Institute of Bioinformatics
Johannes Kepler University Linz

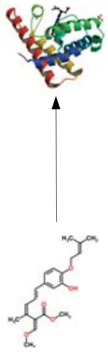
QSTAR
(Fast) Analoging in large databases with structural fingerprint features

Martin Heusel, Andreas Mayr, Günter Klambauer,
Andreas Mitterecker, Ulrich Bodenhofer,
Djork-Arné Clevert, and Sepp Hochreiter

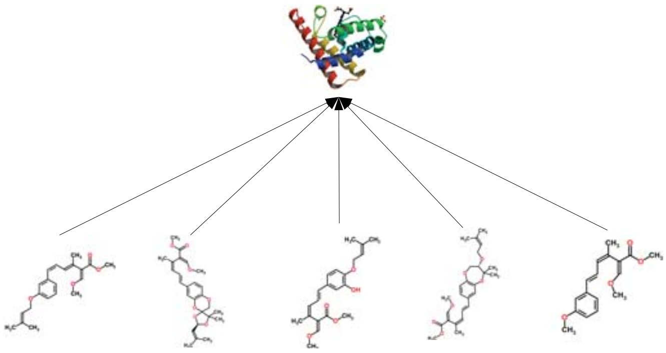
Institute of Bioinformatics
Johannes Kepler University, Linz, Austria

NCS2012, Potsdam, September 25th 2012

Analogs



Analogs share a common bioactivity

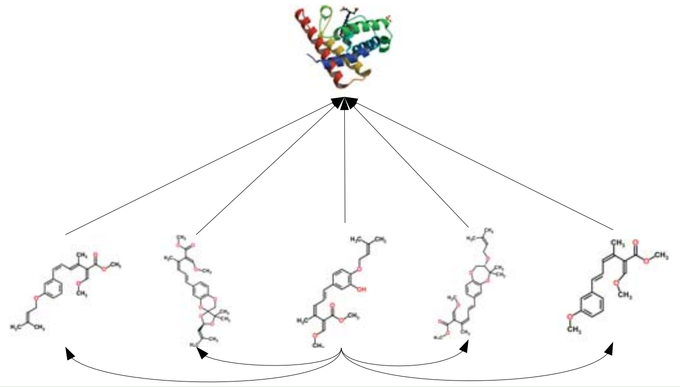


Analogs

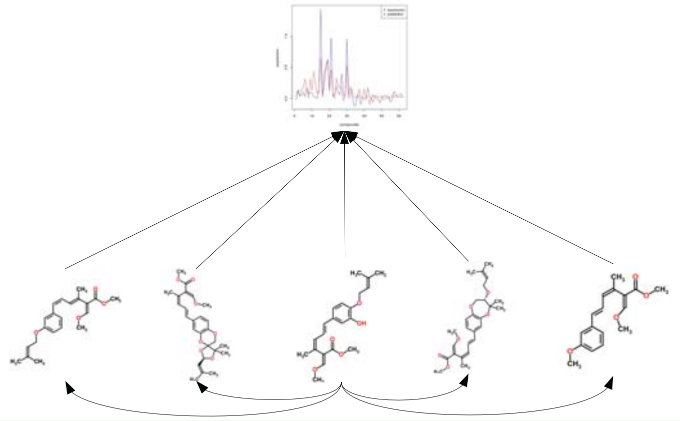
Analogs are vital for drug design helping to improve the final product in terms of

- Effectivity
- Toxicity
- Side effects
- Bacterial resistance
- other limitations or optimizations
- Absorption, Distribution, Metabolism, Excretion/elimination

Structural-activity Relationship (SAR)



Analogs defined by similar gene expression



How to find analogs?

Structural-activity Relationship (SAR)

“Structural similar molecules have similar activities”

⇨

Molecules/compounds are represented by structural fingerprints

⇨

ECFP (Extended-Connectivity Fingerprints)

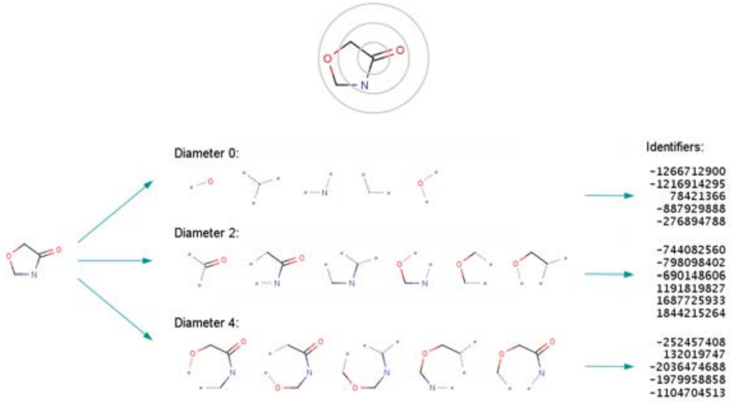
⇨

Potential Support Vector Machine
- Gene expression classification
- Robust Feature Selection

⇨

Search analogs in e.g. ChEMBL with trained P-SVM model and features

Extended-Connectivity Fingerprints (ECFP)



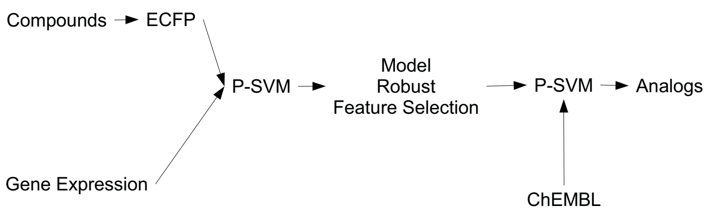
Extended-Connectivity Fingerprints (ECFP)

are circular topological fingerprints designed for molecular characterization, similarity searching, and structure-activity modeling.

