Elaborate Graph Mining: Exploiting Structural Invariants and Latent Information in Graph Databases to predict REACH-relevant Endpoints

Andreas Maunz

Physikalisches Institut, Albert-Ludwigs-Universität, Freiburg,
Institut für Informatik / I12, Technische Universität München.

Abstract

In computational chemistry, Frequent Subgraph Mining has been widely applied to databases of compounds to identify functional groups for drug design or hazard detection. However, the result set is typically too large to be of use to most statistical learners, let alone humans. Moreover, many very similar fragments are retrieved this way.

Two methods are presented that reduce the set of substructures by structural compression and correlation to the endpoint under investigation which leads to tremendous speedup in computation, very high compression while retaining good coverage of the database, and high predictive accuracy.

Several classification models have been produced within Opentox for REACH-relevant endpoints. Predictions can be derived using the well-defined Opentox REST interface, routinely providing an estimation of Applicability Domain.

Correlated Subgraph Mining

Graph Mining with minimum frequency and correlation constraints.

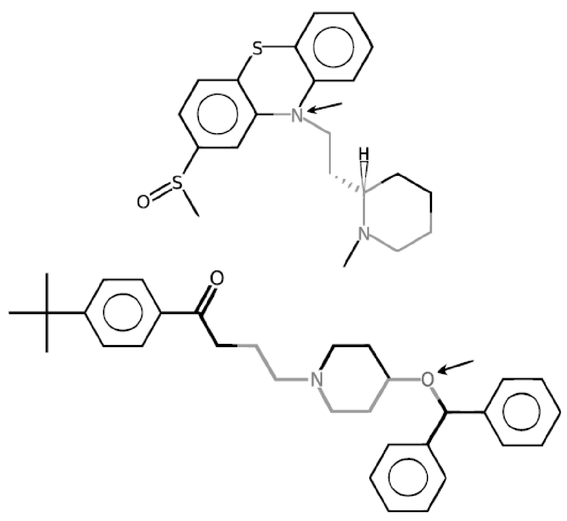

Figure 1: Latent Structure Discovery: Two compounds sharing a similar (non-identical) pattern.

Backbone Refinement Class Mining [4, 5]
BBRC builds a robust collection of structurally diverse descriptors.

- High compression potential (by structural invariant)
- Datasets of $> 20,000$ compounds can be processed in a few minutes.

Latent Structure Pattern Mining [3]
LAST-PM extracts latent (hidden) motifs from a graph database.

- Produces elaborate patterns, integrating structural ambiguities.
- Compares favorably to highly optimized physicochemical descriptors.

Modelling

Lazar (Lazy Structure- Activity Relationships) implements automatic similarity search / categorization by finding compounds similar to the query structure in terms of structure *and* endpoint activity [1, 2]. It uses fingerprints based on SMARTS patterns for feature representation.

```
compound/InChI=1S/C6H7N3O/c1-9(8-10)1H3:
- "[#7&A]-[#6&a](:[#6&a]:[#6&a])(:[#7&a])"
- "[#8&A]=[#7&A]-[#7&A]-[#6&A]"
 ...
```

Figure 2: Fingerprints describe compounds.

The derived similarity is based on fingerprints and weighted by significance of the features.

Applicability Domain Estimation

Any Lazar prediction has an associated confidence value. Confidence values are general, uncalibrated scores (not probabilities), describing neighbor similarity. It only holds that a higher score indicates a higher probability for a correct prediction.

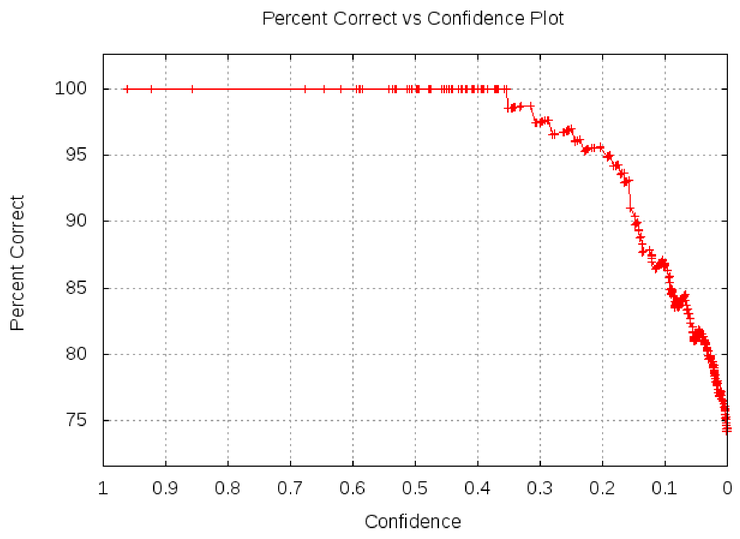

Figure 3: Applicability Domain estimation using confidence values.

Learning Modules

Lazar algorithms include:

- Weighted Tanimoto Kernel SVM
- RBF Kernel SVM
- Weighted Multilinear Regression
- Weighted Majority Classification

Models use BBRC or LAST-PM descriptors by default, but includes support for numeric features as well.

REACH-Relevant Endpoints

Within Opentox, the following endpoints have been modelled:

- ISSCAN Carcinogenicity
- Micronucleus Data from Rat/Mouse
- *Salmonella* Mutagenicity (CPDB)

Results are averages over five times 10-fold crossvalidation.

	n	Accuracy	Weighted Accuracy
ISSCAN Canc	1069	0.69	0.74
Micronucleus	136	0.54	0.56
Mutagenicity	808	0.75	0.76

Weighted accuracy is the (normalized) product of accuracy and confidence.

REST web services

Download the OpenTox Virtual Appliance from

`http://opentox.org/downloads`

Now BBRC, LAST-PM and Lazar are available as OpenTox API compliant REST web services:

	URI
BBRC	`http://localhost/fminer/bbrc`
LAST-PM	`http://localhost/fminer/last`
Lazar	`http://localhost/lazar`

Source code is available via Github:

`http://github.com/opentox`

References

[1] Christoph Helma. Lazy Structure-Activity Relationships (lazar) for the Prediction of Rodent Carcinogenicity and Salmonella Mutagenicity. *Molecular Diversity*, pages 147–158, 2006.

[2] Andreas Maunz and Christoph Helma. Prediction of Chemical Toxicity With Local Support Vector Regression and Activity-specific Kernels. *SAR and QSAR in Environmental Research*, 19(5-6):413–431, July 2008.

[3] Andreas Maunz, Christoph Helma, Tobias Cramer, and Stefan Kramer. Latent Structure Pattern Mining. In José Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 6322 of *Lecture Notes in Computer Science*, pages 353–368. Springer Berlin / Heidelberg, 2010.

[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] Andreas Maunz, Christoph Helma, and Stefan Kramer. Efficient Mining for Structurally Diverse Subgraph Patterns in Large Molecular Databases. *Machine Learning*, 83:193–218, 2011.

OpenTox - An Open Source Predictive Toxicology Framework, funded under the EU Seventh Framework Program: HEALTH-2007-1.3-3 Promotion, development, validation, acceptance and implementation of QSARs (Quantitative Structure-Activity Relationships) for toxicology, Project Reference Number Health-F5-2008-200787 (2008-2011).