ChemProspector: Advanced Mining and Searching of Chemical Content in Patent Documents

Josef Eiblmaier (InfoChem), Hans Kraut (InfoChem), Larisa Insenko (InfoChem), Heinz Saller (InfoChem), Peter Loew (InfoChem)

ICIC 24 – 27 October, Vienna

Supported by:

Federal Ministry of Economics and Technology

on the basis of a decision by the German Bundestag

Outline

- **Introduction**
  - ChemProspector, a THESEUS project
  - Markush in a nutshell
- **Major Goals and Approach**
- **First Results**
- **Outlook**

© cora / PIXELIO, www.pixelio.de

Supported by:
Federal Ministry of Economics and Technology
on the basis of a decision by the German Bundestag

THESEUS
Neue Technologien
für das Internet der Dienste

» “New Technologies for the Internet of Services”

» Research program initiated by the Federal Ministry of Economy and Technology (BMWi)

» Supported with approx. 100 million Euros

» Facilitate access to information, combine data to form new kinds of knowledge and lay the groundwork for new services on the Internet

» Duration: five years (2007 - 2011)

» Divided into six application scenarios combined through core technology cluster

» Phase one: development of core technologies (2007 - 2008)

» Phase two: THESEUS SME (2009 - 2011)

Supported by:
Federal Ministry of Economics and Technology
on the basis of a decision by the German Bundestag

ChemProspector: Basic data

» Main emphasis:

'The automatic extraction of Markush Structures from patent documents'

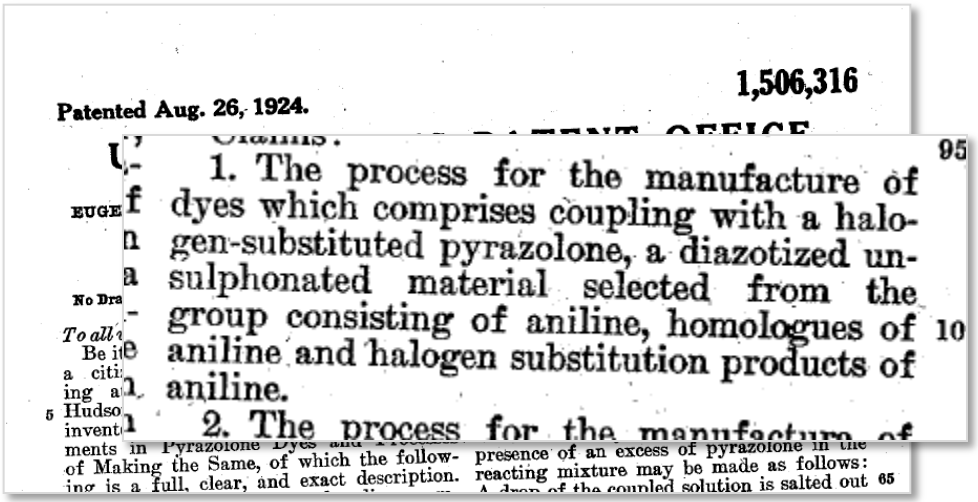
» Research SME-project within the THESEUS research program

» Application scenario ORDO ("Ordnung Digitaler Information")

» Duration: July 2009 – end of 2011

What is a ‘Markush Structure‘?

Quelle: http://www.colorantshistory.org



Patented Aug. 26, 1924. 1,506,316

1. The process for the manufacture of dyes which comprises coupling with a halogen-substituted pyrazolone, a diazotized unsulphonated material selected from the group consisting of aniline, homologues of aniline and halogen substitution products of aniline.

