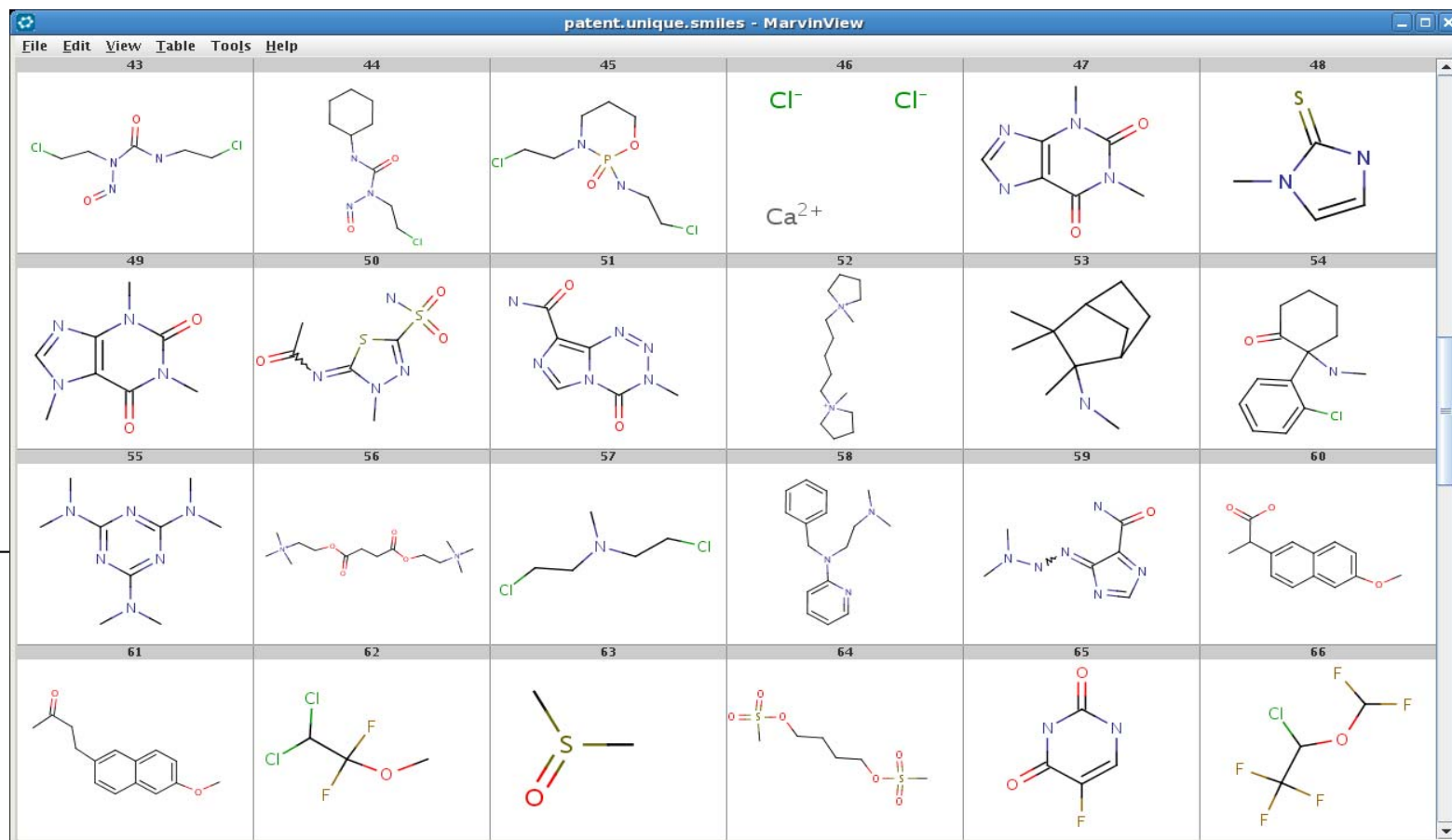
What is the “gain of knowledge” if we combine text analysis and image analysis?

