Latent Structure Pattern Mining

Andreas Maunz*, Christoph Helma†, Tobias Cramer*, and Stefan Kramer‡

*FDM, Universität Freiburg (D), †in-silico Toxicology Basel (CH), ‡Institut für Informatik, TU München (D)
email: {maunza}@fdm.uni-freiburg.de

Introduction

Graph mining algorithms have focused almost exclusively on ground features so far, such as frequent or correlated substructures. In the biochemical domain, Kazius *et al.* have demonstrated the use of more elaborate patterns that can represent several ground features at once. Such patterns bear the potential to reveal latent information which is not present in any individual ground feature. To illustrate the concept of non-ground features, Figure 1 shows two molecules,

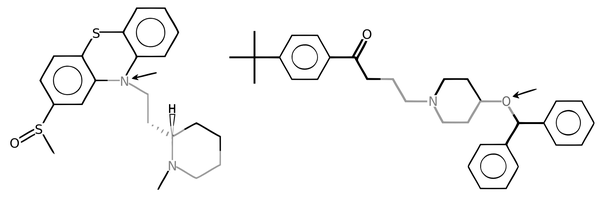

Figure 1: Two molecules with strong polarity, induced by similar fragments (gray).

taken from a biochemical study investigating the ability of chemicals to cross the blood-brain barrier, with similar gray fragments in each of them (in fact, due to symmetry of the ring structure, the respective fragment occurs twice in the second molecule). Note that the fragments are not completely identical, but differ in the arrow-marked atom (nitrogen vs. oxygen). However, regardless of this difference, both atoms have a strong electronegativity, resulting in a decreased ability to cross membranes in the body, such as the blood-brain barrier. So far, the identification of such patterns requires expert knowledge or extensive pre-processing of the data (annotating certain nodes or edges by wildcards or specific labels).

Compression Pipeline

We assume a graph database $R = (\mathbf{r}, a)$, where $\mathbf{r}$ is a set of undirected, labeled graphs, and $a : \mathbf{r} \to \{0, 1\}$ is a function that assigns a class value to every graph (binary classification). Graphs with the same classification are collectively referred to as *target classes*. Every graph is a tuple $r = (V, E, \Sigma, l)$, where $l : V \cup E \to \Sigma$ is a label function for nodes and edges. An *alignment* of a graph r is a bijection $\phi_r : (V, E) \to P$, where P is a set of distinct, partially ordered, identifiers of size $n = |V| + |E|$, such as natural numbers. Thus, the alignment function applies to both nodes and edges.

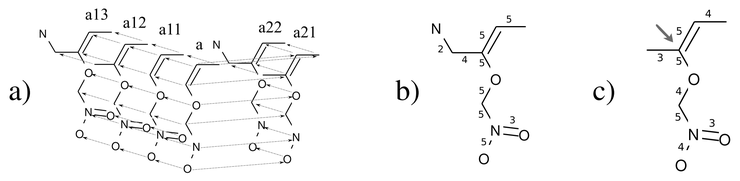

Figure 2: Illustration of the pipeline with the three steps (a) alignment in the partial order, (b) stack, and (c) compress.

For several ground features, alignments can be visualized by overlaying or *stacking* the structures. It is possible to count the occurrences of every component (identified by its position), inducing a weighted graph. Our approach aligns ground features significantly correlated with the target class, builds a weighted edge graph from that, and extracts the latent structure via Singular Value Decomposition (SVD). To obtain the alignments, it exploits the canonical ordering of ground features induced by a depth-first search algorithm.

Theory

Let r and s be graphs. A *maximum refinement* m of r and s is defined as $(r \preceq m) \wedge (s \preceq m) \wedge (\forall n \succeq r : m \succeq n) \wedge (\forall o \succeq s : m \succeq o)$.

Lemma 1: Let r and s be two aligned graphs. Then the following two configurations are equivalent:

1. There is no maximum refinement m of r and s with alignment ϕ_m induced by ϕ_r and ϕ_s, i.e. $\phi_m \supseteq \phi_r \cup \phi_s$.
2. A conflict occurs between r and s, i.e. either
   (a) $v_i \neq v_j$ for nodes $v_i \in r$ and $v_j \in s$ with $\phi_r(v_i) = \phi_s(v_j)$, or
   (b) $e_i \neq e_j$ for edges $e_i \in r$ and $e_j \in s$ with $\phi_r(e_i) = \phi_s(e_j)$.

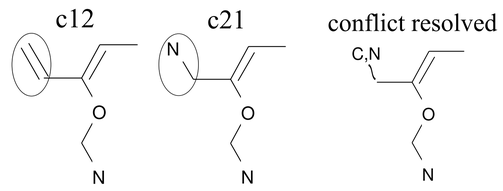

Figure 3: Conflict resolution by logical OR.

As a consequence of Lemma , conflicts prove to be barriers when we wish to merge several features to patterns, especially in case of patterns that stretch beyond the conflict position. A way to resolve conflicts is by logical OR, i.e. any of the two labels may be present (see Figure 3).

Building Complex Patterns

To extract the latent information we use Singular Value Decomposition (SVD, see Figure 2 c). Latent Structure Graphs are subsequently parsed depth-first to generate a SMARTS pattern[a] allowing for optional parts. "Standing" on a node forms weight-levels with the edges, leaving options how much branches to require:

1. nls (next level size) requires |highest level| edges.
2. msa (maximum size of all) requires |largest level| edges.
3. nop prohibits optional edges.

With nls and msa variants, we hope to better capture the information in the latent patterns.

Example: In Figure 2 c), the arrow-marked atom would have nls:2, msa:2, nop:3.