2. The process for the manufacture of

» Dr. Eugene A. Markush (1888-1968), Pharma Chemical Corporation (1917)

» USP No. 1,506,316 (1924), first usage of generic structures in a patent

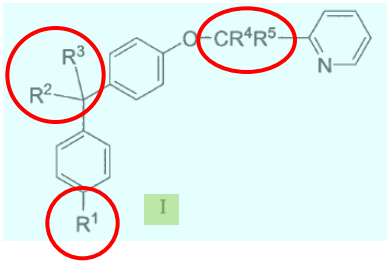
Supported by:

Federal Ministry of Economics and Technology

on the basis of a decision by the German Bundestag

DETAILED DESCRIPTION OF THE INVENTION

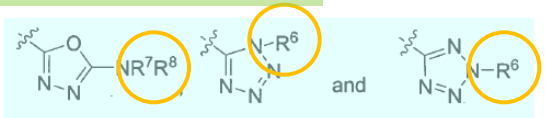
The instant invention provides a compound represented by structural formula I



I

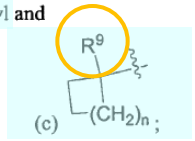
the pyridyl-N-oxide analog of formula I, and the pharmaceutically acceptable salts, esters and solvates thereof wherein:

R^1 is selected from the group consisting of:



and

R^2 is selected from the group consisting of (a) $-C_{1-6}$alkyl optionally substituted with 1-3 of fluoro, (b) $-C_{3-6}$ cycloalkyl and



(c)

n is an integer selected from 0, 1, 2 and 3;

R^3 is selected from the group consisting of –H, -F, -OH, $-CH_3$ and $-CF_3$;

R^4 is selected from the group consisting of –H and $-C_{1-4}$alkyl;

R^5 is selected from the group consisting of –H and $-CH_3$; and

R^6 is selected from the group consisting of –H, $-C_{1-6}$alkyl optionally substituted with 1-3 fluoro, $-C_{3-6}$ cycloalkyl optionally substituted with 1-3 fluoro and $-CH_2-R^{10}$;

R^7 is selected from the group consisting of –H, $-C_{1-6}$alkyl optionally substituted with 1-3 fluoro, $-C_{3-6}$ cycloalkyl optionally substituted with 1-3 fluoro, $-COC_{1-6}$alkyl and $-COC_{3-6}$ cycloalkyl;

Supported by:

Federal Ministry of Economics and Technology

on the basis of a decision by the German Bundestag

Markush: The challenge

» **The Information is contained in the text …**

» Substituents variation

» Homology variation

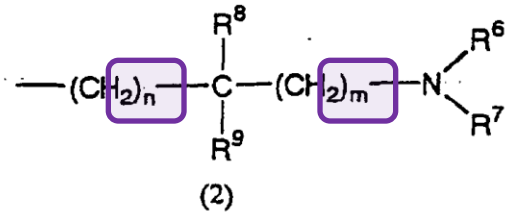
» Topology variation

» … **in the images ...**

» Position variation

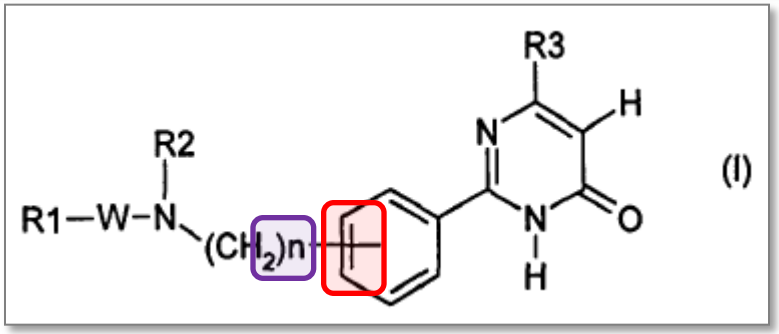
» Frequency variation

R^4 is hydrogen, hydroxy, cyano, fluorine or chlorine
q is 0 or 1;
R^5 is optionally substituted aryl;
W^1 is a group represented by formula (2):



(2)

wherein R^{11} represents a hydrogen atom or a methyl group and R^{12} represents a single bond or a linear or branched alkylene group having 1 to 5 carbon atoms.



