A Polynomial-time Metric for Outerplanar Graphs

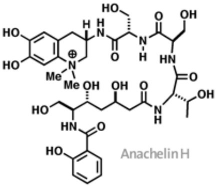
Leander Schietgat
Jan Ramon
Maurice Bruynooghe

Mining and Learning with Graphs
August 1-3 2007, Florence

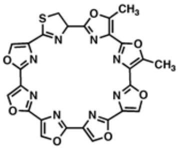
Introduction

- Drug discovery
  - find new drug molecules that are active against some disease
  - need for automatic techniques that select *interesting* molecules
- How to find interesting molecules
  - similarity measure: which molecules are *close* to known drug molecules?
  - observation: molecules with the same structure tend to have the same activity
- Problem
  - how to represent molecules?

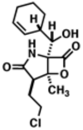
Examples of molecules



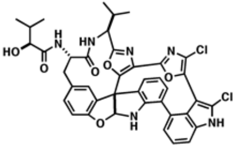
Anachelin H



Telomestatin



Salinosporamide A



Diazonamide A

Graphs

- Very suitable to represent (binary) relational data
  - vertices are entities, edges are relationships between entities
  - molecules: vertices are atoms, edges are bonds
  - graphs can be labeled
    - atoms: C, O, Cu, Cl, H, ...
    - bonds: single, double, aromatic, ...
- Problem
  - operations on graphs are computationally expensive!
  - hence: algorithms that handle graphs directly are avoided

Related work

- Feature-based distances (fingerprints)
  - defining of some features
  - molecule is represented by a vector
  - **advantages:** efficiently computable, use of existing machine learning techniques
  - **disadvantages:** loss of information, feature selection
- Cost-based distances aka. graph edit distances
  - approximation algorithms
  - exact algorithms
  - **advantage:** original graph structure preserved
  - **disadvantage:** efficiency

The problem

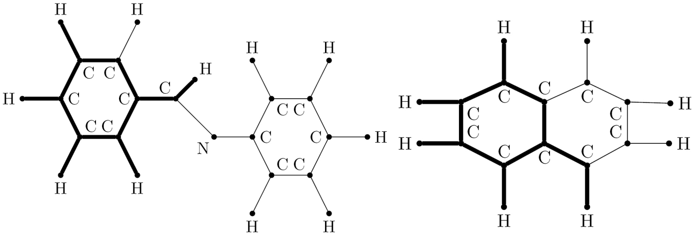
- **Goal of this work:** to develop an efficiently computable metric on graphs representing molecules
- Bunke & Shearer (1998) proposed a distance function on graphs based on the maximum common subgraph (MCS):

$$d_{bs}(G, H) = 1 - \frac{|MCS(G, H)|}{\max(|G|, |H|)},$$

  with $|G|$ equal to the number of vertices in G.
- d_{bs} is a metric
- Other size functions can be used too

Maximum Common Subgraph (MCS)

- Given two graphs G and H
- The MCS is the graph I
  - which is subgraph isomorphic to G and H
  - there exists no other graph J which is also subgraph isomorphic to G and H and $|J| > |I|$



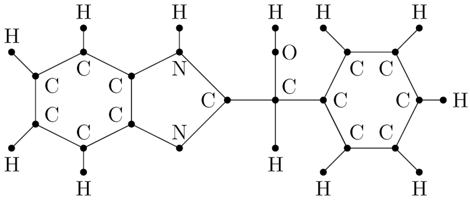
- $d_{bs}(G,H) = 1 - \frac{|MCS(G,H)|}{\max(|G|,|H|)} = 1 - \frac{12}{\max(26,18)} = 0.54$

However...

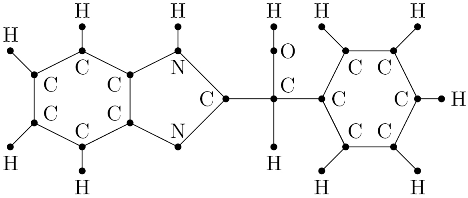
- Problem: the computation of the MCS is not easy
  - the subgraph isomorphism problem is NP-hard for general graphs (unless $P = NP$)
- Previous work on graphs has shown that the complexity of some problems can be reduced by imposing some constraints on the graph structure
  - sequences
  - trees
  - planar graphs
  - graphs of bounded degree
  - graph of with treewidth at most k
  - k-connected graphs
- Task: find an "easier" class of graphs to represent molecules?

