Chemistry and reactions from non-US patents

Daniel Lowe and Roger Sayle

NextMove Software

Cambridge, UK
TOPICS

• Coverage of USPTO vs EPO patents

• Extraction of non-chemical-compound data

• Can text-mining provide insights into reaction informatics
WHAT AM I MISSING BY ONLY USING THE USPTO DATA?
EPO AND USPTO BACKGROUND

• USPTO: 303k grants published in 2013

• EPO: 67k grants published in 2013

• EPO patents must be in one or more of English, French or German

• USPTO documents are all in English
**EPO/USPTO Compounds**

(using StdInChI)

- **USPTO** = 1976-2013 patent grants + 2001-2013 patent applications
- **EPO** = 1978-2013 patent grants and applications

- USPTO: 3,345,877
- EPO: 445,588
- USPTO and EPO overlap: 2,674,069

248th ACS National Meeting, San Francisco CA, USA 10th August 2014
EPO/USPTO REACTIONS
(using separately sorted StdInChIs for reactants/agents and products)

USPTO  = 1976-2013 patent grants + 2001-2013 patent applications
EPO    = 1978-2013 patent grants and applications
EFFECT OF DIFFERENT LANGUAGES

Translation performed by Lexichem v2014.Jun
FIRST DISCLOSURE OCCURRENCE BY LANGUAGE

All compounds

Excluding compounds disclosed by USPTO (1976-2013) patents
TIMELINESS

• Competitive intelligence requires up to date information

• Methodology:
  – Consider compounds present in both EPO and USPTO date not disclosed by either prior to 2006
  – Compare the earliest publication date for each compound
AVERAGE LAG BETWEEN EPO/USPTO

Mean: US 540 days earlier

Compounds first disclosed between 2006-2013
Applications vs Applications

Mean: US 223 days earlier

Compounds first disclosed between 2006-2013
GRANTS VS GRANTS

Mean: US 287 days earlier

Compounds first disclosed between 2006-2013
Oxindolylquinazoline derivatives as angiogenesis inhibitors

US 6294532 B1

ABSTRACT

The invention relates to compounds of formula (I) wherein: ring Z is, for example, a 6-membered aromatic heterocyclic ring containing 1 or 2 nitrogen atoms; R<1> represents hydrogen, C1-alkyl, C1-alkoxymethyl, di(C1-alkoxy)methyl or C1-alkanoyl, R<2> represents, for example, halogen, trifluoromethyl, 2,2,2-trifluoroethyl, cyano, nitro, C1-alkylsulphonyl, carbamoyl, <u>N</u>≡<u>C</u₁-<i>alk</i>yl carbamoyl, <u>N</u>≡<u>C</u₁-<i>alk</i>ylsulphonyl or <u>N</u>≡<u>N</u>≡<u>C</u₁-<i>alk</i>ylaminosulphonyl, <i>n</i> is an integer from 0 to 3, <i>m</i> is an integer from 0 to 4, and R<3> represents, for example, hydroxy, halogen, nitro, trifluoromethyl, C1-alkyl, or substituted C1-alkyl.
EPO/US 1976-2013 AND CHEMBL19 OVERLAP

1,155,034

187,486

6,278,048

Patents

CheMBL19
Average lag between patents/Chembl19

Mean: Patents 345 days earlier

Compounds first disclosed between 2006-2013
OTHER DATA IN PATENTS

- Genes/proteins
- Reactions including role of reagents
- Physical quantities
- Spectra
- Diseases
- Organisms
- Companies
- ...
GENE/PROTEIN IDENTIFICATION

• Identify trends in drug target popularity

• Count number of patents (in a year) mentioning a given gene or its gene products and map to the HGNC symbol

• Many genes are referenced by short ambiguous terms hence great care is required to have good precision
GENE/PROTEIN IDENTIFICATION

**EGFR**

(Epidermal growth factor receptor)

**MET**

(Hepatocyte growth factor receptor)

- **US grants**
- **US applications**
- **ChEMBL 19**
GENE/PROTEIN IDENTIFICATION

(Mammalian target of rapamycin)  (Tissue plasminogen activator)

US grants
US applications
ChEMBL 19
MELTING/BOILING POINT EXTRACTION

2,3,5,6-Tetrafluoro-4-((2-nitrophenyl)ethynyl)pyridine (12c)

[0128]

