Kernels for kernel-based machine learning

Matthias Rupp

Berlin Institute of Technology, Germany

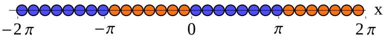
Institute of Pure and Applied Mathematics
Navigating Chemical Compound Space for Materials and Bio Design
Los Angeles, California, USA, March 18, 2011

Kernels: Definition

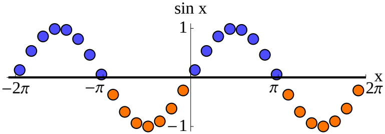
A kernel is a function that corresponds to an inner product

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}, \qquad k(x, z) = <\phi(x), \phi(z)> \text{ with } \phi : \mathcal{X} \to \mathcal{H}$$

Example:



$\overset{\phi}{\mapsto}$



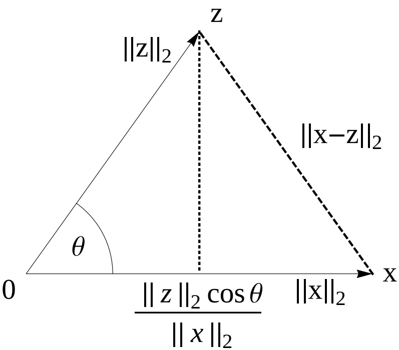
Input space

Feature space

Kernels: Geometric aspects

Kernels describe the geometry of the data

- Correspond to inner products: $k(x,z) = <\phi(x), \phi(z)>$
- Encode information about length and angle in feature space: $k(x,z) = ||\phi(x)||_2 \; ||\phi(z)||_2 \cos\theta$
- Correspond to Euclidean distance: $||\phi(x) - \phi(z)||_2^2 =$
  $$k(x,x) - 2k(x,z) + k(z,z)$$



Kernels: Characterization

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if and only if
for all $n \in \mathbb{N}, x_1, \ldots, x_n \in \mathcal{X}$ the matrix $\mathbf{K} = k(x_i, x_j)$ is symmetric and

- positive semi-definite, $\forall\, \mathbf{v} \in \mathbb{R}^n : \mathbf{v}^T \mathbf{K} \mathbf{v} \geq 0$
- has only non-negative eigenvalues
- all principal minors are non-negative
- $\exists\, \mathbf{L} \in \mathbb{R}^{n \times n} : \mathbf{K} = \mathbf{L}\mathbf{L}^T \wedge \operatorname{diag}(\mathbf{L}) \geq 0$ (Cholesky decomposition)
- $f(\mathbf{v}) = \frac{1}{2}\mathbf{v}^T \mathbf{K} \mathbf{v} + \mathbf{a}^T \mathbf{v} + c$ is convex

Similar characterization for symmetric, positive definite k.

Kernel trick for distances leads to conditionally positive semi-definite kernels, where $\sum_{i=1}^{n} \mathbf{v}_i = 0$.

Kernels: Closure properties

Kernels are closed under

- Addition: $k(x,z) = k_1(x,z) + k_2(x,z)$
- Multiplication with non-negative scalar: $k(x,z) = \gamma k_1(x,z)$, $\gamma \geq 0$
- Point-wise product: $k(x,z) = k_1(x,z) \cdot k_2(x,z)$
- Tensor product $k_1 \otimes k_2$, direct sum $k_1 \oplus k_2$

Linear combination of kernels is a kernel:

$$k(x,z) = \gamma_1 k_1(x,z) + \gamma_2 k_2(x,z) + \ldots + \gamma_m k_m(x,z)$$

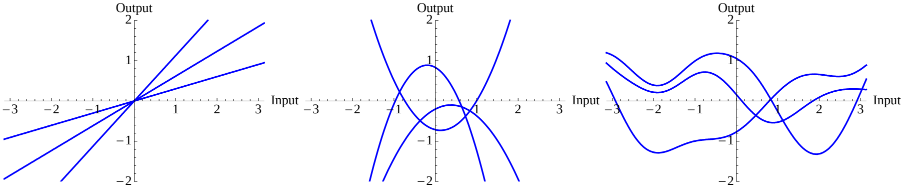
Specific kernels: Vector data

Let $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$.

