Simple Project - Current Activities
Current activity: “in line” entity tagging & classification of chemical names

1. Identify every chemical name.
2. Convert all chemical names into their chemical structures [SMILES] - then convert these SMILES into inchi’s & Inchkeys (a unique identifier for the chemical).
3. Annotate/augment all chemical names with the term “inchikey & the unique inchikey” for that chemical. The InChiKeys are now indexed as-if they were words (text) in the document.
4. Re-index the augmented text [inchikeys] w SOLR.

Text = Chemical
= Target
= Disease
= Assay data

Add the derived structures + annotations (& meta data) to our master database.

Text Index + Annotation Index

SOLR index
Current activity: “in line” entity tagging & classification of chemical names

1. Identify every chemical name.
2. Convert all chemical names into their chemical structures [SMILES] - then convert these SMILES into inchi’s & Inchkeys (a unique identifier for the chemical).
3. Annotate /augment all chemical names with the term “inchikey & the unique inchikey” for that chemical. The InChiKeys are now indexed as if they were words (text) in the document.
4. Re-index the augmented text [inchikeys] with SOLR.
5. Add the derived structures + annotations (& meta data) to our master database.

ChemAxon
[Name= Structure]
Chem Libraries
J Chem Base

- = Chemical
- = Target
▲ = Disease
= Assay data

= cc(=O)OC1=CC=CC=C1C(=O)O

= BSYNRYMUTXBXSQ-UHFFFAOYSA-N

Text Index + Annotation Index

SOLR index
Current activity: “in line” entity tagging & classification for targets (=geneid’s)

- Identify all targets [Gene names & their synonyms]
- Augment all target names with a “tag = geneid” & the NCBI unique Identifier # for that target
- Re-index the augmentented text + geneid identifiers w SOLR

Add the derived annotations (& meta data) to our master database
Current activity: “in line” entity tagging & classification for MeSH terms

Identify all known MeSH terms [for example, diseases (C01) or signs & symptoms (C23)]

Identified & augment every occurrence of every MeSH term with a ‘tag = MeSH & the specific MeSh code Identifier’

Re-index the augmented text + the MeSh tags w SOLR

Text = Headache

Text + Annotation

= Headache +
= MeSH term +
= C23 sign or symptom

New index of original text plus all of it’s associated annotated information

Add the derived annotations (& meta data) to our master database

SOLR index
“Interactions of ibogaine and D-amphetamine: in vivo microdialysis and motor behavior in rats. Ibogaine, an indolalkylamine, has been proposed for use in treating stimulant addiction. In the present study we sought to determine if ibogaine had any effects on the neurochemical and motor changes induced by D-amphetamine that would substantiate the anti-addictive claim. Ibogaine (40 mg/kg, i.p.) injected 19 h prior to a D-amphetamine challenge (1.25 mg/kg, i.p.) potentiated the expected rise in extracellular dopamine levels in the striatum and in the nucleus accumbens, as measured by microdialysis in freely moving rats. Using ... ”
Examples - why this is important and what it enables us to do that we could not easily do before -
Batch Analysis

For Example: You are about to file a patent application – that contains ~300 – 400 chemical compounds. How do you know if any of these (400+) compounds has been patented before?
Batch Search Demo

Paste a list of InChI keys to be batch searched here!
Input list of InChIkeys to be batch searched here!

Click run search!
One can readily search hundreds or even thousands of compounds at at time – to see if any of the compounds have already been patented - & by whom & for what purpose.
Exploring co-table analysis of Molecules with Gene ID’s

For example – show me all of the co-occurrences of these (x) molecules with these (any / all) gene’s !
From the main menu select the Analyze tab
2. From the analyze menu select the Cotable tab!
Now Enter the Inchi keys for the molecules of interest -

Click here to enter a sample (test) set of molecules
Now select - patent field – to explore “patents”!