Molecules can be Found in Text

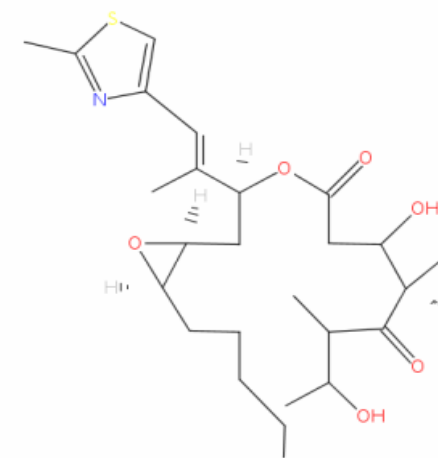
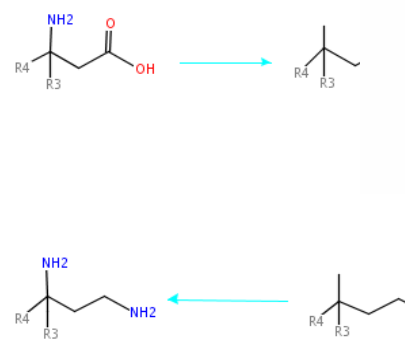
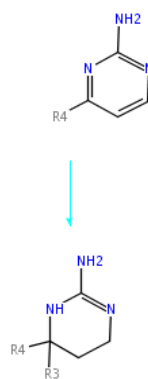
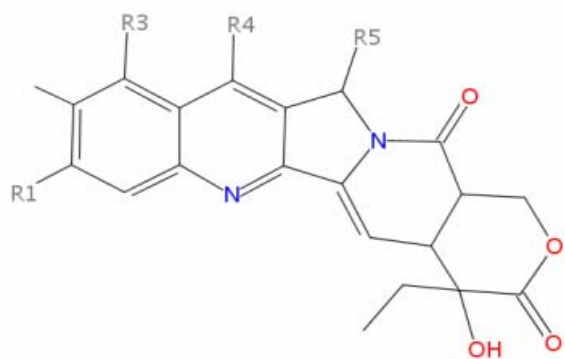
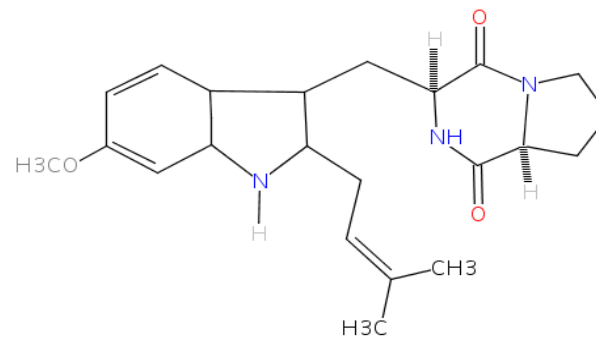
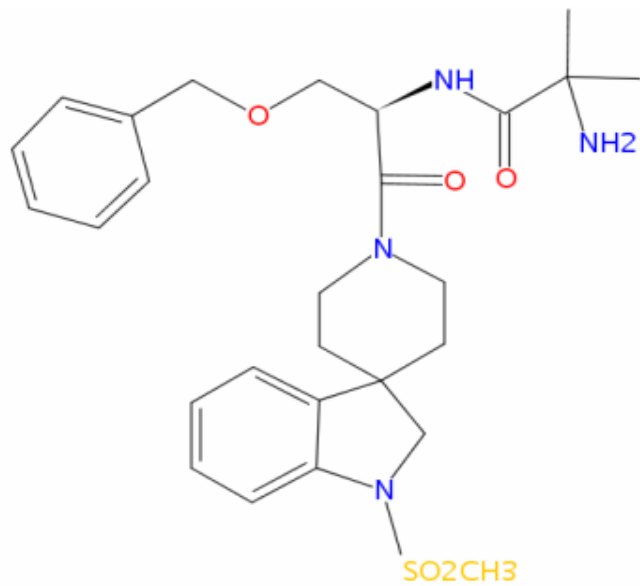
560 different molecules (fragments) identified in text

Mapping to PubChem via dictionary (name to InChI)

Mostly **known** structures



Structure Depictions Frequently Contain *Novel* Structures



Reconstruction of Synthesis Pathways from Patents: an Example



► 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** **10K**

Documents

Enter full text search phrase

- Anatomy
- Disease
- DrugNames
- ProteinGene
- IUPAC

Show Patent

All | Anatomy | Disease | DrugNames | ProteinGene | IUPAC

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



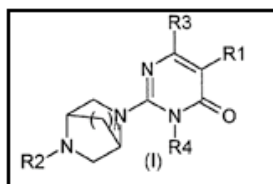
Date: 2004-09-08
File Identifier: EP03290571A1
EP Identifier: 1454908

Statistics

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

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

The invention relates to a 2-(diaz-bicyclo-alkyl)-pyrimidinone derivative represented by formula (I) or a salt thereof.



wherein:

- R1 represents a hydrogen atom, a C₁₋₆ 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 **GSK3 β** , 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 **GSK3 β** .

Background Art

GSK3 β (**glycogen synthase kinase 3 β**) 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 **GSK3 β** 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 tauopathies. Interestingly, protein **kinase B (AKT) phosphorylation** of **GSK3 β** 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 **GSK3 β** of **β -catenin**, a protein involved in cell survival, results in its degradation by an ubiquitination dependent **proteasome pathway**.

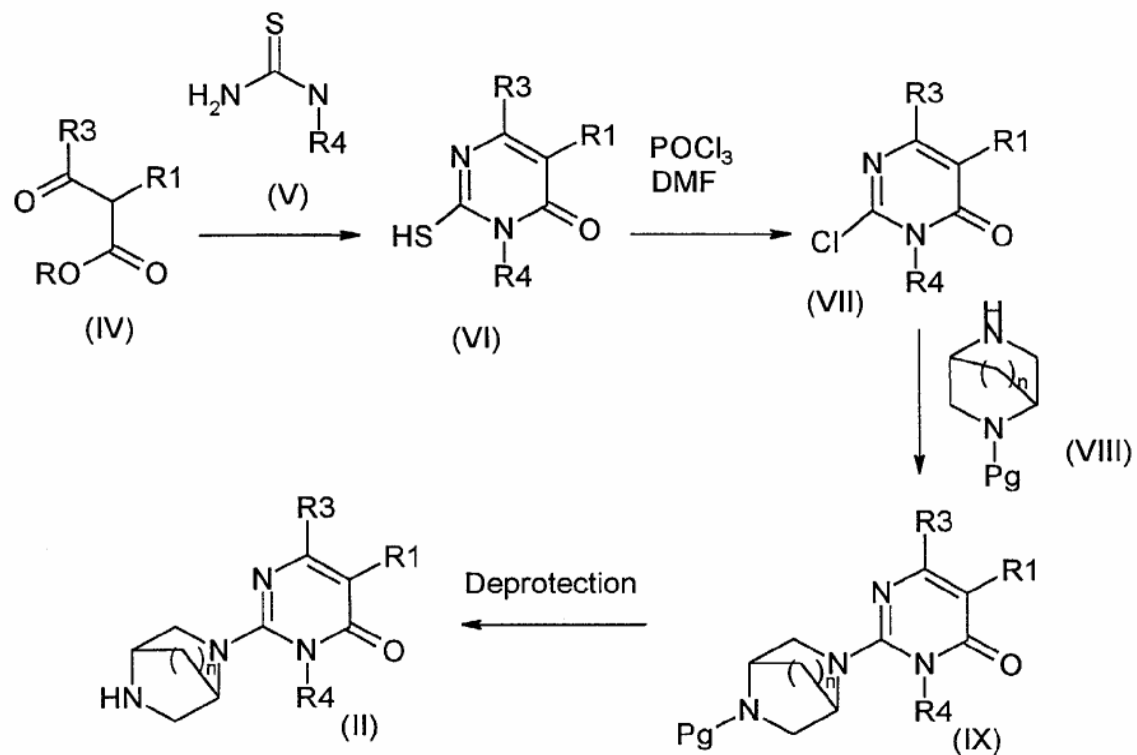
Thus, it appears that inhibition of **GSK3 β** activity may result in neurotrophic activity. Indeed there is evidence that lithium, an uncompetitive inhibitor of **GSK3 β** , enhances neurogenesis in some models and also increases neuronal survival, through the induction of survival factors such as **Bcl-2** and the inhibition of the expression of proapoptotic factors such as **P53** and **Bax**.

