

Text Mining for Chemical Information: The ChiKEL Project

David Milward & Peter Corbett, Linguamatics
Daniel Bonniot de Ruisselet, ChemAxon

ChemAxon UGM, 28th Sept 2011

Overview



- The ChiKEL Project
- Existing ChemAxon & Linguamatics Integration
- Phase 1
 - Name-to-Structure Improvements/Evaluation
 - Integration
 - Clustering
- Phase 2
 - Chemical Drawing and Visualization
 - Research Themes

ChiKEL

Chemically Informed Knowledge Extraction from Literature

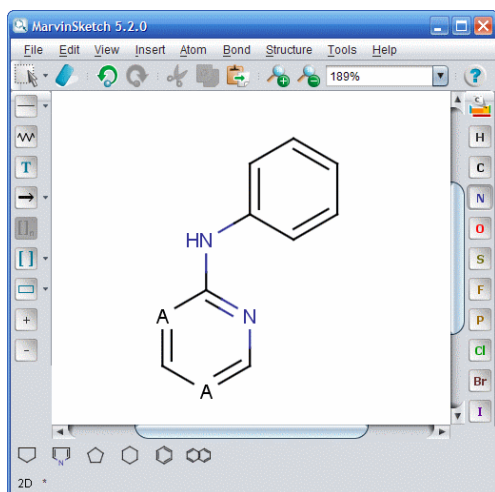


- Motivations
 - Text mining is well established in drug discovery e.g.
 - protein-protein interactions, gene disease associations, biomarker discovery
 - Growing use in other applications
 - Widening document sources
 - Patents being used for:
 - Scientific knowledge
 - Intellectual property
 - Competitive intelligence
 - Want to be able to combine NLP-based text mining with chemical search and visualisation based on structure
- Partners
 - Linguamatics, ChemAxon
- 2yr project supported by European Union EUREKA's Eurostars Programme

Existing ChemAxon & Linguamatics Integration



- Finds mentions of chemicals using dictionary matching
- ChemAxon substructure and similarity search used to find chemicals within terminologies via their SMILES representations



N(C1=CC=CC=C1)C1=*C=*C=N1

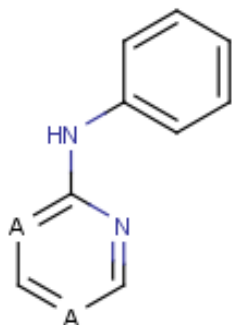
The image is a screenshot of the "Class Properties" dialog box in ChemAxon software. The dialog has two tabs: "Chooser" and "Advanced". The "Advanced" tab is active, showing a list of chemical classes. The list includes "molecular entity", "trial product", "phenylpropanoid", "stilbenoid", "stilbene", "4,4'-bis(4-anilino-6-[bis(2-hydroxyethyl)amino]-1,3,5-triazin-2-yl)amino stilbene-2,2'-disulfonic acid", "inorganic entities", "organohalogen compound", "organofluorine compounds", "nilotinib", "molecular entities", "molecular entities", "oxide salt", "organic chloride salt", "hydrochloride", and "pirenzepine hydrochloride". The "nilotinib" class is highlighted. At the bottom, there is a search bar with the text "Look for: N(C1=CC=CC=C1)C1=*C=*C=N1" and a dropdown menu set to "Substructure". There are also dropdown menus for "in:" and "SMILES", and a "Search" button.

Chemical Text Mining



ChemAxon

- Ability to efficiently answer precise queries e.g.
 - What chemicals with this substructure act as inhibitors



Class1	Relation	Class2	Doc	Hit
▶ imatinib	▶ inhibit	ABL1	▶ 2 16303243	1 Imatinib, an inhibitor of BCR-ABL tyrosine kinase, also inhibits BCRP-mediated drug transport.
▶ imatinib methanesulfonate	▶ inhibit	ABL1	▶ 3 15803362	1 BACKGROUND: Imatinib mesylate is a potent inhibitor of Abl, KIT, and PDGFR tyrosine kinases.
▶ gefitinib	▶ inhibit	EGFR	▶ 2 15692759	▶ 2 The clinical benefit and safety of the EGFR tyrosine kinase inhibitor gefitinib ('Iressa') ¹ was evaluated in this Phase II, multicentre study of patients with taxane and anthracycline pretreated, metastatic breast cancer.
▶ lapatinib	▶ inhibit	ERBB2	▶ 3 16452223	▶ 2 Alternatively, inhibition of ErbB2 signaling using lapatinib (GW572016), a reversible small-molecule inhibitor of ErbB1/ErbB2 tyrosine kinases, at pharmacologically relevant concentrations, leads to marked inhibition of survivin protein with subsequent apoptosis.