These are the molecules of interest – (Inchi keys to explore)

Select Patent field here
Now select - facet = patent field + Gene then click analyze
These are the NCBI Gene ID #’s

To transpose the charts or export the data – click here

This shows the “cotable” results = co-occurrences of molecules + NCBI –Gene ID’s
This shows the transposed chart – of co-occurrences of molecules + NCBI –Gene ID’s

Click here to see the patents containing this molecule + this particular gene
Co-table Analysis

For example: Show me all documents where imitrex was Mentioned with “any” .....sign and / or symptoms

(note: these are terms such as headache, vomiting, nausea ..etc ..there are > 680 of them).
1. Draw a compound of interest
2. Click – view compound in co-table
Draw a compound of interest

ChemAxon
Marvin
[Name= Structure]
Chem Libraries
J Chem Base
Chemical Search
Select a MeSH category for Co-occurrence analysis

Click analyze
This shows the number of documents that contained the source molecule and ANY of the MeSH – C23 terms.

Click on the numbers to “link to” the documents.
Type in a new MeSH code to change the analysis from ‘signs & symptoms’ (C23) to diseases (C01).
This shows the number of documents that contained the source molecule and ANY of the MeSH – disease (C01) terms.
This shows the comparison of 2 drugs and the co-occurrence of MeSH Symptoms (C23) terms.
### Chemical Structures vs. Signs and Symptoms

This shows the comparison of different statins and the co-occurrence of MeSH terms.

<table>
<thead>
<tr>
<th>Class name</th>
<th>Class Size</th>
<th>Compound 1</th>
<th>Compound 2</th>
<th>Compound 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stroke</td>
<td>51</td>
<td>5</td>
<td>33</td>
<td>23</td>
</tr>
<tr>
<td>Subarachnoid Hemorrhage</td>
<td>25</td>
<td>8</td>
<td>7</td>
<td>19</td>
</tr>
<tr>
<td>Muscular Diseases</td>
<td>21</td>
<td>2</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td>Brain Ischemia</td>
<td>19</td>
<td>7</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>Pain</td>
<td>17</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Vasospasm, Intracranial</td>
<td>16</td>
<td>7</td>
<td>5</td>
<td>13</td>
</tr>
<tr>
<td>Alzheimer Disease</td>
<td>13</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Muscle Weakness</td>
<td>12</td>
<td>2</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Carotid Artery Diseases</td>
<td>10</td>
<td>7</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Carotid Stenosis</td>
<td>10</td>
<td>3</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Cerebral Hemorrhage</td>
<td>10</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Ischemic Attack, Transient</td>
<td>9</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Memory Disorders</td>
<td>9</td>
<td>3</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Hypoxia-Ischemia, Brain</td>
<td>7</td>
<td>2</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Cerebral Infarction</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Intracranial Aneurysm</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Cerebral Infarction</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Encephalomyelitis, Autoimmune, Experimental</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Brain Injuries</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Infarction, Middle Cerebral Artery</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Encephalomyelitis, Autoimmune, Experimental</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Learning Disorders</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Brain Injuries</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Multiple Sclerosis</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Myasthenia Gravis</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Brain Edema</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Dementia, Vascular</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Encephalitis</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Dementia, Vascular</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Dementia, Vascular</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Demyelinating Diseases</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Back Pain</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Headache</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Vision Disorders</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
Computer Curation Process Overview & integration with our collaborators -

Services Hosted at IBM Almaden

Data Sources
- U.S. Patents (1976—2009)
- U.S. Pre-Grants (All)
- PCT & EPO Apps
- Medline Abstracts (>18 M)

Selected Internet Content

Annotation Factory
- Parse & Extract data
- (Semantic Associations)
- e Classifier & Other Data Associations
- Database + computed Meta Data
- Computational Analytics
- Annotator 1
- Annotator 2
- ADU*

User Applications
- ChemVerse
- Knime or Pipeline Pilot
- BIW
- Cognos/DDQB/Other Apps
- Chem Axon Search
- SIMPLE

* ADU = Automated Data Update

View selected Documents & Reports
Screen shoots from our SIMPLE / SIIP Web application
Clustering

BioTerm Analysis

Claims Originality

Discovery
Landscape Analysis

Visualization

Networks
I would like to acknowledge the IBM Almaden Research – team

Jeff Kreulen
Ying Chen
Scott Spangler
Alfredo Alba
Tom Griffin
Eric Louie
Su Yan
Issic Cheng
Prasad Ramachandran
Bin He
Ana Lelescu
Qi He
Linda Kato
Ana Lelescu
Brad Wade
John Colino
Meenakshi Nagarajan
Timothy J Bethea
German Attanasio
Laura Anderson
Robert Prill

+ a host of folks from IBM China Labs -