A Polynomial-time Metric for Outerplanar Graphs (Extended Abstract)

Leander Schietgat, Jan Ramon, Maurice Bruynooghe
Dept. of Computer Science, Katholieke Universiteit Leuven,
Celestijnenlaan 200A, 3001 Leuven, Belgium
{leander.schietgat,jan.ramon,maurice.bruynooghe}@cs.kuleuven.be

Abstract

In the chemoinformatics context, graphs have become very popular for the representation of molecules. However, a lot of algorithms handling graphs are computationally very expensive. In this paper we focus on outerplanar graphs, a class of graphs that is able to represent the majority of molecules. We define a metric on outerplanar graphs that is based on finding a maximum common subgraph and we present an algorithm that runs in polynomial time. Having an efficiently computable metric on molecules can improve the virtual screening of molecular databases significantly.

1 Introduction

Because of the increasing costs of drug development, pharmaceutical companies have shown great interest in virtual screening techniques, which automatically select from a database of molecules a set of candidates expressing a desired function. It is widely known that molecules with a similar structure tend to have the same function e.g., the binding capacity to some protein, so usually screening techniques will compare molecules by defining some kind of similarity measure between them. The task is then to search the database for molecules that are closest to a given molecule.

Graphs are excellent representations for molecules: each atom is then represented by a vertex and each atomic bond by an edge. However, there exist few efficient algorithms that can deal with graphs. Therefore, typical approaches try to describe a molecule based on a chemical "fingerprint", which consists of a bit-string listing the occurrence of a set of predefined features, such as a ring structure or a functional group. This reduces the computational complexity, but it ignores the underlying information about the molecule topology.

Still, previous work has shown that a lot of algorithms can be simplified as constraints are introduced on the structure of the graph. An example of a subclass of general graphs are outerplanar graphs. Horváth et al. [4] have

observed that 95% of the molecules in the NCI[1] database, which is used extensively for screening purposes, are outerplanar. In this paper we present a metric that computes a distance between outerplanar graphs in polynomial time. Having an efficient method to compare outerplanar graphs will result in significant speed-ups of molecule screening.

2 A Metric based on the Maximum Common Subgraph

All graphs in this text are assumed to be simple, undirected graphs. For an introduction to graph-related concepts and notation, the reader is referred to an introductory text on graph theory [2].

Bunke and Shearer [1] proposed a distance measure on graphs based on the maximum common subgraph (MCS):

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

with $|\cdot|$ determined by a function $size$. They proved that d_{bs} is a metric, i.e. it is a distance measure for which the properties of reflexivity, symmetry and triangle inequality hold. To understand this metric, some terminology is needed. A pair of graphs is said to be isomorphic if there is a one-to-one correspondence between their vertices and an edge only exists between two vertices in one graph if an edge exists between the two corresponding vertices in the other graph. A graph G is subgraph isomorphic to H if G is isomorphic to a subgraph of H. A maximum common subgraph (MCS) of two graphs G and H is then a graph I which is subgraph isomorphic to G and H and there exists no other graph J which is also subgraph isomorphic to G and H and $|J| > |I|$.

Unfortunately, the above metric is not practical, since computing the size of the MCS of two general graphs is not possible in polynomial time, unless $P = NP$ [3]. For this reason, we limit the space of input graphs to outerplanar graphs. These are graphs which can be embedded in the plane in such a way that all of their vertices lie on the boundary of the outer face, i.e. Fig. 1(a).

Still, even for two outerplanar graphs, finding an MCS remains NP-hard [7]. Therefore, we use a variant of the subgraph isomorphism, which only maps blocks to blocks and bridges to bridges. For a graph G, a block or biconnected component is a maximum subgraph of G, for which there is a cycle between every pair of its vertices. A bridge is an edge that does not belong to a block. It was proven that this so-called block-and-bridge preserving (BBP) subgraph isomorphism runs in polynomial time when considering outerplanar graphs [4]. From an application point of view, the BBP subgraph isomorphism can be motivated from the fact that in molecules cyclic structures and linear fragments usually behave differently, and hence treating them separately is not necessarily a bad thing.

[1] National Cancer Institute (`http://cactus.nci.nih.gov/`).

