Challenges in Chemical Literature Mining

Vidyendra Sadanandan, Shashikala G, Anirban Mudi, Lokanath Khamari, Jignesh Bhate

Molecular Connections, India

ACS National Meeting Philadelphia 2012
Outline

- Application of chemical text mining
- Evolution of text mining
- Challenges in mining chemical literature
- Chemical text mining approaches
- Hybrid approach to mining chemical information
Application of Chemical Text Mining

- Publishing
  - Custom databases

- Pharma
  - Drug Discovery, Drug Safety, Drug Repurposing

http://www.3dchem.com
Key Driver of Text Mining - Information Growth

Number of articles (diamonds) and patents (open boxes) abstracted annually by Chemical Abstracts Services


By IPC classification for patent applications in 2010, the following technologies were represented as such:

- 33.7% Electrical engineering
- 22.3% Mechanical engineering
- 21.4% Other
- 14.9% Chemistry
- 7.7% Instruments
- 5% Other

2012 WIPO IP Facts and Figures
- Manual curation - indexing, classification and abstraction

- Automated methods to reduce the lag between publication and abstraction

- Commercial tools (Lexicrem, TEMIS, SureChem, Linguamatics, IBM, Accelrys, Inforsense, Infochem, ChemAxon, SCAI....)

- Growth of open source initiatives (Cambridge based groups, NACTEM,....)

Source: R Van Noorden, Nature 2012
Chemical Information Sources

- Journal Articles
- Patents
- Meeting abstracts
- Books
- Theses
- Chemical safety reports
- Websites
- Project reports
### Chemical Entity Recognition Applications

<table>
<thead>
<tr>
<th>Name identification</th>
<th>Name to Structure</th>
<th>Structure identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>EbiMed</td>
<td>ChemAxon’s name-to-structure</td>
<td>OSRA (NCI)</td>
</tr>
<tr>
<td>OSCAR</td>
<td>CambridgeSoft’s name=struct</td>
<td>Imago (GGA)</td>
</tr>
<tr>
<td>ChemSpot</td>
<td>OPSIN</td>
<td>ChemReader (UMich)</td>
</tr>
<tr>
<td>SureChem</td>
<td>OpenEye Scientific Software’s Lexichem</td>
<td>SimBioSys, CLiDE™</td>
</tr>
<tr>
<td>IBM, SIIP</td>
<td>InfoChem’s ICN2S</td>
<td>SCAI Fraunhofer Institute, ChemoCR</td>
</tr>
<tr>
<td>SCAI Fraunhofer, ProMiner</td>
<td>ChemInnovation’s NameExpert</td>
<td></td>
</tr>
</tbody>
</table>
Named Entity Recognition Approaches

- Dictionaries based
- Rules (regular expressions) based
- Training corpus (machine learning) based
Dictionaries

- Commercial Databases:
  - CAS, Reaxys, Thomson Reuters Pharma, SureChem

- Free Resources:
  - PubChem, DrugBank, CheMBL, ChEBI, MeSH, Chemspider, BindingDB
Rules (regular expressions) based

OSCAR4: a flexible architecture for chemical text-mining
David M Jessop, Sam E Adams, Egon L Willighagen, Lezan Hawizy and Peter Murray-Rust*
J Chem Inf. 2011
Machine Learning Approaches

- NER:
  CRF, Markov models, Graphical models

- NLP:
  Open NLP, LingPipe, Stanford NLP

- Text Classification:
  WEKA, MALLET (Naïve Bayes, SVM, neural networks,...)
Text Mining : Biology Vs Chemistry

- Biology text mining is more mature

- No proper evaluation of available chemical text mining technologies

- Difficulty in adapting biology text mining tools to chemistry

- Growth of text mining in Biology fuelling text mining in Chemistry

e.g. finding drugs to target a protein of interest
Current Challenges in Chemical Text Mining

- Typographical errors
  - Whitespace, hyphens, commas, parenthesis, brackets, braces, apostrophes, superscripts, greek characters, digits and periods.
  - Hyphenation, linefeeds, XML tags, line and page numbers and similar noise
  - Human errors

- Determining the start and end of IUPAC like names in free text is a tricky business.

