Extract Chemical Information from Patents Using Chemicalize and D2S (Document to Structure)

Wei Deng (David), Daniel Bonniot, Andras Stracz

International Conference and Exhibition on Computer Aided Drug Design & QSAR
Oct 30th, 2012
Chicago, IL, USA
ChemAxon’s Naming Technology

- Structure to Name
- Name to Structure
- Document to Structure
- Document to Database
- Chemicalize
ChemAxon’s Naming Technology

• Structure to Name
  – IUPAC Name, traditional names
  – Reaching maturity
  – Still upcoming: peptides, some fused systems

• Name to structure
  – IUPAC, CAS and systematic names
  – Dictionary of common names and drug names
  – Support CAS Registry number (webservice)
  – Homology group: alkyl, aryl …
  – Future: Biological names, polymers, …

• Accuracy and coverage constantly improving

• Also available from command-line
Name to Structure Internals

• Dictionary of common and drug names
  – Uses nine different source dictionaries
  – Harmonized using weighted consensus method
  – 150K names for 40K unique structures
  – Custom dictionary and plug-in system

• Systematic names
  – Proprietary algorithm
  – ”Understands” systematic names
  – Example:
    (2R)-2-methylsulfanyl-3-hydroxybutanedioate
**Systematic Name Example**

(2R)-2-methylsulfanyl-3-hydroxybutanedioate

<table>
<thead>
<tr>
<th>(2R)</th>
<th>Stereo Locant list</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Locant list</td>
</tr>
<tr>
<td>meth</td>
<td>Structure token: C</td>
</tr>
<tr>
<td>yl</td>
<td>Suffix</td>
</tr>
<tr>
<td>sulfan</td>
<td>Structure token: S</td>
</tr>
<tr>
<td>yl</td>
<td>Suffix</td>
</tr>
<tr>
<td>3</td>
<td>Locant list</td>
</tr>
<tr>
<td>hydroxy</td>
<td>Structure token: O</td>
</tr>
<tr>
<td>but</td>
<td>Structure token: CCCC</td>
</tr>
<tr>
<td>ane</td>
<td>Suffix</td>
</tr>
<tr>
<td>di</td>
<td>Multiplier</td>
</tr>
<tr>
<td>oate</td>
<td>Suffix: C(=O)[O-]</td>
</tr>
</tbody>
</table>
Systematic Name: Parsing

(2R)-2-methylsulfanyl-3-hydroxybutanedioate
Systematic Name: Parsing

(2R)-2-methylsulfanyl-3-hydroxybutanedioate
(2R)-2-methylsulfanyl-3-hydroxybutanedioate
Systematic Name: Parsing

(2R)-2-methylsulfanyl-3-hydroxybutanedioate
Systematic Name: Parsing

(2R)-2-methylsulfanyl-3-hydroxybutanedioate
Systematic Name: Parsing

(2R)-2-methylsulfanyl-3-hydroxybutanedioate
Systematic Name: Structure Generation

(2R)-2-methylsulfanyl-3-hydroxybutanedioate
(2R)-2-methylsulfany1-3-hydroxybutanediolate
(2R)-2-methylsulfanyl-3-hydroxybutanedioate
(2R)-2-rnethylsulfany1-3-hydroxybutanedi0ate

(2R)-2-methylsulfanyl-3-hydroxybutanedioate

Λr-benzyl-Λr-[3-(IH-tetrazol-5-yl)phenyl]propanamide

?-benzyl-?-[3-(?H-tetrazol-5-yl)phenyl]propanamide
(2R)-2-riothylsulfanyl-1-3-hydroxybutanediolate

\[ \downarrow \]

(2R)-2-methylsulfanyl-1-3-hydroxybutanedioate

\[ \downarrow \]

\[ \Lambda r\text{-benzyl}-\Lambda r\text{-}[3\text{-}(1H\text{-tetrazol}-5\text{-yl})\text{phenyl}]\text{propanamide} \]

\[ \downarrow \]

\[ N\text{-benzyl}-N\text{-}[3\text{-}(1H\text{-tetrazol}-5\text{-yl})\text{phenyl}]\text{propanamide} \]
ChemAxon’s “Document to Structure”

• Extract chemical information from documents
  – Names: powered by the Naming Technology
  – Also import smiles, InChI, CAS number …
  – Images: OSRA …
  – Works with **scanned non-searchable PDF**
  – Returns structures and their **location** in the document

• Supported formats:
  – MS Office document: doc, docx, ppt, pptx, xls, xlsx, odt …
  – Embedded structure objects (ChemDraw, Symyx, Marvin, …)
  – PDF, text, XML, HTML
From Document to Structures

Non-searchable patent (50 pages)  Structure (text + image) + location
In Instant JChem, search for
## Search by Structure or Text