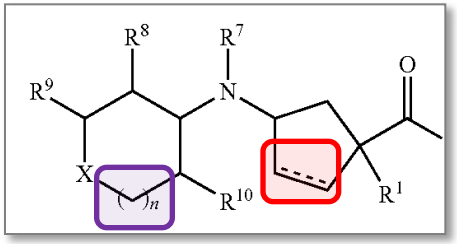
(I)

Supported by:
Federal Ministry of Economics and Technology
on the basis of a decision by the German Bundestag

» **... and both, text and images!**

» Frequency variation

» Bond variation



alkyl are unsubstituted or substituted with 1-3 substituents where the substiuents are independently selected from: halo, hydroxy, —CO_2R^{20}, C_{1-3}alkyl, and C_{1-3}alkoxy;

n is selected from 0, 1 and 2;

the dashed line represents a single or a double bond;

wherein n is an integer of 0 to 30,

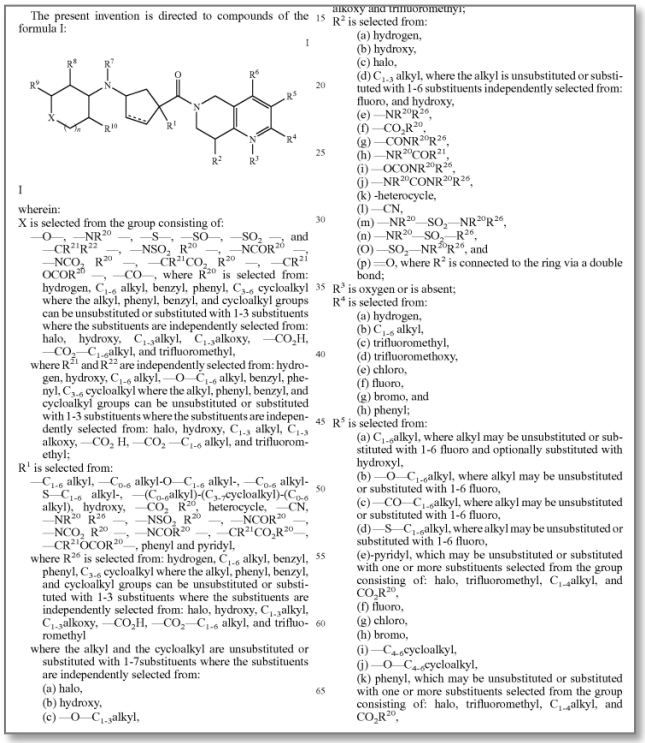
Supported by:

Federal Ministry of Economics and Technology

on the basis of a decision by the German Bundestag

ChemProspector: General assumptions

- **Markush notations follow particular grammar rules**
- **Definition of four complexity levels for Markush-structures**
  - Level 0: fully defined structures
  - Level 1: simple standard notations
  - Level 2: complex standard notations
  - Level 3: complex notations, singletons
- **Reference set: 70 patent documents (Training)**
- **Test set: 70 patent documents (Evaluation)**



The present invention is directed to compounds of the formula I:

I

I

wherein:

X is selected from the group consisting of: —O—, —NR^{20}—, —S—, —SO—, —SO_2—, and —$CR^{21}R^{22}$—, —NSO_2R^{20}—, —$NCOR^{20}$—, —NCO_2R^{20}—, —$CR^{21}CO_2R^{20}$—, —$CR^{21}OCOR^{20}$—, —CO—, where R^{20} is selected from: hydrogen, C_{1-6} alkyl, benzyl, phenyl, C_{3-6} cycloalkyl where the alkyl, phenyl, benzyl, and cycloalkyl groups can be unsubstituted or substituted with 1-3 substituents where the substituents are independently selected from: halo, hydroxy, C_{1-3}alkyl, C_{1-3}alkoxy, —CO_2H, —CO_2—C_{1-6}alkyl, and trifluoromethyl, where R^{21} and R^{22} are independently selected from: hydrogen, hydroxy, C_{1-6} alkyl, —O—C_{1-6} alkyl, benzyl, phenyl, C_{3-6} cycloalkyl where the alkyl, phenyl, benzyl, and cycloalkyl groups can be unsubstituted or substituted with 1-3 substituents where the substituents are independently selected from: halo, hydroxy, C_{1-3} alkyl, C_{1-3} alkoxy, —CO_2 H, —CO_2—C_{1-6} alkyl, and trifluoromethyl;

