CHEMICAL STRUCTURE INDEXING AND CLASSIFICATION: WHY BOTHER?

The use of reverse text-mining to establish whether indexing and classification of chemical patents is still necessary

11th August 2014
OUTLINE

• CHEMICAL NAME TO STRUCTURE WORK
  – Steve Boyer
  – ChemAxon
  – Others

• SEARCHING FOR CHEMICALS

• BACK TO THE FUTURE
  – Markush DARC
  – Fragmentation coding

• EXAMPLE

• CONCLUSIONS
OUTLINE

• CHEMICAL NAME TO STRUCTURE WORK
  – Steve Boyer
  – ChemAxon
  – Others

• SEARCHING FOR CHEMICALS

• BACK TO THE FUTURE
  – Markush DARC
  – Fragmentation coding

• EXAMPLE

• CONCLUSIONS
CHEMICAL NAME STRUCTURE RECOGNITION

**entity identification**

a) (2P/4S)-4-[4-Amino-5-(4-benzyl-oxo-phenyl)pyrrolo[2,3-d]pyrimidin-7-yl]-2-hydroxymethyl-pyrrolidine-1-carboxylic acid tert-butyl ester prepared analogously to Example 18 starting from (2R/4S)-4-[4-amino-5-(4-benzyl-oxo-phenyl)pyrrolo[2,3-d]pyrimidin-7-yl]-pyrrolidine-1,2-dicarboxylic acid 1-tert-butyl ester 2-ethyl ester (Example 20a). 1 H-NMR (CDCl3, ppm): 8.52 (s, 1H), 7.52-7.32 (m, 7H), 7.1 (d, 2H), 6.95 (d, 1H), 5.50 (m, 1H), 5.13 (s, 2H), 4.62-4.44 (m, 3H), 4.10 (m, 1H), 3.95-3.70 (m, 1H), 2.75 (m, 1H), 2.50 (m, 1H), 1.85-1.60 (m, 4H), 1.30-1.10 (m, 9H), 0.90 (d, 3H), 0.70 (d, 3H).

b) (2R/4S)-4-[4-Amino-5-(4-benzyl-oxo-phenyl)pyrrolo[2,3-d]pyrimidin-7-yl]-2-hydroxymethyl-pyrrolidine-1-carboxylic acid tert-butyl ester prepared analogously to Example 18 starting from (2R/4S)-4-[4-amino-5-(4-benzyl-oxo-phenyl)pyrrolo[2,3-d]pyrimidin-7-yl]-pyrrolidine-1,2-dicarboxylic acid 1-tert-butyl ester 2-ethyl ester (Example 20a). 1 H-NMR (CDCl3, ppm): 8.42 (s, 1H), 7.50-7.25 (m, 7H), 7.0 (d, 2H), 6.85 (d, 1H), 5.45 (m, 1H), 5.00 (s, 2H), 4.50-4.30 (m, 3H), 4.00 (m, 1H), 3.85-3.60 (m, 1H), 2.75 (m, 1H), 2.50 (m, 1H), 1.85-1.60 (m, 4H), 1.30-1.10 (m, 9H), 0.90 (d, 3H), 0.70 (d, 3H).

c) 0.100 g of (2R/4S)-4-[4-Amino-5-(4-benzyl-oxo-phenyl)pyrrolo[2,3-d]pyrimidin-7-yl]-2-hydroxymethyl-pyrrolidine-1-carboxylic acid tert-butyl ester was dissolved in methanol and evaporated to dryness. The crude residue was triturated with ethyl ether, filtered off and dried under vacuum. 1 H-NMR (CD3OD, ppm): 8.40 (s, 1H), 7.50-7.25 (m, 7H), 7.0 (d, 2H), 6.85 (d, 1H), 5.45 (m, 1H), 5.00 (s, 2H), 4.50-4.30 (m, 3H), 4.00 (m, 1H), 3.85-3.60 (m, 1H), 2.75 (m, 1H), 2.50 (m, 1H), 1.85-1.60 (m, 4H), 1.30-1.10 (m, 9H), 0.90 (d, 3H), 0.70 (d, 3H).

d) (2R/4S)-4-[4-Amino-5-(4-benzyl-oxo-phenyl)pyrrolo[2,3-d]pyrimidin-7-yl]-2-hydroxymethyl-pyrrolidine-1-carboxylic acid tert-butyl ester prepared analogously to Example 18 starting from (2R/4S)-4-[4-amino-5-(4-benzyl-oxo-phenyl)pyrrolo[2,3-d]pyrimidin-7-yl]-pyrrolidine-1,2-dicarboxylic acid 1-tert-butyl ester 2-ethyl ester (Example 20a). 1 H-NMR (CDCl3, ppm): 8.52 (s, 1H), 7.52-7.32 (m, 7H), 7.1 (d, 2H), 6.95 (d, 1H), 5.50 (m, 1H), 5.13 (s, 2H), 4.62-4.44 (m, 3H), 4.10 (m, 1H), 3.95-3.70 (m, 1H), 2.75 (m, 1H), 2.50 (m, 1H), 1.85-1.60 (m, 4H), 1.30-1.10 (m, 9H), 0.90 (d, 3H), 0.70 (d, 3H).
Document to Structure