- Linear kernel: $k(\mathbf{x}, \mathbf{z}) = <\mathbf{x}, \mathbf{z}>$
- Polynomial kernel: $k(\mathbf{x}, \mathbf{z}) = \left(<\mathbf{x}, \mathbf{z}> + c\right)^d$
- Squared exponential kernel (also radial basis function kernel):

$$k(\mathbf{x}, \mathbf{z}) = \exp\left(-\frac{1}{2}\frac{||\mathbf{x}-\mathbf{z}||^2}{2\sigma^2}\right)$$

Local approximator; free parameter σ (length scale)



Specific kernels: Structured data

- Convolution kernels

$$(k_1 \times \cdots \times k_d)(x, x') = \sum_{R} \prod_{i=1}^{d} k_i(x_i, x'_i)$$

- String kernels
- Kernels between probability distributions
- Kernels between graphs

Specific kernels: String kernels

Spectrum kernel

- Compare substrings of length k (k-mers)
- Position-independent
- Kernel is sum of products of counts

Example: $k = 3$

x $\boxed{\text{AAA}}$C$\boxed{\text{AAA}}$TAAGTAACTAATCTTTTAGAACTTTCAACCATTT...

z TACCTAATTATG$\boxed{\text{AAA}}$TT$\boxed{\text{AAA}}$TTTCAGCTGTGG$\boxed{\text{AAA}}$CGGAGA..

3-mer	AAA	AAC	...	TTT
# in x	2	4	...	3
# in z	3	1	...	1

$$k(x, z) = 2 \cdot 3 + 4 \cdot 1 + \ldots + 3 \cdot 1$$

Specific kernels: String kernels

Weighted degree kernel

- Compare matches at each position
- Position-dependent
- $k(x,z) = \sum_{k=1}^{d} \beta_k \sum_{i=1}^{|x|-k} \mathbf{I}\big(u_{k,l}(x) = u_{k,l}(z)\big)$

Example: $d = 3$

```
       x   AAACAAATAAGTAACTAATCTTTTAG
# 1-mers   .|.|.|||.|..||.|.|..|||.||
# 2-mers   .....||.....|.......||..|.
# 3-mers   .....|..............|.....
       z   TACCTAATTATGAAATTAAATTTCAG
```

$k(x,z) = \beta_1 \cdot 15 + \beta_2 \cdot 6 + \beta_3 \cdot 2$

Specific kernels: Graph kernels (introduction)

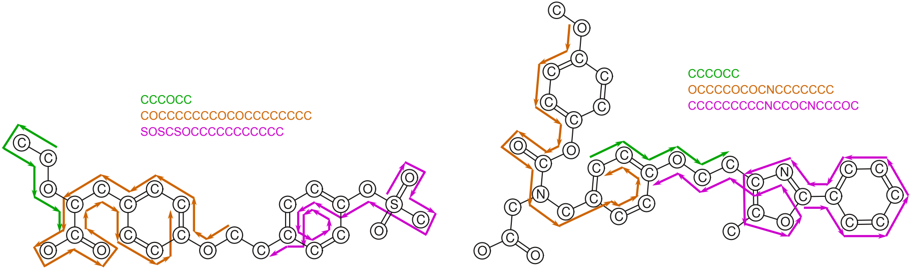
Idea: Define k directly on graphs

- Application to chemical structure graphs
- Allows direct measurement of graph similarity
- Rigorous way to combine graph theory and kernel learning

- Caveat: Complete graph kernels are computationally hard

Specific kernels: Graph kernels (overview)

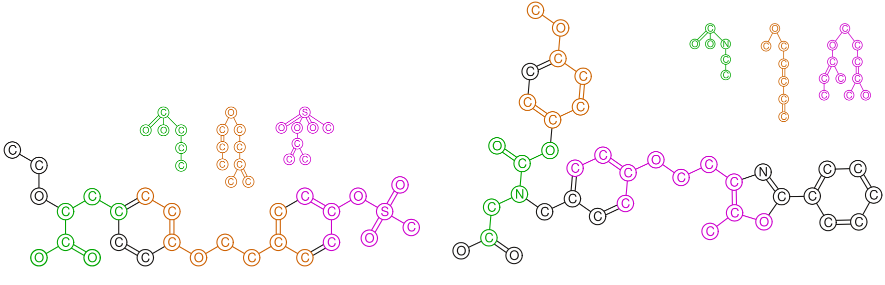
- Random walk graph kernels, path-based graph kernels
- Tree pattern graph kernels, cyclic pattern graph kernels
- Graphlet kernels
- Optimal assignment graph kernels