- Molecular structures are represented by means of circular atom neighborhoods.
- Features represent the presence of particular substructures.
- Not predefined
- Can represent a huge number of different molecular features (including stereochemical information).
- Designed to represent both the presence and the absence of functionality

Calculated with jCompoundMapper (http://jcompoundmapper.sourceforge.net/)

Pipeline



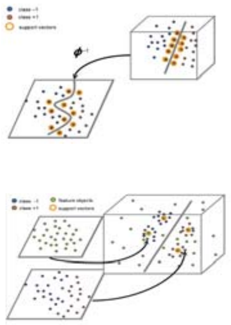
Compounds → ECFP

Gene Expression

P-SVM → Model Robust Feature Selection → P-SVM → Analogs

ChEMBL

Potential Support Vector Machine



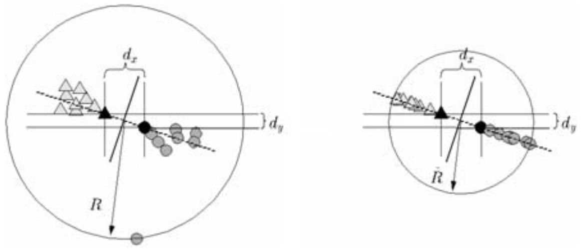
$$K = X^T Z$$

K: Data (ECFP x compounds)

An ECFP feature (substructure) is a scalar product of some ECFP vector and some labeled object (compound) vector.

- Feature weighting (SVM weights feature objects)
- Feature Selection

Scale invariant objective



new	SVM
$\|\mathbf{X}^T\mathbf{w}\|_2^2$	$\|\mathbf{w}\|_2^2$

$$\|\mathbf{X}^T\mathbf{w}\|_2^2 = \mathbf{w}^T\mathbf{X}\,\mathbf{X}^T\mathbf{w}$$ (X is matrix of vectors $\mathbf{x}^i$):

Potential Support Vector Machine

Algorithm

Dual

$$\min_{\boldsymbol{\alpha}^+, \boldsymbol{\alpha}^-} \frac{1}{2}\left(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-\right)^T \mathbf{K}^T\mathbf{K}\left(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-\right) - \mathbf{y}^T\mathbf{K}\left(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-\right) + \varepsilon\,\mathbf{1}^T\left(\boldsymbol{\alpha}^+ + \boldsymbol{\alpha}^-\right)$$

$$\text{s.t.} \quad \mathbf{1}^T\mathbf{K}\left(\boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-\right) = 0 \;, \quad C\,\mathbf{1} \geq \boldsymbol{\alpha}^+, \boldsymbol{\alpha}^- \geq 0$$

$\boxed{\mathbf{w} = \mathbf{Z}\,\boldsymbol{\alpha}}$, where $\boldsymbol{\alpha} = \boldsymbol{\alpha}^+ - \boldsymbol{\alpha}^-$.

$\mathbf{K}^T\mathbf{K}$ is (features x features) and optimization would be computational expensive: Sequential Minimal Optimization (SMO)

Searching Analogs in ChEMBL

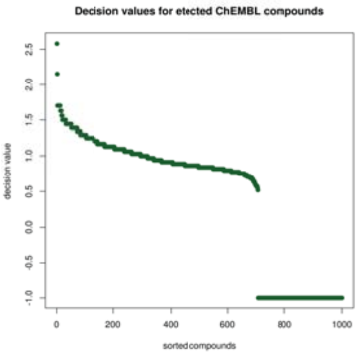
- Robust feature selection with P-SVM and ECFP fingerprint features
- Building PSVM model with selected ECFP fingerprint features from robust feature selection
- Search in 1 million ChEMBL for analogs with model
- Search time ~10s

Potential Support Vector Machine

Characteristica

- Works with data matrix (e.g. ECFP features, substructures by compounds) not necessarily positive definite nor square
- Feature selection: Identification of relevant features

ChEMBL Results (example)



Sorted decision values (P-SVM output) of first 1000 detected compounds.

700 possible analogs found.

References

P-SVM: Sepp Hochreiter, Klaus Obermayer; Support Vector Machines for Dyadic Data, Neural Computation, 18, 1472-1510, 2006

ECFP: Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints.
J. Chem. Inf. Model. 2010, 50(5): 742-754

jCompoundMapper: Hinselmann G. ET AL jCompoundMapper : An Open Source Java Library and Command-Line Tool for Chemical Fingerprints.
http://jcompoundmapper.sourceforge.net/

Robust Feature Selection: S. Hochreiter, K. Obermayer; Gene Selection for Microarray Data, Kernel Methods in Computational Biology, pp. 319-355, MIT Press, 2004
http://www.bioinf.jku.at/publications/bioinf/older/0604.ps