Current Challenges in Mining Chemical Patents

- Difficulty in understanding patent document formats
- Difficult to OCR patents (e.g. PDFs, scanned TIFFs etc.)
- Lot of chemical information exist as images and not in characters in patents
- Classical NLP tools take a long time to mine patents with hundreds of pages
Current Challenges: Terminology Barriers

Multiple ways of naming a simple molecule

- Systematic
  - Acetylsalicylic acid
  - Salicylic acid, acetate
  - 2-acetyloxybenzoate
  - 2-carboxyl phenylacetate

- Abbreviations
  - ASA

- Chemical Structures

- Generic & fragmented
  - salicylates

- Compound 10...

- Common/generic
  - aspirin
  - AG5230
  - Bayer

- Company Codes

- Trade names

- Anaphors
  - CAS RN 50-78-2
  - Index & Ref.
Current Challenges: Chemical Name Recognition

FIGURE 1. Chemical name recognition can be a challenging problem. In this simple example, the placement or misplacement of spaces between methyl, ethyl, and malonate can result in four different chemical structures that correspond to: methyl ethyl malonate, methylethyl malonate, methyl ethyl malonate, and methylethylmalonate.

Mining chemical structural information from the drug literature
Debra L. Banville, Drug Discovery Today, 2006
Current Challenges: Formatting & Copyright/Licensing Barriers

Textual and/or images mixed together

Some are images only!

Diverse document formats

Access to full text restricted
Other barriers

- Legal uncertainty
- Access costs
- Transaction costs
- Entry costs
- Staff costs
- Infrastructure costs

http://www.jisc.ac.uk/media/documents/publications/reports/2012/value-text-mining.pdf
Benefits of text mining...

- Support literature review
- Expedite research
- Increase accessibility and relevance of scholarly content
- Reusable models & curation
- New services and business models

http://www.jisc.ac.uk/media/documents/publications/reports/2012/value-text-mining.pdf
Chemical reading capability:

- Recognition ethylvinyl ether
- Extraction ...ethylvinyl ether was added to the solution...
- Conversion to searchable form
- Annotation/tagging of text to entity
- ...ethylvinyl ether was added to the solution...
- Semantic tagging
Why Mine Chemistry Patents?

- Patents could be a source of information of high commercial value
- Sometimes patents have information which cannot be found in other sources
- Information of interest may be published in patents earlier than in literature

http://infochim.u-strasbg.fr/CS3_2012/Lectures/Muresan.pdf
A. (E and Z)-5-chloro-2-phenoxy-phenyl-3-hydroxy-acrylic acid methyl ester
A solution of (5-chloro-2-phenoxy-phenyl)-acetic acid methyl ester (167 g, 605 mmol) and methyl formate (173 mL, 2.8 mol) in f-butyl methyl ether (1 L) was cooled to -10°C. To this solution was added potassium f-butoxide (115 g, 1.03 mol) portionwise so that the temperature remained below 0°C. After 30 minutes, the reaction was quenched by addition of 4N aqueous hydrochloric acid (400 mL). The organic layer was separated and washed with water (3 x 350 mL) and saturated brine (350 mL). The organic fractions were combined and concentrated to dryness and used without purification. The product is 90% E as shown by NMR. $^1$H NMR (400 MHz, CDCl$_3$) 3.6 (s, 3H, E methyl group; Z methyl is at 3.7), 6.8 (d, 1 H), 6.9 (m, 2H), 7.0 (t, 1 H), 7.2 (m, 3H), 7.3 (m, 2H), 11.8 (d, 1 H, E hydroxy proton; Z hydroxy proton is at 4.7).

B. 8-Chloro-dibenz[a]floxepin-10-carboxylic acid methyl ester
A slurry of 2-(5-chloro-2-phenoxy-phenyl)-3-hydroxy-acrylic acid methyl ester (210 g, 691 mmol) was heated in pyrophosphoric acid (420 g) to 60°C. After two hours, the reaction was cooled to 20°C and hexane (250 mL), toluene (500 mL), and water (500 mL) were added to quench the reaction. The organic layer was separated and washed with saturated aqueous sodium bicarbonate (500 mL) and saturated brine (500 mL). The organic fraction was concentrated and the resulting oil was crystallized from n-butanol (500 mL, 20°C for one hour and -10°C for two hours), which provided the titled compound as off-white crystals (172 g, 85% yield for steps A and B). $^1$H NMR (400 MHz, CDCl$_3$) 3.9 (s, 3H), 7.1-7.2 (m, 3H), 7.3 (m, 2H), 7.4 (t, 1 H), 7.6 (d, 1 H), 8.0 (s, 1 H).
Example 5. Preparation of asenapine, trans-5-chloro-2,3,3a,12b-tetra-hydro-2-methyl-1H-dibenz[2,3;6,7]oxepino[4,5-c]pyrrole (A)