Planar and outerplanar graphs

- Planar graph
  - can be drawn in the plane in such a way that no two edges intersect except at a vertex in common
- Outerplanar graph
  - planar graph with all the vertices adjacent to the outer face



A molecule

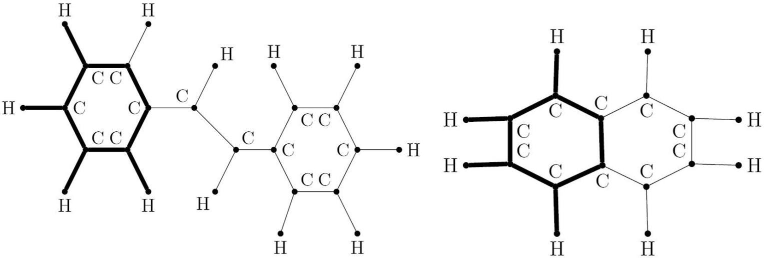


- 95% of the molecules in the NCI database can be represented by outerplanar graphs [Horváth et al. 2006]
- Problem: the subgraph isomorphism problem for outerplanar graphs is still NP-hard [Syslo 1982]

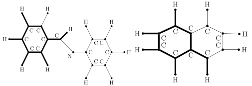
The subgraph isomorphism revisited

- New terminology
  - block: maximal subgraph for which every pair of vertices is involved in a cycle
  - bridge: edge that does not belong to a block
- Block-and-bridge preserving (BBP) subgraph isomorphism
  - variant of the general subgraph isomorphism
  - blocks are mapped onto blocks
  - bridges are mapped onto bridges
- Motivation
  - the BBP subgraph isomorphism for outerplanar graphs is computable in polynomial time [Horváth et al. 2006]
  - chemist viewpoint: ring structures and linear fragments usually behave differently

The maximum common subgraph revisited



- $d_{bs}(G,H) = 1 - \frac{|MCS(G,H)|}{\max(|G|,|H|)} = 1 - \frac{10}{\max(26,18)} = 0.62$



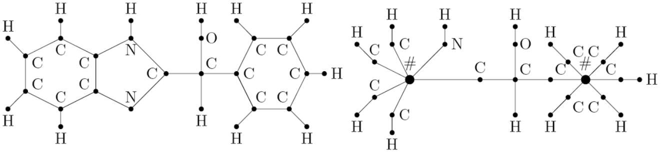
$d_{bs}(G,H) = 0.54$

Sketch of the algorithm

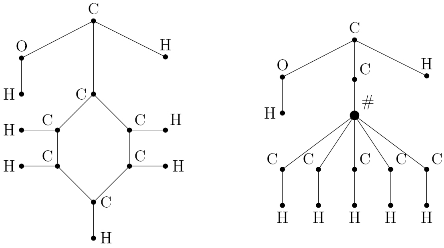
- Dynamic programming approach
  - Generate subgraphs
    - non-block-splitting subgraphs
    - half-graphs
  - Order the subgraphs by ascending "size"
  - Solve them (bottom-up)
    - simple subgraphs (1 node): trivial solution
    - difficult subgraphs (multiple nodes): combine the earlier computed solutions of parts of the subgraphs
- Results in polynomial time complexity

Non-block-splitting subgraphs

- Based on block-bridge trees

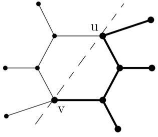


- Example of a non-block-splitting subgraph



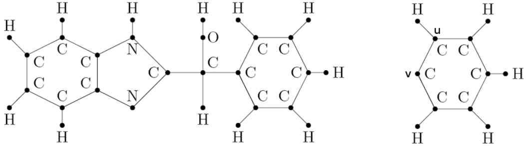
Half-graphs

- Half-graph $G|_{o[u,v]}$: maximal connected subgraph of G containing all vertices of $o[u, v]$ but none of the vertices $V(B) \setminus o[u, v]$ and none of the edges adjacent to v, which do not belong to the block B

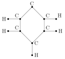
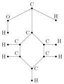
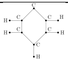
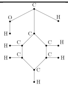


$(o = \curvearrowright)$

- Example of a half-graph



Finding the size of the MCS of two outerplanar graphs

	H	H	...	C–H	O–H	...		
H	1	1		0	0		0	0
H	1	1		0	0		0	0
...								
C–H	0	0		2	0		1	1
O–H	0	0		0	2		0	0
...								
	0	0		1	0		11	11
	0	0		2	0		11	15

