Presentation of a Structurally Diverse and Commercially Available Drug Data Set for Correlation and Benchmarking Studies

Anders Karlén
Uppsala University
Aim of study

• Derive a “benchmark data set“
  - Drug-like
  - Physicochemically diverse
  - Commercially available and inexpensive
  - Amenable to analytical measurements

• Start the generation of benchmark data
  - Derive good-quality data from the same lab
Possible use of the data set

- General description of drugs
- Developing ADME/TOX filters (permeability, solubility, plasma protein binding etc.)
- To validate novel experimental techniques
Generation of a “benchmark” data set based on the list of drugs in Sweden (FASS 2001)

Remove compounds
- Molecular weight >900
- Polymers, polypeptides
- Inorganic and metal containing

799 cpds -> 691 cpds -> 450 cpds

Select commercially available < $800/g

370 cpds -> 332 cpds -> 284 cpds

Remove “odd” ATC classes e.g. A01(Mouth and teeth), A05(Bile acids), A06 (Laxative)...

24-compound data set
Cost and availability of the 691-compound data set

Histogram

450 of the 691 compounds can be bought. Price range $0.03/gram - $3,228,000/gram (2001)

- Methenamine
- Calcitrol
Principal component analysis

- General descriptors
- General hydrogen bonding descriptors
- Hydrogen bond donor descriptors
- Hydrogen bond acceptor descriptors

$\Sigma 28$ molecular descriptors
Principal component analysis
The factorial design
“A face-centered central composite design”
24-compound data set

Thiamazole (− + +)
Amantadine (− + +)
Carbamazepine (− − −)
Chlorzoxazone (− + −)
Flupenthixol (− − −)
Fenofibrate (− − −)
Meclizine (− − −)
Terfenadine (− − −)

Captopril (0 + 0)
Sulindac (0 0 −)
Chlorprothixene (0 0 0)
Prednisone (0 0 0)
Metoclopramide (0 0 0)
Tetracycline (0 0 0)
Carisoprodol (0 0 +)
Fenofibrate (− − −)
Erythromycin (0 − +)

Levodopa (− + +)
Amiloride (+ + +)
Hydrochlorothiazide (− + −)
Terfenadine (− − +)
Levothyroxine (0 − 0)

The cost of buying the entire data set (at least 1 gram of each compound) is less than $1,500
Comparison of the data sets with respect to some common molecular descriptors

<table>
<thead>
<tr>
<th></th>
<th>691-compound data set</th>
<th>24-compound data set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>MW</td>
<td>60</td>
<td>854</td>
</tr>
<tr>
<td>PSA</td>
<td>0</td>
<td>373</td>
</tr>
<tr>
<td>log(P_{\text{Mor}})</td>
<td>−6.4</td>
<td>7.6</td>
</tr>
<tr>
<td>log(D_{\text{ACD,6.5}})</td>
<td>−10.6</td>
<td>12.3</td>
</tr>
<tr>
<td>HBD</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>HBA</td>
<td>0</td>
<td>19</td>
</tr>
</tbody>
</table>

Neomycin
HBD = 19

Candesartan cilexetil
\(\log P_{\text{Mor}} = 7.6\)
Comparison of the data sets with respect to functional groups
## Comparison of the data sets with respect to ATC classes

The Anatomical Therapeutic Chemical (ATC) classification system is the most commonly used classification system for drug substances.

