SMALLWORLD : EFFICIENT MAXIMUM COMMON SUBSTRUCTURE SEARCHING OF LARGE DATABASES

Roger Sayle, Jose Batista and Andrew Grant

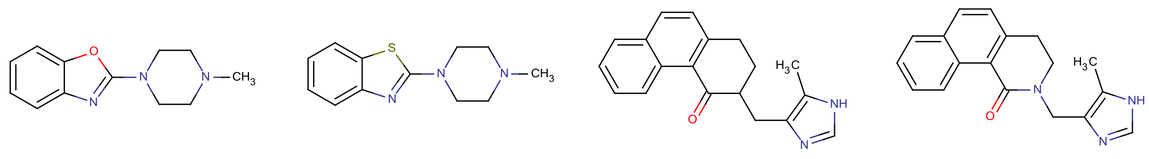
NextMove Software Ltd, Cambridge, UK and AstraZeneca R&D, Alderley Park, UK

1. Abstract

We report a novel chemical database search method based upon explicit representation of chemical space. A pre-computed index allows the exact size of the maximum common edge subgraph (MCES) between a query molecule and molecules in the index to be calculated rapidly. In practice, this allows the 100 nearest neighbours having the largest MCES to a query molecule to be determined in a few seconds, even for target databases containing millions of molecules. This work builds upon the previous efforts of Wipke and Rogers in the late 1980s and of Messmer and Bunke in the 1990s, but takes advantage of the dramatic advances in parallel processing and storage technology now available to researchers.

2. 2D Chemical Similarity

Most 2D chemical similarity methods, including fingerprints, may be considered "local" methods, where a molecular graph is reduced to a vector of features, and similarity between these vectors treated as a surrogate of graph similarity. This reduction to 1D similarity results in a number of artefacts; binary fingerprints such as ECFP, MACCS keys and Daylight fingers saturate and fail to distinguish alkanes, peptides or nucleic acids. "All the right notes, not necessarily the right order". These artefacts mean existing method don't always match chemist's intuition.



Daylight Tanimoto: 0.44 Daylight Tanimoto: 0.39

3. Analogy to Evolution of Bioinformatics

The mathematic concept of string edit distance as a similarity measure between strings was introduced by Vladimir Levenshtein in 1965, as the minimum number of insertions, deletions and substitutions to transform one string into another. Thanks to the efficient dynamic programming algorithms of Needleman and Wunsh, and Smith and Waterman, edit distance now underpins modern biological sequence comparison. Prior to this, methods were restricted to "local" heuristic methods, comparing vectors of k-tuples (short runs of characters).

4. Graph Edit Distance (GED) and Maximum Common Subgraph (MCS)

Graph Edit Distance (GED) as an analogous concept for similarity between graphs was introduced by Sanfeliu and Ku in 1983 [1], and was shown to be related to the Maximum Common Subgraph (MCS) by Bunke in 1997 [2,3]. Given the size of the MCS between two molecules one can determine the edit distance, and likewise given the edit distance it is possible to calculate the size of the MCS. Although intuitive to chemists, the NP-Hard computational complexity of calculating the MCS of two molecules has traditionally prohibited its practical use in chemical database searching.