Chemical Text Mining (2)



- Ability to efficiently answer precise queries e.g.
 - Find IC50 values from unstructured text

IC50	< Numerics to + / - Nu..	Doc	Hit
IC50	< 500 nM	▶ 2 US7628989	▶ 2 ... for HLA Class I an IC 50 of 500 nM or less, often 200 nM or ...
IC50	< 1000 nM	▶ 2 US7628989	▶ 2 ... and for Class II an IC 50 of 1000 nM or less.
IC50	< 5000 nM	▶ 2 US7628989	1 ... i.e., bind at an IC 50 of 5000 nM or less, to three of more ...

- Find EC50 and IC50 values from tables within documents

Dengue virus	Vero	IL-29	0.032	µg/ml	>10	µg/ml	Dengue virus Vero IL-29 0.032 µg/ml >10 µg/ml
Dengue virus	Vero	MetIL-29C172S-	0.0075	µg/ml	>10	µg/ml	Dengue virus Vero MetIL-29C172S- 0.0075 µg/ml >10 µg/ml
Venezuelan	Vero	IL-28A	0.01	µg/ml	>10	µg/ml	Venezuelan Vero IL-28A 0.01 µg/ml >10 µg/ml
Venezuelan	Vero	IL-29	0.012	µg/ml	>10	µg/ml	Venezuelan Vero IL-29 0.012 µg/ml >10 µg/ml
Venezuelan	Vero	MetIL-29C172S-	0.0065	µg/ml	>10	µg/ml	Venezuelan Vero MetIL-29C172S- 0.0065 µg/ml >10 µg/ml

How is this achieved?



- Wider set of tools than traditional search:
 - Co-occurrence at different levels
 - Document, Paragraph, Sentence, N-words apart ...
 - Abstract, Methods, Description, Background, Claim ...
 - Natural language processing
 - Precise patterns that can determine precise roles e.g. products of a reaction, inhibitor of an enzyme
 - Regular expression matching to find amounts, dosages, concentrations etc.
- Wider set of possible outputs:
 - Lists, Assertions, Visualizations
 - Standardized IDs
 - to allow integration with knowledge stored in databases, or semantic web

Dictionary Matching



- Uses a terminology such as ChEBI or Jochem containing chemicals and their synonyms e.g.

Cyclosporine

- Ciclosporin
- CSA
- Neoral
- OL 27-400

- Shown value for a variety of drug discovery projects, but unsuitable for finding novel compounds

- can classify systematic names as chemicals e.g.

... 2-chloro-p-phenylenediamine ... 1-methyl-2-hydroxy-4-aminobenzene ...

- but often only understands part of the name

2-**chloro**-p-phenylenediamine

Chloro group

1-4-phenylenediamine

1-**methyl**-2-hydroxy-4-aminobenzene

Methyl group

aniline

ChiKEL: Phase 1



- Evaluate and Improve Name-to-Structure
- Integrate ChemAxon Name-to-Structure with I2E
 - ensures can find and understand novel chemicals e.g. in patent documents
- Investigate Clustering Techniques

Types of chemical names supported:



- IUPAC
 - 4-hydrazino-5-methyl-1H-pyridin-2-one
- CAS names
 - pyridin-2-one, 4-hydrazino-5-methyl-, 1H-
- Common names
 - Creatine
- Drug names (generic and proprietary)
 - Paracetamol, doliprane
- Abbreviations
 - ATP

Areas of improvement



- Name to Structure:
 - Systematic names
 - Dictionary
- Document to Structure:
 - OCR error correction
 - Punctuation, brackets
 - Eg: amoxicillin-resistant

