In silico chemogenomics with Support Vector Machines

Jean-Philippe Vert
Institut Curie - U900 INSERM - Mines ParisTech

Ligand-based virtual screening / QSAR

From http://cactus.nci.nih.gov
Represent each molecule as a vector...
...and discriminate with machine learning

- LDA
- PLS
- Neural network
- Nearest neighbour
- SVM, …
Support Vector Machine (SVM)

- Large margin
- Nonlinear
- Need pairwise distance / similarity as input instead of vectors / fingerprints
From fingerprints to similarities

**Representation**
- Vectors / Fingerprints
- Tanimoto
- Inner product « Kernel »
- Pairwise distance / similarity

**Discrimination**
- Neural Net
- LDA
- Decision trees
- PLS, …
- SVM
- Kernel PLS
- Kernel LDA
- …
Example: 2D fragment kernel

« All linear fragments »


« All subtree patterns »

Example: 3D pharmacophore kernel

\[ K(x, y) = \sum_{p_x \in \mathcal{P}(x)} \sum_{p_y \in \mathcal{P}(y)} \exp(-\gamma d(p_x, p_y)) . \]

<table>
<thead>
<tr>
<th>Kernel</th>
<th>BZR</th>
<th>COX</th>
<th>DHFR</th>
<th>ER</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D (Tanimoto)</td>
<td>71.2</td>
<td>63.0</td>
<td>76.9</td>
<td>77.1</td>
</tr>
<tr>
<td>3D fingerprint</td>
<td>75.4</td>
<td>67.0</td>
<td>76.9</td>
<td>78.6</td>
</tr>
<tr>
<td>3D not discretized</td>
<td>76.4</td>
<td>69.8</td>
<td>81.9</td>
<td>79.8</td>
</tr>
</tbody>
</table>

Summary so far…

• SVM is an algorithm for supervised classification
• SVM can be used with any « classical » vector or fingerprint description (often giving state-of-the-art performance)
• SVM can also be used with more general measures of similarity (like many related kernel methods)
• Much effort recently to define such kernels in bio- and chemo-informatics
Chemogenomics

Chemical space

Target family
In silico Chemogenomics

Chemical space

Target family
Fingerprint for a (target, molecule) pair?

\[ \Phi_{tar}(t) = \begin{cases} 
- \text{Sequence} \\
- \text{Structure} \\
- \text{Evolution} \\
- \text{Expression} \\
- \ldots 
\end{cases} \]

\[ \Phi_{lig}(c) = \begin{cases} 
- \text{2D} \\
- \text{3D} \\
- \text{Pharmacophore} \\
- \text{logP, \ldots} 
\end{cases} \]

\[ \Phi(c, t) = ??? \]
Fingerprint for a (target,molecule) pair?

\[ T = \Phi_{\text{tar}}(t) = \begin{cases} \text{Sequence} \\ \text{Structure} \\ \text{Evolution} \\ \text{Expression} \\ \ldots \end{cases} \]

\[ c = \Phi_{\text{lig}}(c) = \begin{cases} \text{-2D} \\ \text{-3D} \\ \text{-Pharmacophore} \\ \text{-logP} \end{cases} \]

\[ \Phi(c,t) = \Phi_{\text{lig}}(c) \otimes \Phi_{\text{tar}}(t) \]

\[ 10^6 \quad 10^3 \quad 10^3 \]
Similarity for (target,molecule) pairs

\[ K_{\text{target}}(t, t') = \begin{cases} 
-\text{Sequence} \\
-\text{Structure} \\
-\text{Evolution} \\
-\text{Expression} \\
\ldots
\end{cases} \]

\[ K_{\text{ligand}}(c, c') = \begin{cases} 
-\text{2D} \\
-\text{3D} \\
-\text{Pharmacophore} \\
-\text{logP}, \ldots
\end{cases} \]

\[ K \left( (c, t), (c', t') \right) = K_{\text{target}}(t, t') \times K_{\text{ligand}}(c, c') \]
Summary: SVM for chemogenomics

1. Choose a kernel (similarity) for targets
2. Choose a kernel (similarity) for ligands
3. Train a SVM model with the product kernel for (target/ligand) pairs
Application: virtual screening of GPCR

**Data:** GLIDA database filtered for drug-like compounds
- 2446 ligands
- 80 GPCR
- 4051 interactions
- 4051 negative interactions generated randomly

**Ligand similarity**
- 2D Tanimoto
- 3D pharmacophore

**Target similarities**
- 0/1 Dirac (no similarity)
- Multitask (uniform similarity)
- GLIDA’s hierarchy similarity
- Binding pocket similarity (31 AA)

(Jacob et al., *BMC Bioinformatics*, 2008)
Results (mean accuracy over GPCRs)

