1. MCS definitions
2. MCS similarity searching and applications
3. Methodology
   1. Algorithms
   2. Virtual Screening
4. Results
   1. Recall
   2. Speed
5. Conclusions
Maximum Common Substructure

Subgraphs and Common Substructure

Induced and Edge MCS

• MCS – largest common substructure
• Maximum Common Induced Substructure (MCIS)
  • Common atoms with edges in between
• Maximum Common Edge Substructure (MCES)
  • Only considers common edges
Definitions

Connected and Disconnected MCS

cMCS

dMCS
MCS Applications

Examples
- Similarity searching (and clustering)
- Reaction mapping
- Coordinate stamping
- Hyperstructure construction

Similarity Searching against Fingerprints
- Raymond and Willett (2002) – disconnected MCS had comparable efficacy to BCI fingerprints
- van Berlo et al (2009) – connected MCS competitive with ECFP4 for predicting gene transcript levels
- However, what about comparisons of several MCS types?

Methodology - dMCS Variants

Topologically constrained dMCS (tdMCS):

- Absent of topologically undesirable matched edges
- Given 2 pairs of aligned edges, what is their difference in path distance per molecule?

\[ \Delta D = 5 \] in this example – large difference in distances.

- Apply constraint “\( \theta \)” – the maximum topological distance
- Strict \( \theta \) value reduces search space -> faster MCS calculation

Methodology - dMCS Variants

John Raymond Modular Product Heuristics (hmcs):

- Reduces search space by simplifying modular product (clique detection methods) via exploits:
  - Benzene ring symmetry
  - Topological distance constraints
    (on ring systems)

Methodology – MCS algorithms and types

- 5 MCS algorithms
- Each algorithm suitable for different MCS types:
  - cMCS
  - dMCS (and the related hMCS)
  - tdMCS ($\theta = 0, 1, 2$)

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Precision</th>
<th>cMCS</th>
<th>dMCS</th>
<th>hMCS</th>
<th>hMCS (no rings)</th>
<th>tdMCS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChemAxon</td>
<td>Approx</td>
<td>🎉</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>consR</td>
<td>Approx</td>
<td>🎉</td>
<td>🎉</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kcombu</td>
<td>Approx</td>
<td>🎉</td>
<td>🎉</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fMCS</td>
<td>Exact</td>
<td>🎉</td>
<td></td>
<td>🎉</td>
<td>🎉</td>
<td></td>
</tr>
<tr>
<td>MaxCliqueSeq</td>
<td>Exact</td>
<td>🎉</td>
<td>🎉</td>
<td>🎉</td>
<td>🎉</td>
<td>🎉</td>
</tr>
</tbody>
</table>
Methodology – Virtual Screening

Dataset
• Subset of 50 ChEMBL activity classes identified by Riniker and Landrum
  • Subset of 6 – 2 homogeneous, 2 heterogeneous, 2 of intermediate diversity
  • 100 actives, 10000 inactives (inactives recycled per class)

Similarity Searching
• 5 diverse actives selected using MaxMin algorithm as references
• 10 replicates (different random seed for MaxMin) per class
• Similarity – MAX fusion of Tanimoto similarity of 5 references, to database compound

\[ S_T = \frac{c}{a + b - c} \]

a = 15 - Bonds in molecule 1
b = 16 - Bonds in molecule 2
c = 14 - Bonds in MCS
\[ S_T = 0.824 \]

(Riniker and Landrum, 2013, *J. Cheminform.*, 5)
Methodology – Virtual Screening

Similarity search for an activity class:

- mECFP4 Fingerprint
- MAX group fusion
- MCS group fusion
- Repeat 10 times (change MaxMin seed)
virtual screening recall performance

- mECFP4 - RDKit Morgan Fingerprints at radius 2 (similar to Pipeline Pilot ECFP4)
- Bold borders – method significantly different to mECFP4 (p < 0.05)

