Virtual Screening with Support Vector Machines

Jean-Philippe Vert
Jean-Philippe.Vert@ensmp.fr

Center for Computational Biology
Ecole des Mines de Paris

Pierre Fabre, Institute of Drug Sciences and Technologies of Toulouse, May 22, 2006

CBIO overview

- The newest research center of Ecole des Mines
- Started in 2002, became an autonomous research center in 2006
- Objective: develop mathematical approaches and computational tools to process and analyze biological and chemical data
- http://cbio.ensmp.fr

CBIO research

1. Machine learning and statistics
   - theory
   - algorithms
2. Analysis of post-genomic data and systems biology
   - focus on cancer
   - focus on malaria
3. Data analysis methods for new technologies
   - DNA chips
   - cell chips
   - high-throughput microscopy
4. Virtual screening
   - ligand-based
   - docking

Outline

1. Virtual screening
2. Support Vector Machines
3. 2D Kernel
4. 3D Pharmacophore Kernel
5. Conclusion

Outline

1. Virtual screening

2. Support Vector Machines

3. 2D Kernel

4. 3D Pharmacophore Kernel

5. Conclusion

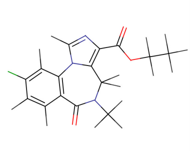
Ligand-Based Virtual Screening

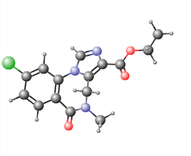
Objective

Build models to predict biochemical properties of small molecules from their structures.

Structures

$C_{15}H_{14}ClN_3O_3$





Properties

- binding to a therapeutic target
- pharmacokinetics (ADME)
- toxicity

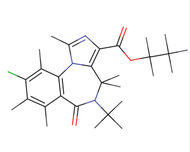
Ligand-Based Virtual Screening

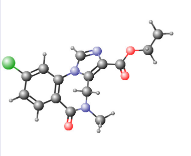
Objective

Build models to predict biochemical properties of small molecules from their structures.

Structures

$C_{15}H_{14}ClN_3O_3$





Properties

- binding to a therapeutic target
- pharmacokinetics (ADME)
- toxicity

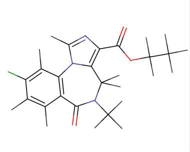
Ligand-Based Virtual Screening

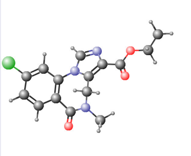
Objective

Build models to predict biochemical properties of small molecules from their structures.

Structures

$C_{15}H_{14}ClN_3O_3$





Properties

- binding to a therapeutic target
- pharmacokinetics (ADME)
- toxicity

Classical approaches

Two important steps

1. Define a feature map to represent each molecule as a vector of fixed dimension
2. Apply an algorithm for regression or pattern recognition to learn from a training set of molecules with labels.

Difficulties

- Expressivity of the features
- Dimension of the vector

MINES PARIS ARMINES

Classical approaches

Two important steps

1. Define a feature map to represent each molecule as a vector of fixed dimension
2. Apply an algorithm for regression or pattern recognition to learn from a training set of molecules with labels.

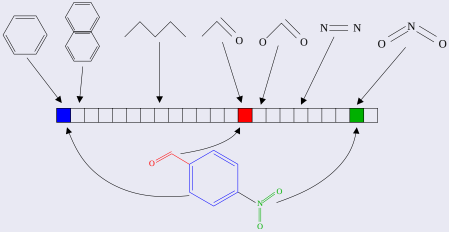
Difficulties

- Expressivity of the features
- Dimension of the vector

Example: 2D Structural Keys

Features

A vector indexed by a limited set of informative stuctures



Pros

- Fine description
- Prior knowledge is included
- interpretability

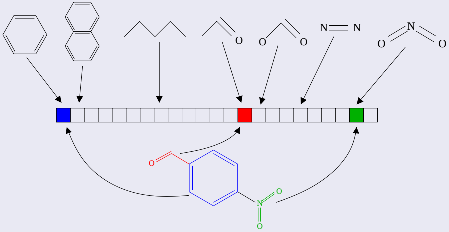
Cons

- Limited number of features
- How to choose the features?