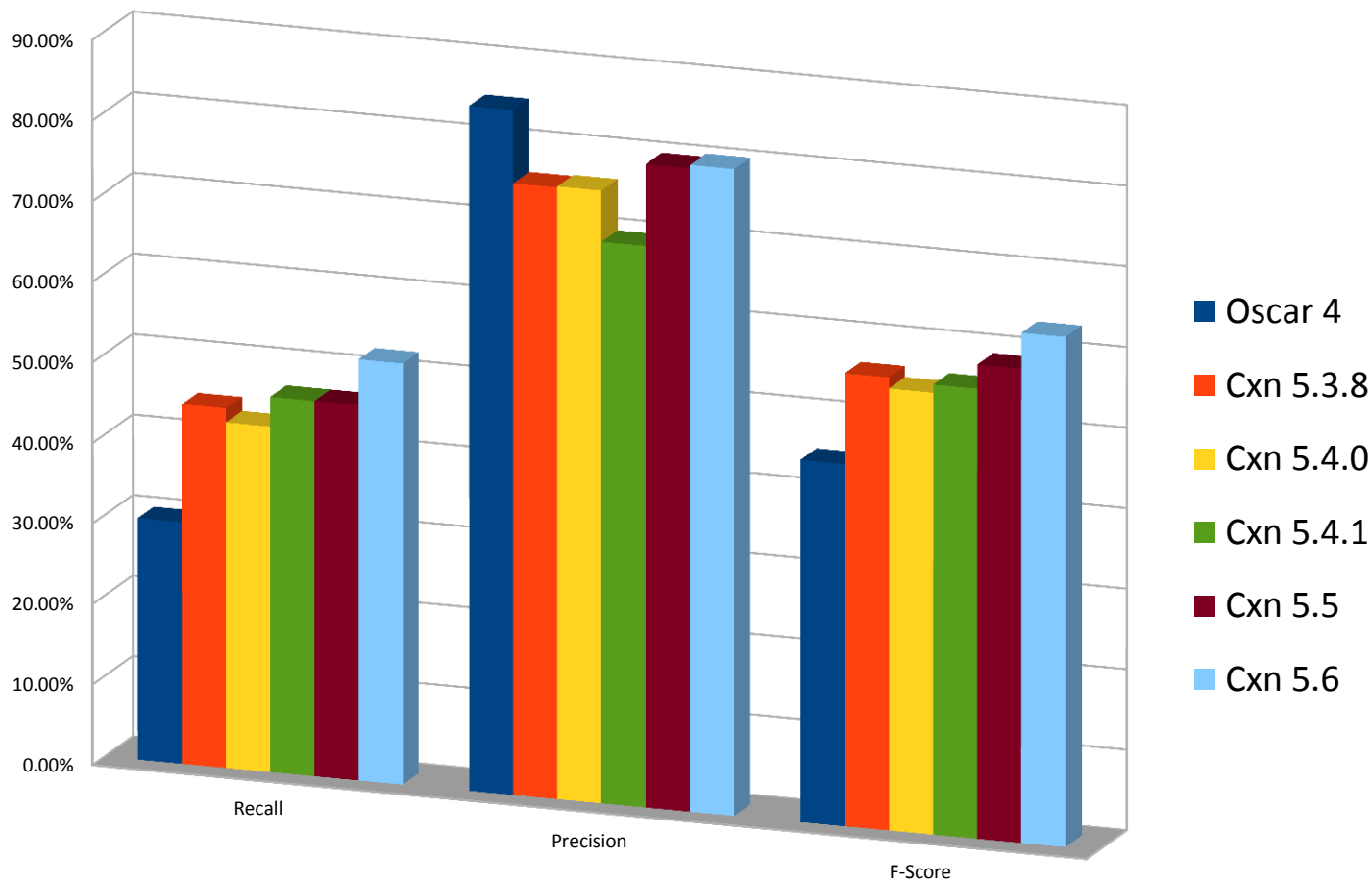
The efficacy of the clinically available beta-lactam/beta-lactamase inhibitor combinations (amoxicillin/clavulanic acid (CA), ticarcillin/CA, amoxicillin/sulbactam, and piperacillin/tazobactam) was evaluated on amoxicillin-resistant Escherichia coli isolates having the main patterns of beta-lactam resistance.

