Development of a Software Tool to Mine Emerging Patterns for Identification of Toxicophores

Richard Sherhod¹, Thierry Hanser¹, Philip N. Judson¹, Jonathan D. Vessey¹, Valerie J. Gillet²

¹Lhasa Limited, Leeds, UK
²Information School, University of Sheffield, Sheffield, UK
Contents

Introduction
• Background
• Goals

Design of the Tool
• Our approach
• Emerging pattern mining
• Our implementation
• Using the tool

Results – Discovering new Alerts for In Vitro Mutagenicity

Conclusions
Contents

Introduction
  • Background
  • Goals

Design of the Tool
  • Our approach
  • Emerging pattern mining
  • Our implementation
  • Using the tool

Results – Discovering new Alerts for In Vitro Mutagenicity

Conclusions
Background

A collaborative project between Lhasa Limited and the University of Sheffield:

- Knowledge transfer partnership
- Joint funded by Lhasa Limited, Technology Strategy Board and the Engineering and Physical Sciences Research Council
- Continuing my PhD research to build a prototype software tool
Background

Toxicity prediction

Potential toxicity of compounds is a crucial consideration in:

- Avoiding late-stage failures – reducing costs
- Gaining regulator approval – reducing animal testing

Toxicity prediction systems are important decision support tools

- e.g. knowledge based prediction systems which encode expert understanding and form reasoned arguments

Structural alerts are the basis of many knowledge-based toxicity prediction systems
**Background**

**Alerts:** Structural features known to result in a toxic endpoint

Alkylating agent alert description:

\[
R_1 = \text{Cl, Br, I, OS(=O)_nR_4} \\
R_2, R_3 = \text{any} \\
R_4 = \text{not OH} \\
n = 1, 2
\]

**Toxicophores:** Substructures responsible for toxic activity

Alkylating agent toxicophores:
Alert development

- Developers expand a toxicity knowledge base by incorporating new knowledge as structural alerts.
- Construction of new and improved structural alerts can be a slow process:

  - Data curation
  - Data preparation
  - Feature identification
  - Feature assessment
  - Alert composition
  - Alert testing

- Increasing the rate of alert formation is desirable
  - *We aim to help expedite the knowledge discovery steps*
Alert development

- Developers expand a toxicity knowledge base by incorporating new knowledge as structural alerts.
- Construction of new and improved structural alerts can be a slow process:
  - Data curation
  - Data preparation
  - Feature identification
  - Feature assessment
  - Alert composition
  - Alert testing

Increasing the rate of alert formation is desirable
- We aim to help expedite the knowledge discovery steps
Contents

Introduction
  • Background
  • Goals

Design of the Tool
  • Our approach
  • Emerging pattern mining
  • Our implementation
  • Using the tool

Results – Discovering new Alerts for In Vitro Mutagenicity

Conclusions
Goals

To expedite alert discovery:

- To develop a method and software tool that automatically identifies potential toxicophores from toxicity data
  - i.e. extracts significant features within toxic compounds
- To present results in an interpretable form and provide automated and interactive analytical techniques
- Use feedback from knowledge base developers to build a tool that will fit into their alert development workflows
Goals

To expedite alert discovery:

- Data curation
- Data preparation
- Feature identification
- Feature assessment
- Alert composition
- Alert testing
To expedite alert discovery:

Our tool

- Pre-processing tools
- Emerging pattern mining
- Post-processing tools

Data curation → Data preparation → Feature identification → Feature assessment → Alert composition → Alert testing
Contents

Introduction

• Background
• Goals

Design of the Tool

• Our approach
• Emerging pattern mining
• Our implementation
• Using the tool

Results – Discovering new Alerts for In Vitro Mutagenicity

Conclusions
Our Approach

Our approach to feature identification:

1. To mine emerging patterns (EPs) from toxicity data
2. To cluster molecules based on their support for emerging patterns
3. To group EPs and their support clusters into hierarchies
4. To present the hierarchies, EPs and clusters in an easily interpretable way
Contents

Introduction
• Background
• Goals

Design of the Tool
• Our approach
• Emerging pattern mining
• Our implementation
• Using the tool

Results – Discovering new Alerts for In Vitro Mutagenicity

Conclusions
Emerging Pattern Mining

**Emerging pattern**¹:

- A form of association rule
  - i.e. associates patterns of co-occurring properties with a data class
- Multiple patterns represent multiple SARs
- A supervised method – differs from Galois lattice approaches²
- Emerging pattern methods have been used for classification³
  - Our implementation helps to *explain* toxic activity