Example: 2D Structural Keys

Features

A vector indexed by a limited set of informative stuctures



Pros

- Fine description
- Prior knowledge is included
- interpretability

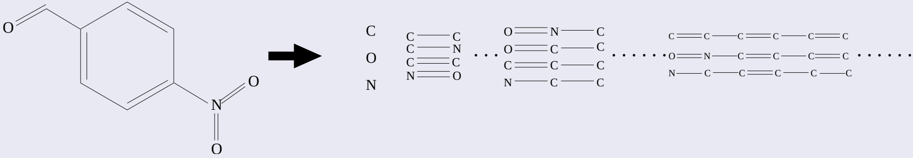
Cons

- Limited number of features
- How to choose the features?

Example: 2D Fingerprints

Features

A vector indexed by a large set of molecular fragments



Pros

- Many features
- Easy to detect

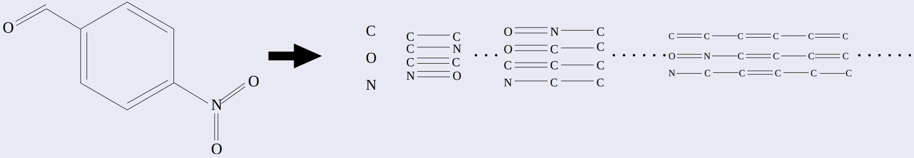
Cons

- Too many features?
- Hashing $\Longrightarrow$ clashes

Example: 2D Fingerprints

Features

A vector indexed by a large set of molecular fragments



Pros

- Many features
- Easy to detect

Cons

- Too many features?
- Hashing $\Longrightarrow$ clashes

Example: 3D Fingerprints

Features

- A collection of all possible combinations of the three/four features (hydrophobic, hydrogen bond donor and acceptor) in the 3D space.
- Discretized to form a vector

Pros

- 3D information
- Pharmacophore detection

Cons

- Discretization
- Size limitation

Example: 3D Fingerprints

Features

- A collection of all possible combinations of the three/four features (hydrophobic, hydrogen bond donor and acceptor) in the 3D space.
- Discretized to form a vector

Pros

- 3D information
- Pharmacophore detection

Cons

- Discretization
- Size limitation

Outline

1. Virtual screening
2. Support Vector Machines
3. 2D Kernel
4. 3D Pharmacophore Kernel
5. Conclusion

The Machine Learning Paradigm

Objective

Predict a property y for objects x

- x = molecule, gene sequence, picture, ...
- y is continuous (regression) or discrete (pattern recognition)

A two-step approach

1. Training: observe a set

$$\mathcal{S} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

of labeled objects, and learn a function $f : \mathcal{X} \rightarrow \mathcal{Y}$

2. Test: Given a new object x, predict its label by $f(x)$.

MINES PARIS ARMINES

The Machine Learning Paradigm

Objective

Predict a property y for objects x

- x = molecule, gene sequence, picture, ...
- y is continuous (regression) or discrete (pattern recognition)

A two-step approach

1. Training: observe a set

$$\mathcal{S} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$$

   of labeled objects, and learn a function $f : \mathcal{X} \rightarrow \mathcal{Y}$

2. Test: Given a new object x, predict its label by $f(x)$.

Examples

In biomedical research..

- Virtual screening : x is the description of a molecule, y is the activity / toxicity / drugability ...
- Medical diagnosis and prognosis: x is a set of features (age, weight, transcriptome...), y is the risk / type of tumor / expected evolution of disease.
- Functional genomics : x is a set of gene features (sequence, expresssion...), y is the function of the gene
- ...

What is a SVM?

