Drug discovery

Modern therapeutic research
From serendipity to rationalized drug design

Ancient Greeks treat infections with mould

Biapenem in PBP-1A
Drug discovery process

1. Find a target
   - Protein that we want to inhibit so as to interfere with a biological process

2. Identify hits
   - Compounds likely to bind to the target

3. Hit-to-lead: characterize hits
   - Can they be drugs? (ADME-Tox)

4. Lead optimization and synthesis
   - bioactivity
   - pharmacokinetics
   - synthetic pathway

5. Assay
   - in vitro
   - in vivo
   - clinical
Drug discovery process

1. Find a target
2. Identify hits
3. Hit-to-lead: characterize hits
4. Lead optimization and synthesis
5. Assay
Drug discovery process

1. Find a target
2. Identify hits
3. Hit-to-lead: characterize hits
4. Lead optimization and synthesis
5. Assay

$500,000,000 to $2,000,000,000
52 months 90 months
How can computer science help?

→ Chemoinformatics!

“...the mixing of information resources to transform data into information, and information into knowledge, for the intended purpose of making better decisions faster in the arena of drug lead identification and optimisation.” – F. K. Brown

“... the application of informatics methods to solve chemical problems.”

– J. Gasteiger and T. Engel
1. Find a target
2. Identify hits
3. Hit-to-lead: characterize hits
4. Lead optimization and synthesis
5. Assay
The chemical space

- $10^{60}$ possible small organic molecules
- $10^{22}$ stars in the observable universe

(Slide courtesy of Matthew A. Kayala)
Drug discovery process

1. Find a target
2. Identify hits
3. Hit-to-lead: characterize hits
4. Lead optimization and synthesis
5. Assay

**QSAR:** Qualitative Structure-Activity Relationship
i.e. classification

**QSPR:** Quantitative Structure-Property Relationship
i.e. regression
Representing chemicals *in silico*

Expert knowledge molecular descriptors
→ hard, potentially incomplete

Molecules are...
Similar Property Principle
Molecules having similar structures should exhibit similar activities.

→ Structure-based representations
Compare molecules by comparing substructures
Molecular graph

Undirected labeled graph
Define **feature vectors** that record the presence/absence (or number of occurrences) of particular patterns in a given molecular graph

\[
\phi(A) = (\phi_s(A))_{s \text{ substructure}}
\]

where

\[
\phi_s(A) = \begin{cases} 
1 & \text{if } s \text{ occurs in } A \\
0 & \text{otherwise}
\end{cases}
\]

| 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | ...

**Extension of traditional chemical fingerprints**
Learning from fingerprints

Classical machine learning and data mining techniques can be applied to these vectorial feature representations.

- Any distance / kernel can be used
- Classification
- Feature selection
- Clustering
Fingerprints compression

- Systematic enumeration $\rightarrow$ long, sparse vectors
  e.g. 50,000 random compounds from ChemDB
  $\rightarrow$ 300,000 paths of length up to 8
  $\rightarrow$ 300 non-zeros on average

- “Naive” Compression
  - List the positions of the 1s
  - $2^{19} = 524,288$
  - average encoding: $300 \times 19 = 5,700$ bits
Fingerprints compression

Modulo Compression (lossy)
MOLFEA [Helma et al., 2004]

Data Mining and Machine Learning Techniques for the Identification of Mutagenicity Inducing Substructures and Structure Activity Relationships of Noncongeneric Compounds

Christoph Helma,*† Tobias Cramer,† Stefan Kramer,†‡ and Luc De Raedt†

Institute for Computer Science, Machine Learning Lab, University Freiburg, Georges Köhler Allee 79, D-79110 Freiburg/Br., Germany, and Institute for Computer Science, Technical University Munich, Boltzmannstrasse 3, D-85748 Garching, Germany

- $\mathcal{P} =$ positive (mutagenic) compounds
- $\mathcal{N} =$ negative compounds
- features: fragments (= patterns) $f$ such that both $freq(f, \mathcal{P}) \geq t$ and $freq(f, \mathcal{N}) \geq t$
- Limited to frequent linear patterns
- ML algorithm: SVM with linear or quadratic kernel
MOLFEA [Helma et al., 2004]

- CPDB – Carcinogenic Potency DataBase
- 684 compounds classified in 341 mutagens and 343 non-mutagens according to Ames test on *Salmonella*

![Chemical structures](image)