Recent studies have demonstrated that β -amyloid increases the **GSK3 β** 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 **GSK3 β** antisense mRNA. These observations strongly suggest that **GSK3 β** 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 **GSK3 β** 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 **GSK3 β** 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 **GSK3 β** 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

ChemoCR GUI

Input:
Bitmaps

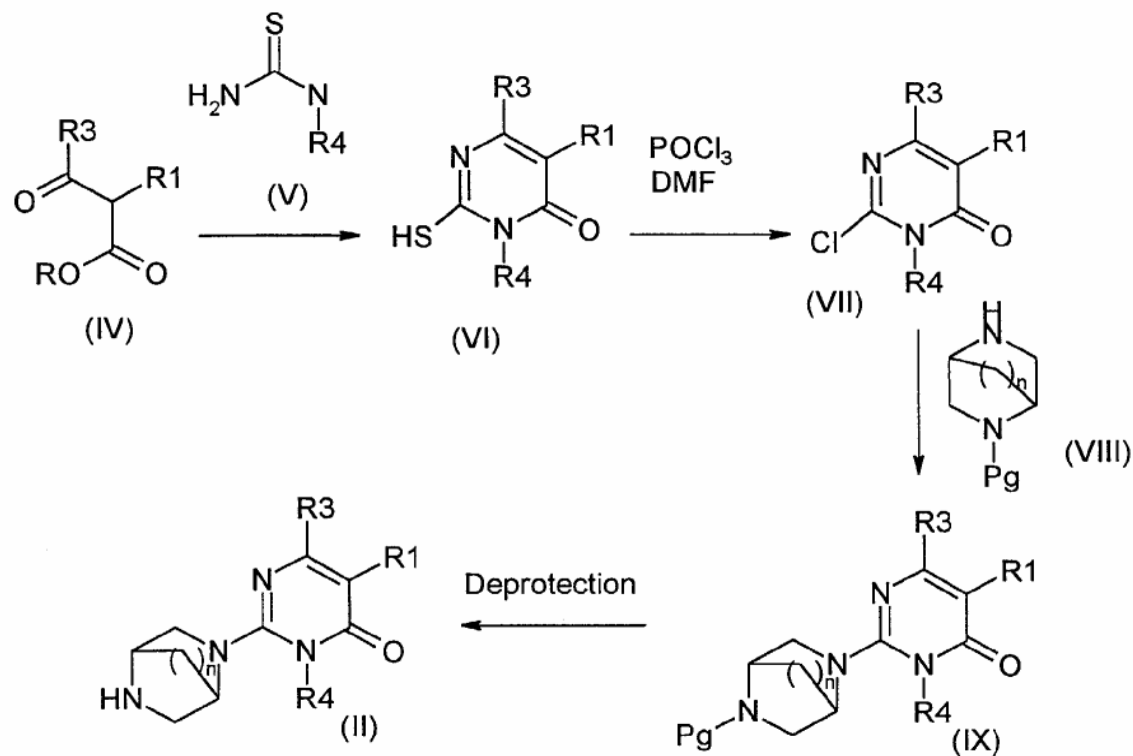
The screenshot displays the ChemoCR software interface. The main window shows a chemical reaction scheme labeled "Scheme 3". The scheme starts with a reactant (IV) reacting with a thioamide (V) to form a thioamide intermediate (VI). This intermediate then reacts with POCl₃ and DMF to form a chloroimide (VII). Finally, a deprotection step is shown, converting a protected intermediate into a final product (II).

On the right side of the interface, there is a table titled "Reconstructed Molecule (23)". The table has columns for "Id", "Molecule", "Name", and "Filename". It lists four fragments:

Id	Molecule	Name	Filename
1		fragment(1) of created from /home/marc	file:/home/marc/work...
2		fragment(2) of created from /home/marc	file:/home/marc/work...
3		fragment(3) of created from /home/marc	file:/home/marc/work...
4		fragment(4) of created from /home/marc	file:/home/marc/work...