SMARTS patterns are "regular expressions for molecules", providing a wide range of powerful filters. With the help of *recursive SMARTS*, we defined LAST-SMARTS[a], that allows for recursive nesting of optional parts of the structure.

Example: `[#7][#6;$([#6]([#7])([#7])=[#6]),$([#6]([#7][#6])([#7])=[#6])](~*)=[#6]`

This describes a nitrogen connected to a carbon, double connected to a carbon (bold). The middle carbon's environment is described recursively, allowing optional branching to a nitrogen or to a nitrogen and a carbon. The notation $(\sim *)$ ensures that one of the branches is actually attached. Figure 4 shows the matching for this pattern on several molecules. Indeed, it does not match when the branch is missing.

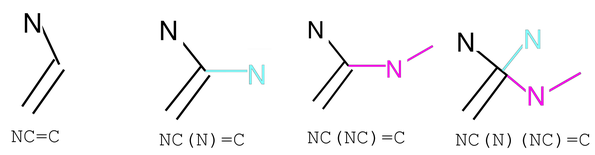

Figure 4: Matching of a complex pattern (red).

Experiments

We compared our method to other substructural descriptors that compress the search space, namely MOSS [1] and BBRC [4]. Furthermore, we compared against the set of all ground features, from which LAST-PM descriptors were derived (baseline comparison), see Figure 5.

We related feature count and runtime of LAST-PM and ALL (median of 20 folds). FCR is the feature count ratio, RTR the runtime ratio between LAST-PM and ALL. Since $1/FCR$ always exceeds RTR, we conclude that the additional computational effort is justified. Profiling showed that most CPU time is spent on alignment calculation, while SVD can be neglected.

External Test Sets

We predicted the external *bioavailability* test set by Yoshida and Topliss [5], comprising 40 compounds,

[a] See supporting website at http://last-pm.maunz.de for additional information about SMARTS language.

as well as the external *blood-brain barrier* test set by Hu and coworkers [3], comprising 64 positive and 32 negative compounds. Results are shown in Figure 6.

Dataset	LAST-PM			ALL	BBRC	MOSS
	Variant	%Train	%Test	%Test	%Test	%Test
bloodbarr	nls+nls	84.19	72.20	70.49[a]	68.50[a]	67.49[a]
nctrer	nls+msa	88.01	80.22	79.13	80.22	77.17[a]
yoshida	nop+msa	82.43	69.81	65.19[a]	65.96[a]	66.46[a]

[a] significant difference to LAST-PM. [b] result from the literature, no testing possible.

Dataset	LAST-PM	ALL	FCR/RTR
bloodbarr	249 (1.23s)	1613 (0.36s)	0.15 /3.41
nctrer	193 (12.49s)	22942 (0.13s)	0.0084 /96.0769
yoshida	124 (0.28s)	462 (0.09s)	0.27 /3.11

Figure 5: Repeated ten-fold crossvalidation

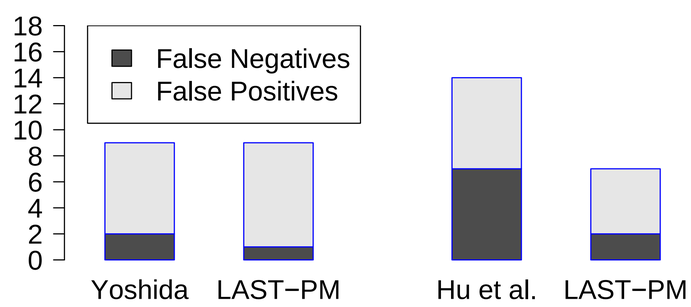

Figure 6: Classification performance on external test sets.

We conducted further experiments with another 110 molecule *blood-brain barrier* dataset (46 active and 64 inactive compounds) by Hou and Xu [2], that we obtained together with precomputed physico-chemical descriptors. Here, we were able to improve their results by combining physico-chemical descriptors with LAST-PM features.

Conclusions

LAST-PM is a novel approach to directly mine latent structures not present in any individual ground feature. Those features significantly outperform state-of the art substructural descriptors, including the ones from which they were derived. The key experimental results were obtained by comparison to physico-chemical descriptors on blood-brain barrier (BBB) and bioavailability data, which have been hard for substructure-based approaches so far.

Acknowledgements

The research was supported by the EU seventh framework programme under contract no Health-F5-2008-200787 (OpenTox).

References

[1] Heiko Hofer, Christian Borgelt, and Michael R. Berthold. Large scale mining of molecular fragments with wildcards. *Intell. Data Anal.*, 8(5):495–504, 2004.

[2] T. J. Hou and X. J. Xu. Adme evaluation in drug discovery. 3. modeling blood-brain barrier partitioning using simple molecular descriptors. *Journal of Chemical Information and Computer Sciences*, 43(6):2137–2152, Oct 2003.

[3] Hu Li, Chun Wei Yap, Choong Yong Ung, Ying Xue, Zhi Wei Cao, and Yu Zong Chen. Effect of selection of molecular descriptors on the prediction of blood-brain barrier penetrating and nonpenetrating agents by statistical learning methods. *Journal of Chemical Information and Modeling*, 45(5):1376–1384, Aug 2005.

[4] Andreas Maunz, Christoph Helma, and Stefan Kramer. Large-scale graph mining using backbone refinement classes. In *KDD '09: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 617–626, New York, NY, USA, 2009. ACM.

[5] Fumitaka Yoshida and John G. Topliss. Qsar model for drug human oral bioavailability1. *Journal of Medicinal Chemistry*, 43(13):2575–2585, Jun 2000.

*Address for correspondence: Zentrum für Datenanalyse und Modellbildung, Universität Freiburg, Hermann-Herder Str. 3a, D-79104 Freiburg