To a first reaction vessel containing tetrahydrofuran (2.4 mL) at 0°C was added aluminum chloride (167 mg, 1.26 mmol). To the resulting suspension was slowly added lithium aluminum hydride (3.77 mmol). To a second reaction vessel containing THF (4 mL) was added trans-5-chloro-10,11-dihydro-11-[(methylamino)methyl]-dibenzo[b,j]oxepin-10-carboxylic acid hydrochloride (1)HCl (0.4 g, 1.26 mmol). At -10°C, the contents of the first reaction vessel were added to the second reaction vessel. After 30 minutes the reaction mixture was quenched with water and extracted with toluene. The toluene was dried with magnesium sulfate and evaporated. This provided crude fraA7s-5-chloro-2,3,3a, 12b-tetrahydro-2-methyl-1H-dibenz[2,3;6,7]-oxepino[4,5-c]pyrrole (A) in nearly quantitative yield with a purity of 63%; 1H-NMR (400.13 MHz in CDCl3 relative to TMS) δ 2.54 (s, 3H), 3.15 (m, 2H), 3.25 and 3.61 (2 x m, 4H), 7.01-7.36 (m, 7H).

Example 6. Preparation of asenapine, trans-5-chloro-2,3,3a,12b-tetra-hydro-2-methyl-1H-dibenz[2,3;6,7]oxepino[4,5-c]pyrrole (A)

Aluminum chloride (6.9 kg) is added in portions to tetrahydrofuran (100 L) at 0°C. Stirring is continued and a 10% solution of lithium aluminum hydride in tetrahydrofuran (35.0 L) is added keeping the temperature below 10°C. The mixture is cooled to 0°C and stirred for 15 minutes. A solution of trans-5-chloro-2,3,3a,12b-tetrahydro-2-methyl-1H-dibenzoxepino[4,5-c]pyrrol-1-one (IV) (10.0 kg) in tetrahydrofuran (100 L) is added to the mixture while keeping the temperature below 15°C. Tetrahydrofuran (5 L) is added, stirring is continued for 1 hour at 10°C, and 0.6 N sodium hydroxide solution (100 L) is added keeping the temperature below 15°C. Stirring is continued for 15 minutes at 20°C and the salt layer is separated. The salt layer is extracted with toluene (2 x 50 L). The combined filtrates are separated and the aqueous layer is extracted with of toluene (50 L). The combined organic layers are evaporated in vacuum at 70°C to give trans-5-chloro-2,3,3a,12b-tetrahydro-2-methyl-1H-dibenz[2,3;6,7]oxepino[4,5-c]pyrrole (A) (3.4 kg, 99%); 1H-NMR (400.13 MHz in CDCl3 relative to TMS) δ 2.54 (s, 3H), 3.15 (m, 2H), 3.25 and 3.61 (2 x m, 4H), 7.01-7.36 (m, 7H).
D is 5-chloro-2,3-dihydro-2-methyl-1H-dibenz[2,3:6,7]oxepino[4,5-c]pyrrol-1-one
In an alternative method for the ring closure, the amino acid derivative (I) is converted into an activated acid form, such as an acid chloride, which upon neutralization with a base, spontaneously cyclizes to the desired lactam (IV).