---

Emerging Pattern Mining

Emerging pattern$^1$: A pattern of binary properties that occurs more in the data entries of one class compared to another

<table>
<thead>
<tr>
<th>Molecules</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Molecules</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. This emerging pattern is characterized by a specific pattern of binary properties (X) that occurs more frequently in the data entries of one class compared to another.
Emerging Pattern Mining

Emerging pattern\(^1\):
A pattern of binary properties that occurs more in the data entries of one class compared to another

<table>
<thead>
<tr>
<th>Molecules</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>4</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Actives

<table>
<thead>
<tr>
<th>Molecules</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

Inactives
Emerging pattern support:

- The set of data entries that support an emerging pattern may be considered a cluster
  - The clustering is supervised i.e. with reference to both classes

<table>
<thead>
<tr>
<th>Molecules</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>5</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Actives:

- {a, c}
- [1, 2, 3, 6]

<table>
<thead>
<tr>
<th>Molecules</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Inactives:
Emerging Pattern Mining

**EP mining algorithm: Contrast pattern tree mining**¹:

- An extension of frequent pattern growth (FP-growth) mining²
- Rearranges a dataset into a tree representation
  - Encodes all combinations of binary properties and their frequency in each class

---

1. Fan, H. IEEE Transactions on Knowledge and Data Engineering 2006, 18, 721–737.
Emerging Pattern Mining

**EP mining algorithm: Contrast pattern tree mining:**

<table>
<thead>
<tr>
<th>Molecules</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>4</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>6</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

```
<table>
<thead>
<tr>
<th>b</th>
<th>a</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>a</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>b</th>
<th>a</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>a</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
</tbody>
</table>
```
Emerging Pattern Mining

EP mining algorithm: Contrast pattern tree mining:

<table>
<thead>
<tr>
<th>Molecules</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>6</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Emerging Pattern Mining

EP mining algorithm: Contrast pattern tree mining:

a

\[ \overset{\text{N}^+}{\overset{\text{O}}{\text{N}}} \overset{\text{O}^-}{\text{O}} \]

b

\[ \text{Cyclic compound} \]

c

\[ \text{Cyclic compound} \]

d

\[ \text{Cyclic compound} \]
Emerging Pattern Mining

EP mining algorithm: Contrast pattern tree mining:

- Searches the tree (depth first) and grows patterns of properties that:
  - Are frequent in the data
  - Are more frequent in one data class compared to another
  - Significantly change in frequency as they are grown
- Continuously builds and prunes the tree while searching
  - Reduces memory footprint
Emerging Pattern Mining

EP mining algorithm: Emerging pattern selection:

“Generalised noise-tolerant emerging patterns”¹

- Noise tolerant
- Tolerant of imbalanced datasets
- Can be configured for different datasets

¹ Fan, H. IEEE Transactions on Knowledge and Data Engineering 2006, 18, 721–737.
Contents

Introduction
  • Background
  • Goals

Design of the Tool
  • Our approach
  • Emerging pattern mining
  • Our implementation
  • Using the tool

Results – Discovering new Alerts for In Vitro Mutagenicity

Conclusions
Our Implementation

Accepts toxicity data described using any binary fingerprints
Our Implementation

- Extracts patterns that are more frequent in active molecules
  - i.e. patterns that are characteristic of toxicants
- The set of toxicants that support an emerging pattern represent a potentially toxic cluster
  - i.e. a cluster formed around a potential toxicophore
Our Implementation

The hierarchies represent families of structures.

Fingerprints of actives → Emerging pattern mining → Fingerprints of inactives

Arranges patterns and support clusters into hierarchies.

Active emerging patterns → Support hierarchy formation

The hierarchies represent families of structures.
Our Implementation

- The patterns are arranged by the subset-superset relationships between their support sets.