R^1 is selected from: —C_{1-6} alkyl, —C_{0-6} alkyl-O—C_{1-6} alkyl-, —C_{0-6} alkyl-S—C_{1-6} alkyl-, —(C_{0-6}alkyl]-(C_{3-6}-cycloalkyl)-(C_{0-6} alkyl), hydroxy, —CO_2R^{20}, heterocycle, —CN, —$NR^{20}R^{26}$, —NSO_2R^{20}, —$NCOR^{20}$, —NCO_2R^{20}, —$NCOR^{20}$, —$CR^{21}CO_2R^{20}$, —$CR^{21}OCOR^{20}$—, phenyl and pyridyl, where R^{26} is selected from: hydrogen, C_{1-6} alkyl, benzyl, phenyl, C_{3-6} cycloalkyl where the alkyl, phenyl, benzyl, and cycloalkyl groups can be unsubstituted or substituted with 1-3 substituents where the substituents are independently selected from: halo, hydroxy, C_{1-3}alkyl, C_{1-3}alkoxy, —CO_2H, —CO_2—C_{1-6} alkyl, and trifluoromethyl where the alkyl and the cycloalkyl are unsubstituted or substituted with 1-7substituents where the substituents are independently selected from:
(a) halo,
(b) hydroxy,
(c) —O—C_{1-3}alkyl,
alkoxy and trifluoromethyl;

R^2 is selected from:
(a) hydrogen,
(b) hydroxy,
(c) halo,
(d) C_{1-8} alkyl, where the alkyl is unsubstituted or substituted with 1-6 substituents independently selected from: fluoro, and hydroxy,
(e) —$NR^{20}R^{26}$,
(f) —CO_2R^{20},
(g) —$CONR^{20}R^{26}$,
(h) —$NR^{20}COR^{21}$,
(i) —$OCONR^{20}R^{26}$,
(j) —$NR^{20}CONR^{20}R^{26}$,
(k) -heterocycle,
(l) —CN,
(m) —NR^{20}—SO_2—$NR^{20}R^{26}$,
(n) —NR^{20}—SO_2—R^{26},
(O) —SO_2—$NR^{20}R^{26}$, and
(p) =O, where R^2 is connected to the ring via a double bond;

R^3 is oxygen or is absent;