### Distribution in ATC

<table>
<thead>
<tr>
<th>ATC</th>
<th>Description</th>
<th>24-set</th>
<th>691-set</th>
<th>24-set</th>
<th>691-set</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>GI</td>
<td>1</td>
<td>69</td>
<td>4,2%</td>
<td>9,99%</td>
</tr>
<tr>
<td>B</td>
<td>Blood</td>
<td>0</td>
<td>21</td>
<td>0,0%</td>
<td>3,04%</td>
</tr>
<tr>
<td>C</td>
<td>Cardio</td>
<td>2</td>
<td>89</td>
<td>8,3%</td>
<td>12,88%</td>
</tr>
<tr>
<td>D</td>
<td>Topical</td>
<td>0</td>
<td>36</td>
<td>0,0%</td>
<td>5,21%</td>
</tr>
<tr>
<td>G</td>
<td>Gen.hormones</td>
<td>1</td>
<td>38</td>
<td>4,2%</td>
<td>5,50%</td>
</tr>
<tr>
<td>H</td>
<td>Hormones</td>
<td>3</td>
<td>14</td>
<td>12,5%</td>
<td>2,03%</td>
</tr>
<tr>
<td>J</td>
<td>Infection</td>
<td>5</td>
<td>89</td>
<td>20,8%</td>
<td>12,88%</td>
</tr>
<tr>
<td>L</td>
<td>Tum.,immuno</td>
<td>1</td>
<td>53</td>
<td>4,2%</td>
<td>7,67%</td>
</tr>
<tr>
<td>M</td>
<td>Muscle,mov.</td>
<td>3</td>
<td>37</td>
<td>12,5%</td>
<td>5,35%</td>
</tr>
<tr>
<td>N</td>
<td>Nervous</td>
<td>6</td>
<td>134</td>
<td>25,0%</td>
<td>19,39%</td>
</tr>
<tr>
<td>P</td>
<td>Antiparasite</td>
<td>0</td>
<td>13</td>
<td>0,0%</td>
<td>1,88%</td>
</tr>
<tr>
<td>R</td>
<td>Respiration</td>
<td>1</td>
<td>52</td>
<td>4,2%</td>
<td>7,53%</td>
</tr>
<tr>
<td>S</td>
<td>Eye,ear</td>
<td>1</td>
<td>24</td>
<td>4,2%</td>
<td>3,47%</td>
</tr>
<tr>
<td>V</td>
<td>Various</td>
<td>0</td>
<td>22</td>
<td>0,0%</td>
<td>3,18%</td>
</tr>
</tbody>
</table>
Start the generation of benchmark data. Derive good-quality data from the same lab

1. Measurment of pKa by pH-metric or pH-UV technique (n=20)

2. Measurment of lipophilicity
   (a) pH-metric logP (n=18)
   (b) capacity factors by RP-HPLC (n=21)

3. Measurment of intrinsic and kinetic solubility
   pH-metric solubility (CheqSol technique) or shake-plate solubility (n=17)

4. Measurment of permeability across Caco-2 Cells. A to B direction (n=22)
2. Lipophilicity
pH-metric measurement of logP and logD

logP missing for:
• Folic acid
• Carbamazepin
• Prednisone
• Carisoprodol

logP (neutral)

logD (pH 7.4)
2. Lipophilicity

Experimental logP vs calculated logP

- Crippen logP: $R^2 = 0.70$
- ACD/LogP: $R^2 = 0.88$
- ClogP (BioByte): $R^2 = 0.89$
- Moriguchi logP: $R^2 = 0.80$
2. Lipophilicity

Correlation between the measured HPLC capacity factor ($k$) and pH-metric log $D$ (pH 6.8)

- Compounds from the 8 corner points have different colors
- The 2 compounds at each corner point have the same color
- The axis points are colored black
- Center point pink

$$R^2 = 0.92$$
3. Solubility

Measurement of intrinsic solubility using CheqSol
(24-compound data set)