M. Rupp, G. Schneider, Mol. Inf. 29(4): 266–273, 2010.

Specific kernels: Graph kernels (overview)

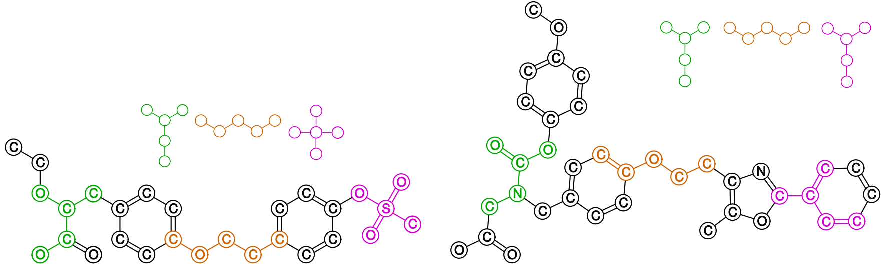
- Random walk graph kernels, path-based graph kernels
- Tree pattern graph kernels, cyclic pattern graph kernels
- Graphlet kernels
- Optimal assignment graph kernels



M. Rupp, G. Schneider, Mol. Inf. 29(4): 266–273, 2010.

Specific kernels: Graph kernels (overview)

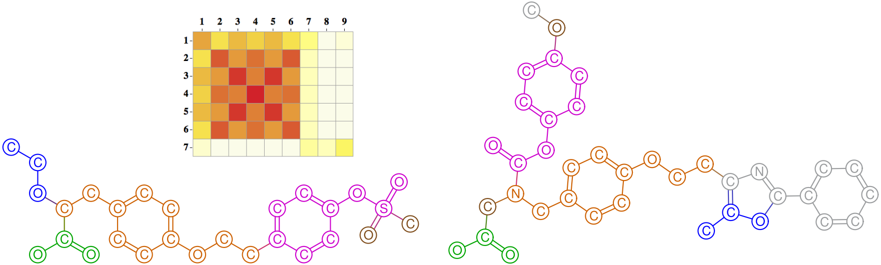
- Random walk graph kernels, path-based graph kernels
- Tree pattern graph kernels, cyclic pattern graph kernels
- Graphlet kernels
- Optimal assignment graph kernels



M. Rupp, G. Schneider, Mol. Inf. 29(4): 266–273, 2010.

Specific kernels: Graph kernels (overview)

- Random walk graph kernels, path-based graph kernels
- Tree pattern graph kernels, cyclic pattern graph kernels
- Graphlet kernels
- Optimal assignment graph kernels



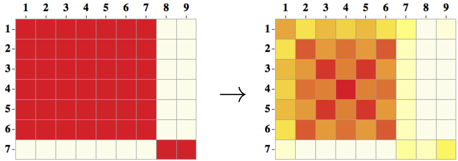
M. Rupp, G. Schneider, Mol. Inf. 29(4): 266–273, 2010.

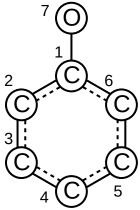
Specific kernels: Graph kernels (example)

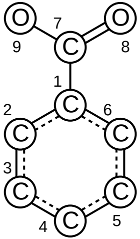
ISOAK = iterative similarity optimal assignment kernel

$$\mathbf{X}_{v,v'} = (1-\alpha)k_v(v,v') + \alpha \max_{\pi} \frac{1}{|v'|} \sum_{\{v,u\}\in E} \mathbf{X}_{u,\pi(u)} k_e\big(\{v,u\},\{v',\pi(u)\}\big)$$

α controls recursiveness; π assigns neighbors of v to neighbors of v'







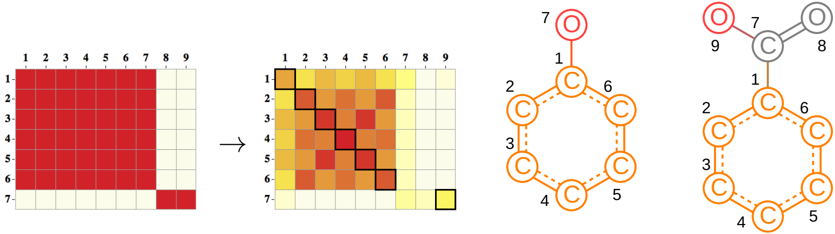
Rupp et al, J. Chem. Inf. Mol. Model. 47(6): 2280, 2007.