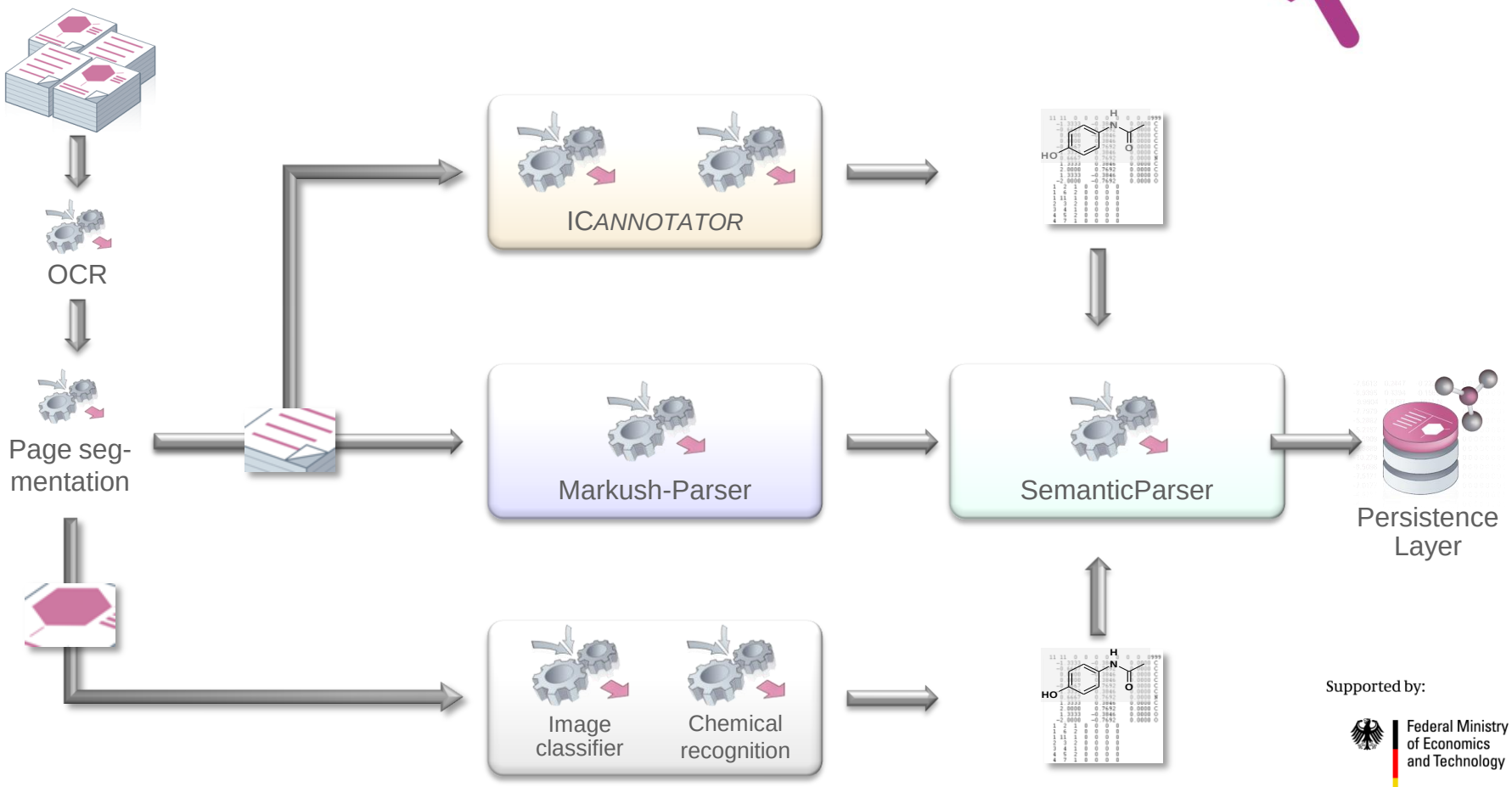
R^4 is selected from:
(a) hydrogen,
(b) C_{1-6} alkyl,
(c) trifluoromethyl,
(d) trifluoromethoxy,
(e) chloro,
(f) fluoro,
(g) bromo, and
(h) phenyl;

R^5 is selected from:
(a) C_{1-6}alkyl, where alkyl may be unsubstituted or substituted with 1-6 fluoro and optionally substituted with hydroxyl,
(b) —O—C_{1-6}alkyl, where alkyl may be unsubstituted or substituted with 1-6 fluoro,
(c) —CO—C_{1-6}alkyl, where alkyl may be unsubstituted or substituted with 1-6 fluoro,
(d) —S—C_{1-6}alkyl, where alkyl may be unsubstituted or substituted with 1-6 fluoro,
(e)-pyridyl, which may be unsubstituted or substituted with one or more substituents selected from the group consisting of: halo, trifluoromethyl, C_{1-4}alkyl, and CO_2R^{20},
(f) fluoro,
(g) chloro,
(h) bromo,
(i) —C_{4-7}cycloalkyl,
(j) —O—C_{4-6}cycloalkyl,
(k) phenyl, which may be unsubstituted or substituted with one or more substituents selected from the group consisting of: halo, trifluoromethyl, C_{1-4}alkyl, and CO_2R^{20},

Federal Ministry of Economics and Technology

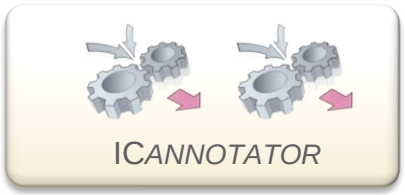
on the basis of a decision by the German Bundestag

ChemProspector: Approach



Supported by:
Federal Ministry
of Economics
and Technology
on the basis of a decision
by the German Bundestag

ICANNOTATOR



» Extracts chemical named entities

The aromatic hydrocarbon residue may preferably be an aryl group having a carbon number of 6 to 15. Examples thereof include phenyl, cresyl, xylyl, 2,6-dimethylphenyl, 2,4,6-trimethylphenyl, butylphenyl, nonylphenyl and the like.