Output:
Molecules

From Picture to Reaction: Pre-Processing



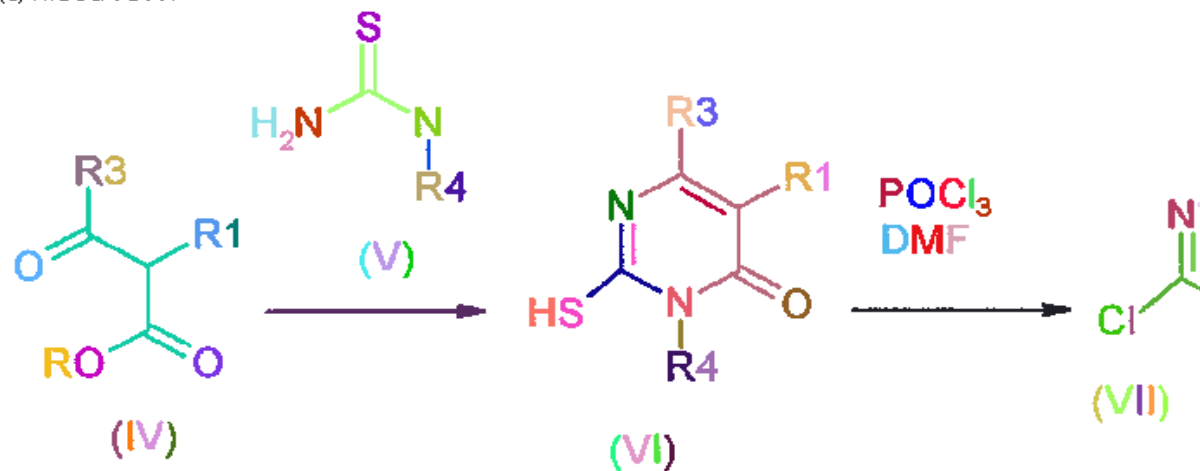
0: Picture

Scheme 3

BMP after scaling
& binarization

From Picture to Reaction: Identification of Connected Components

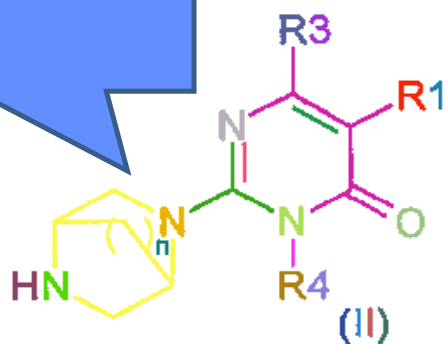
(C) FhG SCAI 2007



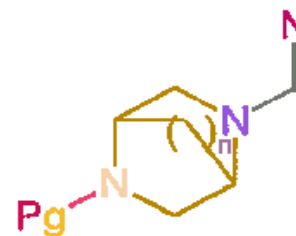
0: Picture

1: Connected Components

1 component
- Bondset
- Letter



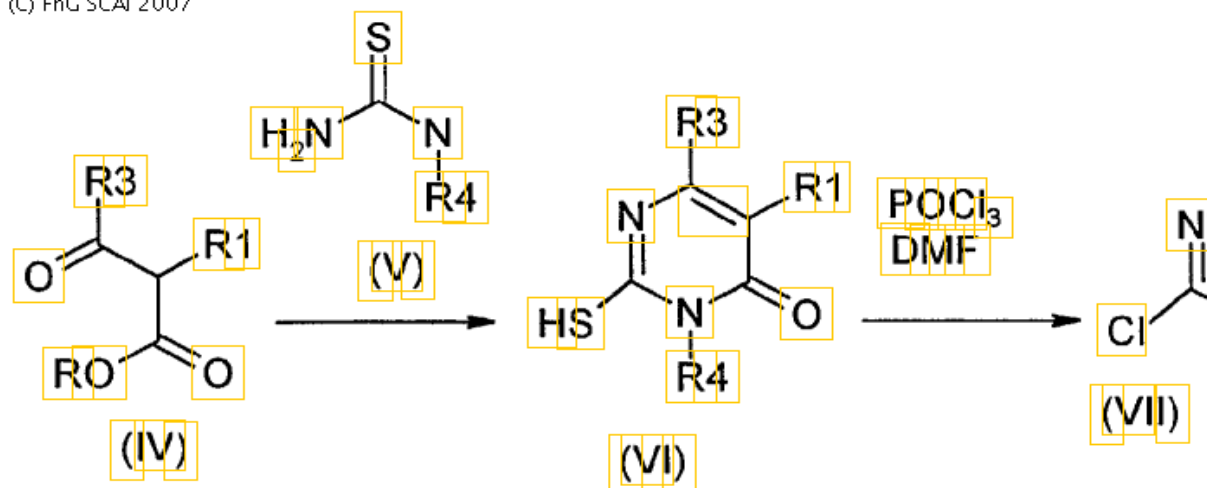
Deprotection



Scheme 3

From Picture to Reaction: Tagging of Characters

(C) FhG SCAI 2007

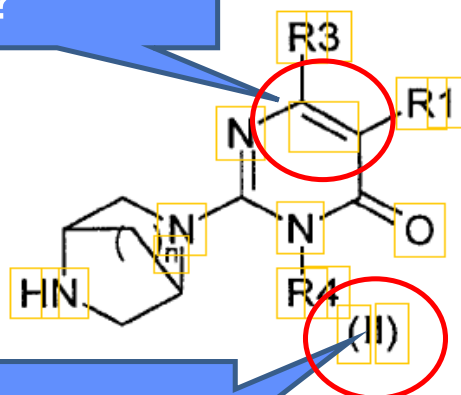


0: Picture

1: Connected Components

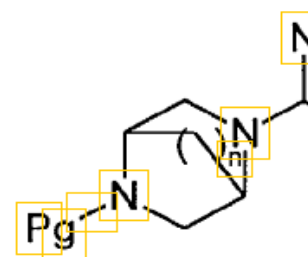
2: Tag Text

Text or
Bond?