<table>
<thead>
<tr>
<th>CddId</th>
<th>Structure</th>
<th>Mol Weight</th>
<th>Formula</th>
<th>type</th>
<th>confidence</th>
<th>page</th>
<th>document</th>
<th>text</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><img src="image1" alt="Structure" /></td>
<td>414.89</td>
<td>C21H23ClN4O3</td>
<td>systematic</td>
<td></td>
<td></td>
<td></td>
<td>1-(4-chloro-2-([1,2]oxazin-8-yl)urea)</td>
</tr>
<tr>
<td>2</td>
<td><img src="image2" alt="Structure" /></td>
<td>414.89</td>
<td>C21H23ClN4O3</td>
<td>systematic</td>
<td></td>
<td></td>
<td></td>
<td>1-(4-chloro-2-(pyridin-3-yl)benzyl)-3-(3,4-dihydro-3-oxo-2H-benzo[8][1,4] oxazin-8-yl)urea</td>
</tr>
</tbody>
</table>
Example 52 (scheme 1)

1-(4-chloro-2-(piperidin-1-yl)benzyl)-3-(3,4-dihydro-3-oxo-2H-benzo[b][1,4]oxazin-8-yl)urea

[0233] Preparation of 1-(4-chloro-2-(piperidin-1-yl)benzyl)-3-(3,4-dihydro-3-oxo-2H-benzo[b][1,4]oxazin-8-yl)urea

Amine 2hn (1.7 g, 7.59 mmol) was dissolved in 50 ml of ACN and at 0°C triphosgene (2.26 g, 7.59 mmol) was added to the solution. The mixture was warmed at 80°C for 4 hours then evaporated and dissolved in 5 ml of DMF. The solution of the isocyanate was added dropwise to a solution in DMF (20 ml) of compound 11 (1.5 g, 7.51 mmol) and the mixture was warmed at 80°C for 8 hours. (TLC ACOEt 1 / petroleum ether 1). The solvent was evaporated and the crude was dissolved in AcOEt (50 ml) and washed with water (1 X 30 ml) and brine. The organic phase was dried over sodium sulfate and concentrated under vacuum. The purification of the crude residue by chromatographic column gave 210 mg of a white solid. Yield = 7% 1H NMR (DMSO, 200 MHz) δ 1.54 (2H, m), 1.66 (4H, m), 2.78 (4H, m), 4.29 (2H, d, J = 5.6 Hz), 4.36 (2H, d), 4.81 (2H, s), 6.47 (1H, dd, J = 7.8 Hz, J' = 1.2 Hz), 6.61 (1H, t, J = 8 Hz), 7.13 (4H, m), 7.75 (1H, dd, J = 8.2 Hz, J' = 1.4 Hz), 8.14 (1H, bs), 10.86 (1H, bs); [M+H]+ 414.9 (C27H32CN4O5 requires 414.89).

Example 53: 1-(4-methyl-2-(piperidin-1-yl)benzyl)-3-(3,4-dihydro-3-oxo-2H-benzo[b][1,4]oxazin-8-yl)urea (scheme 1)

[0235] Preparation of 1-(4-methyl-2-(piperidin-1-yl)benzyl)-3-(3,4-dihydro-3-oxo-2H-benzo[b][1,4]oxazin-8-yl)urea

Amine 2nbh (1.08 g, 5.5 mmol) was dissolved in 40 ml of ACN and at 0°C triphosgene (1.56 g, 5.4 mmol) was added to the solution. The mixture was warmed at 80°C for 4 hours then evaporated and the residue was dissolved in 15 ml of DMF. The solution of the isocyanate was added dropwise to a solution in DMF (10 ml) of compound 11 (1 g, 6.46 mmol) and the mixture was warmed at 80°C for 8 hours. (TLC ACOEt 1 / petroleum ether 9). The solvent was evaporated and the crude was dissolved in AcOEt (50 ml) and washed with water (1 X 30 ml) and brine. The organic phase was dried over sodium sulfate and concentrated under vacuum. The purification of the crude residue by chromatographic column gave 150 mg of a white solid. Yield = 7% 1H NMR (DMSO, 200 MHz) δ 1.63 (6H, m), 2.25 (3H, s), 2.76 (4H, m), 4.28 (2H, d, J = 6.4 Hz), 4.61 (2H, s), 6.46 (1H, dd, J = 7.8 Hz, J' = 1.4 Hz), 6.81 (3H, m), 7.12 (2H, m), 7.75 (1H, dd, J = 8.2 Hz, J' = 1.6 Hz), 8.11 (1H, s), 10.86 (1H, bs); [M+H]+ 395.0 (C26H30N4O4 requires 394.5).
“Document to Database”
ChemAxon’s “Document to Database”

• Data in DB:
  – Structures
  – Source (name, smiles, embedded, …) and location
  – Documents, Authors, Metadata...

• Questions:
  – What structures appear in a specific document?
  – What documents contain a structure/substructure/...?
  – What documents written since 2010 in location X contain substructure S?
  …
ChemAxon’s “Document to Database”

• Customizable:
  – Metadata extracted from documents
  – Interface (IJC forms, webapp)

• Demo:
  – One month of US patents
  – 85K unique structures from systematic names
  – 1M occurrences
Summary

- Extensive, improving naming technologies (n2s, s2n)
- Increasing support for Document mining (d2s, d2db, SharePoint)
- Putting it all together → going large scale, extract and use valuable chemical information
- Still component-based and responding to our users requests
Free Online Service Chemicalize.org

- Extract
- Interactively display
- Calculate
- Search

Webpage - Chemicalized

Aspirin
From Wikipedia, the free encyclopedia

"Aspirin" redirects here. For the author, see Robert.