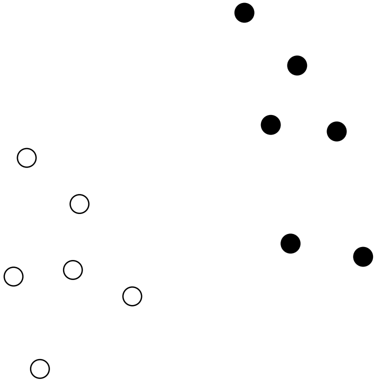
Main features

- an algorithm for pattern recognition and regression
- robust in high dimension (e.g., images, texts, microarrays, fingerprints)
- handles vectorial or structured data (e.g., sequences, graphs)
- allows easy integration of heterogeneous data (e.g., gene sequence and expression, docking score and molecule structure...)
- state-of-the-art performance on many real-world applications.

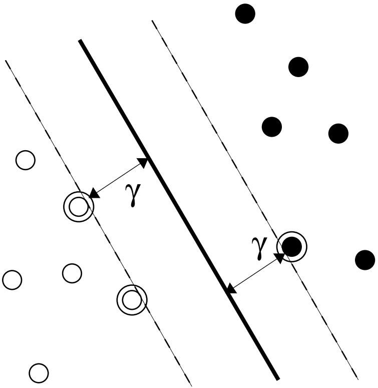
Related approaches

- least-square regression
- neural networks
- decision trees
- ...

Simplest SVM



Simplest SVM



Properties

Performance

- State-of-the-art in many real-world applications
- Resistant to large dimensions

Data representation

- Data do not need to be explicitly vectors
- A similarity function $K(x, x')$ between data is enough
- K must be symmetric and positive definite

Properties

Performance

- State-of-the-art in many real-world applications
- Resistant to large dimensions

Data representation

- Data do not need to be explicitly vectors
- A similarity function $K(x, x')$ between data is enough
- K must be symmetric and positive definite

Kernel examples

For vectors

- The linear kernel

$$K_{lin}\left(\mathbf{x},\mathbf{x}'\right) = \mathbf{x}^\top \mathbf{x}' \ .$$

- The polynomial kernel

$$K_{poly}\left(\mathbf{x},\mathbf{x}'\right) = \left(\mathbf{x}^\top \mathbf{x}' + a\right)^d \ .$$

- The Gaussian RBF kernel:

$$K_{Gaussian}\left(\mathbf{x},\mathbf{x}'\right) = \exp\left(-\frac{\parallel \mathbf{x} - \mathbf{x}' \parallel^2}{2\sigma^2}\right) \ .$$

Outline

1 Virtual screening

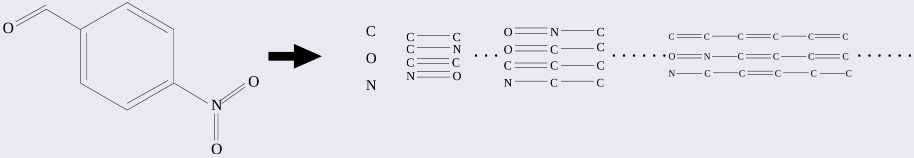
2 Support Vector Machines

3 2D Kernel

4 3D Pharmacophore Kernel

5 Conclusion

Motivations



Let $\Phi(x)$ the vector of fragment counts:

- Long fragments lead to large dimensions :
  SVM can learn in high dimension
- $\Phi(x)$ is too long to be stored, and hashes induce clashes:
  SVM do not need $\Phi(x)$, they just need the kernel

$$K(x,x') = \phi(x)^\top \phi(x') \; .$$

2D fingerprint kernel

Definition

For any $d > 0$ let $\phi_d(x)$ be the vector of counts of all fragments of length up to d:

$$\phi_d(x) = (\quad \texttt{\#(C),\#(O),\#(N),} \quad \ldots \quad \texttt{,\#(C-C),\#(C=O),\#(C-N),} \\ \ldots\texttt{,\#(C-C-C=O-C-N),\#(C=C-C=C-C=C),} \quad \ldots)^\top$$