Text or
Bond?

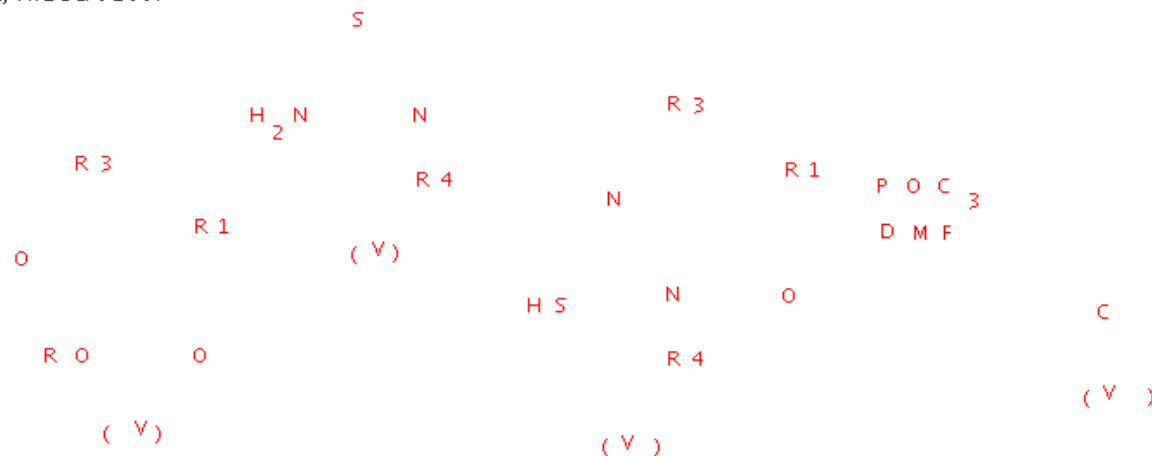
Deprotection



Scheme 3

From Picture to Reaction: OCR of Identified Characters

(C) FhG SCAI 2007



0: Picture

1: Connected Components

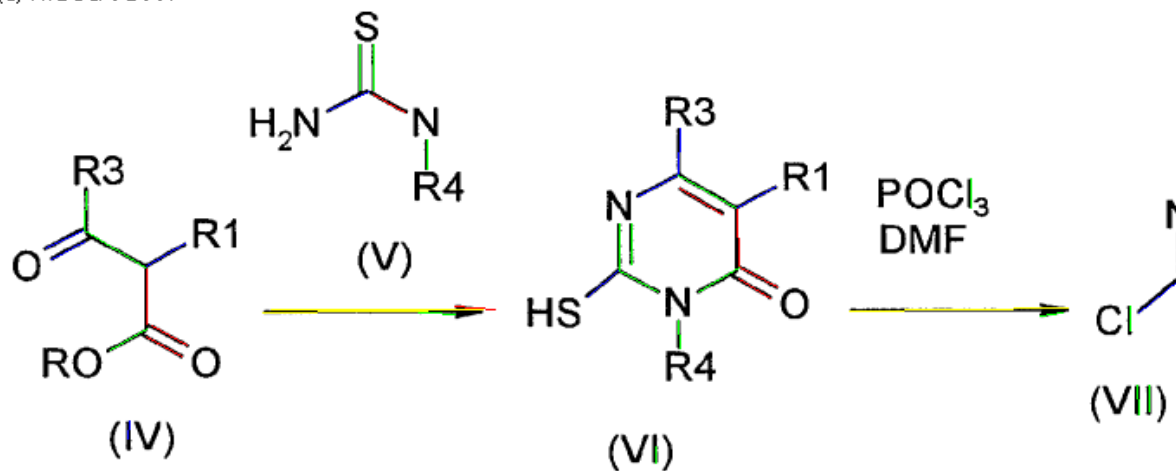
2: Tag Text

3: OCR



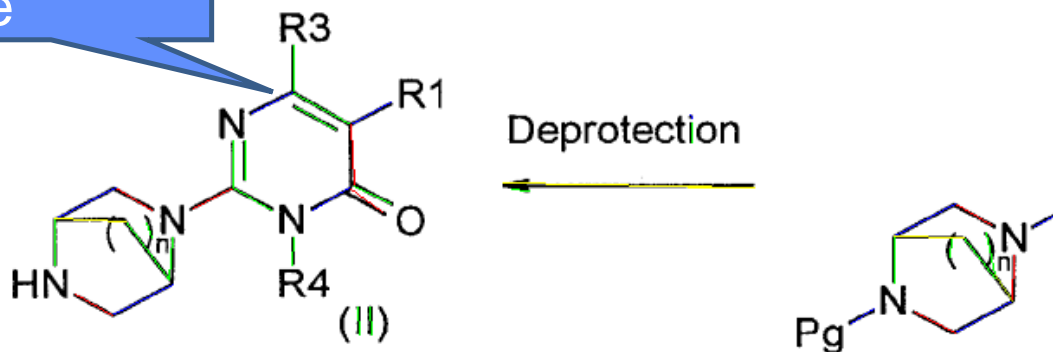
From Picture to Reaction: Identification of Lines (Bonds)