Observations

- All methods apart from tdMCS (θ = 0 or 1) are generally beaten by fingerprints
- Approximate methods for dMCS competitive with exact hMCS
Similar observations can be made across all the classes
Virtual Screening Recall Performance (Data Fusion)

- SUM Fusion of fingerprint and MCS method (add the ranks of the 2 similarities)
- tdMCS ($\theta = 0$) generally the best, usually outperforms fingerprints alone!
- cMCS-fingerprint fusion at worst competitive with fingerprints

Observations
Virtual Screening Time Performance

- tdMCS ($\theta = 0$) fastest, even faster than the inexact methods
- For ~50 000 compound pairs evaluated, tdMCS ($\theta = 0$) never exceeds 60 seconds total.
- Potentially viable for alternative large database searching?

(Note: no significance tests shown)
Conclusions and Future Work

Outlook

• Exact algorithms are not necessary for similarity screening (given the usefulness of current approximate algorithms)

• Topological restrictions are highly desirable for improving recall and speed

• Fusion with fingerprints improves results, and with tdMCS brings recall over fingerprints alone

• Aim to release the MCS algorithms via KNIME

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>consR</td>
<td>Zhu et al. (2013), <em>VLDB Journal</em>, 22(3), 345-368</td>
</tr>
<tr>
<td>fMCS</td>
<td><a href="https://bitbucket.org/dalke/fmcs">https://bitbucket.org/dalke/fmcs</a></td>
</tr>
</tbody>
</table>

Questions?
Maximum Common Substructure (MCS)

Subgraphs and Common Subgraphs

Connected Maximum Common Substructure (cMCS)
Maximum Common Substructure (MCS)

disconnected Maximum Common Substructure (dMCS)

- All the fragments common between two graphs, as opposed to the largest common fragment
- Advantageous for comparing molecules with different scaffolds

![Diagram showing examples of dMCS and cMCS]
Introducing Hyperstructures

Minimum Common Supergraph (MiCS) and the Hyperstructure

- A supergraph is the opposite of a subgraph, same for common supergraphs
- Although there is a finite set of subgraphs, there are infinite possible subgraphs of a graph
- A Hyperstructure is a common supergraph between multiple molecules
- Typically the MCS is involved in hyperstructure construction of 2 molecules

Several possible hyperstructures could exist:

![Example 1](image1.png)  ![Example 2](image2.png)

G₁  G₂
Plan

1) Title & Contents
2) MCS – what is it? Emphasis on dMCS (contrast with cMCS)   MCES, not MCIS. Largely arbitrary reason though it often covers more than MCIS does.
   - NP completeness of problem? Modular product definition might help but I'm afraid of confusing people here (1 slide hopefully)
   - Illustrate importance of tdMCS alignments by talking about hyperstructures. This presentation will NOT be about hyperstructures, but I believe they serve a useful purpose for demonstrating "alignment quality" (max 2 slides)
   - Applications of MCS. Stress little academic knowledge on similarity searching. Previous comparisons of MCS algorithms? I'm thinking of Eleanor and Andrew's work, as well as 2 more recent (though less exhaustive) examples I've found (max 1 slide). Comparisons against Fingerprints – MCS never better than FP

.MCS algorithm and type benchmark

- what algorithms and MCS types am I comparing for virtual screening? The work that I've done on benchmarking the algorithms I'll mention in passing, but again I don't think most people will actually care about this if I've already told them which are the fastest ones I've found anyway. The algorithms I've used for virtual screening is a small subset of the ones I've benchmarked for time/size performance (max 1 slide)
- Introduce benchmark dataset of Riniker and Landrum (1 slide)
- BEDROC results
- Time results (VS only)
- Data Fusion with fingerprints – BEDROC only
  Why do fingerprints perform so well? Matching of redundant common features (benzene rings)?
- Summary and conclusions (what's the best MCS, best algorithm? Exact vs approximate methods)?
- Maybe as final slide show a comparison table of MCS algorithms from Chapter 6 of my thesis