The 2D fingerprint kernel is defined by

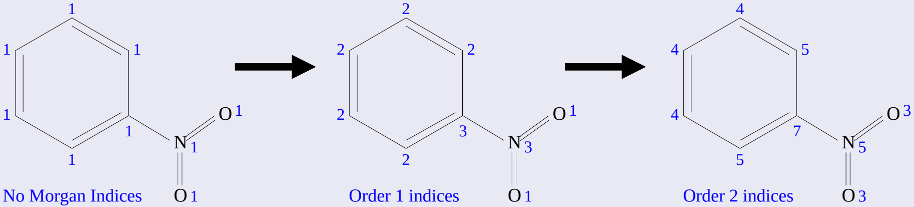
$$K_d(x, x') = \phi_d(x)^\top \phi_d(x') \, .$$

Extensions

Infinite fragments

- $d = +\infty$ is possible, if the contribution of a fragment of length p is weighted, e.g., by λ^p with $0 < \lambda < 1$.
- Worst-case complexity: $O(\mid x \mid \times \mid x' \mid)$ (faster in practice)

Atom relabebling with the Morgan index



No Morgan Indices

Order 1 indices

Order 2 indices

- compromise between fingerprints and structural keys features

Experiments

MUTAG dataset

- aromatic/hetero-aromatic compounds
- high mutagenic activity /no mutagenic activity
- 188 compouunds: 125 + / 63 -

Results

10-fold cross-validation accuracy

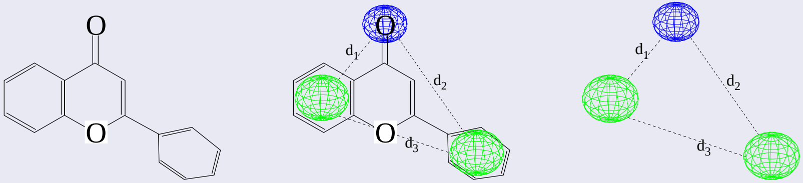
Method	Accuracy
Progol1	81.4%
2D kernel	91.2%

Outline

1. Virtual screening
2. Support Vector Machines
3. 2D Kernel
4. 3D Pharmacophore Kernel
5. Conclusion

Space of pharmacophore

3-points pharmacophores



A set of 3 atoms, and 3 inter-atom distances:

$$\mathcal{T} = \{((x_1, x_2, x_3), (d_1, d_2, d_3)), x_i \in \{\text{atom types}\}; d_i \in \mathbb{R}\}$$

3D fingerprint kernel

Pharmacophore fingerprint

1. Discretize the space of pharmacophores $\mathcal{T}$ (e.g., 6 atoms or groups of atoms, 6-7 distance bins) into a finite set $\mathcal{T}_d$
2. Count the number of occurrences $\phi_t(x)$ of each pharmacophore bin t in a given molecule x, to form a pharmacophore fingerprint.

3D kernel

A simple 3D kernel is the inner product of pharmacophore fingerprints:

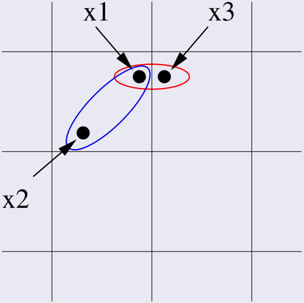
$$K(x, x') = \sum_{t \in \mathcal{T}_d} \phi_t(x)\phi_t(x') .$$

Discretization of the pharmacophore space

Common issues

1. If the bins are too large, then they are not specific enough
2. If the bins are too large, then they are too specific

In all cases, the arbitrary position of boundaries between bins affects the comparison:



$\rightarrow d(x_1, x_3) < d(x_1, x_2)$

BUT $\text{bin}(x_1) = \text{bin}(x_2) \neq \text{bin}(x_3)$

Kernels between pharmacophores

A small trick

$$
\begin{aligned}
K(x,y) &= \sum_{t \in \mathcal{T}_d} \phi_t(x)\phi_t(y) \\
&= \sum_{t \in \mathcal{T}_d} \Big(\sum_{p_x \in \mathcal{P}(x)} \mathbf{1}(\mathrm{bin}(\mathbf{p_x}) = \mathbf{t}) \Big) \Big(\sum_{p_y \in \mathcal{P}(y)} \mathbf{1}(\mathrm{bin}(\mathbf{p_y}) = \mathbf{t}) \Big) \\
&= \sum_{p_x \in \mathcal{P}(x)} \sum_{p_y \in \mathcal{P}(y)} \mathbf{1}(\mathrm{bin}(\mathbf{p_x}) = \mathrm{bin}(\mathbf{p_y}))
\end{aligned}
$$