![Graph showing mutagenicity prediction](image)

Mutagenicity prediction [Hema04]

- Linear kernel
- Quadratic kernel

Cross-validated sensitivity

Frequency threshold

1% 3% 5% 10%

50 60 70 80 90 100
Spectrum kernels

\[ \phi(A) = (\phi_s(A))_{s \in S} \]

\[ K_{\text{spectrum}}(A, A') = k(\phi(A), \phi(A')) \]

\[ k \in \mathbb{R}^{|S| \times |S|} \] can be

- Dot product (linear kernel)
- RBF kernel
- Tanimoto kernel: \[ k(A, B) = \frac{A \cap B}{A \cup B} \]
- MinMax kernel: \[ \frac{\sum_{i=1}^{N} \min(A_i, B_i)}{\sum_{i=1}^{N} \max(A_i, B_i)} \]
Spectrum kernels

Tanimoto and MinMax

Both Tanimoto and Minmax are kernels.


- Proof for MinMax:

  \[
  \text{MinMax}(x, y) = \frac{\langle \phi(x), \phi(y) \rangle}{\langle \phi(x), \phi(x) \rangle + \langle \phi(y), \phi(y) \rangle - \langle \phi(x), \phi(y) \rangle}
  \]

  with \( \phi(x) \) of length: \# patterns \( \times \) max count

  \( \phi(x)_i = 1 \) iff. the pattern indexed by \( \lfloor i/q \rfloor \) appears more than \( i \mod q \) times in \( x \)
All patterns fingerprints

Paths fingerprints

Labeled sub-paths (walks)

Some sub-paths of length 3
Circular fingerprints

- Labeled sub-trees - Extended-Connectivity (or Circular) features

Example of a circular substructure of depth 2
All patterns fingerprints

2D spectrum kernels [Azencott et al., 2007]

One- to Four-Dimensional Kernels for Virtual Screening and the Prediction of Physical, Chemical, and Biological Properties

Chloé-Agathe Azencott,‡ Alexandre Ksikes,‡ S. Joshua Swamidass,‡ Jonathan H. Chen,‡
Liva Ralaivola,§ and Pierre Baldi*,∥

School of Information and Computer Sciences, University of California—Irvine, Irvine, California 92697-3435

Received September 13, 2006

- Systematically extract paths / circular fingerprints, for various maximal depths
- SVM with Tanimoto / Minmax
All patterns fingerprints

2D spectrum kernels [Azencott et al., 2007]

- Mutagenicity (Mutag): 188 compounds
- Benzodiazepine receptor affinity (BZR): 181 + 125 compounds
- Cyclooxygenase-2 inhibitors (COX2): 178 + 125 compounds
- Estrogen receptor affinity (ER): 166 + 180 compounds

<table>
<thead>
<tr>
<th>Data</th>
<th>SVM</th>
<th>Previous best</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mutag</td>
<td>90.4%</td>
<td>85.2% (gBoost)</td>
</tr>
<tr>
<td>BZR</td>
<td>79.8%</td>
<td>76.4%</td>
</tr>
<tr>
<td>COX2</td>
<td>70.1%</td>
<td>73.6%</td>
</tr>
<tr>
<td>ER</td>
<td>82.1%</td>
<td>79.8%</td>
</tr>
</tbody>
</table>
Goal: scalability

Compute a sequence that captures topological and label information of graphs in a runtime linear in the number of edges

→ sub-tree kernel
Weisfeiler-Lehman kernel

[Shervashidze et al., 2011]

Given labeled graphs $G$ and $G'$

Result of steps 1 and 2: multiset-label determination and sorting

Result of step 3: label compression

Result of step 4: relabeling

End of the 1st iteration

Feature vector representations of $G$ and $G'$

$$
\phi^{(1)}_{WLsubtree}(G) = \begin{pmatrix} 2, 1, 1, 1, 1, 2, 0, 1, 0, 1, 0, 1 \end{pmatrix}
$$

$$
\phi^{(1)}_{WLsubtree}(G') = \begin{pmatrix} 1, 2, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1 \end{pmatrix}
$$

Counts of original node labels

Counts of compressed node labels

$$
k^{(1)}_{WLsubtree}(G, G') = \langle \phi^{(1)}_{WLsubtree}(G), \phi^{(1)}_{WLsubtree}(G') \rangle = 11.
$$
Convolution kernels