Exact chemical entities	*methyl, ethyl, n-propyl, phenyl, chloro, nitro, amino, hydroxy, hydrogen, carbon, 1-naphthyl, 2-pyridyl, tosyl, piperidyl ...*
Generic and homology groups, fragments	*alkyl, alkoxy, aryl, halogenid, hydrocarbon ...*
Combinations	*alkylamino, 4-aryl-phenyl, ...*

upported by:
Federal Ministry of Economics and Technology
on the basis of a decision by the German Bundestag

Markush-Parser

» **Extracts Markush specific entities**

Formula definitions	*formula 1, general formula (I), derivatives represented by (3), …*
Variable definitions	*R, R^1, R^2, R', A, X, Y, Z, Ar, …*
Wherein definitions	*where, wherein, in which, …*
Link group	*represents, may be, one of, is selected from, …*
Chain lengths	*3-20 carbon atoms, …*
Topologic definitions	*branched or unbranched, …*
Bond types	*may contain double bonds, …*
References	*as defined above, …*
Substitutions	*optionally substituted by, …*

upported by:
Federal Ministry of Economics and Technology
n the basis of a decision by the German Bundestag

SemanticParser

» **Finds patterns of entities/ reassembles the components**

Grammar rule 1	FORMULA DEFINITION	LIGAND LIST				
Grammar rule 2	FORMULA DEFINITION	CHEMICAL STRUCTURE	WHERIN DEFINITION	LIGAND LIST		
Grammar rule 3	CHEMICAL STRUCTURE	CHEMICAL STRUCTURE	CHEMICAL STRUCTURE	WHERIN DEFINITION	FORMULA DEFINITION	LIGAND LIST
Grammar rule n						

on the basis of a decision by the German Bundestag

One simple sample

US 20030180237A1

(19) **United States**

(12) **Patent** ... **Lersch et** ...

(54) **COMPOSIT... AND TONI... DISORDER... AND/OR CE... ACTIVE SU...**

(75) Inventors: **P... C... U...**

Correspondence: **SCULLY SC... 400 GARDE... GARDEN C...**

(73) Assignee: **G...**

(21) Appl. No.:

(22) Filed:

(30) **Fore...**

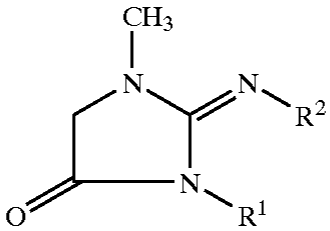
Feb. 15, 2002

Pub...

(51) **Int. Cl.**[7] .....

at least one of the compounds of the general formula (I) and/or salts thereof

(I)