EXTRACT CHEMICAL INFORMATION FROM DOCUMENTS

Powered by ChemAxon's Naming technology, Document to Structure (D2S) is a versatile application to extract chemical information from documents. Various types of chemical information in different document formats can be recognized. D2S also applies text OCR and image OSR technologies to extract information from non-searchable PDF documents. Once the structure is converted, the location of that structure is also returned. All these features make D2S an excellent choice for text mining, patent analysis and internal document management.

PRODUCT TYPE: component
INTERFACES: CLI

Retrieve Chemical Information from Documents
WORK FROM OTHER GROUPS

• Text extraction
  (a) Extractors (IUPAC names)
    - TEMIS Chemical Entity Relationships Skill Cartridge
    - Accelrys Pipeline Pilot extractor
    - Fraunhofer
    - Oscar
    - SureChem
  (b) Converter
    (Names ➔ connection table)
    - CambridgeSoft name=struct
    - Openeye Lexicchem

• Image extraction
  - OSRA (NIH)
  - Clide Pro (Keymodule Ltd.)
  - Fraunhofer chemoCR
  - ChemReader
OUTLINE

• CHEMICAL NAME TO STRUCTURE WORK
  – Steve Boyer
  – ChemAxon
  – Others

• SEARCHING FOR CHEMICALS

• BACK TO THE FUTURE
  – Markush DARC
  – Fragmentation coding

• EXAMPLE

• CONCLUSIONS
CHEMICAL NOMENCLATURE

Valium  =  Diazepam  =  CAS # 439-14-5  =

- 7-CHLORO-1-METHYL-5-PHENYL-2H-1,4-BENZODIAZEPIN-2-ONE
- 7-CHLORO-1-METHYL-5-PHENYL-3H-1,4-BENZODIAZEPIN-2(1H)-ONE
- 7-CHLORO-1-METHYL-5-PHENYL-1,3-DIHYDRO-2H-1,4-BENZODIAZEPIN-2-ONE
- 7-CHLORO-1-METHYL-2-OXO-5-PHENYL-3H-1,4-BENZODIAZEPINE
- 1-METHYL-5-PHENYL-7-CHLORO-1,3-DIHYDRO-2H-1,4-BENZODIAZEPIN-2-ONE
- 7-CHLORO-1,3-DIHYDRO-1-METHYL-5-PHENYL-2H-1,4-BENZODIAZEPIN-2-ONE
- 7-CHLORO-1-METHYL-5-3H-1,4-BENZODIAZEPIN-2(1H)-ONE
SEARCHING FOR CHEMICALS

• Chemicals have a wide variety of synonyms and systematic names
• Text searching is insufficient to retrieve all references
• Expanding synonyms is insufficient
• Structure search is the best solution
SUCCESSES AND CHALLENGES

• Work to translate chemical names to chemical structures for specific molecules is increasingly successful with increasing conversion ratios and high accuracy

• Conversion of images to searchable chemical structures for specific compounds is also achieving success

• Challenges still remain with translation of description and images of Markush structures to searchable forms
• Prepn. of 4-aryl-butane-2-one derivs. of formula ArCH2CH2COMe (I) comprises (a) reacting ArY (IV) with a 4-substd.-2-butanone of formula XCH2CH2COMe (IIIa) Ar = opt. substd. phenyl…

![Chemical Structures]

1. \( \text{CH}_3\text{CH}_2\text{CH}(\text{X})\text{CH}_2\text{COMe} \)
2. \( \text{CH}_3\text{O} \text{Ar}\text{CH}_2\text{CH}(\text{Y})\text{CH}_2\text{COMe} \)
3. \( \text{CH}_3\text{O} \text{ArCH}_2\text{CH}(\text{Ar})\text{CH}_2\text{COMe} \)
OUTLINE

• CHEMICAL NAME TO STRUCTURE WORK
  – Steve Boyer
  – ChemAxon
  – Others

• SEARCHING FOR CHEMICALS

• BACK TO THE FUTURE
  – Markush DARC
  – Fragmentation coding

• EXAMPLE

• CONCLUSIONS
BACK TO THE FUTURE

• Traditional forms of rendering Markush structures into searchable forms by chemical structure:
  – Markush DARC
  – Fragmentation coding
MARKUSH DARC

CHK Superatom with NT attribute will retrieve all specific alkyl groups (methyl, ethyl, propyl…)