<table>
<thead>
<tr>
<th>~K_{tar} \backslash K_{lig} ~</th>
<th>~2D Tanimoto ~</th>
<th>~3D pharmacophore</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dirac</td>
<td>86.2 ± 1.9</td>
<td>84.4 ± 2.0</td>
</tr>
<tr>
<td>multitask</td>
<td>88.8 ± 1.9</td>
<td>85.0 ± 2.3</td>
</tr>
<tr>
<td>hierarchy</td>
<td>93.1 ± 1.2</td>
<td>88.5 ± 2.0</td>
</tr>
<tr>
<td>binding pocket</td>
<td>90.3 ± 1.9</td>
<td>87.1 ± 2.3</td>
</tr>
</tbody>
</table>

5-fold cross-validation

<table>
<thead>
<tr>
<th>~K_{tar} \backslash K_{lig} ~</th>
<th>~2D Tanimoto ~</th>
<th>~3D pharmacophore</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dirac</td>
<td>50.0 ± 0.0</td>
<td>50.0 ± 0.0</td>
</tr>
<tr>
<td>multitask</td>
<td>56.8 ± 2.5</td>
<td>58.2 ± 2.2</td>
</tr>
<tr>
<td>hierarchy</td>
<td>77.4 ± 2.4</td>
<td>76.2 ± 2.2</td>
</tr>
<tr>
<td>binding pocket</td>
<td>78.1 ± 2.3</td>
<td>76.6 ± 2.2</td>
</tr>
</tbody>
</table>

Orphan GPCRs setup

(Jacob et al., *BMC Bioinformatics*, 2008)
Influence of the number of known ligands

Number of ligands / GPCR

Performance improvement (hierarchy vs Dirac)

(Jacob et al., BMC Bioinformatics, 2008)
Screening of enzymes, GPCRs, ion channels

Data: KEGG BRITE database, redundancy removed

<table>
<thead>
<tr>
<th>Enzymes</th>
<th>GPCRs</th>
<th>Ion channels</th>
</tr>
</thead>
<tbody>
<tr>
<td>-675 targets</td>
<td>-100 targets</td>
<td>-114 targets</td>
</tr>
<tr>
<td>-524 molecules</td>
<td>-219 molecules</td>
<td>-462 molecules</td>
</tr>
<tr>
<td>-1218 interactions</td>
<td>-399 interactions</td>
<td>-1165 interactions</td>
</tr>
<tr>
<td>-1218 negatives</td>
<td>-399 negatives</td>
<td>-1165 negatives</td>
</tr>
</tbody>
</table>

(Jacob and V., *Bioinformatics*, 2008)
## Results (mean AUC)

<table>
<thead>
<tr>
<th>$K_{tar}$ \ Target</th>
<th>Enzymes</th>
<th>GPCR</th>
<th>Channels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dirac</td>
<td>0.646 ± 0.009</td>
<td>0.750 ± 0.023</td>
<td>0.770 ± 0.020</td>
</tr>
<tr>
<td>Multitask</td>
<td>0.931 ± 0.006</td>
<td>0.749 ± 0.022</td>
<td>0.873 ± 0.015</td>
</tr>
<tr>
<td>Hierarchy</td>
<td><strong>0.955 ± 0.005</strong></td>
<td><strong>0.926 ± 0.015</strong></td>
<td><strong>0.925 ± 0.012</strong></td>
</tr>
<tr>
<td>Mismatch</td>
<td>0.725 ± 0.009</td>
<td>0.805 ± 0.023</td>
<td>0.875 ± 0.015</td>
</tr>
<tr>
<td>Local alignment</td>
<td>0.676 ± 0.009</td>
<td>0.824 ± 0.021</td>
<td>0.901 ± 0.013</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$K_{tar}$ \ Target</th>
<th>Enzymes</th>
<th>GPCR</th>
<th>Channels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dirac</td>
<td>0.500 ± 0.000</td>
<td>0.500 ± 0.000</td>
<td>0.500 ± 0.000</td>
</tr>
<tr>
<td>Multitask</td>
<td>0.902 ± 0.008</td>
<td>0.576 ± 0.026</td>
<td>0.704 ± 0.026</td>
</tr>
<tr>
<td>Hierarchy</td>
<td><strong>0.938 ± 0.006</strong></td>
<td><strong>0.875 ± 0.020</strong></td>
<td><strong>0.853 ± 0.019</strong></td>
</tr>
<tr>
<td>Mismatch</td>
<td>0.602 ± 0.008</td>
<td>0.703 ± 0.027</td>
<td>0.729 ± 0.024</td>
</tr>
<tr>
<td>Local alignment</td>
<td>0.535 ± 0.005</td>
<td>0.751 ± 0.025</td>
<td>0.772 ± 0.023</td>
</tr>
</tbody>
</table>

10-fold CV

Orphan setting

(Jacob and V., *Bioinformatics*, 2008)
Influence of the number of known ligands

- Enzymes
- GPCRs
- Ion channels

Relative improvement: hierarchy vs Dirac

(Jacob and V., Bioinformatics, 2008)
Conclusion

• SVM offer state-of-the-art performance in chemo- and bio-informatics
• Much work recently to define « kernels » for small molecules and proteins
• Combining them provides a theoretically sound and computationally efficient framework for in silico chemogenomics
• Promising results on several benchmarks for important target families
References: http://cbio.ensmp.fr/~jvert/