Datasets

- NCI cancer dataset
  - publicly available (National Cancer Institute)
  - screening results for the ability of more than 70,000 compounds to suppress or inhibit the growth of a panel of 60 human tumour cell lines
- 60 datasets from Swamidass et al. (2006)
  - for each cell line: two-class classification problem
  - more or less balanced datasets
  - $\sim$3500 examples, $\sim$90% outerplanar

kNN-classification

- find the nearest neighbour(s) according to the defined distance measure
- parameters
  - $k = 5$
  - distance measure:

$$d_{bs}(G,H) = 1 - \frac{|MCS(G,H)|}{\max(|G|,|H|)}$$

  - $|G|$: number of nodes in G
- prediction for molecule m:
  - majority voting
  - weighted voting: e.g., $|MCS(G,H)| * class(H)$

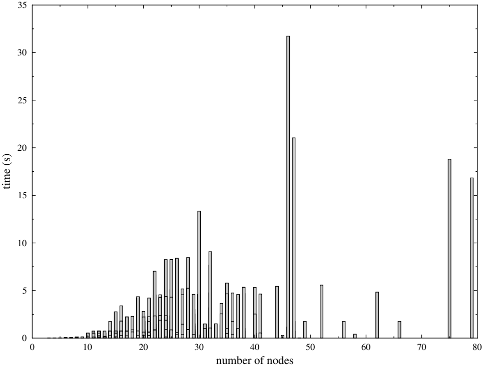
Preliminary results

- Evaluation method:
  - leave-one-out crossvalidation

Dataset	#examples	#positives	#negatives	Acc	AUROC
1	3085	1572	1513	69	0.75
2	3047	1520	1527	70	0.76
3	3278	1624	1654	70	0.76
4	3105	1545	1560	70	0.76
5	2426	1190	1236	70	0.76
6	3136	1607	1529	70	0.76
7	3049	1903	1146	69	0.73
8	3191	1648	1543	68	0.75
9	1053	701	352	70	0.72
10	1072	768	304	74	0.72

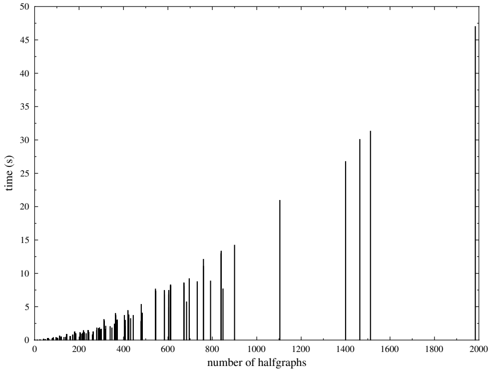
Time complexity

- molecule NCI 76026, #nodes = 30

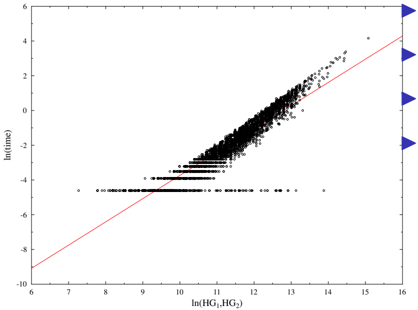


Time complexity

- molecule NCI 76026, #nodes = 30, #halfgraphs = 1104



Time complexity



- $y = 1.3379 * x - 17.1147$
- $O(HG) = (HG_G * HG_H)^{1.34}$
- $O(V) \sim (V_G^2 * V_H^2)^{1.34}$
- $O(V) \sim V_G^{2.68} * V_H^{2.68}$
- $O(V) \sim V^{5.36}$

Conclusions

- We introduced
  - a polynomial algorithm to find the size of the maximum connected common subgraph between two outerplanar graphs under the block and bridge preserving subgraph isomorphism
  - which can be used to construct a metric on outerplanar graphs and have a similarity measure between molecules
- Preliminary results
  - predictive performance
  - running time

Further work

- Full-scale experiments
  - investigating other distance measures, size functions, ...
  - comparison with similar algorithms and metrics
    - Swamidass et al. (2006)
    - Ceroni et al. (2007)
    - ...
- Investigation of other subclasses of graphs
  - 10% of molecules in this dataset are not outerplanar
  - look for other graph properties for which we can develop polynomial algorithms
    - e.g., graphs with bounded treewidth

Questions?