(C) FhG SCAI 2007



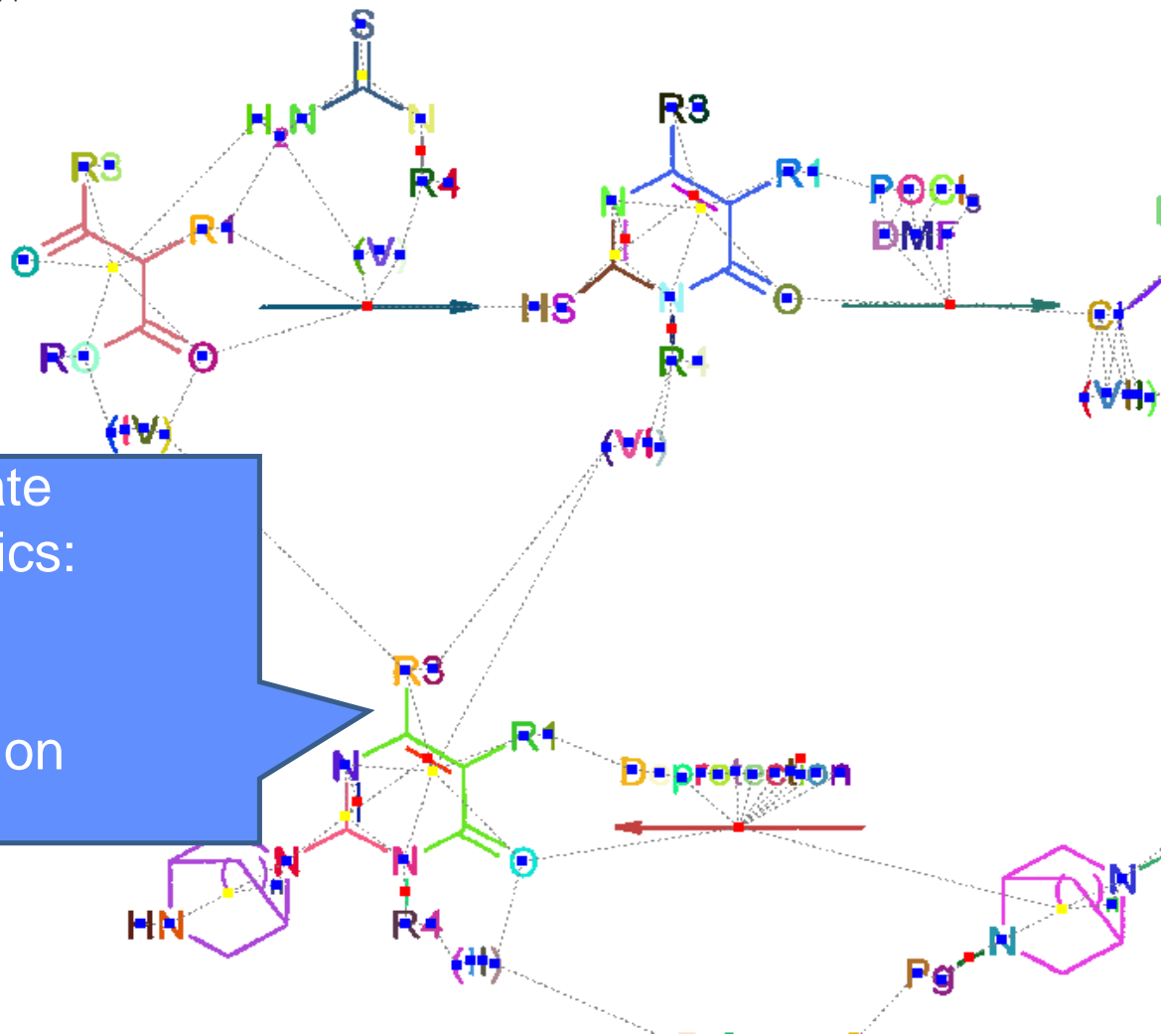
- 0: Picture
- 1: Connected Components
- 2: Tag Text
- 3: OCR
- 4: Vectorizer

Only the
bonds,
please



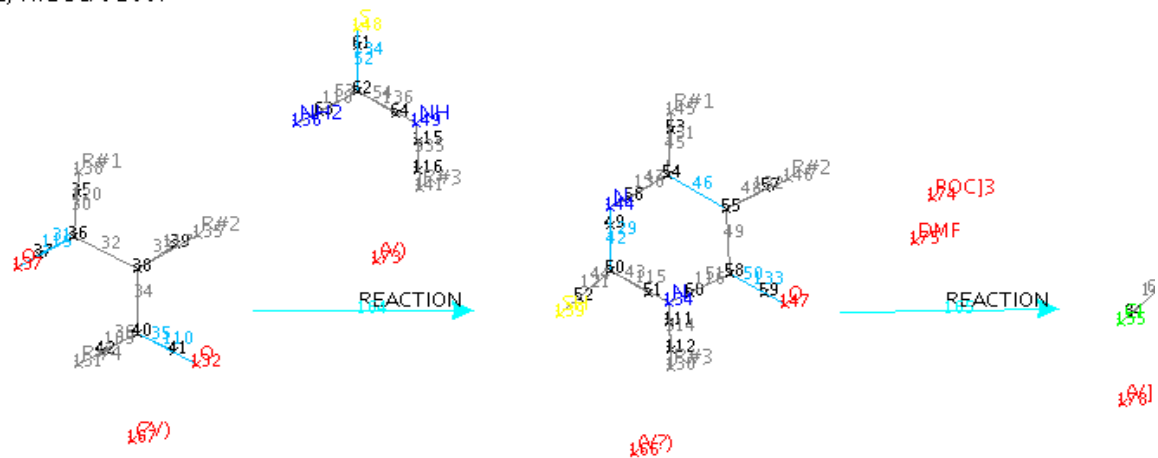
From Picture to Reaction: Applying Chemical Knowledge