in which

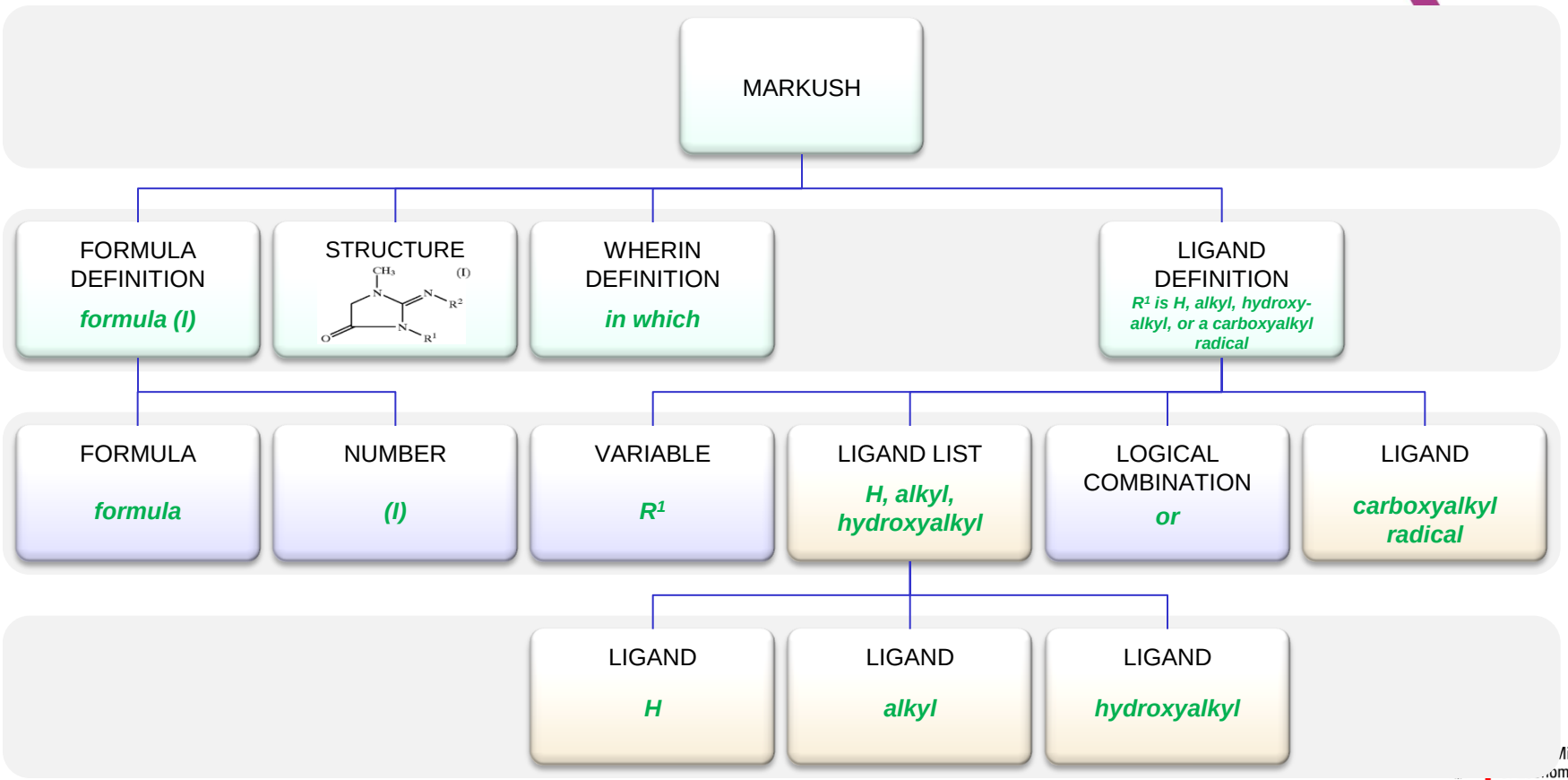
R^1 is H, alkyl, hydroxyalkyl, or a carboxyalkyl radical having 2 to 30 carbon atoms; and

Supported by:

Federal Ministry of Economics and Technology

on the basis of a decision by the German Bundestag

Markush Grammar: sample



Ministry ...omics and Technology

on the basis of a decision by the German Bundestag

First Results

» **Sample test set of four USPTO-Patents**

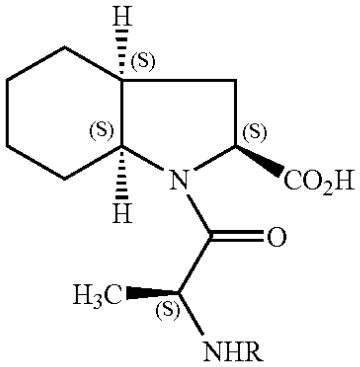
Patent	Markush Structures	Markush Structures extracted	Comment
US20070123581	8	6	One structure level 2 One structure has no ligands
US20030180237	4	4	
US20040171886	16	12	Four structures level 2
US20040030160	7	4	Three structures level 2
Sum	**35**	**26**	

Supported by:
Federal Ministry of Economics and Technology
on the basis of a decision by the German Bundestag

First Results (Samples)

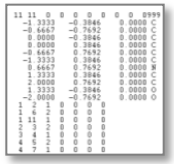
Process for the synthesis of compounds of formula (I):

(I)



wherein R represents a hydrogen atom or a protecting group for the amino function. Application in the synthesis of

FORMULA *formula (I)*



R = *hydrogen* **OR PROTECTING_GROUP**

Supported by:

