Introduction

Maximum common substructure (MCS) search is a computationally hard problem that is of great importance in multiple aspects of chemoinformatics. Its diverse applications include:

- Similarity search
- Clustering
- Reaction mapping
- Molecule alignment
- Matched molecular pair (MMP) search

Examples:

![Connected MCS](image1)

![Disconnected MCS](image2)

Large example (disconnected MCS)

Aim

The aim of our research was to develop highly efficient heuristic MCS algorithms. Our implementations are optimized for both running time and accuracy. We also consider the chemical relevance of the result, which includes customizable matching rules, keeping rings together, and providing correct mappings for reaction mapping and molecule alignment.

Heuristics

Apart from comparing existing algorithms and heuristics, for example the connectivity heuristic, which we have found very powerful; new heuristics and methods have also been developed.

One of the most important ones is based on ECFP hash codes. We assign hash values to atom neighborhoods, and prefer matching atoms which have large neighborhoods whose hash values match.

Representation

A new representation for the compatibility graph in max-clique based methods has been developed, which greatly improves both memory usage and running time.

![Graphical representation](image3)

Measurements of the memory usage of the new representation on graphene structures of varying size. The empirical results nicely support the theoretical improvement from $O(m^4)$ to $O(m^3)$ where $m$ denotes the bond count of the input molecules.

Results

The effectiveness of the mentioned heuristics was thoroughly measured and compared.

![Graphical representation](image4)

Typical results when comparing the size of the found MCS with and without using heuristics.

![Graphical representation](image5)

Running times for the same test set.

Mapping optimization

Another improvement is that the one-to-one atom correspondence defined by the found common substructure is also taken into account. This is especially important in applications like molecule alignment or reaction mapping.

Examples:

![Non-preferred mapping](image6)

![Preferred mapping](image7)

![Non-preferred mapping](image8)

![Preferred mapping](image9)

The developed solution handles these simple cases and handles many other more complex cases as well.

Customizability

Methods have been developed to provide options like keeping rings together, enumerating multiple common substructures, or providing fully customizable matching options. The possible non-local and non-transitive nature of these matching rules require special considerations.

Examples:

![Default MCS](image10)

![MCS where atoms in rings can only match atoms in rings of similar size](image11)

![MCS keeping rings](image12)

Conclusion

Effective heuristics have been developed to improve both the performance and the accuracy of MCS algorithms, as well as to provide chemically more relevant results for several applications.

The resulting implementations are included in the 6.1 version of ChemAxon's JChem software suite.