

A Kernel Approach to Molecular Similarity Based on Iterative Graph Similarity

Matthias Rupp

Beilstein Endowed Chair for Chem- and Bioinformatics
Johann Wolfgang Goethe-University
Frankfurt am Main, Germany

2007-07-03, University of Frankfurt, Germany

modlab

JOHANN WOLFGANG  GOETHE
UNIVERSITÄT
FRANKFURT AM MAIN

Outline

- Introduction Molecular similarity, graph-based methods
- Method Optimal assignments, iterative graph similarity
- Results Retrospective virtual screening
- Conclusions Assessment, future work

Molecular similarity

- ▶ Applications in drug development:
 - ▶ Enrichment / focused libraries
 - ▶ Quantitative structure-activity relationships
 - ▶ *De novo* design
 - ▶ Virtual screening
- ▶ Quantum methods are computationally infeasible on this scale
- ▶ Similarity principle (Johnson & Maggiora, 1990)
“Similar molecules tend to exhibit similar properties”
- ▶ Abundancy of specialized similarity measures

The vectorization-based approach

- ▶ Uses established vector-based methods
- ▶ Uses descriptors to represent molecules as vectors
- ▶ Many molecular descriptors
- ▶ Descriptor selection is NP-hard

Advantages:

- ▶ simple & works
- ▶ uses existing techniques

Disadvantages:

- ▶ Interpretation of results unintuitive
- ▶ Loss of information, introduction of noise

Non-vector based similarity measures

Alternative: Direct comparison of non-vector based models

Example: Use methods from graph theory on molecular graphs

- ▶ Several approaches
 - ▶ Spectrum-based
 - ▶ Subgraph matching
 - ▶ Random walks
 - ▶ Optimal assignments
 - ▶ ...
- ▶ Separating all non-isomorphic graphs is NP-complete

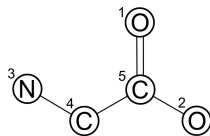
Optimal assignments

$G = (V, E)$, $G' = (V', E')$ are two molecular graphs.

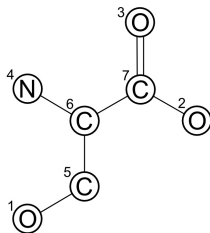
Idea:

- ▶ Compute matrix $X \in [0, 1]^{|V| \times |V'|}$ of pairwise vertex similarities
- ▶ Match vertices so that sum of similarities is maximal

Example:



Glycine



Serine

X	1	2	3	4	5	6	7
1	.50	.50	.98	.00	.00	.00	.00
2	.89	.98	.50	.34	.17	.16	.11
3	.38	.33	.00	.91	.20	.13	.14
4	.20	.24	.00	.17	.77	.81	.67
5	.13	.11	.00	.14	.78	.68	.96
$\Sigma = 4.64$					(.78 normalized)		

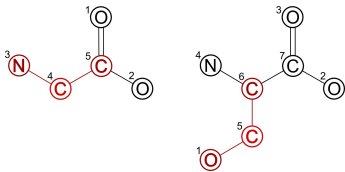
How to compute X ?

Iterative graph similarity

- ▶ Problem: Compute a pairwise atom similarity matrix X
- ▶ Idea: Vertices are similar if their neighbours are similar.
- ▶ Recursive definition leads to a non-linear system of equations
- ▶ Solved by iteration

$$X_{i,j}^{(n)} = (1-\alpha)k_v(v_i, v_j) + \alpha \max_{\pi} \frac{1}{|V_j'|} \sum_{v \in n(v_i)} X_{v, \pi(v)}^{(n-1)} k_e(\{v_i, v\}, \{v_j', \pi(v)\})$$

Example:



$$X_{4,5} = \frac{1}{4}1 + \frac{3}{4} \max \frac{1}{2} (X_{3,1}1 + X_{5,6}1, X_{3,6}1 + X_{5,1}1)$$

$$\text{for } \alpha = \frac{3}{4}, k_v(a, b) = k_e(a, b) = 1_{a=b}$$

Retrospective results

Virtual screening using support vector machines for binary classification. 10 runs of 10-fold stratified cross-validation.

Comparison against “standard” descriptor/kernel combinations:

Dataset	Standard	cc	ISOAK	cc
Drug	rbf/gc	0.745 ± 0.04	dppp/dbond	0.777 ± 0.04
AChE	rbf/gc	0.874 ± 0.13	delem/none	0.926 ± 0.09
COX-2	poly/gc	0.861 ± 0.09	dppp/dbond	0.858 ± 0.09
DHFR	rbf/cats2d	0.983 ± 0.05	none/none	0.994 ± 0.03
FXa	poly/cats2d	0.945 ± 0.05	echarge/none	0.973 ± 0.03
PPAR	rbf/cats2d	0.822 ± 0.12	dppp/none	0.989 ± 0.09
Thrombin	poly/cats2d	0.891 ± 0.07	dppp/dbond	0.930 ± 0.06

rbf = radial basis function kernel, poly = polynomial kernel
gc = Ghose-Crippen descriptor, cats2d = CATS2D descriptor
ISOAK = iterative similarity optimal assignment kernel
cc = correlation coefficient

Conclusions

Summary:






- ▶ “Direct” comparison of molecules (no vectorization) is possible
- ▶ Introduction of a novel molecular similarity measure based on iterative graph similarity and optimal assignments
- ▶ Encouraging results.

Future work:

- ▶ Directly solving the underlying non-linear system of equations
- ▶ Making the similarity measure positive semidefinite
- ▶ Obtaining prospective results.

Thank you for your attention.

References

-  Johnson, M. & Maggiora, G. (editors).
Concepts and Applications of Molecular Similarity.
Wiley, 1990.
-  Todeschini, R. & Consonni, V.
Handbook of Molecular Descriptors.
Wiley, 2000.
-  Munkres, J.
Algorithms for the assignment and transportation problems.
J. Soc. Ind. Appl. Math., **5**(1), 1957, 32–38.
-  Fröhlich, H., Wegner, J., Sieker, F., & Zell, A.
Optimal assignment kernels for attributed molecular graphs.
Proceedings of ICML 2005, 225–232.
-  Zager, L.
Graph similarity and matching.
Master's thesis, Massachusetts Institute of Technology.