[0129] A solution of Trimethyl((2-nitrophenyl)ethynyl) silane, (11c) (1.10 g, 5.02 mmol) in DMF (10 mL) was added to the mixture of pentafluoropyridine (1.08 g, 6.53 mmol) and CsF(1.10 g, 7.52 mmol) in DMF (10 mL) slowly. The reaction mixture was stirred overnight. Brine (30 mL) and dichloromethane (60 mL) were added. Organic phase was separated and washed with water (30 mL×3). Solvent was evaporated by rotary evaporation and the residue was chromatographed (Hexane, then EtOAc:Hexane =1:30) to provide the desired product in 77% yield: $^1$H-NMR (300 MHz, CDCl$_3$) δ 8.24 (dd, J=7.8, 1.5 Hz, 1H), 7.85 (dd, J=7.5, 1.5 Hz, 1H), 7.73 (ddd, J=7.5, 7.5, 1.5 Hz, 1H), 7.66 (ddd, J=7.8, 7.8, 1.5 Hz, 1H); $^{13}$C-NMR (150 MHz, CDCl$_3$) δ 149.7, 143.7 (dm, J=234.1 Hz), 142.1 (dm, J=264.7 Hz), 135.5, 133.5, 131.2, 125.3, 116.7 (t, J=16.3 Hz), 116.2, 100.9 (t, J=3.2 Hz), 79.8 (t, J=4.1 Hz); HRMS (Cl+): calcd for C$_{13}$H$_2$F$_4$N$_2$O$_2$ [M+H]$^+$ 297.02872, found 297.0291 F. m.p. 109-110°C.
# Melting/Boiling Point Extraction

<table>
<thead>
<tr>
<th>#</th>
<th>Structure</th>
<th>SMILES</th>
<th>Paragraph</th>
<th>Patent</th>
<th>Quantity Type</th>
<th>Unit</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>93359</td>
<td><img src="image1.png" alt="Structure Image" /></td>
<td>(9S)-2,5-Diamino-4-(3-((3-(2-fluorophenyl)amino)ethyl)indol-4-yl)ethyl</td>
<td>0126</td>
<td>US20120069940A1</td>
<td>melting</td>
<td>g/mol</td>
<td>&gt;200</td>
</tr>
<tr>
<td>93360</td>
<td><img src="image2.png" alt="Structure Image" /></td>
<td>3-((3-(2-fluorophenyl)amino)ethyl)indol-4-yl)</td>
<td>0129</td>
<td>US20120069940A1</td>
<td>melting</td>
<td>g/mol</td>
<td>108-110</td>
</tr>
</tbody>
</table>
Unleashing over a million reactions into the wild

Posted on February 27, 2014 by daniel

Unlike with small molecules, there are currently no large sets of publically available reaction data.

To remedy this situation, we have extracted over a million reactions from United States patent applications (2001-2013) and the same again from patent grants (1976-2013). This contrasts to the original data release of “only” 420 thousand (from 2008-2011 applications) whilst I was in the PMR group.

The reactions are available as reaction SMILES or CML from here, as 7zip archives. The CML representation includes quantities and yields where these were found. A documentation zip provides further information on the format of the data. This data is made available under CC-Zero i.e. without copyright.
PREDICTING YIELD

• Features to consider:
  • 2D fingerprints (especially around the reaction centers)
  • Reaction type
  • Temperature
  • Time
  • Change in complexity e.g. chiral centres
DATA EXTRACTION

- Can pull out yields, quantities, times and temperatures
- Can sanity check text-mined yield by calculating from amounts (or even masses)

\[
\frac{\text{mass of compound (g)}}{\text{molar mass [calc from structure] (gmol}^{-1})} = \text{mols of compound}
\]

\[
\frac{\text{mols of product}}{\text{mols of limiting reactant}} = \% \text{yield}
\]
Preparation of 3-fluoro-4-morpholinyl aniline

[0049] 10% Pd—C 4.0 g was added to 3-fluoro-4-morpholinyl nitrobenzene (40 g, 177 mmol), ammonium formate (50 g, 793 mmol) in 200 mL of ethyl acetate and stirred at 4550°C for 8 h until the completion of the reaction. The mixture was then filtrated and separated by water. The organic layer was washed with brine and dried over anhydrous magnesium sulfate, filtered, and the solvent was evaporated to provide 33 g of solid in 95% yield.
SCALE VS YIELD

2001-2013 US applications, Suzuki couplings
TEMPERATURES
IDENTIFY SYNTHETIC ROUTES