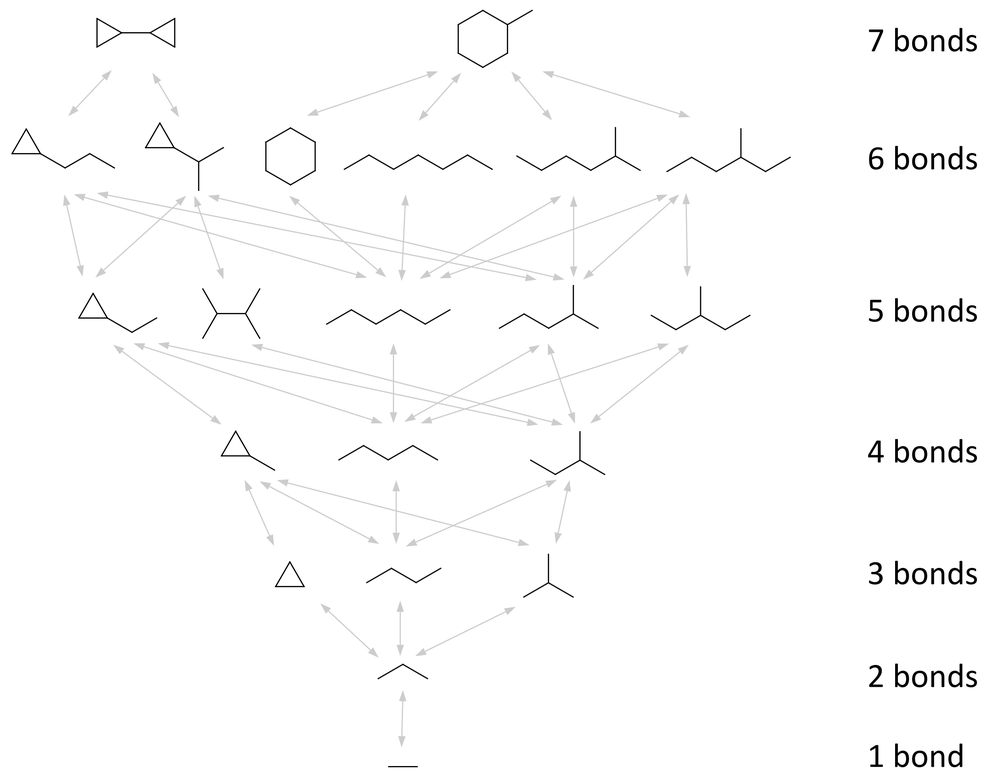
5. SmallWorld

NextMove Software's **SmallWorld** finesses these issues by taking advantage of recent advances in computer science and technology, using pre-computation to make MCS/GED searching faster than fingerprint based searches. Conceptually, SmallWorld explicitly encodes the relationship between molecular graphs and all of their possible subgraphs. Edit Distance and MCS queries become "Big Data" graph neighbour problems much like those encountered in social network sites (like LinkedIn or Facebook), or Graph500 supercomputer benchmarks. Advanced algorithms and current storage technology allow management of databases of many billions of subgraphs, orders of magnitude larger than current chemical databases.

Name	Atoms	MW	Anon SGs	Elem SGs
Aspirin	13	180	127	332
Ranitidine	21	314	436	1,207
Clopidrogel	21	322	10,071	22,170
Amlodipine	28	409	58,139	147,128
Lisinopril	29	405	24,619	34,496
Gefitinib	31	447	190,901	337,174
Atorvastatin	41	559	3,638,523	6,019,427

Unique subgraph counts (SGs) of drug-like molecules, supressing bond orders and either preserving (elem) or suppressing (anon) atomic numbers [i.e. topologies].

6. SmallWorld Chemical Universe



7. SmallWorld Database Index Statistics

Here we describe some features of a small SmallWorld index used as a proof-of-concept and to prototype alternative database construction, incremental update, maintenance and search implementations.

This "anonymous" (element and bond order suppressed) database was seeded with the contents of the NCBI PubChem and Accelrys MDDR databases. The resulting (incompletely enumerated) index contained 69 million nodes and 557 million edges, requiring around 40 Gbytes of disk space. Projections based on these experiments indicate a 95% indexed pharmaceutical registry database of a few million compounds should fit in a few terabytes of storage [costing a few hundred dollars at current prices].

The average degree of each node is between 16 and 17 edges. Of the 69 million nodes, 27 million represent acyclic graphs, and 21 million represent graphs with a single ring. Of the 557 million edges, 205 million are ring deletion edges and 352 are terminal bond deletion edges.

A Briem & Lessel benchmark run, searching for the (up to) 10 nearest neighbours of each of 380 drug-sized query molecules takes 8 minutes 42 seconds (less than 1.4 seconds per query) on a single thread on an Intel i7-2600 with 16 Gbytes RAM.

The median distance for each search to search was 8 edits, with the shortest search finishing after just two edits and the furthest search reaching 24 edits. Run times can be reduced to 7 minutes, by limiting searches to 10 edits. The worst-case "wave-front" size was 6,624,624 nodes which occurred at distance 9.

8. Conclusions

- The sub-linear behaviour of SmallWorld's nearest neighbour search makes it faster than fingerprint-based similarity methods for sufficiently large data sets.
- This is thanks to the "blessing of dimensionality"
- With continual advances in computer hardware, SmallWorld-like approaches are likely to become the basis of most chemical similarity calculations within a decade.

10. Bibliography

1. Alberto Sanfeliu and K.S. Fu, "**A Distance Measure between Attributed Relational Graphs for Pattern Recognition**", *IEEE Transactions of Systems, Man and Cybernetics (SMC)*, Vol. 13, No. 3, pp. 353-363, 1983.
2. Horst Bunke, "**On a Relation between Graph Edit Distance and Maximum Common Subgraph**", *Pattern Recognition Letters*, Vol. 18, No. 8, pp. 689-694, 1997.
3. Horst Bunke and Kim Shearer, "**A Graph Distance Metric Based on the Maximal Common Subgraph**", *Pattern Recognition Letters*, Vol. 19, pp. 255-259, 1998.
4. Bruno T. Messmer and Horst Bunke, "**Subgraph Isomorphism in Polynomial Time**", Technical Report, University of Berlin, 1995.
5. Bruno T. Messmer and Horst Bunke, "**Subgraph Isomorphism Detection in Polynomial Time on Preprocessed Model Graphs**", In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 151-155, 1995.

NEXTMOVE SOFTWARE

www.nextmovesoftware.co.uk
www.nextmovesoftware.com

NextMove Software Limited
Innovation Centre (Unit 23)
Cambridge Science Park
Milton Road, Cambridge
England CB4 0EY