Solubility ranges from 0.009 $\mu$g/ml to 2119 $\mu$g/ml
3. Solubility

19 of the compounds studied also present in the 691-compound data set

CheqSol solubility ranges from 0.9 μg/mL to 3500 μg/mL in these 19 compounds

In the 24-compound data set the solubility ranges from 0.009 μg/ml to 2119 μg/ml

<table>
<thead>
<tr>
<th>Name</th>
<th>Equilibrium solubility</th>
<th>Kinetic Solubility</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CheqSol</td>
<td>Shake-Flask</td>
</tr>
<tr>
<td>1 Phthalic Acid</td>
<td>5330</td>
<td>5950</td>
</tr>
<tr>
<td>2 Quinine</td>
<td>363</td>
<td>201</td>
</tr>
<tr>
<td>3 Trazodone</td>
<td>134.6</td>
<td>138.0</td>
</tr>
<tr>
<td>4 Nitrofurantoin</td>
<td>112.5</td>
<td>109.5</td>
</tr>
<tr>
<td>5 Nortriptyline</td>
<td>27.0</td>
<td>49.3</td>
</tr>
<tr>
<td>6 Verapamil</td>
<td>48.5</td>
<td>48.5</td>
</tr>
<tr>
<td>7 Niflumic Acid</td>
<td>9.53</td>
<td>29.5</td>
</tr>
<tr>
<td>8 Imipramine</td>
<td>17.2</td>
<td>21.7</td>
</tr>
<tr>
<td>9 Flumequine</td>
<td>34.2</td>
<td>20.7</td>
</tr>
<tr>
<td>10 Furosemide</td>
<td>19.7</td>
<td>20.4</td>
</tr>
<tr>
<td>11 Maprotiline</td>
<td>5.80</td>
<td>8.05</td>
</tr>
<tr>
<td>12 Piroxicam</td>
<td>5.92</td>
<td>5.95</td>
</tr>
<tr>
<td>13 Warfarin</td>
<td>5.30</td>
<td>5.25</td>
</tr>
<tr>
<td>14 Chlorpromazine</td>
<td>2.70</td>
<td>2.41</td>
</tr>
<tr>
<td>15 Lidocaine</td>
<td>3500</td>
<td>3810</td>
</tr>
<tr>
<td>16 Famotidine</td>
<td>740</td>
<td>1100</td>
</tr>
<tr>
<td>17 Hydrochlorothiazide</td>
<td>630</td>
<td>700</td>
</tr>
<tr>
<td>18 Chlorpheniramme</td>
<td>608.3</td>
<td>615.2</td>
</tr>
<tr>
<td>19 Sulfamerazine</td>
<td>200.3</td>
<td>203.0</td>
</tr>
<tr>
<td>20 Ketoprofen</td>
<td>130.6</td>
<td>178.0</td>
</tr>
<tr>
<td>21 Propranolol</td>
<td>81.0</td>
<td>70.0</td>
</tr>
<tr>
<td>22 Ibuprofen</td>
<td>50.0</td>
<td>49.0</td>
</tr>
<tr>
<td>23 Pindolol</td>
<td>41.7</td>
<td>32.7</td>
</tr>
<tr>
<td>24 Miconazole</td>
<td>1.00</td>
<td>0.67</td>
</tr>
<tr>
<td>25 Diclofenac</td>
<td>0.90</td>
<td>0.80</td>
</tr>
<tr>
<td>26 Amodiaquin</td>
<td>0.41</td>
<td></td>
</tr>
<tr>
<td>27 Pamoic acid</td>
<td>0.0003</td>
<td></td>
</tr>
</tbody>
</table>

Compound not present in the 691 data set
24-compound data set is structurally diverse
4. Permeability/absorption

![Graph showing permeability comparison between Caco-2 and human jejunum cells.]
4. Permeability/absorption

In vitro $P_{\text{app}}$ values in human Caco-2 cells
Suggestions on the “Uppsala diverse data set” usage

- The 24 compounds can be used
  - as a test set for testing already derived models of permeability, lipophilicity, solubility etc.
  - as a validation set for new experimental techniques
  - on its own for building and validating models by dividing it into a training set and a test set

We hope that other groups are willing to help us to supplement the herein-started characterization

“Bench mark data set”

*J. Med. Chem.; (ASAP); 2006; 49(23); 6660-6671*
Acknowledgements

Faculty of Pharmacy
Uppsala University
Christian Sköld
Torbjörn Lundstedt
Anders Hallberg
Hans Lennernäs

Sirius Analytical Instruments Ltd
John Comer
Karl Box
Ruth Allen
Jon Mole

AstraZeneca R&D Mölndal
Susanne Winiwarter
Anna-Lena Ungell
Johan Wernevik
Fredrik Bergström
Leif Engström