General pharmacophore kernel

$$
K(x,y) = \sum_{p_x \in \mathcal{P}(x)} \sum_{p_y \in \mathcal{P}(y)} K_P(p_x, p_y)
$$

New pharmacophore kernels

- Discretizing the pharmacophore space is equivalent to taking the following kernel between individual pharmacophores:

$$K_P(p_1, p_2) = \mathbf{1}\left(\mathrm{bin}(\mathbf{p_x}) = \mathrm{bin}(\mathbf{p_y})\right)$$

- For general kernels, there is no need for discretization!
- For example, is $d(p_1, p_2)$ is a Euclidean distance between pharmacophores, take:

$$K_P\left(p_1, p_2\right) = \exp\left(-\gamma d\left(p_1, p_2\right)\right) \ .$$

Experiments

4 public datasets

- BZR: ligands for the benzodiazepine receptor
- COX: cyclooxygenase-2 inhibitors
- DHFR: dihydrofolate reductase inhibitors
- ER: estrogen receptor ligands

	TRAIN		TEST	
	Pos	Neg	Pos	Neg
BZR	94	87	63	62
COX	87	91	61	64
DHFR	84	149	42	118
ER	110	156	70	110

Experiments

Results (accuracy)

Kernel	BZR	COX	DHFR	ER
2D (Tanimoto)	71.2	63.0	76.9	77.1
3D fingerprint	75.4	67.0	76.9	78.6
3D not discretized	**76.4**	**69.8**	**81.9**	**79.8**

Outline

1 Virtual screening

2 Support Vector Machines

3 2D Kernel

4 3D Pharmacophore Kernel

5 Conclusion

Summary

- SVM is a powerful and flexible machine learning algorithm. The kernel trick allows the manipulation of non-vectorial objects at the cost of defining a kernel function.
- The 2D kernel for molecule extends classical fingerprint-based approches. It solves the problem of bit clashes, and allows infinite fingerprints.
- The 3D kernel for molecule extends classical pharmacophore fingerprint-based approaches. It solves the problems of bit clashes and of discretization.
- Both kernels improve upon their classical counterparts, and provide competitive results on benchmark datasets.

Ongoing works

- Further validation of the kernel approach on larger datasets.
- Learning from multiple conformers.
- Combination of ligand-based virtual screening with docking approaches.

Acknowledgements

- Pierre Mahé (CBIO)
- Tatsuya Akutsu, Nobuhisa Ueda, Jean-Luc Perret (Kyoto University)
- Liva Ralaivola (U Marseille)

References

- Kashima, H., Tsuda, K., and Inokuchi, A. *Marginalized kernels between labeled graphs*. Proceedings of the 20th ICML, 2003, pp. 321-328.
- L. Ralaivola, S. J. Swamidass, H. Saigo, and P. Baldi. *Graph kernels for chemical informatics*. Neural Netw., 18(8):1093-1110, Sep 2005.
- P. Mahé, N. Ueda, T. Akutsu, J.-L. Perret, and J.-P. Vert. *Graph kernels for molecular structure-activity relationship analysis with SVM*. J. Chem. Inf. Model., 45(4):939-951, 2005.
- P. Mahé, L. Ralaivola, V. Stoven, and J-P Vert.*The pharmacophore kernel for virtual screening with SVM*. Technical Report Technical Report HAL:ccsd-00020066, http://hal.ccsd.cnrs.fr/ccsd-00020066, march 2006.
- Open-source kernels for chemoinformatics: `http://chemcpp.sourceforge.net/`