Evaluation



- Existing gold standard annotation:
 - Manually annotated MEDLINE abstracts (SCAI corpus)
- New gold standard for patents being developed:
 - Develop annotation guidelines:
 - Annotate polymers, copolymers ...
 - Do not annotate chemical groups e.g. methyl ...
 - Manually annotate sampled paragraphs to create a gold standard
- Evaluate the software against the gold standard
- Evaluate quality (recall and precision)
- Drive development
- Measure progress

Evaluation on SCAI Corpus

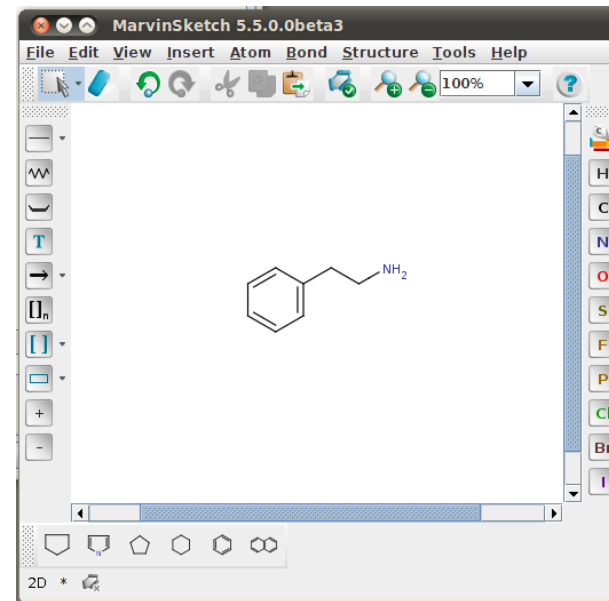


Integration with I2E



- Substructure & similarity search (as before)
- Chemicals found by name-to-structure as well as by dictionary matching
- Terminology created on the fly, with different matches brought together as a single chemical concept via an ID, either
 - SMILES
 - InChi
 - InChi Key

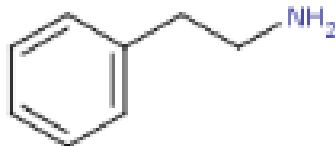
Preferred names for each chemical concept use either the common name (if available), or traditional name via structure-to-name



Finding Chemicals in Patents



ChemAxon



(2S)-2-(methylamino)-3-phenylpropanoic acid	1 US 07202269 B2	2 ... added to a suspension of (S)-2-methylamino-3-phenyl-propionic acid (0.2 g, 1.12 ...
1-[(1S)-1-[4-(benzyloxy)phenyl]-2-(4-methylpiperazin-1-yl)ethyl]cyclohexan-1-ol	1 US 07550456 B2	1 1-[(1S)-1-[4-(benzyloxy)phenyl]-2-(4-methylpiperazin-1-yl)ethyl]cyclohexanol;
2-(3,5-dichlorophenyl)-8a-({4-[2-(hydroxymethyl)phenyl]phenyl)methyl}-tetrahydroimidazolidino[1,5-a]pyridine-1,3-dione	1 US 06897225 B1	1 ...) To a solution of 6-[4-[2-(hydroxymethyl)phenyl]benzyl]-8-(3,5-dichlorophenyl)-1,8-diazabicyclo[4.3.0]nonane-7,9-dione (0.318 g) in ...
2-amino-5-[4-(difluoromethoxy)phenyl]-5-phenyl-3H-imidazol-4-one	1 US 07723368 B2	1 The present invention relates to amino-5-[4-(difluoromethoxy)phenyl]-5-phenylimidazolone compounds, which are inhibitors ...
8a-({4-[2-(chloromethyl)phenyl]phenyl)methyl}-2-(3,5-dichlorophenyl)-tetrahydroimidazolidino[1,5-a]pyridine-1,3-dione	1 US 06897225 B1	1 ... dried under vacuum to give 6-[4-[2-(chloromethyl)phenyl]benzyl]-8-(3,5-dichlorophenyl)-1,8-diazabicyclo[4.3.0]nonane-7,9-dione.
9-(4-bromobutyl)-N-(2,2,2-trifluoroethyl)thioxanthene-9-carboxamide	1 US 07030120 B2	1 (d) 4-[9-(2,2,2-Trifluoroethylcarbamoyl)-9H-thioxanthene-9-yl]butyl bromide was synthesized using the compound ...
(2R)-2-{methyl[(4-phenylphenyl)carbamoyl]amino}-3-phenylpropanoic acid	1 US 07202269 B2	1 A. (R)-2-(3-Biphenyl-4-yl-1-methyl-ureido)-3-phenyl-propionic acid.
codeine	1 US 07772224 B2	1 ... ; an antitussive such as codeine, hydrocodone, caramiphen, ...
dextromethorphan	1 US 07772224 B2	1 ... caramiphen, carbetapentane, or dextromethorphan; a diuretic; a ...
epinephrine	1 US 07772224 B2	1 ... , pseudoephedrine, oxymetazoline, epinephrine, naphazoline, xylometazoline, ...