A further aspect of the invention relates to a process (Scheme IV) for the preparation of fraA7S-2-chloro-10,11-dihydro-11-[(methylamino)methyl]-dibenzo[b,]oxepine-10-carboxylic acid (II), which is a regioisomer of the amino acid derivative (I) depicted in Scheme III. The process comprises the steps of: (a) reduction of an enamide, 5-chloro-2,3-dihydro-2-methyl-1 H-dibenzo[2, 3:6, 7]-oxepino[4,5-c]pyrrol-1-one (VI), to provide a mixture of lactams, trans-5-chloro-2,3,3a, 12b-tetrahydro-2-methyl-1 H-dibenzo[2, 3:6, 7]-oxepino[4,5-c]pyrrol-1-one (VII), and cis-5-chloro-2,3,3a, 12b-tetrahydro-2-methyl-1 H-dibenzo[2, 3:6, 7]-oxepino[4,5-c]pyrrol-1-one (VIII); (b) hydrolysis of the lactams (VII), (VIII) by treatment in an alcoholic solution with an excess of a strong (alkali) base thereby producing a mixture of the amino acid derivative (II) and its corresponding isomer, cis-2-chloro-10,11-dihydro-11-[(methylamino)methyl]-dibenzo[b,]oxepine-10-carboxylic acid (Ma); and (c) purification of the amino acid derivative (II) from the mixture. Cyclization of the amino acid derivative (II) to the desired lactam (VII), which is a regioisomer of the lactam (IV) depicted in Scheme I, above, and subsequent reduction of the lactam amide group yields asenapine (A).
The last steps in a known process for the preparation of asenapine (A) are depicted in Scheme I. In this process, the double bond in the enamide, 11-chloro-2,3-dihydro-2-methyl-1H-dibenzo[2,3;6,7]oxepino[4,5-c]pyrrol-1-one (III), is reduced by treatment with magnesium in methanol/toluene to produce a mixture of a desired trans-isomer, 11-chloro-2,3,3a,12b-tetrahydro-2-methyl-1H-dibenzo[2,3,6,7]oxepino[4,5-c]pyrrol-1-one (IV), and an unwanted cis-isomer, cis-11-chloro-2,3,3a,12b-tetrahydro-2-methyl-1H-dibenzo[2,3,6,7]oxepino[4,5-c]pyrrol-1-one (V), in a 1:4 ratio.

**11-chloro-2,3-dihydro-2-methyl-1H-dibenzo[2,3;6,7]oxepino[4,5-c]pyrrol-1-one.** The label (III) should not be a part of the name.
Synthesis and biological properties of 2-, 5-, and 6-fluoronorepinephines.

Kirk KL, Cantacuzene D, Nimitkitpaisan Y, McCulloh D, Padgett WL, Daly JW, Creveling CR.

Abstract

2-Fluoro-, 5-fluoro- and 6-fluorodimethoxybenzaldehydes were prepared by photochemical decomposition of the corresponding diazonium fluoroborates. The aldehydes were converted to the cyanohydrin trimethylsilyl ethers, which, in turn, were reduced to the dimethoxyphenethanolamines. Boron tribromide demethylation afforded the racemic ring-fluorinated norepinephines. An alternate route, using the dibenzyloxyfluoroaldehyde, was also used to prepare 6-fluoronorepinephrine. The fluorine substituent markedly increases the phenolic acidities of these analogues. The biological properties conferred upon norepinephrine by the fluorine substituents in peripheral and central adrenergically responsive systems clearly demonstrate that 2-fluoronorepinephrine is a nearly a pure beta-adrenergic agonist, while 6-fluoronorepinephrine is an alpha-adrenergic agonist. 5-Fluoronorepinephrine retains both beta- and alpha-adrenergic agonist properties. Receptor-binding studies with specific radiolabeled ligands indicate that the specificity conferred by the site of fluorine substituents results from a change in the affinity of these analogues for the alpha- and beta-adrenergic receptors.
Traditional Chemical Text Mining

Automated Indexing

- Limited accuracy for entity extraction
- Quicker turnaround time
- False positives inevitable (as a result of wrong tags)
- Erroneous search results

Diagram: Tokenization → Stemming → Stop word removal → Named Entity recognition → Ontology Mapping
Manual Indexing

- Accuracy of ~99%
- More specific annotations
- Context specific information
- Time consuming

Diagram:
- Categorizations
  - Entity tagging
  - Ontology Mapping
Hybrid Approach

Step 1

MC proprietary Text mining technology

Step 2

Skilled MS and PhDs

Categorization

Manual Validation
Solutions for Chemical Text Mining

<table>
<thead>
<tr>
<th>MC Offerings</th>
<th>Abstracting &amp; Indexing</th>
<th>Analytics &amp; Visualization</th>
<th>Content Enrichment</th>
<th>Integration with Native KI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tagging resources/expertise</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Ontology development/mapping</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Tools &amp; workflow</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

- Automated, Manual and Hybrid tagging expertise
- Ontology Development and use of existing public ontologies
- Development of Tools and Workflow for deep Indexing with flexibility to incorporate external text-mining tools and ontologies
Thank You