Diagram:
- Higher support
  - [1 2 3 4 5] \{ab\}
  - [1 2 3] \{abc\}
  - [1 2] \{abcd\}

- Lower support
  - [1 2 5] \{abd\}
  - [1 2] \{abcd\}

- More generic patterns
  - [4 5] \{abe\}

- More specific patterns
Our Implementation

Exploring the hierarchies allows relationships between structures to be analysed.
Contents

Introduction
  • Background
  • Goals

Design of the Tool
  • Our approach
  • Emerging pattern mining
  • Our implementation
  • Using the tool

Results – Discovering new Alerts for In Vitro Mutagenicity

Conclusions
The knowledge discovery tool: Mercury

- Uses Lhasa Limited’s CERES chemistry engine
- Based on the Coralie framework
Using the Tool
Using the Tool
Using the Tool

**Importation options**
Select your dataset and define the activity/prediction tags.

- **Source**:
  - SDF source file: C:\Users\richards\CorallieSpace\KBDemoSpace\Demo\benchmark-derek.sdf

- **Define activity**:
  - True activity
    - Activity field: 
    - Activity: 1
      - Active: 1
      - Inactive: 0
      - Value: 2689
    - Activity field: 
    - Activity: 2
      - Active: 1
      - Inactive: 0
      - Value: 2311
  - Test activity
    - Activity field: 
      - Activity: 3
        - Active: 1
        - Inactive: 0
        - Value: 2314
    - Activity field: 
      - Activity: 4
        - Active: 1
        - Inactive: 0
        - Value: 2312

- **Dataset**:
  - Name: benchmark-derek
  - Endpoint: benchmark-derek
  - Description: Built from benchmark-derek.sdf

- **Output**
  - Dataset file name (*.dts): C:\Users\richards\CorallieSpace\KBDemoSpace\Demo\benchmark-derek.dts
Using the Tool

Define experimental activity

Activity classes: active, inactive, neutral
Using the Tool

Define experimental activity

Define test/predicted activity

Activity classes: active, inactive, neutral
Using the Tool
Using the Tool
Using the Tool
Using the Tool
Using the Tool
Using the Tool
Contents

Introduction
  • Background
  • Goals

Design of the Tool
  • Our approach
  • Emerging pattern mining
  • Our implementation
  • Using the tool

Results – Discovering new Alerts for In Vitro Mutagenicity

Conclusions
Discovering New Alerts

Using EP mining to suggest new mutagenicity alerts:

- EPs mined from a large public mutagenicity dataset
- EPs and clusters that suggested gaps in Lhasa Limited’s Derek Nexus knowledge base were identified
- Relevant literature was searched for a mechanistic rationale for the EPs
  - Looking for evidence that the clusters are a genuine class of toxicants
Discovering New Alerts

Dataset:
- Internally curated Hansen\textsuperscript{1} benchmark mutagenicity dataset
- 3489 toxicants : 2981 innocuous compounds

Descriptors:
- 1949 fragments and ring structures
- Very sparse fingerprints

Results:
- 604 EPs in 181 hierarchies
- 6 clusters with a high number of false negatives by Derek Nexus were sent to the knowledge base department

Discovering New Alerts

23 mutagens : 8 non-mutagens : 13 false negatives

c13c(Cc2c(O1)cccc2)cccc3
Discovering New Alerts

23 mutagens : 8 non-mutagens : 13 false negatives

c13c(Cc2c(O1)cccc2)cccc3
Discovering New Alerts

43 mutagens : 19 non-mutagens : 20 false negatives
Discovering New Alerts

43 mutagens : 19 non-mutagens : 20 false negatives

C2Oc1ccccc1C2
Discovering New Alerts

C2Oc1cccccc1C2
Discovering New Alerts

C2Oc1cccc1C2
Discovering New Alerts

Five alerts submitted for addition to the knowledge base

• Four new alerts
• One extension to an existing alert

Reduced the time needed to analyse a dataset of this size by 60%

• Most of the remaining time spent analysing relevant literature
Contents

Introduction
- Background
- Goals

Design of the Tool
- Our approach
- Emerging pattern mining
- Our implementation
- Using the tool

Results – Discovering new Alerts for In Vitro Mutagenicity

Conclusions
Conclusions

- Developed a software tool that automatically identifies potential toxicophores from toxicity data
  - Presents results in an interpretable form
  - Applies a robust feature identification method
- Demonstrated the tool’s ability to identify toxicophores in large datasets
  - Expanded Lhasa Limited’s Derek Nexus knowledge base
- The knowledge discovery tool is being used by toxicologists now to develop structural alerts
Acknowledgments

Supervisors:

• Prof. Valerie Gillet
• Dr Jonathan Vessey

Collaborators:

• Dr Philip Judson
• Dr Thierry Hanser

Lhasa Limited’s research group and developers

• Dr Chris Barber
Development of a Software Tool to Mine Emerging Patterns for Identification of Toxicophores