(C) FhG SCAI 2007



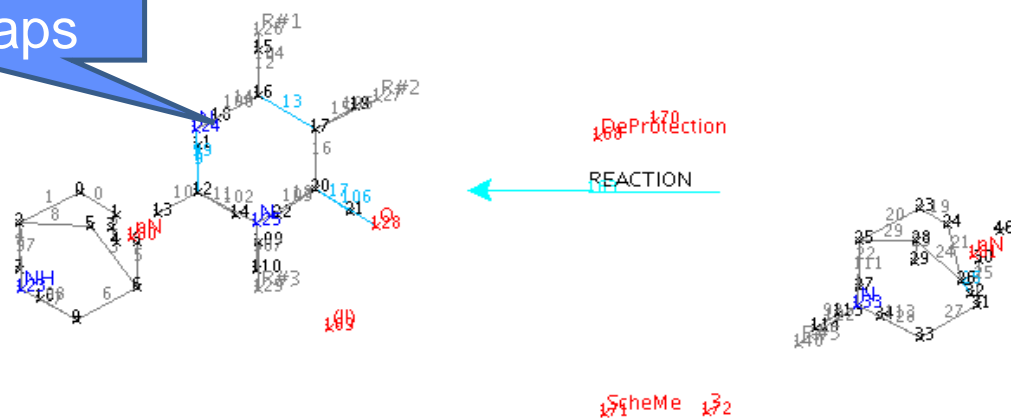
From Picture to Reaction: Interpretation as Chemical Graphs

(C) FHG SCAI 2007

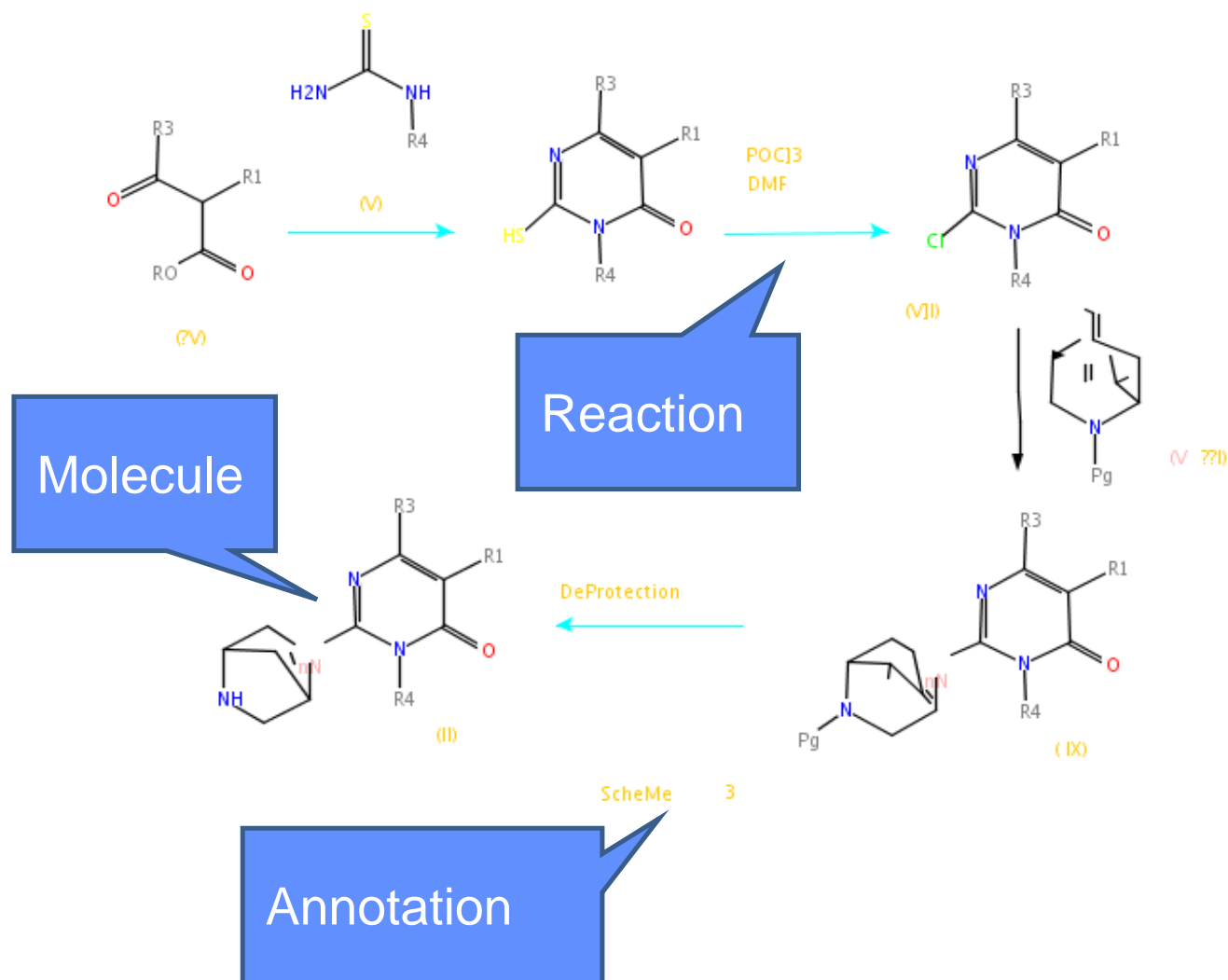


- 0: Picture
- 1: Connected Components
- 2: Tag Text
- 3: OCR
- 4: Vectorizer
- 5: Expert System
- 6: Chemical Graph

