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
Outline

» Introduction
  › ChemProspector, a THESEUS project
  › Markush in a nutshell

» Major Goals and Approach

» First Results

» Outlook
“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)
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‘?

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

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

Quelle: http://www.colorantshistory.org
DETAILED DESCRIPTION OF THE INVENTION

The instant invention provides a compound represented by structural formula I

![Chemical structure of compound I](image)

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

- **R¹** is selected from the group consisting of:
  - ![Possible R¹ options]

- **R²** is selected from the group consisting of (a) -C₁₋₆alkyl optionally substituted with 1-3 of fluoro, (b) -C₃₋₆cycloalkyl and:
  - ![Possible R² options]

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

- **R³** is selected from the group consisting of -H, -F, -OH, -CH₃ and -CF₃;

- **R⁴** is selected from the group consisting of -H and -C₁₋₆alkyl;

- **R⁵** is selected from the group consisting of -H and -CH₃; and

- **R⁶** is selected from the group consisting of -H, -C₁₋₆alkyl optionally substituted with 1-3 fluoro, -C₃₋₆cycloalkyl optionally substituted with 1-3 fluoro and -CH₂-R⁷;

- **R⁷** is selected from the group consisting of -H, -C₁₋₆alkyl optionally substituted with 1-3 fluoro, -COC₁₋₆alkyl and -COC₃₋₆cycloalkyl;
Markush: The challenge

» The Information is contained in the text ...
   » Substituents variation
   » Homology variation
   » Topology variation

» ... in the images ...
   » Position variation
   » Frequency variation

R⁴ is hydrogen, hydroxy, cyano, fluorine or chlorine
q is 0 or 1;
R⁵ is optionally substituted aryl;
W¹ is a group represented by formula (2):

\[
\begin{align*}
& \text{(CH}_2\text{n)} \quad \text{C} \quad \text{(CH}_2\text{m)} \quad \text{N} \\
& \quad \text{R}^8 \quad \text{R}^9 \quad \text{R}^6 \quad \text{R}^7
\end{align*}
\]

wherein R¹¹ represents a hydrogen atom or a methyl group
and R¹² represents a single bond or a linear or branched
alkylene group having 1 to 5 carbon atoms.
... and both, text and images!

» Frequency variation

» Bond variation

alkyl are unsubstituted or substituted with 1-3 substituents where the substituents are independently selected from: halo, hydroxy, $\text{CO}_2\text{R}^{20}$, $C_{1-3}$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 and some or different, are an aliphatic hydro
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)
ChemProspector: Approach

1. OCR
2. Page segmentation
3. Chemical recognition
4. Image classifier
5. Markush-Parser
6. ICANNOTATOR
7. SemanticParser
8. Persistence Layer

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

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, ...

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.
### Markush-Parser

» Extracts Markush specific entities

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

» Finds patterns of entities/
  reassembles the components

Grammar rule 1

Grammar rule 2

Grammar rule 3

Grammar rule n
One simple sample

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

in which

\[ R^1 \text{ is } H, \text{ alkyl, hydroxyalkyl, or a carboxyalkyl radical having 2 to 30 carbon atoms; and} \]
Markush Grammar: sample

FORMULA DEFINITION

\[ \text{formula (I)} \]

STRUCTURE

WHEREIN DEFINITION

\[ \text{in which} \]

LIGAND DEFINITION

- \( R' \) is H, alkyl, hydroxyalkyl, or a carboxyalkyl radical

FORMULA

\[ \text{formula} \]

NUMBER

\( (I) \)

VARIABLE

\( R' \)

LIGAND LIST

- H, alkyl, hydroxyalkyl

LOGICAL COMBINATION

- or

LIGAND

- carboxyalkyl radical

LIGAND

- H

- alkyl

- hydroxyalkyl
First Results

Sample test set of four USPTO-Patents

<table>
<thead>
<tr>
<th>Patent</th>
<th>Markush Structures</th>
<th>Markush Structures extracted</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>US20070123581</td>
<td>8</td>
<td>6</td>
<td>One structure level 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>One structure has no ligands</td>
</tr>
<tr>
<td>US20030180237</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>US20040171886</td>
<td>16</td>
<td>12</td>
<td>Four structures level 2</td>
</tr>
<tr>
<td>US20040030160</td>
<td>7</td>
<td>4</td>
<td>Three structures level 2</td>
</tr>
<tr>
<td>Sum</td>
<td>35</td>
<td>26</td>
<td></td>
</tr>
</tbody>
</table>
First Results (Samples)

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

wherein R represents a hydrogen atom or a protecting group for the amino function. Application in the synthesis of...
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 tetrahydropyranoloxamine salt is subsequently reacted with an alkali to form a tetrahydropyranyl oxamine represented by the general formula (2) shown below.

\[
\begin{align*}
\text{H}_2\text{N} - \text{X} - \text{OH} & \quad \text{(1)} \\
\text{H}_2\text{N} - \text{X} - \text{O} & \quad \text{(2)}
\end{align*}
\]

(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)
First Results (Samples)

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

\[
\text{FORMULA formula (I) SALTOF}
\]

\[
\begin{align*}
\text{R1} &= \text{hydrogen OR alkyl OR hydroxyalkyl OR carboxyalkyl} \\
\text{CARBON_ATOM_COUNT} &\geq 2 - 30
\end{align*}
\]

\[
\begin{align*}
\text{R2} &= \text{hydrogen OR hydrocarbon} \\
\text{CARBON_ATOM_COUNT} &\geq 1 - 30 \\
\text{BRANCHED} &\geq 0 \\
\text{DOUBLEBOND} &\geq 0
\end{align*}
\]

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.
First Results (Samples)

The first results involve the analysis of aldehydes or ketones represented by the following formula (7):

\[
\begin{align*}
\text{R}^1 & = H \text{ OR alkyl} \\
\text{CARBON_ATOM_COUNT} & = 1 - 8 \\
\text{R}^2 & = H \text{ OR alkyl} \\
\text{CARBON_ATOM_COUNT} & = 1 - 8
\end{align*}
\]

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\(_{1-8}\) alkyl groups and the like and \( R^3 \) represents C\(_{1-3}\) alkyl groups.
Next steps/Outlook

» Markush storage and retrieval (Extension of ICCARTRIDGE)

» Extension of grammar rules to level 2 Markush Structures

» Page segmentation, image recognition
Acknowledgements

» The InfoChem ChemProspector team

» The German Federal Ministry of Economy and Technology (BMWi)
Thank you!
Questions?

© edithome / PIXELIO, www.pixelio.de