Filtering to Chemical Products



- Query restricted to chemicals mentioned as the product of a reaction e.g. "gives 2.4g of ..."

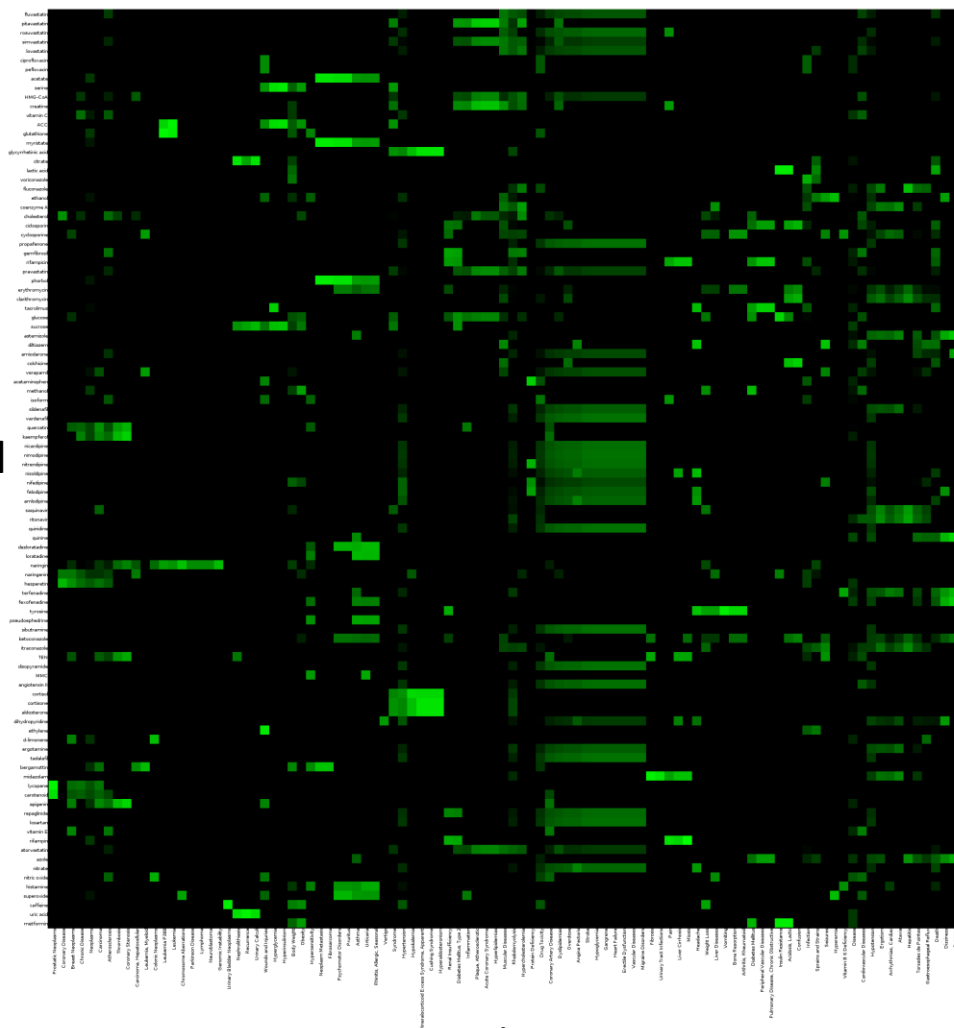
Jchem	Doc	Hit
ethyl 2-(naphthalen-2-yl)-2-(4-nitroimidazol-2-yl)acetate	1 US 06992097 B2	2 ... :3:0.2) giving 30.99 grams of ethyl 2-(2-naphthyl)-2-(4-nitroimidazolyl)acetate which was 90% pure by :3:0.1) to give 0.90 grams (46%) of ethyl 2-(2-naphthyl)-2-(4-nitroimidazolyl)acetate as a tan oil.
2-(4-fluorophenyl)-5-methyl-4-({[3-(prop-2-en-1-yloxy)cyclohexyl]oxy)methyl}-1,3-oxazole	1 US 07335671 B2	1 This gives 2.4 g of 4-(3-allyloxycyclohexyloxymethyl)-2-(4-fluorophenyl)-5-methyloxazole as a yellowish oil.
8a-({4-[2-(chloromethyl)phenyl]phenyl)methyl}-2-(3,5-dichlorophenyl)-tetrahydroimidazolidino[1,5-a]pyridine-1,3-dione	1 US 06897225 B1	1 ... residue was dried under vacuum to give 6-[4-[2-(chloromethyl)phenyl]benzyl]-8-(3,5-dichlorophenyl)-1,8-diazabicyclo[4.3.0]nonane-7,9-dione.
methyl 2-[3-(2-hydroxyethyl)phenoxy]methyl]benzoate	1 US 07521461 B2	1 ... solvent was removed by evaporation to give 1.99 g of methyl 2-[[3-(2-hydroxyethyl)phenoxy]methyl]benzoate (yield 90%).

