On the Co-Occurrence and Diversity of Backbone Refinement Classes

Andreas Maunz
Freiburg Center for Data Analysis and Modelling (FDM), Hermann-Herder-Str. 3a, 79104 Freiburg, Germany

This document analyzes the co-occurrence and diversity of backbone refinement class representatives. Specifically, an euclidean embedding in 2D based on co-occurrences and entropy, and a measure for mean similarity between the features based on maximum common subgraph is given for two molecular datasets.

1. Hannes Schulz, Christian Kersting, Andreas Karwath, ILP, the Blind, and the Elephant: Euclidean Embedding of Co-Proven Queries (Proceedings of the 19th International Conference on Inductive Logic Programming (ILP 2009) (*forthcoming*)).
2. Al Hasan, M., Chaoji, V., Salem, S., Besson, J., & Zaki, M. (2007). Origami: Mining Representative Orthogonal Graph Features. ICDM 2007. Seventh IEEE International Conference on Data Mining.
3. Andreas Maunz, Christoph Helma, 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 (*forthcoming*).

1 Introduction

We use the salmonella mutagenicity (SM) dataset from [3] to visualize and analyze the distribution of bbrc features on molecular data. We investigate their occurrences, their entropy and their diversity.

2 Co-Occurrence and Entropy

In this section, we use the method of Schulz *et.al.* [1] to embed bbrc features and molecules into a 2D-pane. The point coordinates are influenced by:

- Feature-Molecule co-occurrence: Molecules are placed close to the features they instantiate
- Feature-Feature co-occurrence: Co-occuring features are placed close to each other
- The entropy of patterns induced by the target classes of molecules (active/inactive)

For the feature-molecule co-occurrences, an instantiation matrices is used. The negative entropy of the features is multiplied on this matrix to associate molecules with the features that are most significant for the determination of their class, where the entropy is defined as

$$H(q) = - \sum_{c \in \{act., inact.\}} p_{c,q} \, log \, p_{c,q}, \tag{1}$$

where $p_{c,q}$ is the empirical (induced) probability of feature q being instantiated in target class c. Then, all co-occurrence values are combined in a likelihood function. Figure 1 shows the euclidean embedding. The data points consist of features (triangles) and molecules (circles). The distances are found by a local optimization procedure maximizing the log-likelihood of the data.

The lower the entropy of a feature, the brighter its color, indicating its potential to discriminate between classes (green for active, red for inactive). It can be observed that the differently colored features are nearly perfectly separated. Also, there are very few non-discriminating features (mainly in the center). Features are well distributed across the pane, with few clusters. Thus, bbrc features are highly discriminative and diverse.