Connect everything & mind the gaps



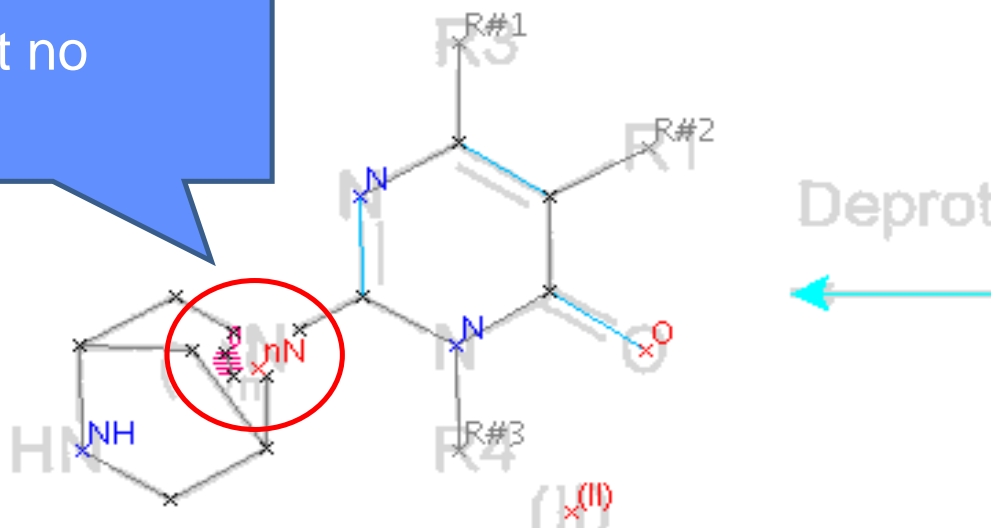
From Picture to Reaction: Representation Format



- 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



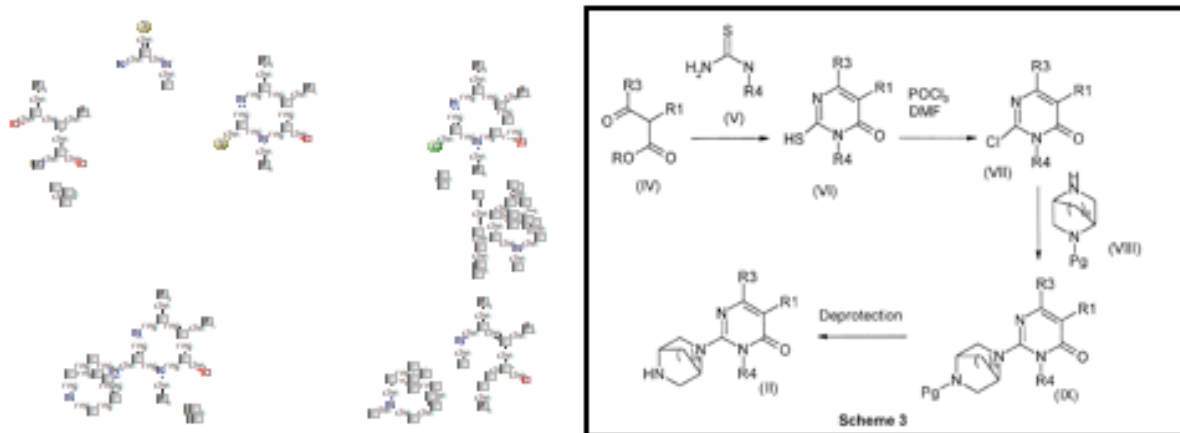
- 0: Picture
- 1: Connected Components
- 2: Tag Text
- 3: OCR
- 4: Vectorizer
- 5: Expert System
- 6: Chemical Graph
- 7: Molecule
- 8: Validation

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

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)).

Lessons Learned from First Encounter in Patent Mining

- Patents are the hardest data source for text miners
- Text and images comprise complementary information
- Chemical search thus has to be combined with text search
- Preprocessing needs a lot of attention (e.g. scaling, rotation, cleaning)
- Image categorization is required (identification of images containing chemistry versus images with mice, cells and different machines)
- Images in patents are sometimes a nightmare; quality ranges from bad to saumaessig, need for adaptive parameter optimization
- Extension for Markush and tables on the wish list; reconstruction of synthesis schemata works reasonably well

Seite 33

Next Steps on our Roadmap

1. Focus on pre-categorization of images (chemistry vs non-chemistry)
2. Generating rapid overview on entity-types in a patent and context („this patent talks about“)
3. Integration of dictionary-based, pattern-based and image-based information in the field of chemistry
4. Bridging from chemistry to biology and medicine: extraction of relationships between chemical and biomedical entities

Summary and Take Home Message

1. ProMiner and ChemoCR work nicely on full text documents
2. ProMiner and ChemoCR can be combined to efficiently annotate full text patent documents, enabling queries that were previously impossible
3. The persistence layer based on @neuLink allows us to store the results of mining approaches and integrates them with other knowledge bases
4. However, we just started to adjust our machinery to the challenge of analyzing patent literature. We already face significant challenges in the area of document segmentation and image categorization
5. But: besides being a scientific challenge, mining of patents is real fun and that is why we will report on our progress in this field next year

Acknowledgement

In alphabetical order:

Holger Dach

Juliane Fluck

Christoph Friedrich

Tobias Gattermayer

Tobias Goecke

Carina Haupt

Roman Klinger

Corinna Kolarik

Peter Kral

Theo Mevissen

Chia-Hao Ou

A. Weihermüller

Marc Zimmermann