Number of steps

Occurrences

<table>
<thead>
<tr>
<th>Number of steps</th>
<th>Intermediates</th>
<th>Terminal Products</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>197702</td>
<td>385149</td>
</tr>
<tr>
<td>2</td>
<td>103114</td>
<td>149445</td>
</tr>
<tr>
<td>3</td>
<td>56611</td>
<td>81837</td>
</tr>
<tr>
<td>4</td>
<td>31403</td>
<td>47579</td>
</tr>
<tr>
<td>5</td>
<td>17268</td>
<td>27670</td>
</tr>
<tr>
<td>6</td>
<td>9230</td>
<td>16619</td>
</tr>
<tr>
<td>7</td>
<td>5057</td>
<td>9320</td>
</tr>
<tr>
<td>8</td>
<td>2701</td>
<td>5263</td>
</tr>
<tr>
<td>9</td>
<td>1256</td>
<td>2511</td>
</tr>
<tr>
<td>10</td>
<td>639</td>
<td>1330</td>
</tr>
<tr>
<td>11</td>
<td>301</td>
<td>678</td>
</tr>
<tr>
<td>12</td>
<td>136</td>
<td>373</td>
</tr>
<tr>
<td>13</td>
<td>58</td>
<td>111</td>
</tr>
<tr>
<td>14</td>
<td>15</td>
<td>63</td>
</tr>
<tr>
<td>15</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>17</td>
<td></td>
<td>5</td>
</tr>
</tbody>
</table>

248th ACS National Meeting, San Francisco CA, USA 10th August 2014
Number of steps vs complexity

Average Complexity Score

\[
\text{ringComplexityScore} = \log_{10}(n\text{RingBridgeAtoms} + 1) + \log_{10}(n\text{SpiroAtoms} + 1)
\]

\[
\text{stereoComplexityScore} = \log_{10}(n\text{StereoCenters} + 1)
\]

\[
\text{macrocyclePenalty} = \log_{10}(n\text{Macrocycles} + 1)
\]

\[
\text{sizePenalty} = \text{natoms}^{1.005} - \text{natoms}
\]

Ertl & Schuffenhauer 2009

doi:10.1186/1758-2946-1-8
Number of steps vs complexity

Average number of heavy atoms
YIELD VS CHANGE IN COMPLEXITY

2001-2013 US applications, all reactions
TRENDS IN SOLVENT USE

Percentage of reactions in that year


Solvents:
- Tetrahydrofuran
- Dichloromethane
- Water
- Dimethylformamide
- Methanol
- Ethyl acetate
- Ethanol
- 1,4-Dioxane
- Toluene
- Acetonitrile
- Acetic acid
- Chloroform
- Acetone
- Benzene
### Are Solvents Getting Greener?

<table>
<thead>
<tr>
<th>Solvent</th>
<th>1976</th>
<th>2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water</td>
<td>(21%)</td>
<td>Tetrahydrofuran (15%)</td>
</tr>
<tr>
<td>Ethanol</td>
<td>(11%)</td>
<td>Dichloromethane (14%)</td>
</tr>
<tr>
<td>Benzene</td>
<td>(8%)</td>
<td>Water (13%)</td>
</tr>
<tr>
<td>Methanol</td>
<td>(7%)</td>
<td>Dimethylformamide (10%)</td>
</tr>
<tr>
<td>Tetrahydrofuran</td>
<td>(5%)</td>
<td>Methanol (8%)</td>
</tr>
<tr>
<td>Dichloromethane</td>
<td>(4%)</td>
<td>Ethyl acetate (7%)</td>
</tr>
<tr>
<td>Dimethylformamide</td>
<td>(4%)</td>
<td>Ethanol (5%)</td>
</tr>
<tr>
<td>Acetic acid</td>
<td>(4%)</td>
<td>1,4-Dioxane (4%)</td>
</tr>
<tr>
<td>Chloroform</td>
<td>(3%)</td>
<td>Toluene (3%)</td>
</tr>
<tr>
<td>Acetone</td>
<td>(3%)</td>
<td>Acetonitrile (3%)</td>
</tr>
</tbody>
</table>

Total for top 10: 71% 82%
CONCLUSIONS

- A significant amount of novel chemistry, from EPO patents, comes from the non-English patents

- Compounds disclosed by both the USPTO and EPO are on average published earlier by the USPTO but for many compounds an EPO patent will be the earlier disclosure

- Gene/protein identification can identify clear changes in patenting behaviour over time

- Text mining provides the tools to answer many reaction informatics questions
ACKNOWLEDGEMENTS

• Funding provided by:

![Vernalis Logo]

![Novartis Logo]
Thank you for your time!

http://nextmovesoftware.com
http://nextmovesoftware.com/blog
daniel@nextmovesoftware.com