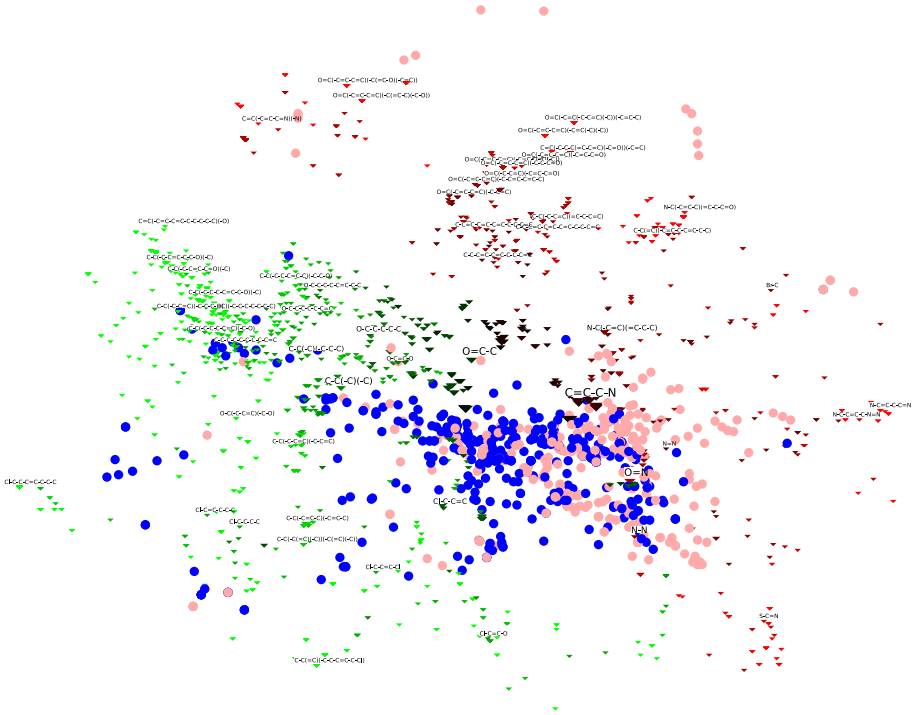

Figure 1: Euclidean embedding based on co-occurrence and entropy

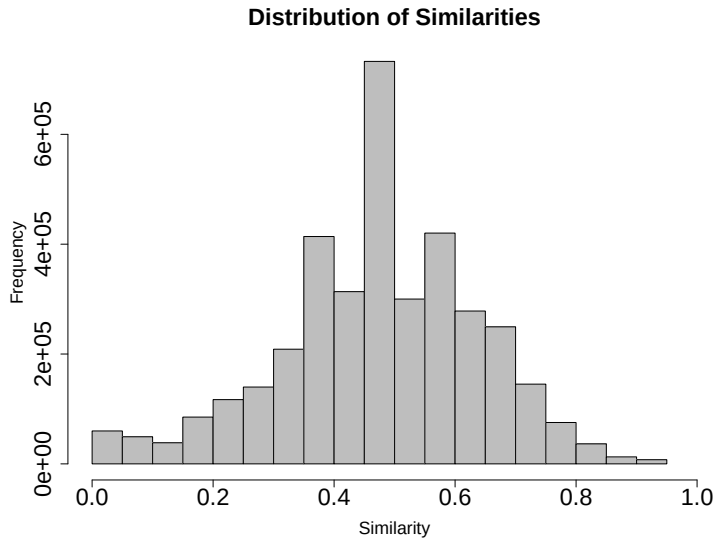

Figure 2: Histogram of similarity values between bbrc features

It can be clearly observed, that the molecule classes (blue for active, salmon for inactive) are also quite well separated, following the feature distribution to some extent. The main agglomeration of molecules (below the center) is widely stretched, especially for the active class (green). There are some remarkably well-distinguished and dense molecule groups in the outer parts, which always consist of members of the same class. Those groups are especially well characterized.

In summary, the analysis generates a 2D-embedding of molecules and features by combining co-occurrences and entropy information. The results indicate suitability of bbrc features for classifiation tasks.

3 Diversity

This section investigates the diversity of bbrc features. As in the ORIGAMI approach by Al Hasan *et.al.* [2], we calculate the mean similarity between two features as

$$sim(p,q) = \frac{|m_{pq}|}{max(|p|,|q|)}, \tag{2}$$

where m_{pq} is the maximum common (induced) subgraph of p and q and $|x|$ is the number of edges of x. The 2715 compounds had a mean (median) similarity of 0.4801 (0.5000), the 25% (75%) quantile was at 0.3846 (0.6000). The running time for mining the bbrc features with minimum frequency 6 were $0.51s$. A histogram is shown in Figure 2.

Al Hasan *et.al.* investigated similarity values between 0.15 and 0.35. They obtained such feature sets patterns by sampling from the positive border (the maximum patterns). Running times reported in their analysis were $\approx 2750s$ ($\approx 45m48s$) for 1000 features.

In summary, for bbrc features the tradeoff between mining time and structural diversity is favorable. The can be derived efficiently enough to be even used in "on demand"-settings, i.e. in ad-hoc situations.