- a.k.a. **decomposition kernels**

- \((x_1, \ldots, x_D)\) is a tuple of parts of \(x\), with \(x_d \in X\) for each part \(d = 1, \ldots, D\)

- \(k_d \in \mathbb{R}^{X_d \times X_d}\): a Mercer kernel

\[
K_{\text{decomposition}}(x, x') = \sum_{x_1 x_2 \ldots x_D = x} \sum_{x_1' x_2' \ldots x_D' = x'} k_1(x_1, x_1') k_2(x_2, x_2') \ldots k_D(x_D, x_D')
\]

- Spectrum kernels are a particular case of convolution kernels
Match atoms and weigh them according to a kernel between subgraphs that include these atoms

\[ K_{WDK}(x, x') = \sum_{(a, \sigma \in D_r(x))} \sum_{(a', \sigma' \in D_r(x'))} \delta(a, a')K_c(\sigma, \sigma') \]

\( r > 0 \in \mathbb{N} \)

\( D_r(x) \): decompositions of the molecular graph of \( x \) in an atom \( a \) and a subpath \( \sigma \) of \( x \) including \( a \) and of depth at most \( r \)
Weighted Decomposition Kernel [Menchetti et al., 2005]

\[ K_c(\sigma, \sigma') = \sum_{l \in \mathcal{L}} \min(f^\sigma(l), f^{\sigma'}(l)) \]

- \( K_c \): contextual kernel, here: histogram intersection kernel
- \( \mathcal{L} \): possible labels for edges and vertices
- \( f^\sigma(l) \): frequency of label \( l \) subgraph \( \sigma \).

<table>
<thead>
<tr>
<th></th>
<th>( l = 1 )</th>
<th>( l = 2 )</th>
<th>( l = 3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>MM</td>
<td>70.5±4.3</td>
<td>70.0±5.5</td>
<td>69.9±6.3</td>
</tr>
<tr>
<td>FM</td>
<td>67.4±6.9</td>
<td>68.1±9.7</td>
<td>69.1±5.8</td>
</tr>
<tr>
<td>MR</td>
<td>63.8±6.4</td>
<td>67.8±7.2</td>
<td>68.4±6.3</td>
</tr>
<tr>
<td>FR</td>
<td>61.5±8.1</td>
<td>61.3±7.4</td>
<td>60.4±5.7</td>
</tr>
</tbody>
</table>
Introducing spatial information

3D Histograms [Azencott et al., 2007]

- Groups of $k$ atoms
- Associated size:
  - Pairwise distances ($k = 2$)
  - Diameter of the smallest sphere that contains all $k$ atoms
Introducing spatial information

**3D Histograms** [Azencott et al., 2007]

One histogram per class of $k$-tuple (e.g. C-C-C, C-C-O)

<table>
<thead>
<tr>
<th>N-O distance (Å)</th>
<th>Frequency of N-O</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>
### 3D Histograms: performance [Azencott et al., 2007]

<table>
<thead>
<tr>
<th>Data</th>
<th>2D kernel</th>
<th>Hist3D kernel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mutag</td>
<td>90.4%</td>
<td>88.8%</td>
</tr>
<tr>
<td>BZR (loo)</td>
<td>82.0%</td>
<td>79.4%</td>
</tr>
<tr>
<td>ER (loo)</td>
<td>87.0%</td>
<td>86.1%</td>
</tr>
<tr>
<td>COX2</td>
<td>76.9%</td>
<td>78.6%</td>
</tr>
</tbody>
</table>
Introducing spatial information

3D Decomposition Kernels [Ceroni et al., 2007]

Classification of Small Molecules by Two- and Three-Dimensional Decomposition Kernels

Alessio Ceroni, Fabrizio Costa, Paolo Frasconi

Machine Learning and Neural Networks Group, Dipartimento di Sistemi e Informatica, Università degli Studi di Firenze, Italy. http://www.dsi.unifi.it/neural/

Associate Editor: Prof. Anna Tramontano

Remember: $K_{WDK}(x, x') = \sum_{a, \sigma \in D_r(x)} \sum_{a', \sigma' \in D_r(x')} \delta(a, a') K_c(\sigma, \sigma')$

$K_{3DDK}(x, x') = \sum_{\sigma \in S_r(x)} \sum_{\sigma' \in S_r(x')} K_s(\sigma, \sigma')$