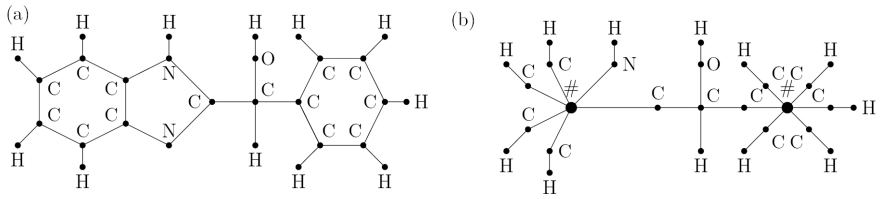

Figure 1: (a) An outerplanar graph. (b) Its corresponding block-bridge tree.

3 Algorithm

We have now reduced the problem to finding the size of the MCS of two outerplanar graphs under the BBP subgraph isomorphism. For this we use an algorithm based on a dynamic programming strategy. First, partial solutions are computed for pairs of smaller structures of the two graphs G and H, and then building on these, the size of the MCS is computed for structures of increasing size, until the size of the MCS of G and H themselves is computed.

We introduce two kinds of partial graph structures: the non-block-splitting subgraphs and the half-graphs. The former are generated by first transforming an outerplanar graph G into its block-bridge tree (Figure 1) [4] and then using this tree to generate all possible subgraphs of G in which blocks are not split. The latter are generated as follows. An embedding of a block of an outerplanar graph has a unique Hamiltonian cycle, going either clockwise ($\curvearrowright$) or counterclockwise ($\curvearrowleft$). In the context of a block B, if $o \in \{\curvearrowleft, \curvearrowright\}$ and $u, v \in V(B)$ are distinct vertices, we will denote with $o[u, v]$ the sequence of vertices along the Hamiltonian path of B between (and including) u and v, following the orientation o. We define the half-graph $G|_{o[u,v]}$ to be the maximum 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 (Figure 2).

The idea of the algorithm is then to consider all couples of non-block-splitting subgraphs and half-graphs for both graphs G and H in the order imposed by the function *size*. The subgraph isomorphism was proven to be polynomial for trees [6] as well as for biconnected outerplanar graphs [5]. Building on this, our algorithm computes the size of the MCS between all these partial structures,

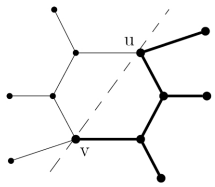

Figure 2: A half-graph $G\,|_{\curvearrowright[u,v]}$ (highlighted in bold).

making use of earlier computed solutions. The time complexity of this algorithm can be roughly bounded by $O(|V(G)|^{7/2}.|V(H)|^{7/2})$.

4 Conclusions and Further Work

We have introduced a polynomial-time algorithm that computes the MCS of two outerplanar graphs under the block and bridge preserving subgraph isomorphism, in order to have a metric on outerplanar graphs. It will be possible to use this metric as a similarity measure between molecules.

The next step is to conduct a series of experiments with the NCI database. As a proof of concept, we will evaluate the performance of an instance-based learner that uses our metric. Next, we plan a full comparison with similar algorithms and metrics, in terms of efficiency as well as predictive performance. Since 5% of the molecules in the NCI database cannot be represented by outerplanar graphs, we also plan to investigate other classes of graphs which would be suitable to represent molecules and for which we can design polynomial-time algorithms.

Acknowledgements

Leander Schietgat is supported by a PhD grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT Vlaanderen). Jan Ramon is a post-doctoral fellow of the Fund for Scientific Research (FWO) of Flanders.

References

[1] H. Bunke and K. Shearer. A graph distance metric based on the maximal common subgraph. *Pattern Recognition Letters*, 19:255–259, 1998.

[2] R. Diestel. *Graph Theory*. Springer-Verlag, 2000.

[3] M.R. Garey and D.S. Johnson. *Computers and Intractability: a Guide to the theory of NP-Completeness*. Freeman and Co., 1979.

[4] T. Horváth, J. Ramon, and S. Wrobel. Frequent subgraph mining in outerplanar graphs. In *Proceedings of the 12th ACM SIGKDD*, 197–206, 2006.

[5] A. Lingas. Subgraph isomorphism for biconnected outerplanar graphs in cubic time. *Theoretical Computer Science*, 63:295–302, 1989.

[6] R. Shamir and D. Tsur. Faster subtree isomorphism. *Journal of Algorithms*, 33(2):267–280, 1992.

[7] M. Syslo. The subgraph isomorphism problem for outerplanar graphs. *Theoretical Computer Science*, 17(1):91–97, 1982.