Clustering



ChemAxon

Chemical



Disease

- Heatmap of chemical-disease associations
- Association using text mining
- Clustering of chemicals via maximal common substructure
- Agglomerative clustering of diseases

ChiKEL: Phase 2



- Continuing improvements to Name-to-Structure
 - Develop gold standard data further
- Integrating:
 - Chemical drawing
 - Chemical structures in results
- Research
 - Image recognition
 - Automated clustering
 - Markush search integration

The screenshot displays a web-based interface for chemical structure clustering. At the top, there are tabs for "All clusters", "No singletons", and "Only singletons". Below these are sorting options: "Natural", "Desc. size", and "Asc. size". There are also icons for "Image size", "Columns", "Rows", and "Show/hide". A navigation bar shows "1-15" and "16-30 31-45 46-57 58-72 73-87 88-102". The main content is a grid of 15 chemical structure groups, labeled (1) through (15). Each group contains one or more chemical structures and a "Size" label with a percentage. For example, Group 1 has a size of 3 (2.3%), Group 2 has a size of 2 (1.5%), and Group 15 has a size of 2 (1.5%).

Group ID	Chemical Structure	Size
(1) Group ID: 1		Size: 3 (2.3 %)
(2) Group ID: 2		Size: 2 (1.5 %)
(3) Group ID: 3		Size: 2 (1.5 %)
(4) Group ID: 4		Size: 2 (1.5 %)
(5) Group ID: 5		Size: 2 (1.5 %)
(6) Group ID: 6		Size: 2 (1.5 %)
(7) Group ID: 7		Size: 2 (1.5 %)
(8) Group ID: 8		Size: 2 (1.5 %)
(9) Group ID: 9		Size: 2 (1.5 %)
(10) Group ID: 10		Size: 2 (1.5 %)
(11) Group ID: 11		Size: 2 (1.5 %)
(12) Group ID: 12		Size: 2 (1.5 %)
(13) Group ID: 13		Size: 2 (1.5 %)
(14) Group ID: 14		Size: 2 (1.5 %)
(15) Group ID: 15		Size: 2 (1.5 %)

Contacts



- If you have interest in:
 - providing requirements
 - early access
 - more information

- Please contact:

chikel@linguamatics.com