$S_r(x)$: subgraphs of $x$ composed of $r$ distinct vertices

$K_s(\sigma, \sigma') = \prod_{i=1}^{r(r-1)/2} \delta(e_i, e_i') e^{-\gamma(l_i - l_i')}$

$l_i =$ length of edge $e_i$ in $x$

$(e_1, e_2, \ldots, e_{r(r-1)/2}$ lexicographically ordered; $\gamma \in \mathbb{R}$
Introducing spatial information

### 3DDK: Performance [Ceroni et al., 2007]

<table>
<thead>
<tr>
<th>Data</th>
<th>2D kernel</th>
<th>Hist3D kernel</th>
<th>3DDK</th>
<th>Circ3DDK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mutag</td>
<td>90.4%</td>
<td>88.8%</td>
<td>86.7%</td>
<td>83.5%</td>
</tr>
<tr>
<td>BZR (loo)</td>
<td>82.0%</td>
<td>79.4%</td>
<td>78.4%</td>
<td>81.4%</td>
</tr>
<tr>
<td>ER (loo)</td>
<td>87.0%</td>
<td>86.1%</td>
<td>82.3%</td>
<td>82.1%</td>
</tr>
<tr>
<td>COX2</td>
<td>76.9%</td>
<td>78.6%</td>
<td>75.6%</td>
<td>75.2%</td>
</tr>
</tbody>
</table>
The pharmacophore kernel [Mahé et al., 2006]

The Pharmacophore Kernel for Virtual Screening with Support Vector Machines

Pierre Mahé,*,† Liva Ralaivola,‡ Véronique Stoven,‡ and Jean-Philippe Vert†

Center for Computational Biology, Ecole des Mines de Paris, 35 rue Saint Honoré, 77305 Fontainebleau, France, and Laboratoire d’Informatique Fondamentale, University Provence/Aix-Marseille, 139 rue Joliot-Curie, F-13453 Marseille Cedex 13, France

Received April 13, 2006

- **pharmacophore** \( p \in \mathcal{P}(x) \): \( p = [(x_1, l_1), (x_2, l_2), (x_3, l_3)] \)
- \( x_i \) 3D coordinates of atom \( i \) of \( x \); \( l_i \) = label of atom \( i \)

\[
K(x, x') = \sum_{p \in \mathcal{P}(x)} \sum_{p' \in \mathcal{P}(x')} K_P(p, p')
\]

- \( K_P(p, p') = K_{dist}(d_1, d'_1)K_{dist}(d_2, d'_2)K_{dist}(d_3, d'_3)K_{feat}(l_1, l'_1)K_{feat}(l_2, l'_2)K_{feat}(l_3, l'_3) \)
- \( K_{dist}: \text{RBF Gaussian} \ K_{dist}(d, d') = \exp\left(\frac{||d-d'||^2}{2\sigma^2}\right) \)
- \( K_{feat}: \text{Dirac} \)
Introducing spatial information

3D LAP kernel [Hinselmann et al., 2010]

$M$: pairwise intramolecular matrix of inter-atomic geometric distances

<table>
<thead>
<tr>
<th>Kernel</th>
<th>ACE</th>
<th>ACHE</th>
<th>BZR</th>
<th>CBG</th>
<th>GPB</th>
<th>HIA*</th>
<th>PGF2A</th>
<th>THR</th>
</tr>
</thead>
<tbody>
<tr>
<td>DotProduct</td>
<td>1.60 ± 0.59</td>
<td>1.24 ± 0.50</td>
<td>0.64 ± 0.40</td>
<td>0.48 ± 0.29</td>
<td>0.62 ± 0.40</td>
<td><strong>86.1</strong> ± <strong>8.69</strong>°</td>
<td>0.71 ± 0.46</td>
<td>0.68 ± 0.39</td>
</tr>
<tr>
<td>2D-LAP(CK)</td>
<td>1.67 ± 0.63</td>
<td>0.95 ± 0.39</td>
<td>0.62 ± 0.43</td>
<td>0.29 ± 0.21</td>
<td><strong>0.46</strong> ± <strong>0.35</strong></td>
<td>81.50 ± 8.70</td>
<td>0.57 ± 0.55</td>
<td><strong>0.48 ± 0.26</strong></td>
</tr>
<tr>
<td>2D-LAP(OA)</td>
<td>1.81 ± 0.65</td>
<td><strong>0.82 ± 0.33</strong></td>
<td>0.60 ± 0.41</td>
<td><strong>0.22 ± 0.18</strong></td>
<td><strong>0.39 ± 0.33</strong>°</td>
<td>85.09 ± 8.33</td>
<td>0.49 ± 0.40</td>
<td><strong>0.41 ± 0.26</strong>°</td>
</tr>
<tr>
<td>3D-LAP(CK)</td>
<td>2.04 ± 0.78</td>
<td><strong>0.76 ± 0.36</strong></td>
<td><strong>0.61 ± 0.40</strong></td>
<td><strong>0.21 ± 0.17</strong></td>
<td>0.53 ± 0.41</td>
<td>82.66 ± 8.78</td>
<td><strong>0.29 ± 0.24</strong></td>
<td>0.58 ± 0.36</td>
</tr>
<tr>
<td>3D-LAP(OA)</td>
<td>1.91 ± 0.71</td>
<td><strong>0.74 ± 0.33</strong></td>
<td>0.60 ± 0.38</td>
<td><strong>0.16 ± 0.16</strong>°</td>
<td><strong>0.49 ± 0.40</strong></td>
<td>83.50 ± 8.34</td>
<td><strong>0.28 ± 0.19</strong></td>
<td>0.59 ± 0.37</td>
</tr>
</tbody>
</table>
Introducing spatial information