Aspirin (USAN), also known as acetylsalicylic acid (IPA: /ˌækətsələˈsɪlɪk/; also known as aspirin, abbreviated ASA), is a salicylate drug, often used as an analgesic to relieve minor aches and pains, as an antipyretic to reduce fever, and as an anti-inflammatory medication. It was first isolated by Arthur Eichengrün, a chemist with the German company Bayer.[1]

Salicylic acid, the main metabolite of aspirin, is an integral part of human and animal metabolism. While much of it is attributable to diet, a substantial part is synthesized endogenously.[2]

Aspirin also has an antiplatelet effect by inhibiting the production of thromboxane, which under normal circumstances binds platelet molecules together to create a patch over damaged walls of blood vessels. Because the platelet patch can become too large and also block blood flow, locally and downstream, aspirin is also used long-term, at low doses, to help prevent heart attacks, strokes, and blood clot formation in people at high risk of developing blood clots.[3] It has also been established that low doses of aspirin may be given immediately after a heart attack to reduce the risk of another heart attack or of the death of cardiac tissue.[4][5]

The main undesirable side-effects of aspirin taken by mouth are gastrointestinal ulcers, stomach bleeding, and tinnitus, especially in higher doses. In children and adolescents, aspirin is no longer indicated to control flu-like symptoms or the symptoms of chickenpox or other viral illnesses, because of the
Customizable report layout for calculation results. Users can move, open, close, expand calculation boxes and this is remembered on the next visit.
Aspirin

From Wikipedia, the free encyclopedia

*Aspirin* redirects here. For the author, see Robert Aspin.

Aspirin (USAN), also known as acetylsalicylic acid (IPA: /əˈsɪljəsəˈlɪsɪk/), abbreviated ASA, is a salicylate drug, often used as an analgesic to relieve minor aches and pains, as an antipyretic to reduce fever, and as an anti-inflammatory medication. It was first isolated by Arthur Eichengrün, a chemist with the German company Bayer.[1]

Salicylic acid, the main metabolite of aspirin, is an integral part of human and animal metabolism. While much of it is attributable to diet, a substantial part is synthesized endogenously.[2]

Aspirin also has an antplatelet effect by inhibiting the production of thromboxane, which under normal circumstances binds platelet molecules together to create a patch over damaged walls of blood vessels. Because the platelet patch can become too large and also block blood flow, locally and downstream, aspirin is also used long-term, at low doses, to help prevent heart attacks, strokes, and blood clot formation in people at high risk of developing blood clots.[3] It has also been established that low doses of aspirin may be given immediately after a heart attack to reduce the risk of another heart attack or of the death of cardiac tissue.[4][5]

The main undesirable side-effects of aspirin taken by mouth are gastrointestinal ulcers, stomach bleeding, and tinnitus, especially in higher doses. In dilutions to control flu-like symptoms or the symptoms of chickenpox or other viral illnesses, because of the

![Aspirin molecule](https://en.wikipedia.org/wiki/Aspirin)
The combination of aspirin, paracetamol (acetaminophen) and caffeine (as found in the OTC brand Excedrin) is even more potent. For the treatment of migraine headache, this formulation works better than any of its three components taken separately \cite{38} better than ibuprofen \cite{40} and better than sumatriptan. As with all other medications for migraine, it is recommended to take aspirin at the first signs of the headache, and it is the way these medications were used in the comparative clinical trials \cite{41}.

Aspirin alleviates pain in 60–75% of patients with episodic tension headaches \cite{42,43}. It is equivalent to paracetamol (acetaminophen) in that respect, except for the higher frequency of gastrointestinal side-effects \cite{43}. Comparative clinical trials indicated metamizole and ibuprofen may relieve pain faster than aspirin, although the difference becomes insignificant after about two hours. The addition of caffeine in a dose of 60–130 mg to aspirin increases the analgesic effect in headache \cite{42,44}. The combination of aspirin, paracetamol (acetaminophen) and caffeine is still more effective, but at the cost of more stomach discomfort, nervousness and dizziness \cite{45}. There is some evidence low-dose aspirin has benefit for reducing the occurrence of migraines in susceptible individuals \cite{46,47,48,49}.
Aspirin: query highlighted in results
Aspirin; web page hits - “show” related structures
• Autosuggest while typing
Everything is Published

- Recent viewed
  - Webpages
  - Structures
    [http://www.chemicalize.org/blog/2012/09/26/1-million-molecules/](http://www.chemicalize.org/blog/2012/09/26/1-million-molecules/)
  - Documents
  - Searched queries (structure and keyword)
Availability and Customization

- Source code available
- Minor changes required on example codes for customization, such as
  - Import extracted structures to other databases
  - Post-process filtering according to properties
  - Batch process of multiple documents