2 Free Sites on the N atom

#Cl
FRAGMENTATION CODING

CMC UPB 20140617
DRN: 1512-U 0277-U
DCR: 104530-U 600-U
M3 *01*  G013 G100 H4 H401 H441 H8 J5 J581 M210 M211 M262 M281 M312 M321
N422 N480 N513 N522 M905 M904
DCN: R17318-K R17318-P
DCR: 6417-K 6417-P
M3 *02*  G013 G100 H5 H541 H8 J5 J581 M210 M211 M262 M272 M281 M311 M321
M342 M372 M391 M414 M510 M520 M531 M540 M730 M905 M904
DCN: RA443J-K RA443J-S
DCR: 26441-K 26441-S
M3 *03*  A119 A940 C101 C108 C550 C730 C801 C802 C804 C805 C807 M411 M781
M905 M904 M910
DCN: R01512-K R01512-U R01512-V
DCR: 104530-K 104530-U 104530-V
M3 *04*  H4 H498 H9 J0 J011 J1 J171 M280 M311 M321 M342 M349 M381 M391
M416 M620 M781 M905 M904 M910
DCN: R00277-K R00277-U R00277-V R08005-K R08005-U R08005-V
DCR: 600-K 600-U 600-V
IS THIS NECESSARY?

• Expensive
• Time-consuming
• Requires highly trained specialist skills
  – Indexing
  – Searching
OUTLINE

• CHEMICAL NAME TO STRUCTURE WORK
  – Steve Boyer
  – ChemAxon
  – Others

• SEARCHING FOR CHEMICALS

• BACK TO THE FUTURE
  – Markush DARC
  – Fragmentation coding

• EXAMPLE

• CONCLUSIONS
RASPBERRY KETONE

• Only 30% of people are now at a healthy weight. Being overweight has become the new norm.

• Among the most popular miracle weight loss products out there is a supplement called Raspberry ketones.

• Raspberry ketones were featured on the Dr. Oz show in 2012 and have since then been very popular all over the world.

• Raspberry ketones are claimed to cause the fat within cells to be broken down more effectively, helping the body burn fat faster.

• They are also claimed to increase levels of adiponectin, a hormone that helps to regulate metabolism.
Specific Compound

Preferred IUPAC Name = 4-(4-hydroxyphenyl)butan-2-one
Synthesis of raspberry ketone involves mixing anisyl acetone, potassium hydroxide and N,N-dimethylacetamide, heating, adding demethylating reagent, heating under inert atmosphere, concentrating, and crystallizing.
FRAGMENTATION CODES

Black 1963-date M900
Red 1970-date M901
Blue 1972-date M902
Green 1981-date M903
SEARCH STRATEGY

L1  16045 S (G100 (S) H401 (S) J581 (S) M414)/M3
L2  15909 S L1 (S) (M510 (S) M520)/M3
L3  8941 S L2 (S) (M210 (S) M211 (S) M281 (S) M312 (S) M321 (S) M332 (S) M342 (S) M391)/M3
L4  3209 S L3 (S) (G013 (S) M262 (S) M372)/M3
L5  440 S L4 NOT (H1 OR H2 OR H3 OR H5 OR H6 OR H7 OR H9 OR J0 OR J1 OR J2 OR J3 OR J4 OR J6 OR J9 OR K0 OR M1)/M3

M3 = General chemicals
REVERSE TEXT ENGINEERING OF RESULTS

• 440 results from STN fragmentation code search loaded into Thomson Innovation

• Text mining performed to identify key text terms
TEXT CLUSTERING
RETRIEVAL

- Specific compounds
- Markush
SPECIFIC COMPOUND EXAMPLE

4-(4-Hydroxyphenyl)-2-butanone, used as perfume is prep'd. by reacting phenol with methyl vinyl ketone.

\[
\text{HO-} + \text{CH}_2=\text{CHCOCOCH}_3 \rightarrow \text{HO-CH}_2\text{CH}_2\text{COCH}_3
\]
MARKUSH RETRIEVAL EXAMPLE

• WO1996040608A2

• Prepn. of 4-aryl-butano-2-one derivs. of formula ArCH2CH2COMe (I) comprises (a) reacting ArY (IV) with a 4-substd.-2-butanone of formula XCH2CH2COMe (IIIA) Ar = opt. substd. phenyl…
OUTLINE

• CHEMICAL NAME TO STRUCTURE WORK
  – Steve Boyer
  – ChemAxon
  – Others

• SEARCHING FOR CHEMICALS

• BACK TO THE FUTURE
  – Markush DARC
  – Fragmentation coding

• EXAMPLE

• CONCLUSIONS
CONCLUSIONS

• Text searching is insufficient to retrieve all references
• Structure search is the best solution for retrieval of chemical information
• Conversion of chemical names and images to structures is largely successful for specific compounds
• Challenges remain with Markush structures
• Intellectual indexing with traditional methods is still absolutely necessary
THANK YOU

Bob Stembridge
Senior IP Analyst
IP & Science

bob.stembridge@thomsonreuters.com