Conclusion

- How relevant is 3D information?
- How good is 3D information?
Drug discovery process

1. Find a target
2. Identify hits
3. Hit-to-lead: characterize hits
4. Lead optimization and synthesis
5. Assay

Docking
Virtual High-Throughput Screening
High-throughput screening

- Assay a large library of potential drugs against their target
- Very costly
  - $\rightarrow$ docking
  - $\rightarrow$ virtual high-throughput screening (vHTS)
Measuring performance

Imbalanced data

- Typically, most compounds are inactive ⇒ many more negative than positive examples

- E.g. DHFR data set:
  99,995 chemicals screened for activity against dihydrofolate reductase; < 0.2% active compounds

- Accuracy is not appropriate:
  predicting all compounds negative ⇒ accuracy = 99.8%

- sensitivity = $\frac{\text{# True Positives}}{\text{# Positives}}$

- specificity = $\frac{\text{# True Negatives}}{\text{# Negatives}}$

- For many methods, the output is continuous 
  ⇒ accuracy, sensitivity and specificity depend on a threshold $\theta$
Measuring performance

Receiver-Operator Characteristic Curves

- For all possible values of $\theta$, report \textit{sensitivity} and $1 - \text{specificity}$.
- AUROC (Area under the ROC Curve) is a numerical measure of performance.
- AUROC(random) = 0.5 and AUROC(optimal) = 1

<table>
<thead>
<tr>
<th>label</th>
<th>prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>0.95</td>
</tr>
<tr>
<td>-</td>
<td>0.94</td>
</tr>
<tr>
<td>+</td>
<td>0.90</td>
</tr>
<tr>
<td>+</td>
<td>0.81</td>
</tr>
<tr>
<td>-</td>
<td>0.73</td>
</tr>
<tr>
<td>-</td>
<td>0.52</td>
</tr>
<tr>
<td>-</td>
<td>0.20</td>
</tr>
<tr>
<td>+</td>
<td>0.17</td>
</tr>
<tr>
<td>-</td>
<td>0.12</td>
</tr>
<tr>
<td>-</td>
<td>0.09</td>
</tr>
</tbody>
</table>
Inhibition of DHFR: ROC Curves [Azencott et al., 2007]

<table>
<thead>
<tr>
<th>method</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRV</td>
<td>0.71</td>
</tr>
<tr>
<td>SVM</td>
<td>0.59</td>
</tr>
<tr>
<td>kNN</td>
<td>0.59</td>
</tr>
<tr>
<td>MAX-SIM</td>
<td>0.54</td>
</tr>
</tbody>
</table>
Measuring performance

Precision-recall curves

- Precision = \( \frac{\# \text{ True Positives}}{\# \text{ Predicted Positives}} \)
- Recall = sensitivity

![Graph showing precision-recall curves with points at (0.17, 0.09), (0.2, 0.0), (0.52, 0.02), (0.73, 0.08), (0.81, 0.04), (0.9, 0.02), (0.94, 0.01), and (0.95, 0.01) for perfect and real cases.]
Other applications of graph mining in chemoinformatics

- Database indexing and search
- Prediction of 3D structures of small compounds and proteins
- Reaction Prediction