Federal Ministry of Economics and Technology

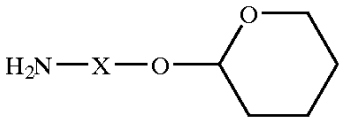
on the basis of a decision by the German Bundestag

First Results (Samples)

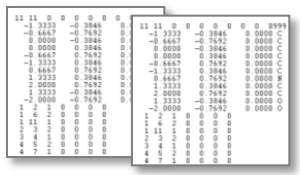
formula (1) shown below is reacted with an acid, the obtained aminoalcohol salt is reacted with 3,4-dihydro-2H-pyran, and the obtained tetrahydropyranyloxyamine salt is subsequently reacted with an alkali to form a tetrahydropyranyloxyamine represented by the general formula (2) shown below.

$H_2N—X—OH$ (1)

(2)



(wherein in said formula (1) and said formula (2), X represents a methylene group, an ethylene group or a straight chain polymethylene group having 3 to 20 carbon atoms)



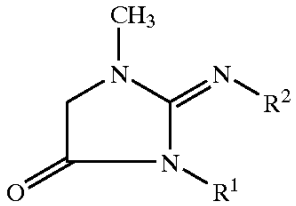
FORMULA *formula (1)* **FORMULA** *formula (2)*
X = *methylene* **OR** *ethylene* **OR** *polymethylene* **BRANCHED** = 0
CARBON_ATOM_COUNT *3 – 20*

Supported by:
Federal Ministry of Economics and Technology
on the basis of a decision by the German Bundestag

First Results (Samples)

at least one of the compounds of the general formula (I) and/or salts thereof

(I)

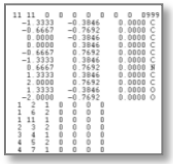


in which

R^1 is H, alkyl, hydroxyalkyl, or a carboxyalkyl radical having 2 to 30 carbon atoms; and

R^2 is H or a hydrocarbon radical having 1 to 30 carbon atoms which may be branched or unbranched and may or may not contain double bonds.

FORMULA *formula (I)* **SALTOF**



R1 = *hydrogen* **OR** *alkyl* **OR** *hydroxyalkyl* **OR** *carboxyalkyl*
CARBON_ATOM_COUNT *2 – 30*

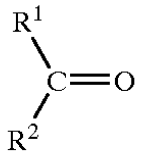
R2 = *hydrogen* **OR** *hydrocarbon*
CARBON_ATOM_COUNT *1 – 30*
BRANCHED >= 0
DOUBLEBOND >= 0

Supported by:
Federal Ministry of Economics and Technology
on the basis of a decision by the German Bundestag

First Results (Samples)

aldehyde or ketone represented by the following formula (7):

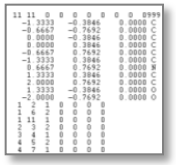
(7)



[0022] (wherein R^1 and R^2 represent the same meanings as defined above).

(wherein n represents **1** or **2**, R^1 and R^2 each represent H, $C_{1\text{-}8}$ alkyl groups and the like and R^3represents $C_{1\text{-}3}$ alkyl groups).

FORMULA *formula (7)*



R1 = *H* **OR** *alkyl*
CARBON_ATOM_COUNT *1 – 8*

R2 = *H* **OR** *alkyl*
CARBON_ATOM_COUNT *1 – 8*

Supported by:

Federal Ministry of Economics and Technology

on the basis of a decision by the German Bundestag

Next steps/Outlook

» **Markush storage and retrieval (Extension of IC*CARTRIDGE*)**

» **Extension of grammar rules to level 2 Markush Structures**

» **Page segmentation, image recognition**

Supported by:

Federal Ministry of Economics and Technology

on the basis of a decision by the German Bundestag

Acknowledgements

» The InfoChem ChemProspector team

» The German Federal Ministry of Economy and Technology (BMWi)

Federal Ministry of Economics and Technology

© P. Storz / PIXELIO, www.pixelio.de

Thank you!

Supported by:

Federal Ministry of Economics and Technology

on the basis of a decision by the German Bundestag

Questions?

Supported by:

Federal Ministry of Economics and Technology

on the basis of a decision by the German Bundestag