Specific kernels: Graph kernels (example)

ISOAK = iterative similarity optimal assignment kernel

$$\mathbf{X}_{v,v'} = (1-\alpha) k_v(v,v') + \alpha \max_{\pi} \frac{1}{|v'|} \sum_{\{v,u\}\in E} \mathbf{X}_{u,\pi(u)} k_e\big(\{v,u\},\{v',\pi(u)\}\big)$$

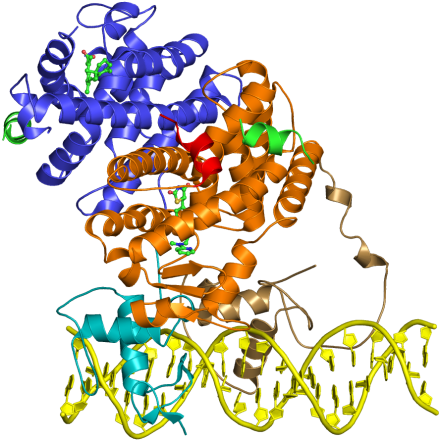
α controls recursiveness; π assigns neighbors of v to neighbors of v'



Rupp et al, J. Chem. Inf. Mol. Model. 47(6): 2280, 2007.

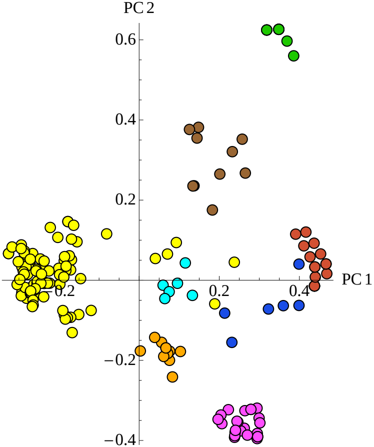
Application example: Virtual screening

Target



Peroxisome proliferator-activated receptor γ (PPARγ)

Data



Kernel principle component analysis with ISOAK graph kernel ($n = 176$)

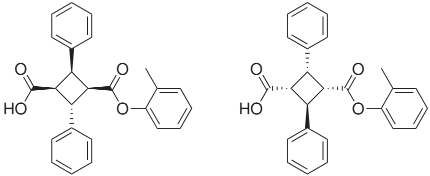
Rücker et al., *Bioorg. Med. Chem.* 14(15): 5178, 2006; Rupp, PhD thesis, 2009.

Application example: Virtual screening

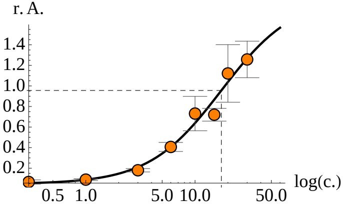
Methods:

- Graph kernel, vector kernels
- Multiple kernel learning
- Gaussian process regression

Results:



Stereochemistry



Dose-response curve

Rupp et al., *ChemMedChem* 5(2): 191, 2009

Summary

- Kernels correspond to inner products
- Well-characterized function class
- Kernels can be defined on structured data

Literature

- K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, B. Schölkopf: *An Introduction to Kernel-Based Learning Algorithms*, IEEE Trans. Neural Network. 12(2): 181–201, 2001.
  22 pages, review, broader introduction
- T. Hofmann, B. Schölkopf, A. Smola: *Kernel Methods in Machine Learning*, Ann. Stat. 36(6): 1171–1220, 2008.
  50 pages, review, mathematically oriented
- O. Ivanciuc: *Applications of Support Vector Machines in Chemistry*, ch. 6, p. 291–400. In K. Lipkowitz, T. Cundari, *Reviews in Computational Chemistry*, vol. 23, Wiley, 2007.
  110 pages, review
- M. Rupp, G. Schneider: *Graph kernels for molecular similarity*, Mol. Inf 29(4): 266–273, 2010.
  8 pages, overview of graph kernels for small structure graphs